INTEGRATIVE STATISTICAL APPROACHES TO GAIN BIOLOGICAL INSIGHTS FROM GENOME-WIDE ASSOCIATION STUDIES

By

Ying Ji

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in
Human Genetics
June 30, 2021
Nashville, Tennessee

Approved:

James S. Sutcliffe, Ph.D.

Douglas M. Ruderfer, Ph.D.

Edwin H. Cook Jr, M.D.

Lea K. Davis, Ph.D.

Nancy J. Cox, Ph.D.

Bingshan Li, Ph.D.

To my family and friends

ACKNOWLEDGMENTS

I would like to first express my sincere appreciation to my advisor, Dr. Bingshan Li, for his endless support, constant guidance, and kindness through the course of my doctoral studies. In addition to guiding me through the completion of this dissertation, Dr. Li also supported me to pursue a Master's degree in Biostatistics. This experience enabled me to better appreciate the beauty of statistical genetics.

I also want to thank my committee, Dr. James Sutcliffe, Dr. Douglas Ruderfer, Dr. Ed Cook, Dr. Lea Davis, Dr. Nancy Cox for their insightful comments and suggestions that helped me tremendously to improve my research.

I am also grateful for Dr. Ran Tao, my advisor for Biostatistics Master program, for his valuable discussion, assistance, and encouragement throughout my graduate study.

I am very thankful to all members of the Li Lab, past and present. The lab is like my second family away from home. Dr. Rui Chen, Dr. Quan Wang, Dr. Qiang Wei, Dr. Yi Jiang, Dr. Hai Yang all have given me tremendous help from brainstorming to troubleshooting during all stages of my research projects.

I would like to thank all the collaborators of my ”polygenic risk score” project, especially Dr. Jirong Long, Dr. Xiao-Ou Shu, and Dr. Wei Zheng for sharing data from Asia Breast Cancer Consortium and for their help in writing the manuscript. This work would not have been possible without their effort and support.

I am thankful to the members of VGI for their continuous warmth and encouragement. This is the place that broadened my horizon into the fascinating field of genetics and introduced me to a group of brilliant scientists from various backgrounds. I feel lucky to be a part.

The path toward this dissertation is sometimes challenging and lonely, and I could not have progressed down the path without my supportive, inspiring friends. Especially during the quarantine, I am thankful for the precious time we spent together hiking and video-chatting.

Lastly, my forever and deepest gratitude is to my dear family, for their unconditional love, caring, and understanding that allows me to freely explore the world.

TABLE OF CONTENTS

[1] This chapter has been previously published in Genetic Epidemiology (Ji et al., 2021)

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Most common human traits such as diabetes, height, Alzheimer's disease, schizophrenia are complex. These traits are called complex because they are not caused by the dysfunction of a single gene but influenced by many genes and environmental factors. The polygenic nature of many complex traits makes it challenging to study their genetic basis. The rapid development of genotyping and sequencing technology, which has enabled the identification of genotype of many single nucleotide polymorphisms (SNPs), has revolutionized the study of complex traits. Leveraging these technologies, genome-wide association studies (GWAS) are designed to map the genetic architecture of complex traits by identifying genetic variants at a significantly different frequency in individuals with the trait and without. Since the first GWAS published in 2005 (Klein et al., 2005), GWASs have grown significantly in sample size and number of SNP-trait associations, with 251401 associations reported on the GWAS catalog to date (Buniello et al., 2019).

However, the path from GWAS to biology is not straightforward. Around 90% of disease-associated loci identified in GWAS are located in the noncoding regions of the genome (Eicher et al., 2015). It is not clear which genes and in which biological contexts does this regulation occurs. Therefore, it is very important to conduct follow-up analyses to bridge the gap from GWAS to biology. Multiple data sources, such as regulatory atlases, rare variants based findings, or GWAS from multiple ancestries can be leveraged in follow-up analyses to aid the interpretation of GWAS findings. Some GWAS follow-up analyses focus on basic research, trying to gain biological insights of traits; some are more translational, aiming to take what's learned and apply that in the development of solutions in clinical care. We are interested in integrating the above-mentioned rich data sources with GWAS findings for both types of follow-up analyses.

The availability of gene regulatory data has enabled researchers to interpret the regulatory impact of a GWAS significant locus or to prioritize the genes that are key to the predisposition of traits through integrative analysis with GWAS data using innovative analytics approaches. Common risk variants identified by GWASs map overwhelmingly to regulatory regions, suggesting that they influence traits via gene regulatory effects. Large-scale international efforts such as the Genotype-Tissue Expression (GTEx) Consortium (Consortium, 2015) have provided a regulatory landscape of gene expression and splicing variation in a broad collection of primary human tissues (Barbeira et al., 2020). While there are lots of efforts in identifying genes that influence traits via expression, genes influencing traits via splicing remain understudied.

Different from the majority of common variants that might influence traits via regulatory effects, lots of rare variants might play an important role in influencing traits by affecting protein-coding regions. A recent study suggests that most of the missing heritability, a phenomenon that GWAS identified variants together did not amount to the genetic contribution predicted by family studies, can be found in rarer gene variants in several complex traits (Wainschtein et al., 2019). Advances in sequencing technology enable focused explorations on low-frequency and rare variants to human traits. These resources provide opportunities to learn from both common and rare variants to understand the roots of many traits. But few studies have focused on using the complementary signals from rare and common variants for gene discovery.

GWAS identified genetic variants associated with a disease is valuable in indicating relevant genes, but there are also expectations that GWAS findings could be used to predict disease risk with potential clinical utility. Polygenic risk score (PRS) is developed to capture part of an individual's susceptibility to diseases through combining GWAS identified variants. Many studies have shown that PRSs can predict disease status in research-based case-control studies (Khera et al., 2018; Mavaddat et al., 2019), population-based cohort studies (Musliner et al., 2019) and in electronic health record-based studies (Lewis and Hagenaars, 2019). There are still many challenges to establish the clinical utility of PRS, one of them is ensuring that they are equally applicable to users across ethnic groups to limit exacerbating health disparities (Martin et al., 2019).

In this dissertation, we introduce new development and applications of integrative statistical approaches of GWAS data with other types of data aforementioned to both basic and more translational aspects of GWAS follow-up analyses. For the more basic research part, I introduce two different frameworks for prioritizing disease risk genes from GWAS signals: one leverages the important regulatory effects of splicing that is often ignored, the other leverages rare variant associations studies (RVASs) results that are complementary to common variants based GWAS findings for gene-based analyses; for the more translational part, I introduce an approach to improve PRS prediction accuracy for minority populations using GWAS information from multiple ancestries.

In detail, Chapter 2 of this dissertation introduces Multidimensional Splicing Gene (MSG) discovery approach to identify genes that influence traits via RNA splicing regulation. Genetic effect on RNA splicing is of comparable importance and often independent of that on expression. However, distinct from the active development and gene discovery on expression data from Predixcan (Gamazon et al., 2015)/TWAS (Gusev et al., 2016) and its multidimensional variants (e.g., S-MultiXcan (Barbeira et al., 2019)/UTMOST (Hu et al., 2019)), there has been a lack of studies on the efficient use of multidimensional splicing information for trait-associated gene discovery. To harness the rich splicing mediated effects, we propose MSG to implicate novel risk genes through integrative modeling of GWAS summary statistics and multidimensional splicing data from GTEx. We demonstrate in real and simulated data that this approach achieves controlled error rate and superior power compared to current state-of-the-art approaches. This work is still in preparation for submission.

In Chapter 3, we present a three-stage pipeline to identify risk genes from both GWASs and RVASs. RVASs offer opportunities to pinpoint genes with clear functional supports, since the ambiguity about the direction and the magnitude of impact on gene function is limited. However, due to the infrequency of rare variants and multiple testing burden, the power to identify genes from RVASs using conventional FDR control procedures like Benjamini Hochberg (BH) approach is usually limited. Hypothesis weighting provides an attractive strategy to make more discoveries while controling false discovery rate (FDR) by incorporating useful prior information about each hypothesis. Here, we first build supervised machine learning models, fed with high-confidence risk genes and local background genes selected near GWAS significant loci through a published framework iRIGS as training set genes, and multiple biological features, to assign each gene (both in and outside genome-wide loci) a prediction score that measures its disease risk. Then we use the prediction scores as covariates to prioritize RVAS results in the independent hypothesis weighting (IHW) framework. We applied the pipeline to SCZ and ASD RVASs and observed sizeable improvements on the number of genes discovered.

In Chapter 4, we develop a rescaled meta-analysis framework that improves the prediction accuracy for PRS in non-European populations. PRS is an estimate of an individual's genetic predisposition to a trait by aggregating the effects of many variants identified from GWAS. While using PRS in clinical care has a long road ahead, it has the potential for precision medicine. Currently, most large-scale GWAS efforts have been conducted in European (EUR) ancestry populations, with only 10% of all GWAS participants being of non-European (non-EUR) descent (Loos, 2020; Mills and Rahal, 2020). One important challenge we need to address before PRS can be used in clinical settings is racial disparity: most current scores are computed based on EUR GWAS studies and show reduced accuracy in other ancestries. We develop a rescaled meta-analysis framework that upweights non-EUR signals over EUR signals, yielding effect size estimates closer to the true effect sizes in non-EUR while taking advantage of the large sample sizes of EUR GWASs. As an application, we constructed PRSs using summary statistics from the rescaled meta-analysis of EUR and East Asian (EAS) breast cancer GWAS data and then evaluated their performances in an independent EAS validation dataset. Our PRS outperforms PRSs derived from the EUR or EAS GWAS alone as well as the conventional meta-analysis of EAS and EUR GWASs.

In Chapter 5, we discuss the conclusions and point out some future directions. The increasing availability of high-throughput genome-scale technologies will make more comprehensive multi-omics databases available. There are ongoing efforts to include minority populations in GWAS and biobank initiatives that have the capacity to yield extensive genetic data, and to connect genetic profiling to electronic health records. The access to such a large amount of information will accelerate the translation of GWAS loci into new biological insights.

CHAPTER 2

Integration of multidimensional splicing data and GWAS summary statistics for risk gene discovery

2.1 Introduction

Over the past two decades, genome-wide association studies (GWAS) have led to the discovery of many trait-associated loci. However, most loci are located in non-coding regions of the genome, whose functional relevance remains largely unclear (Visscher et al., 2017). Recent research suggested that a large portion of GWAS loci might influence complex traits through regulating gene expression levels (Nicolae et al., 2010; Li et al., 2016b). Methods like PrediXcan (Gamazon et al., 2015), FUSION (Gusev et al., 2016), and S-PrediXcan (Barbeira et al., 2018) have been developed to test the mediating role of gene expression variation in complex traits. These methods first build gene expression prediction models using reference transcriptome datasets (e.g., the Genotype-Tissue Expression (GTEx) Project) and then perform transcriptome-wide association studies (TWAS) to infer the genetically regulated component of tissue-specific gene expression using readily-available GWAS individual- or summary-level data. These methods have quickly become popular in practice (e.g., more than 800 citations of (Gamazon et al., 2015) and (Gusev et al., 2016) in five and six years, respectively) as they facilitate the functional interpretation of existing GWAS associations and detection of novel trait-associated genes.

Gene expression is not the only mediator of genetic effects on complex traits. Splicing is of comparable importance and often functions independently of expression (Li et al., 2016b; Barbeira et al., 2018, 2020; Gamazon et al., 2018). The splicing process involves highly context-dependent regulation and other complex mechanisms, which could be prone to errors with potentially pathological consequences (Scotti and Swanson, 2016). Recent studies indicated that at least 20% of disease-causing mutations might affect pre-mRNA splicing (Wang and Cooper, 2007), and splicing QTLs (sQTLs) could account for disproportionately high fractions of disease heritability (Akula et al., 2021; Walker et al., 2019). Despite the importance of splicing regulation, it has been understudied. There is a pressing need to investigate trait-associated genes with effects mediated by splicing.

While gene expression can usually be summarized into one measurement per gene per tissue, there are on average eight RNA splicing events per gene per tissue (Li et al., 2018). A straightforward extension of the TWAS framework for expression to splicing data would test each genetically regulated splicing event separately and then correct for multiple testing (Barbeira et al., 2020; Walker et al., 2019; Gusev et al., 2018; Li et al., 2019; Raj et al., 2018). For example, (Gusev et al., 2018) tested for around 9 times splicing events (99,562) compared to expression data (10,819) and detected a comparable number of significant genes between the two for association with schizophreniaGusev et al. (2018). These results provide support for the importance of splicing as a genotype-phenotype link. Moreover, they suggest that there might be room for appreciative power gain for detecting gene-level associations with reduced multiple testing burden.

One potential solution to reduce multiple testing burden is by analyzing multiple splicing events not individually but integratively. To integrate the multidimensional splicing data, a number of data integration approaches can be considered. These approaches can be characterized by their strategy: (A) Early: Combining data from different sources into a single dataset on which the model is built, (B) Intermediate: Combining data through inference of a joint model, and (C) Late: Building models for each dataset separately and combining them to a unified model (Gligorijević and Pržulj, 2015; Rodosthenous et al., 2019). In expression data analysis space, late integration approaches like S-MultiXcan (Barbeira et al., 2019) and UTMOST (Hu et al., 2019) have been developed to improve power by integrating signals from multiple tissues. They work by first building models for individual tissue expression, and then combining multiple single-tissue association tests within a gene into a joint test. However, these methods are designed for expression analysis and haven't been evaluated in splicing data yet.

Apart from these late integration approaches, intermediate integration approaches might offer improvements by aggregating signals during the prediction step. Here, we propose an intermediate integration approach called multidimensional splicing gene (MSG) approach, to integrate the correlated splicing events directly in the model building process based on sparse canonical correlation analysis (sCCA) (Witten et al., 2009). Previous studies have used sCCA to improve gene expression prediction models, and the authors suggested that a potential advantage of sCCA is its ability to better capture genetic contribution to gene expression shared across multiple tissues (Feng et al., 2021). As it remains unclear whether sCCA can help in splicing analysis, an exploration of its application will provide a valuable reference for future studies.

Here, we applied both established multidimensional expression analysis approaches S-MultiXcan and UTMOST and our proposed MSG approach to simulated data as well as real splicing data, and evaluated their performance. With simulations, we show that MSG provides controlled type I error rate and yields substantial power gain over S-MultiXcan and UTMOST. Real data applications using GTEx data and summary statistics from 15 complex human traits demonstrate that MSG identified on average 2.15 times and 3.23 times significant genes than S-MultiXcan, UTMOST respectively. We showcased the application of MSG to GWAS summary statistics from Alzhimer's disease (AD), low-density lipoprotein cholesterol (LDL-C) and schizophrenia (SCZ), and found the majority of splicing identified significant genes that would have been missed from expression based analysis (75%, 86%, 89%, for AD, LDL-C and SCZ respectively), highlighting the importance of splicing in genetic regulation.

2.2 Results

2.2.1 MSG model building overview

Our proposed approach MSG consists of two main stages: first, build splicing models, and then test the association between the constructed splicing models with trait of interest using GWAS summary statistics data. In the first stage, we use sCCA to construct latent canonical vectors (CVs) by identifying sparse linear combinations of SNPs and splicing events that are highly correlated with each other. In the second stage, we test for the association between genetically regulated splicing and the trait of interest based on a matrix built from CVs and SNP-trait associations from GWAS summary statistics.

There are several steps within the second stage. To integrate single splicing CV-trait relationships into a gene-level statistic, we estimate the joint effect sizes of predicted splicing on trait and compute the covariance matrix of the effect sizes. During correlation matrix estimation, as multiple predicted splicing variables within a gene can be highly correlated and can lead to numerical issues, we use a pseudo-inverse matrix derived via singular value decomposition to keep only k components of large variation following S-MultiXcan (Barbeira et al., 2019). Finally, the single splicing CV-trait tests are combined and association is detected at the gene level using using χ^2_k test. Fig 2.1 displays an overview of the method (see details in the Materials and Methods section).

We trained MSG models in both 1) simulated genotype and splicing data under different scenarios to evaluate theoretical type I error and power and 2) real genotype and splicing data from the GTEx v8 release (Consortium et al., 2020) and summary GWAS statistics for candidate gene discovery. For both types of applications, we also evaluated the performance of existing methods S-MultiXcan and UTMOST.

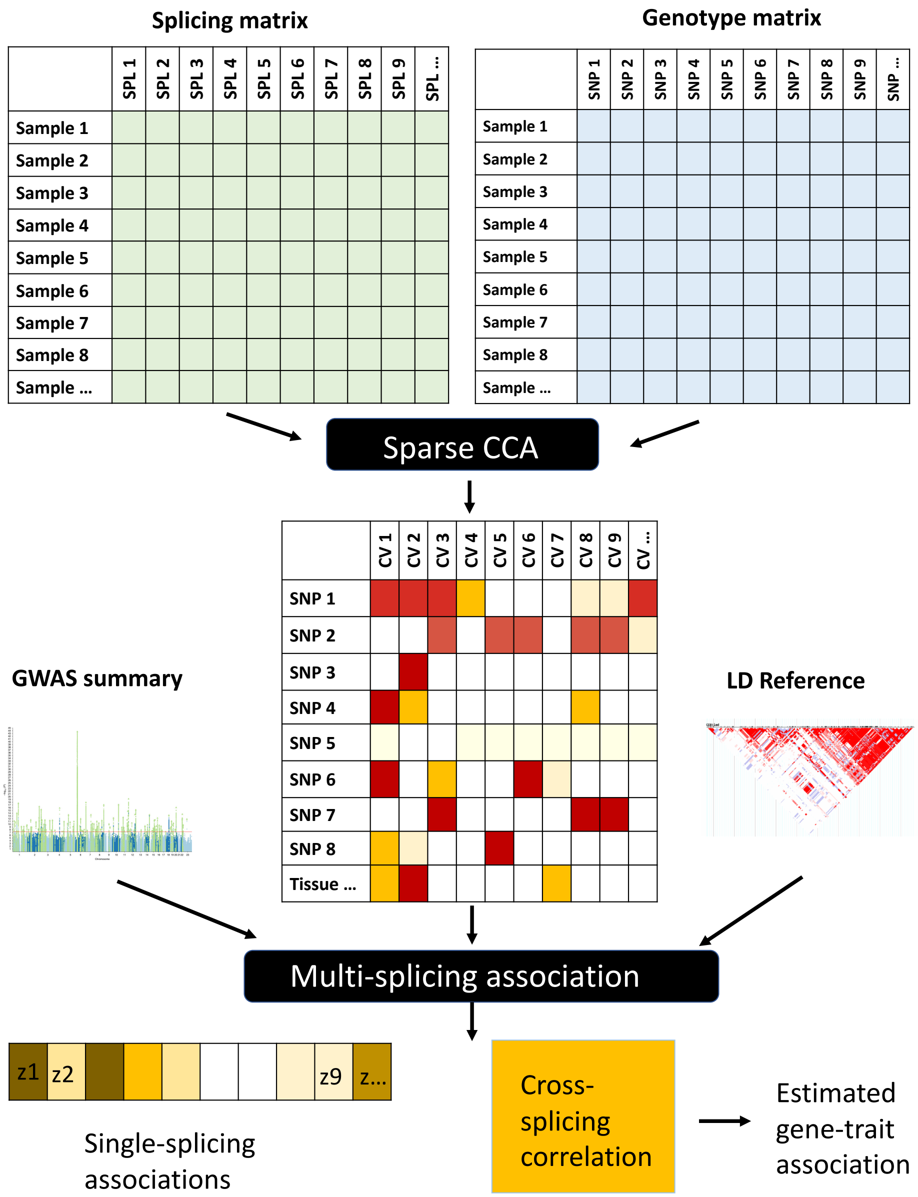


Figure 2.1: Schematic of MSG method. sCCA (Witten et al., 2009) is used to compute sparse canonical variables (CVs) from individual level genotype and splicing events data. Thes splicing CVs are used as explanatory variables in association analysis with trait of interest. We estimated the correlation matrix of predicted splicing based on the weights in CVs and LD information from a reference panel. To avoid multicolinearity, we use the SVD pseudo-inverse of the predicted correlation matrix. We quantify the significance of the inferred multi-splicing gene-level association using the single-splicing associations and the psuedo-inverse matrix in a χ^2 test. More details are provided in the Materials and Methods section.

2.2.2 Simulations: type I error and power analysis

We performed simulations to carefully examine the type I error and power of MSG and compared it with existing approaches S-MultiXcan and UTMOST. We considered several important parameters of splicing and trait genetic architecture in various realistic simulation scenarios. In the first set of simulations, we focused on the impact of different sparsity levels and cellular-level heritability (Wheeler et al., 2016; Yang et al., 2020): sparsity is the proportion of genetic variants that have non-zero effects on splicing; and heritability is the proportion of the variance of splicing events that can be explained by genotype. In the second set of simulations, we examined the impact of genetic effect sharing between splicing events, and the number of splicing events contribute to the trait. We define ”effect-sharing splicing events” as splicing events with shared SNPs with non-zero effects; and ”trait-contributing splicing events” as splicing events with non-zero effects on the trait. For each gene, we present three scenarios: 1) all splicing events are trait-contributing splicing events; 2) only effect-sharing splicing events are trait-contributing splicing events; 3) only non-effect-sharing splicing events are trait-contributing splicing events.

2.2.2.1 Type I error

We found that all three methods can effectively control the type-I error in both sets of simulations. Table 2.1 shows that splicing heritability and sparsity have little impact on the type I error in the first set of simulations. Table 2.2 shows that the number of effect-sharing splicing events and trait-contributing splicing events have little impact on the type I error in the second set of simulations with fixed heritability and sparsity at 0.05 (due to little impact of those parameters and limits on computation time). In both sets of simulations, we found S-MultiXcan and MSG provide comparable level of control while UTMOST shows a slight deflation in most scenarios.

Table 2.1: Type-I error rate in the first simulation analysis*.

shareI[a]	sparsity[b]	h^2_{sp} [c]	S-MultiXcan	UTMOST	MSG
2	0.01	0.01	0.047	0.036	0.049
		0.05	0.051	0.038	0.050
		0.10	0.057	0.043	0.049
	0.05	0.01	0.044	0.035	0.051
		0.05	0.050	0.036	0.051
		0.10	0.055	0.042	0.052
	0.10	0.01	0.041	0.031	0.051
		0.05	0.054	0.042	0.053
		0.10	0.053	0.039	0.052
4	0.01	0.01	0.044	0.034	0.050
		0.05	0.053	0.039	0.051
		0.10	0.056	0.043	0.048
	0.05	0.01	0.043	0.033	0.051
		0.05	0.048	0.034	0.047
		0.10	0.053	0.040	0.051
	0.10	0.01	0.044	0.031	0.047
		0.05	0.047	0.037	0.051
		0.10	0.057	0.045	0.054
8	0.01	0.01	0.044	0.034	0.051
		0.05	0.047	0.037	0.050
		0.10	0.054	0.041	0.049
	0.05	0.01	0.044	0.033	0.047
		0.05	0.051	0.036	0.050
		0.10	0.053	0.040	0.050
	0.10	0.01	0.042	0.032	0.048
		0.05	0.050	0.040	0.050
		0.10	0.044	0.034	0.051

* Type I error was computed as the proportion of significant genes under *p*-value cutoff at 0.05. Each entry is based on 20,000 replicates.

[a] Number of splicing events with shared non-zero effects.

[b] Sparsity: the proportion of genetic variants that have non-zero effects on splicing.

[c] Splicing heritability: the proportion of variance of sQTL that can be explained by genotype

2.2.2.2 Power

Fig 2.2 shows sparsity has little impact on power, yet the splicing heritability increase is associated with power increase in the first set of simulations. Fig 2.3 shows that power

Table 2.2: Type-I error rate in the second simulation analysis*.

Trait-contributing splicing events[a]	shareI[b]	S-MultiXcan	UTMOST	MSG
Only effect-sharing events	2	0.046	0.036	0.049
	4	0.048	0.036	0.051
	8	0.050	0.038	0.051
All events	2	0.045	0.035	0.053
	4	0.045	0.032	0.051
	8	0.050	0.034	0.051
Non effect-sharing events	2	0.048	0.036	0.049
	4	0.051	0.035	0.051
	8	0.049	0.037	0.048

* Type I error was computed as the proportion of significant genes under *p*-value cutoff at 0.05. Each entry is based on 20,000 replicates. Sparsity is fixed at 0.05, splicing heritability is fixed at 0.05.

[a] splicing events that contribute to the trait

[b] Number of splicing events with shared non-zero effects.

increases with the number of trait-contributing splicing events, regardless of which method is used and the number of effect-sharing splicing events in the second set of simulations, with trait heritability fixed at 0.01, and sparsity fixed at 0.05. In both sets of simulations, we found MSG has greater power than S-MultiXcan and UTMOST.

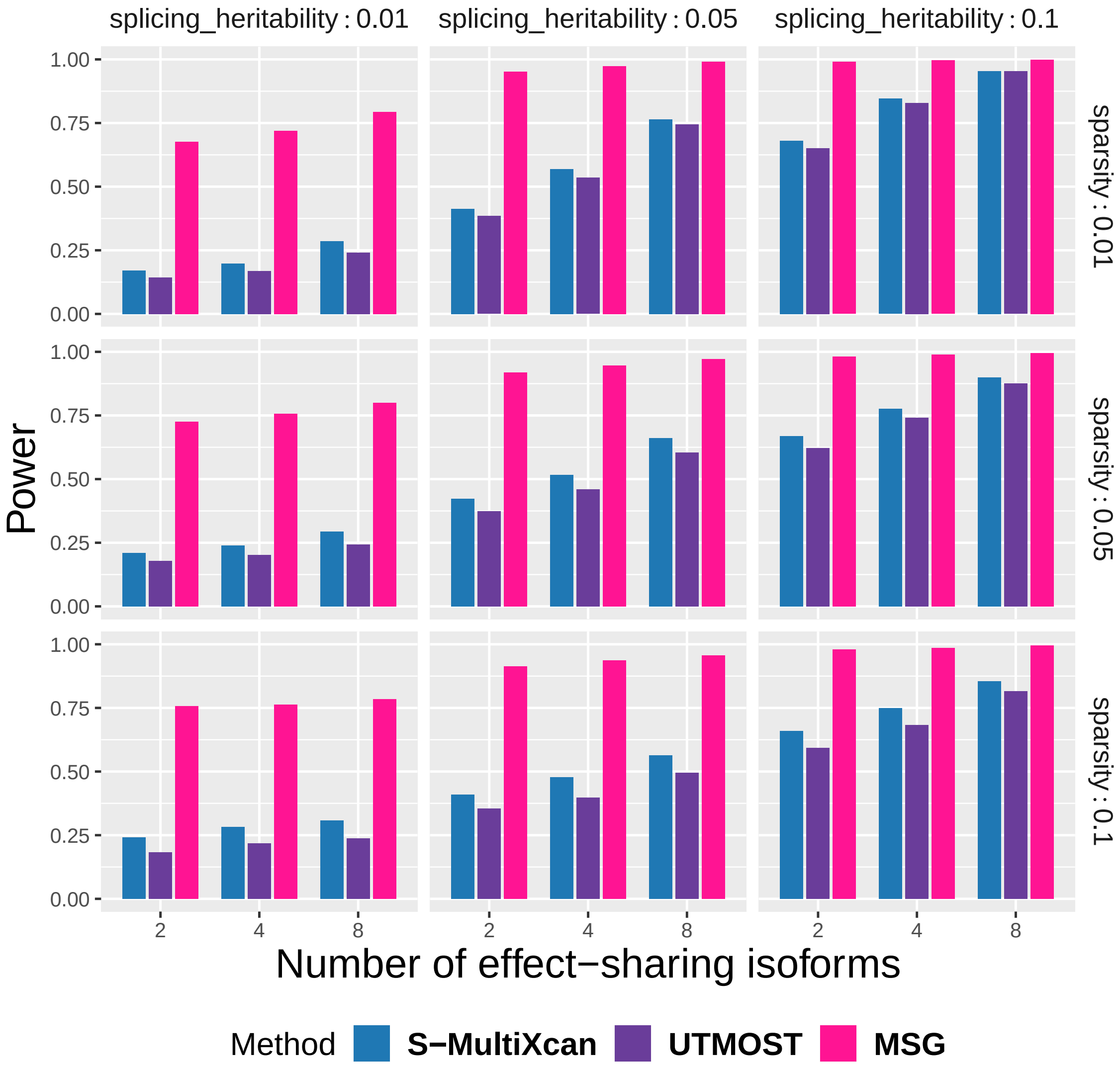


Figure 2.2: Comparison of power for S-MultiXcan, UTMOST, MSG models in simulation with different number of effect-sharing splicing events $(2,4,8)$, sparsity $(0.01,0.05,0.1)$ and splicing heritability $(0.01,0.05,0.1)$. Trait heritability is fixed at 0.01. For each subplot, the x-axis stands for the number of effect-sharing splicing events and the y-axis stands for the proportion of significant genes within 2000 simulations.

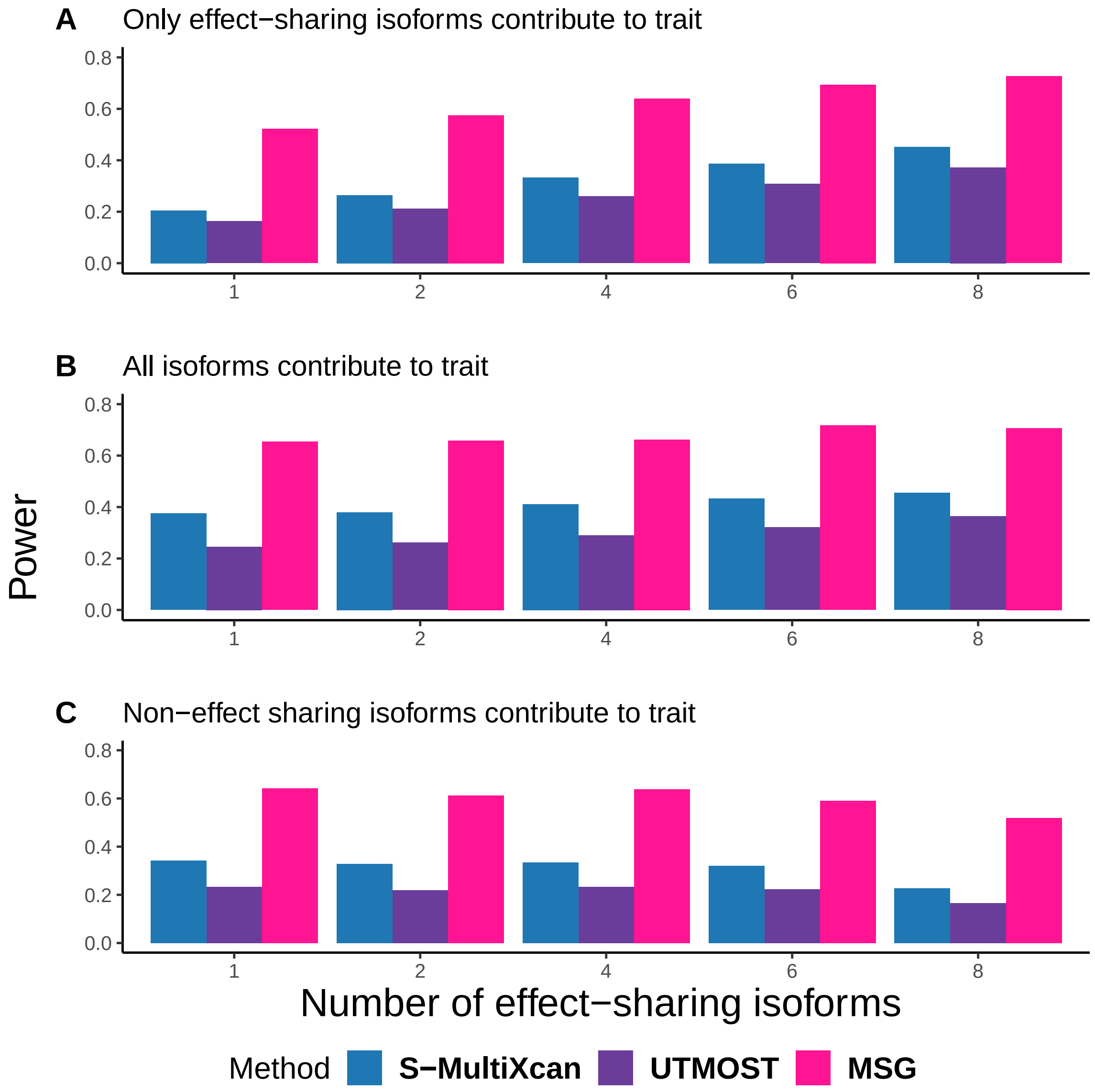


Figure 2.3: Comparison of power for S-MultiXcan, UTMOST, MSG models with different trait-contributing splicing events. Trait heritability is fixed at 0.01. For each subplot, the x-axis stands for the number of effect-sharing splicing events (1, 2, 4, 6, 8, 10) and the y-axis stands for the proportion of significant genes within 2000 simulations under p-value cutoff at 5×10^{-6}, which was chosen to mimic Bonferroni correction in real data application. A) Only effect-sharing splicing events contribute to the trait B) All splicing events contribute to the trait C) Only splicing events without shared effects contribute to trait.

2.2.3 Application to complex human traits

2.2.3.1 Summary of applications to 15 traits

We applied S-MultiXcan, UTMOST and MSG to splicing data from GTEx project to obtain prediction models, then integrated with GWAS summary statistics from 15 complex traits to identify trait-associated genes. For each trait, we chose the tissue with the top trait heritability enrichment in the respective tissue-specific annotation using linkage disequilibrium score regression (Bulik-Sullivan et al., 2015) as previously described (Hu et al., 2019). The sample sizes of these tissues range from 175 (brain frontal cortex BA9) to 706 (muscle skeletal). We extracted cis-SNPs within 500 kb upstream of the transcription start site or 500 kb downstream of the transcription stop site. GWASs for both quantitative traits (e.g., body mass index) and binary traits (e.g. Alzheimer's disease) of relatively large sample size, ranging from 51,710 (Bipolar disorder) to type 2 diabetes (408,953) are included. LD reference panels are required for inference from GWAS summary statistics. For LD reference, we used European subsamples from 1000 Genome (Consortium et al., 2015) for S-MultiXcan and UTMOST as recommended in the original publications; we used 5,000 random selected European subsamples from BioVU for MSG since our simulations and previous literature (Yang et al., 2012) suggested that a larger sample size is required for less sparse models, as our MSG models are (see Table S1 for details). To detect associations, we used a Bonferroni threshold accounting for all genes that were tested (0.05/total number of genes with splicing variation in the selected tissue) for each trait.

Table 2.3 shows that MSG identified on average 2.15 and 3.23 times the significant genes from S-MultiXcan and UTMOST respectively from the tested traits. We examined the results from three disorders (Alzheimer's disorder, LDL-C, and Schizophrenia) and found the majority of splicing informed significant genes would have been missed from expression based analysis, and some of these genes have strong external support, suggesting the complementary roles of splicing to expression and the capture of potential true signals using MSG. The genes significant in MSG for all 15 traits are included in Table S2.

Trait	Tissue	MSG	S.MultiXcan	UTMOST
Alzheimer’s Disease	BA9	32	19	14
Bipolar disorder	BA9	67	23	17
Major depressive disorder	BA9	23	5	3
Body mass index	BA9	1757	786	497
schizophrenia	BA9	458	203	145
Neuroticism	BA9	178	68	46
Type 2 diabetes	Liver	83	48	26
Total cholesterol	Liver	202	109	66
LDL cholesterol	Liver	200	108	69
Serum urate	Liver	87	63	50
HDL cholesterol	Adipose subcutaneous	161	79	53
Triglycerides	Adipose subcutaneous	144	96	69
Type 2 diabetes	Adipose subcutaneous	104	53	41
Waist hip ratio adjusted for BMI	Adipose subcutaneous	860	397	259
Age at Natural Menopause	muscle skeletal	220	118	79

Table 2.3: Numbers of significant gene-trait associations across 15 human traits using S-MultiXcan, UTMOST and MSG. The reference penal is European subsamples from 1000 Genome (for S-MultiXcan and UTMOST) and European subsamples from BioVU (for MSG). The source of GWAS traits and significant genes identified from MSG for all traits can be found in Table S2.

2.2.3.2 Application to Alzheimer's disease

We applied MSG, S-MultiXcan, and UTMOST to brain frontal cortex splicing measurements and stage I GWAS summary statistics from the International Genomics of Alzheimer's Project (IGAP(Lambert et al., 2013); N = 54,162). Fig 2.4 shows that 32, 19, and 14 significant genes in GTEx frontal cortex (BA9) was identified from MSG, S-MultiXcan, and UTMOST respectively (Bonferroni-corrected *p*-value¡ 0.05). To replicate our findings, we applied these three approaches to summary statistics from the GWAS by proxy (GWAX (Liu et al., 2017b); N = 114,564) and we found 6, 1, 0 genome-wide significant genes under Bonferroni correction from MSG, S-MultiXcan and UTMOST respectively. All significant genes from GWAX using MSG are also significant from IGAP, which amounts to 6 out of 32 genes being successfully replicated using MSG (MARK4, ERCC1, RELB, CLASRP, PPP1R37, CEACAM19). We found some well-known AD genes are significant from MSG and labelled those in Fig 2.4.

We observe 26 out of 32 MSG significant genes are within 500 kb distance to 5 GWAS identified lead SNPs, including PTK2B-CLU locus on chromosome (CHR) 1, SPI1 locus on CHR 11, MS4A4A locus on CHR 11, PICALM locus on CHR 11 and APOE locus on CHR 19 (see full list of these genes in Table S2). The observation that the most significant genes are near GWAS loci is consistent with previous reports from application of UTMOST to AD (Hu et al., 2019), and TWAS, S-MultiXcan to other traits (Gusev et al., 2018; Barbeira et al., 2018).

Furthermore, we conducted conventional S-PrediXcan analysis using GTEx prefrontal cortex gene expression data and the same summary statistics, and compared those to MSG identified splicing genes. We found 8 genes are overlapped between expression (S-PrediXcan) and splicing (MSG) significant genes. The remaining 24 out of 32 splicing genes would have been missed in conventional approaches evaluating gene expression levels alone, a few of them are showed in Fig.2.5. For genes that can only be identified via splicing, PICALM (MSG splicing *p*-value=1.93×10^{-9}, expression *p*-value=9.80×10^{-1}) and PTK2B

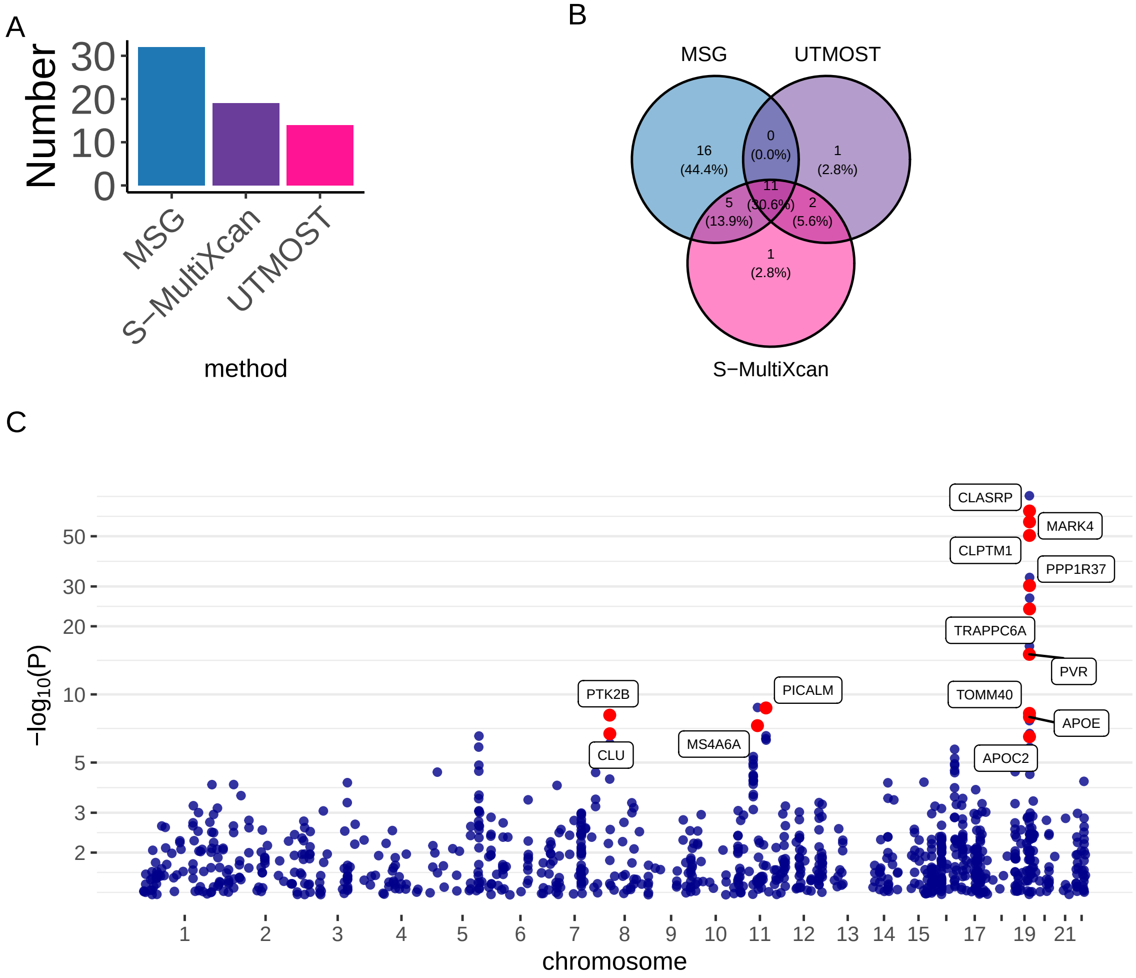


Figure 2.4: Applying various methods on IGAP Alzheimer's disease. A). Bar plots of the number of significant genes using the different training models. B), Venn diagram plots showing the overlap of the genes from different methods. C). Manhattan plot for Alzheimer's disease from the application of MSG model in IGAP stage I GWAs summary statistics. Genes with strong literature support are annotated in red.

(MSG splicing p-value=7.96×10^{-9}, expression p-value=8.98×10^{-1}) are two genes previously shown to be significantly differentially spliced between AD patients and controls from ROSMAP dataset (Raj et al., 2018). MARK4 (MSG splicing p-value=1.31×10^{-58}, expression p-value=8.48×10^{-2}) was shown to change the properties of tau (Oba et al., 2020) and has variants reported to be associated with AD and AD family history (Marioni et al., 2018; Jansen et al., 2019). Several genes in APOE region are also significant from splicing analysis but not from expression, including APOE (MSG splicing p-

value=1.12×10^{-8}, expression p-value=2.49×10^{-3}), a well-known risk gene (Yu et al., 2007) for late-onset AD, with reports that alternative splicing (exclusion of exon 5) is associated with increased beta-amyloid deposition, and affecting tau structure (Love et al., 2015); APOC1 (MSG splicing p-value=4.65×10^{-17}, expression p-value=2.07×10^{-3}) has been reported to be associated with family history of AD and AD (Schwartzentruber et al., 2021; Herold et al., 2016); TOMM40 (MSG splicing p-value=5.67×10^{-9}, expression p-value=4.84×10^{-1}) have previously reported to have intronic variants associated with family history of AD (Marioni et al., 2018) and HDL levels (Zhu et al., 2019); ERCC1 (MSG splicing p-value=2.32×10^{-27}, expression p-value=4.48×10^{-2}), a DNA repair enzyme, has been shown to be associated with quantification of amount of tau and implicated in AD research (Wang et al., 2020a).

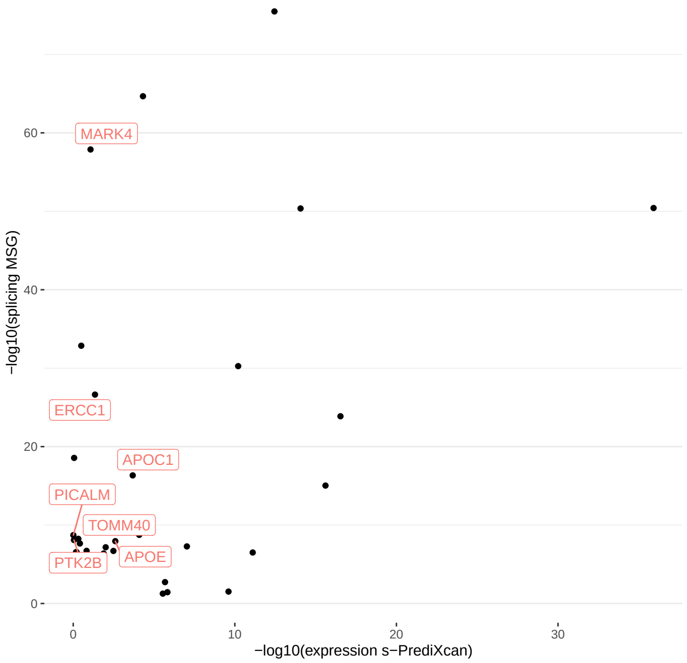


Figure 2.5: AD Genes identified via splicing that would be missed from expression. The full list of significant splicing genes identified using MSG is shown in Table S2.

2.2.3.3 Application to LDL-C

We applied MSG, UTMOST and S-MultiXcan to a continuous trait: low-density lipoprotein cholesterol (LDL-C) from global lipids genetics consort ism (GLSC) GWAS (n=188,578) (Willer et al., 2013). Fig.2.6 shows that 200, 108, 69 significant genes are identified in GTEx liver from MSG, UTMOST, and S-MultiXcan, respectively (Bonferroni-corrected *p*-value¡ 0.05). There are 57 genes shared by all three methods. To replicated our findings, we applied three approaches to summary statistics from the LDL-C UK Biobank GWAS (N = 343,621) and identified 474, 223, and 175 from MSG, S-MultiXcan, UTMOST respectively. The replication rate is high in all three methods: of the significant genes identified from the GLSC GWAS (n=188,578), 161 out of 200 genes (81%), 79 out of 108 genes (73%), 52 out of 69 genes (75%) are replicated in UK Biobank GWAS analysis using MSG, S-MultiXcan, and UTMOST respectively. As shown in Fig 2.6, We found MSG captured some well-known lipid metabolism genes (Zhou et al., 2020) including LPIN3, FADS3, LDLRAP1, FADS1, LDLR, FADS2. We note LDL-C significant genes tend to cluster around known SNP-level significant loci to a lesser extent than AD. 102 out of 200 MSG significant genes are within 500 kb distance to 20 GWAS significant lead SNPs (see the full list of these genes in Table S2).

Furthermore, we conducted conventional S-PrediXcan analysis using GTEx liver gene expression data with the same GWAS summary statistics, and compared those to MSG identified splicing genes. We found 27 of the 56 expression genes overlapped with the splicing significant genes identified by MSG. 173 out of 200 splicing genes identified from MSG would have been missed in conventional approaches evaluating gene expression levels alone. As shown in Fig 2.7, some genes can only be identified via splicing: HMGCR (MSG *p*-value=1.14×10^{-40}, expression *p*-value=3.00×10^{-4}) has variants affecting alternative splicing of exon13 and is associated with LDL-C across populations (Burkhardt et al., 2008); PARP10 (MSG *p*-value=1.51×10^{-8}, expression *p*-value=1.07×10^{-2}) has been prioritized as causal gene from exome-wide association analysis in more than 300,000

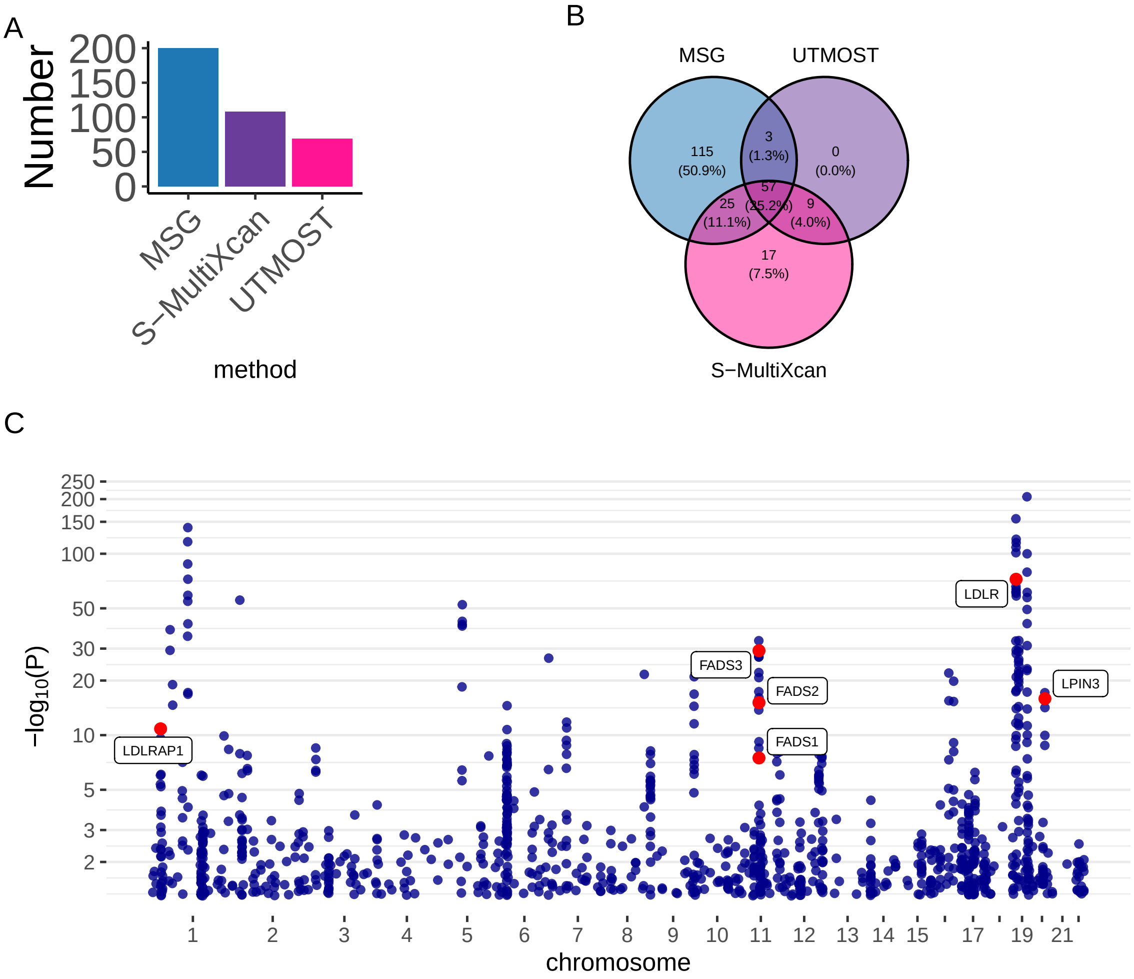


Figure 2.6: Comparison of the performance of the various methods on LDL-C in Teslovich et al, Nature (2010) GWAs summary statistics. a. Bar plots of the number of significant genes using the different training models. b, Venn diagram plots showing the overlap of the genes. c, Manhattan plot for LDL-c from the application of MSG model, six genes previously known to be associated with LDL-C are annotated in red.

individuals (Liu et al., 2017a); SMARCA4 (MSG p-value=3.27×10^{-109}, expression p-value=6.52×10^{-2}) was shown to have variants associated with LDL cholesterol levels (Richardson et al., 2020), coronary heart disease susceptibility (Guo et al., 2017; Dichgans et al., 2014) and myocardial infarction (Nikpay et al., 2015); LDLR (MSG p-value=4.49×10^{-73}, expression p-value=1.43×10^{-1}) has been reported to be associated with statin use from UK Biobank studies, and has intronic variants identified in Familial Hypercholesterolemia cases (Reeskamp et al., 2018); CARM1 (MSG p-value=1.05×10^{-66}, expression p-value=3.06×10^{-1}) has been reported to have intronic variants associated LDL-C and total cholestrol (Hoffmann et al., 2018).

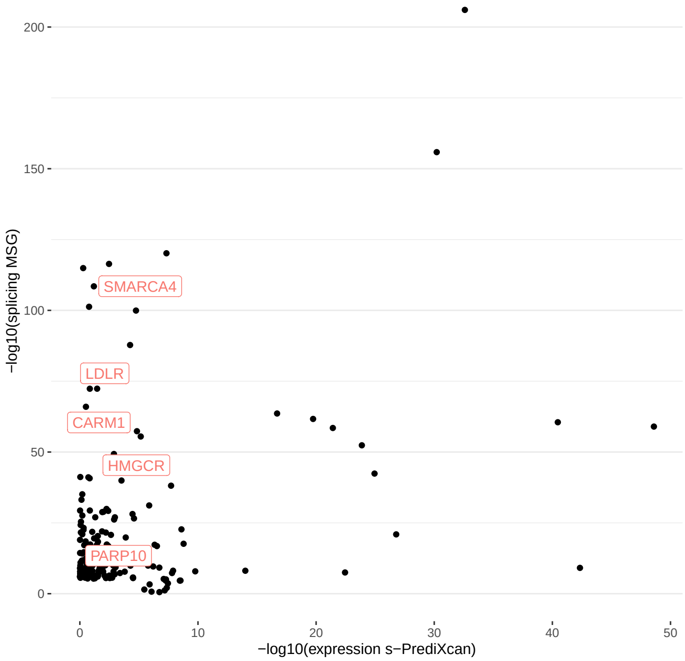


Figure 2.7: LDL-C Genes identified via splicing that would be missed from expression. The full list of significant splicing genes identified using MSG is shown in Table S2.

2.2.3.4 Application to Schizophrenia

We applied MSG, UTMOST and S-MultiXcan to a schizophrenia (SCZ) GWAS (N=105,318) (Pardiñas et al., 2018). Fig.2.8 shows that 501, 222, 153 significant splicing genes are identified in GTEx frontal cortex (BA9) using MSG, S-MultiXcan, and UTMOST respectively (Bonferroni-corrected *p*-value¡ 0.05). There are 116 genes shared by all three methods. Current available large-scale SCZ GWAS usually have sample overlaps, so we weren't able to replicate with a independent GWAS. We found a few genes previously reported to influence SCZ risk via splicing, including SNX19 (Ma et al., 2020b), AS3MT (Li et al., 2016a) and CYP2D6 (Ma et al., 2020b). We note SCZ significant genes tend to cluster around known SNP-level significant loci to a lesser extent than AD. 376 out of 501 MSG significant genes are within 500 kb distance to 76 GWAS significant lead SNPs (see full list of these genes in Table S2).

Furthermore, we conducted conventional S-PrediXcan analysis using GTEx prefrontal cortex gene expression data and the same summary statistics, and compared those to MSG identified splicing genes. We found 55 expression genes overlapped with the significant splicing genes identified by MSG. Due to the complex haplotype and LD structure of the major histocompatibility complex (MHC) locus, we further separate the results into genes in MHC region and genes not in MHC region. In the MHC region, 30 genes overlapped between 33 expression genes (S-PrediXcan) and 101 splicing genes (MSG). Some well-known genes can only be identified from splicing: NOTCH4 (MSG *p*-value=8.35×10^{-29}, expression *p*-value=8.19×10^{-2}) (Aberg et al., 2013), TRIM26 (MSG *p*-value=4.64×10^{-14}, expression *p*-value=4.40×10^{-1}) (Consortium et al., 2012), ZSCAN9 (MSG *p*-value=4.64×10^{-14}, expression *p*-value=4.40×10^{-1}) (aut, 2017). Outside of the MHC region, 25 genes overlapped between 58 expression genes (S-PrediXcan) and 400 splicing genes (MSG). Fig 2.9 shows that some genes can only be identified via splicing: SNX19 (MSG *p*-value=2.27×10^{-10}, expression *p*-value=2.38×10^{-3}) is known to have risk-associated transcripts defined by an exon-exon splice junction between exons 8 and 10

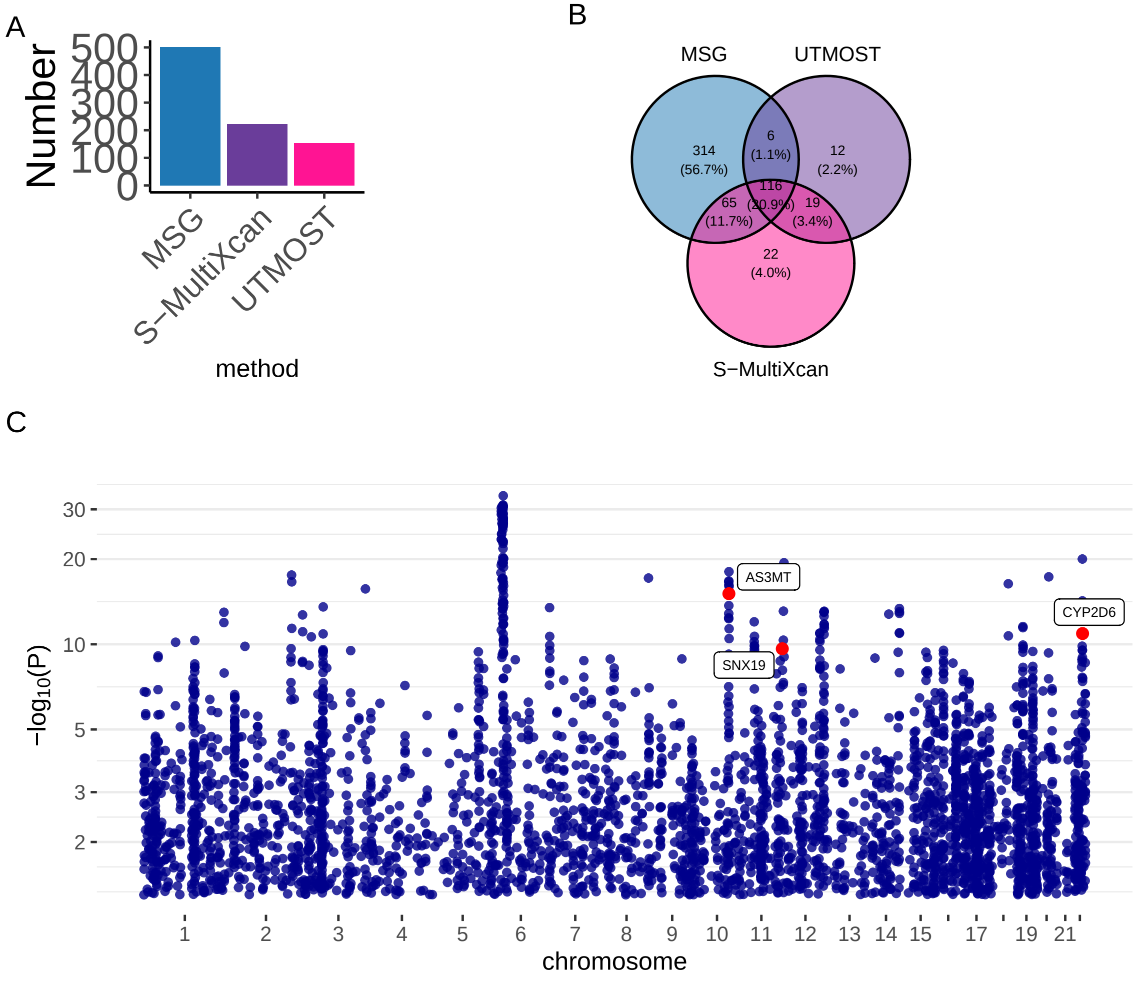


Figure 2.8: Comparison of the performance of the various methods on SCZ. a) Bar plots of the number of significant genes using the different training models. b) Venn diagram plots showing the overlap of the genes. c) Manhattan plot for scz from the application of MSG model, genes previously known to be associated with SCZ by splicing are annotated in red.

(junc8.10) that is predicted to encode proteins that lack the characteristic nexin C terminal domain (Ma et al., 2020a); GRIA1 (MSG p-value=1.28×10^{-8}, expression p-value=1.62×10^{-5}) has been reported to be associated with SCZ (Ripke et al., 2013); CACNA1C (MSG p-value=9.35×10^{-10}, expression p-value=5.44×10^{-1}) and CACNA1G (MSG p-value=2.45×10^{-6}, expression p-value=7.15×10^{-1}) encode calcium voltage-gated channel subunit and has been implicated in multiple studies to be associated with SCZ (Lam et al., 2019a; Ripke et al., 2013); PPP1R16B (MSG p-value=4.86×10^{-18}, expression p-value=8.62×10^{-1}) has been reported to be associated with SCZ in several populations (Ripke et al., 2013;

Goes et al., 2015) and multiple psychiatric disorders (Wu et al., 2020; Lam et al., 2019b).

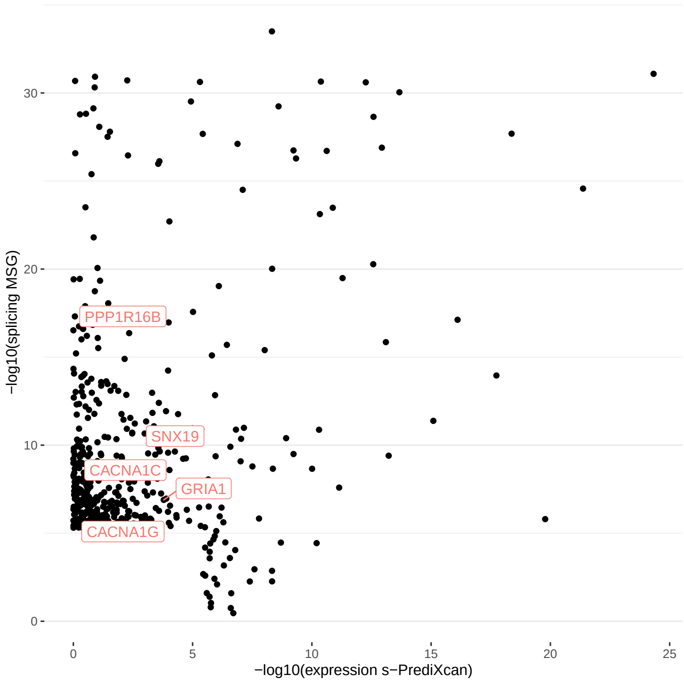


Figure 2.9: SCZ genes identified via splicing that would be missed from expression. The full list of significant splicing genes identified using MSG is shown in Table S2.

2.3 Discussion

Distinct from the active development and gene discovery using expression data from Predixcan/TWAS and its multidimensional variants (e.g., S-MultiXcan/UTMOST), there is a lack of studies on trait-associated gene discovery using splicing data. Splicing data presents some unique challenges due to its multidimensional nature, which prompts the development of efficient analytic approaches. In this paper, we proposed a method (MSG) to construct cross-splicing event models using sCCA to boost power in identifying genes influencing traits via splicing. Through simulations, we showed MSG has controlled type I error rate and superior power compared to current state-of-the-art approaches S-MultiXcan/UTMOST. In real data applications, we identified on average 2.15 times and 3.23 times significant genes from MSG than from S-MultiXcan or UTMOST across 15 complex traits. We highlighted studies on AD, LDL-C and SCZ, and found independent literature support for MSG-identified genes, indicating MSG captures some true signals. Additionally, the majority of genes identified from MSG are not implicated in traditional expression-based studies, consistent with the complementary roles between genetic regulation of splicing and expression.

Through MSG, we found a considerable number of trait-associated splicing genes that were not identified from expression data, demonstrating the important role of RNA splicing on trait susceptibility. Besides, the number of splicing genes is usually larger than that from expression. A few factors might contribute to this. One is that splicing is highly prevalent, affecting over 95% of human genes (Wang and Cooper, 2007). It provides the possibility of cell type- and tissue-specific protein isoforms, and the possibility of regulating the production of different proteins through specific signalling pathways (Kornblihtt et al., 2013). Another factor is that the rich multi-dimensional splicing information provides higher power to detect association compared to single dimension expression information. It was shown that the power of S-PrediXcan/TWAS approaches an maximum when sample size reachs 1000 (Gusev et al., 2016). As most tissues in GTEx have a sample size less than

1000, the sample size from a target tissue maybe too small to have enough power for expression data-based S-PrediXcan/TWAS analysis, but enough to detect some associations for multidimensional splicing data analysis. Thus, we believe splicing data analysis may offer great opportunities to study complex traits, and we view our method as an important early step toward using sQTLs for GWAS interpretation and gene discovery.

We observed a 2- to 3- fold increase in the number of trait-associated splicing genes from MSG compared to established methods S-MultiXcan and UTMOST. The relative increase of power using MSG can be attribute to several factors. First, we found MSG to increase the number of ”testable” genes compared to the alternatives. For example, for SCZ, there are 1041 genes not testable in S-MultiXcan or UTMOST but testable by MSG. We found sCCA models tend to be less sparse (i.e., include more SNPs with estimated non-zero effects) than S-MultiXcan and UTMOST and explain more variability in splicing variation, and in turn are more likely to be testable for association with traits. Another potential strength of MSG is that sCCA might be able to capture the correlated genetic effects more effectively. We speculate the CVs in MSG might tend to capture genetic variance. Meanwhile, predictions in S-MultiXcan/UTMOST might tend to capture the total phenotypic variance which includes both genetic and non-genetic variation (Aschard et al., 2014), and thus are less powerful than MSG in the application to multidimensional splicing data. These desirable properties of sCCA on correlated data have also been suggested in previous applications to multi-tissue expression data (Feng et al., 2021).

We note MSG models tend to be less sparse compared to alternative methods like S-MultiXcan and UTMOST. Accordingly, simulations show our method require a larger reference panel than the commonly used 1000 Genomes European samples to have controlled type I error. Model sparsity seems to have influenced the granularity level needed for LD reference, which is partially driven by the reference sample size. Intuitively, the more SNPs in the model, the less standard deviation of each SNP is needed in the LD matrix, since the sum of many errors will make the model unreliable. In fact, previous studies have recommended choosing a reference with a large sample size (i.e, at least 2,000) to estimate the LD correlations with little error in analyzing GWAS summary statistics (Yang et al., 2012). Through simulations, we explored using LD reference panels of different sample sizes (i.e., 400; 1000; 5,000; 10,000; 50,000) when conducting analysis with a GWAS of 50,000 samples (see Table S1 for details). We found a reference panel of 5,000 individuals is adequate for MSG. To construct this large reference sample, we randomly selected 5,000 samples of European population in BioVU and have made the reference LD correlation matrices available. However, while not ideal, we think 1000 Genomes samples could also be used for initial screening purpose, but the more stringent validation will be needed for the genes identified.

There are several limitations to our approach. First, we focused on single-gene, single-trait analysis on splicing data in our method, while there are exciting opportunities for method development and gene discovery if we transfer the knowledge of multi-tissue, multi-trait, multi-gene, cis and trans from expression analysis to splicing studies (Liu et al., 2021; Luningham et al., 2020; **?**). Second, when combining signals from multiple splicing canonical vectors, we used SVD approach following S-MultiXcan (Barbeira et al., 2019). This application was effective in solving collinearity and numerical issues, but there are other approaches to combine signals like ACAT (Liu et al., 2019) might lead to further power gain as suggested by applications in expression data analysis (Feng et al., 2021). Third, our models were derived from GTEx transcriptome sequencing from adult bulk tissue, so findings driven by differences in cellular composition or developmental stages cannot be fully resolved. As splicing must be tightly regulated, the association of splicing implicated genes with traits in different cell types or developmental stages remains to be studied. Fourth, like other S-PrediXcan/TWAS-type approaches, results from our method need to be interpreted with caution: they do not implicate causality. Further causal analysis using methods like FOCUS (Mancuso et al., 2019) and experimental validation are needed to determine causal genes.

By integrating multidimensional splicing information with GWAS, we were able to pinpoint candidate genes associated with common traits via splicing. This approach can potentially be extended to integrate molecular data beyond splicing, such as epigenetic data. With the increasing availability of summary statistics and molecular data, we believe we will have a better understanding of how genes influence complex traits through diverse regulatory effects.

2.4 Software and resources

The genotype data for the GTEx project are available on AnVIL. Processed GTEx gene expression and splicing data (fully processed, filtered, and normalized splice phenotype matrices (in BED format)) is available from the GTEx portal (https://gtexportal.org). The source of the summary statistics datasets of all GWAS meta-analyses analyzed in this paper can be found in Table S2. The significant MSG genes from the 15 human traits are provided in Table S2. The LD correlation matrices for cis-SNPs of each gene from a reference panel of 5,000 BioVU samples of European ancestry will be available on https://zenodo.org/. The LD reference panel from 1000 Genomes is available at https://data.broadinstitute.org/alkesgroup/FUSION/LDREF.tar.bz2.

The code for MSG is available at Github https://github.com/yingji15/MSG_public. We used R package PMA (Witten and Tibshirani, 2020) to implement sCCA for splicing analysis. Part of the code is modified from previous work: S-MultiXcan at https://github.com/hakyimlab/MetaXcan; TWAS at https://github.com/gusevlab/fusion_twas/; UTMOST at https://github.com/Joker-Jerome/UTMOST; TisCoMM at https://github.com/XingjieShi/TisCoMM; JTI at https://github.com/gamazonlab/MR-JTI.

2.5 Materials and methods

2.5.1 MSG framework

To identify candidate trait-associated genes, suppose we have both the splicing and genotype data, and GWAS summary statistics from a study that measure both the trait of interest and the genotype. The two studies are independent with no sample overlap. Our proposed method consists of two main stages. In the first stage, we construct sparse latent canonical vectors (CVs) by identifying sparse linear combinations of SNPs and splicing events that are highly correlated with each other from data provided in GTEx v8 release. In the second stage, we test the association between genetic regulated splicing and the trait of interest based on the obtained SNP-splicing CVs and summary-level data from GWAS to perform association analysis of estimated splicing with a phenotype.

In the first stage, for a given gene, we denote X_1 as a $n_1 \times p$ genotype matrix for p SNPs reside in the cis-region of the gene (i.e., 1-Mb window around a gene) among n_1 samples, and Y as a $n_1 \times t$ matrix of measured t splicing events from GTEx. Given X_1 and Y, sCCA seeks sparse latent CVs written in matrix forms B and V such that $Cor(X_1B, YV)$ is maximized based on constraints. The objective function is:

$$\begin{aligned} &\max_{B,U} B^T X_1^T YV \\ &\text{subject to} ||B||^2 \leq 1, ||V||^2 \leq 1, ||B||_1 \leq c_1, ||V||_1 \leq c_2, \end{aligned} \tag{2.1}$$

where c_1 and c_2 are parameters chosen to yield sparse B and V, with $||B||_1$ and $||V||_1$ denote L_1 (or *lasso*) penalties. The solutions for B and V are obtained via an iterative algorithm (Witten et al., 2009; Witten and Tibshirani, 2009).

In the second stage, we test the association between genetic-regulated splicing CVs and the trait of interest. If individual level data from GWAS study is available, we denote X_2 as a $n_2 \times p$ genotype matrix for the same p SNPs reside in the cis-region of the gene among n_2 samples, and z_2 as a n_2-vector of the phenotypic value. Then we can use PCA regularization on B by decompose the predicted matrix X_2B into principal components and keep only the k eigenvectors of non-negligible variance, following S-MultiXcan (Barbeira et al., 2019) to avoid collinearity issues. As individual level GWAS studies are usually not available, our MSG approach focus on using GWAS summary statistics and a reference panel to conduct the association test following previous work (Gusev et al., 2016; Barbeira et al., 2019; Hu et al., 2019). It consists of the following steps:

(1) Computation of single splicing CV association results following S-PrediXcan (Barbeira et al., 2018). Let S denote the matrix of imputed genetic regulated splicing, X_3 denote a $n_3 \times p$ genotype matrix from reference panel of n_3 individuals and p SNPs. We predict genetic regulated splicing variables using the splicing CV matrix B obtained in the previous

stage:

$$S = X_3 B. \tag{2.2}$$

Let T denote the trait of interest in GWAS, X denote the genotype matrix of GWAS samples of the p cis-SNPs and S denote the predicted splicing variables in GWAS samples, γ denote the effect sizes of predicted splicing variables within this gene on the trait. We assume a linear regression model, and we are interested in estimating regression-coefficient-vector γ:

$$T = S\gamma + \varepsilon_T, \tag{2.3}$$

We also have GWAS estimates vector β_{GWAS} of the p cis-SNPs for this gene, and the linear regression model of the trait on the genotype of these SNPs is:

$$T = X\beta_{GWAS} + \varepsilon_{GWAS}, \tag{2.4}$$

The ordinary least-square estimator $\hat{\gamma}$ and its variance is as follows:

$$\hat{\gamma} = (S^T S)^{-1} S^T T = (S^T S)^{-1} B^T X^T T = (S^T S)^{-1} B^T X^T X \beta_{GWAS}, \tag{2.5}$$

$$se(\hat{\gamma})^2 \approx Var(T)(S^T S)^{-1} = Var(\beta_{GWAS})(X^T X)(S^T S)^{-1} = se(\beta_{GWAS})^2 (X^T X)(S^T S)^{-1}. \tag{2.6}$$

Plug in both equations 2.5 and 2.6, let $z_{GWAS} = \frac{\beta_{GWAS}}{se(\beta_{GWAS})}$ be a vector of z-scores from GWAS summary statistics, the estimated z scores for γ is

$$\hat{z} = \frac{\hat{\gamma}}{se(\hat{\gamma})} = \frac{\sqrt{X^T X}}{\sqrt{S^T S}} B^T z_{GWAS} \tag{2.7}$$

Since individual level data from GWAS samples is usually not available, we estimate $X^T X$ and $S^T S$ from the reference population, not the actual GWAS population. Then we obtain $\hat{z}$ by plugging in these estimates.

(2) Estimation of the correlation matrix of predicted splicing using the linkage disequilibrium (LD) information from a reference panel. As multiple predicted splicing events within a gene can be highly correlated and can lead to numerical issues caused by collinearity, we obtain a pseudoinverse for correlation matrix $S^T S = (X_3 B)^T X_3 B$ by discarding the components of the smallest variation via singular value decomposition (SVD), analogous to PC analysis discussed above in individual-level data (Barbeira et al., 2019). We denote the resulting matrix with k components surviving the SVD peudo-inverse as Σ_k^+.

(3) Quantification of the predicted multidimensional splicing gene-level association. We use the fact that the regression coefficients follow $\hat{\gamma} \sim N(\gamma, \sigma^2 (S^T S)^{-1})$. Under the null hypothesis of no association, it follows that $\hat{\gamma}^T \frac{S^T S}{\sigma^2} \hat{\gamma} \sim \chi^2$. As previously shown (Barbeira et al., 2019), $\hat{\gamma}^T \frac{S^T S}{\sigma^2} \hat{\gamma} \approx \hat{z}^T Cor(S)^{-1} \hat{z}$ and we can use the pseudo-inverse matrix Σ_k^+ in place of $Cor(S)^{-1}$ in practice. Then we combine all the association statistics $\hat{z}$ within a gene $\hat{z}^T \Sigma_k^+ \hat{z}$ in a χ^2 test. We estimate for all genes using the above procedure. For each trait and tissue combination, we use Bonferroni correction to determine the genome-wide significance threshold by dividing 0.05 with the number of genes with at least 2 splicing events in that tissue. This value varies between trait-tissue combinations, usually approximately at $0.05/10000 = 5 \times 10^{-6}$.

2.5.2 Simulations

To evaluate the type I error rate and power of the association tests, we simulate a training dataset with both genetic and splicing data, a GWAS dataset with both individual-level data and summary statistics, and an LD reference panel. Then we conduct gene level association tests using our proposed approach MSG and alternative methods (S-MultiXcan and UTMOST) in a wide variety of different controlled scenarios.

We assume a linear regression model to simulate the training genotype data X_1 and splicing events data Y of a gene: $Y = X_1 B + E$. Here, X_1 is a matrix with 200 rows representing samples and 300 columns representing the cis-SNPs residing in a 1Mb window around the gene, simulated with values drawn from a multivariate normal distribution with autoregressive covariance structure determined by $\rho_X = 0.1$. Y is a matrix with 200 rows representing samples and 10 columns representing splicing events within the gene. The effect size matrix B, with rows representing SNPs and columns representing splicing events, is factored into SNP-dependent and splicing event-dependent components represented by $B = diag(b)W$, with b determining the magnitude of shared effect on each SNP and W specifying the effects of SNPs on splicing events following previous work (Shi et al., 2020). To model the structure of W, we determine the location of non-zero elements in W through the following parameters: we denote $shareTi$ ($shareTi = 1, 2, 4, 6, 8$) to be the number of ”effect-sharing splicing events”, which are splicing events which have non-zero effect SNPs shared between them; $shares$ ($shares = 0.3$) to be the fraction of shared SNPs in non-zero effect SNPs for these effect-sharing splicing events; and sparsity level s ($s = 0.01, 0.05, 0.1$) to be the overall fraction of non-zero effect SNPs out of the total 300 SNPs for each splicing event (see illustration in supplementary materials Fig 2.10). The value of these non-zero elements in W are randomly generated from a uniform distribution. We describe the splicing heritability h_c ($h_c = 0.01, 0.05, 0.1$) as the proportion of splicing variability explained by SNPs. E is a matrix of random Gaussian noise term with a scale such that it accounts for $(1 - h_c^2)\%$ of the expected variance in Y with autoregressive covariance structure determined by $\rho_E = 0.5$. The values are chosen to be comparable to the observed value in real data applications.

To simulate the GWAS dataset, we assume $T = X_2 B\alpha + \varepsilon_T$. We first generated an individual level GWAS set and obtained summary statistics from marginal linear regression to this simulated individual-level GWAS dataset, then we provided the summary statistics to all prediction models. Here, the genotype matrix X_2 is a matrix of 50,000 rows representing samples and 300 columns representing cis-SNPs, and is simulated similar to X_1. Quantitative GWAS trait T is generated using B from the simulated splicing training dataset and a vector α representing the effects of 10 splicing events on the trait. The non-zero entries in α denote the "trait-contributing splicing events", which are splicing events with nonzero effects on the trait, with values generated from a uniform distribution. The heritability of the trait h_T is set to be 0.01. ε_T is a random Gaussian noise term chosen such that it accounts for $(1-h_T^2)\%$ of the expected variance in T to keep h_T at 0.01.

To aid using summary level GWAS data for inference, we also simulated the LD reference panel X_3 with the same cis-SNPs using similar settings as X_1 and X_2. We consider two reference panels, one with 400 samples (mimic 1000 Genomes European reference samples) and another with 5,000 samples (mimic random selected BioVU samples). Our simulation shows MSG requires a larger reference sample size, a comparison between the two reference panels is included in the Table 2.4.

We performed simulations of diverse scenarios through varying splicing sparsity, splicing heritability, effect-sharing splicing events, and trait-contributing splicing events. We estimated the type I error control of different methods (i.e., S-MultiXcan, UTMOST, MSG) under the null ($h_z = 0$) and alternative hypothesis ($h_z = 0.01$). For type I error evaluation, we repeated simulations 2×10^5 times for each scenario and report the proportion of simulations with p-value < 0.05. For power evaluation, we repeated the scenario for 2,000 times and report the proportion of simulations with p-value $< 5 \times 10^{-6}$ which was chosen to mimic Bonferroni correction in real data application.

2.5.3 Datasets

GTEx data Genotype and splicing data from RNA sequencing data were obtained from the Genotype-Tissue Expression Project (GTEx v8p). To select relevant tissue for each trait, we used the top tissues enriched for trait heritability provided by UTMOST supplementary materials (Hu et al., 2019). The resulting tissues for our selected traits are frontal cortex (BA9), liver, adipose, and muscle skeletal tissues (see Table S2 for sample size per tissue).

GWAS summary statistics We obtained GWAS summary statistics from 15 traits of relatively large sample size ($N \geq 50,000$) from categories like metabolites (e.g., HDL-C, LDL-

C) and psychiatric/neurodegenerative disorders (e.g., Alzheimer's disease, Schizophrenia). Details of these traits can be found in Supplementary Table S2. In the main text, we discussed Alzheimer's disease, LDL-C and Schizophrenia in detail. A summary of analysis results for all traits can be found in Table 2.3.

LD Reference panel Due to the absence of GWAS genotype data using summary statistics, we use reference samples to estimate the LD structures among SNPs in the study samples. We used two European reference panels since diseases and traits considered in our real data application are for European population cohorts: one from random selected 5,000 samples in BioVU and another from 1000 Genome Project (Consortium et al., 2015). Our simulation established that a reference LD correlation matrix constructed from 5,000 individuals is recommended for our MSG method (Table 2.4). The LD correlation matrices from 5,000 random BioVU samples will be provided at https://zenodo.org/.

2.5.4 Alternative methods

We conducted simulations and real data analysis to evaluate the performance of different methods by performing gene-trait association tests. For multidimensional splicing data analysis, we compared the performance of our MSG approach with two current state-of-the-art methods in the main text: S-MultiXcan (Barbeira et al., 2019) and UTMOST (Hu et al., 2019). Specifically, for splicing event i, let X_i and y_i be the genotype and splicing data. Both S-MultiXcan and UTMOST assume an elastic net model that combines L_1 and L_2 penalty as a variable selection approach following the assumptions made in PrediXcan (Gamazon et al., 2015) to select a sparse set of SNPs with non-zero effects on gene splicing and estimate their effects in the linear regression: $y_i = X_i\beta + E$. They assume that $\beta \propto exp(\lambda_1||\beta||_1 + \lambda_2||\beta_2||)$ where $||.||_1$ and $||.||_2$ denote the L_1 and L_2 norms, respectively. The penalty λ_1 and λ_2 is selected via cross validation. A group-lasso penalty on the effect size of one SNP across all isoforms is also used in UTMOST to integrate information across multiple dimensions. To combine gene-trait associations across multidimensional data, S-MultiXcan first uses PCA regulation and then performs χ^2 test, while UTMOST leverages GBJ test (Sun et al., 2019). For single dimension expression data analysis, we used S-PrediXcan (Barbeira et al., 2018).

2.5.5 Compilation of well-known trait associated gene lists

We obtained AD-associated genes from a previously curated list (Lin et al., 2021). The authors performed intensive hand-curation to identify confident AD-associated genes (positives) from various disease gene resources, including AlzGene, AlzBase, OMIM, DisGenet, DistiLD, and UniProt, Open Targets, GWAS Catalog, differentially expressed genes (DEGs) in ROSMAP, and published literature. We obtained LDL-C related genes from previous curated list (Zhou et al., 2020) that includes genes from literature and KEGG pathways. We obtained SCZ-associated genes via splicing from literature (Takata et al., 2017; Ma et al., 2020b; Cai et al., 2021; Glatt et al., 2009). The full lists are provided in supplementary materials.

2.6 Supplementary Materials

2.6.1 Illustration of W matrix

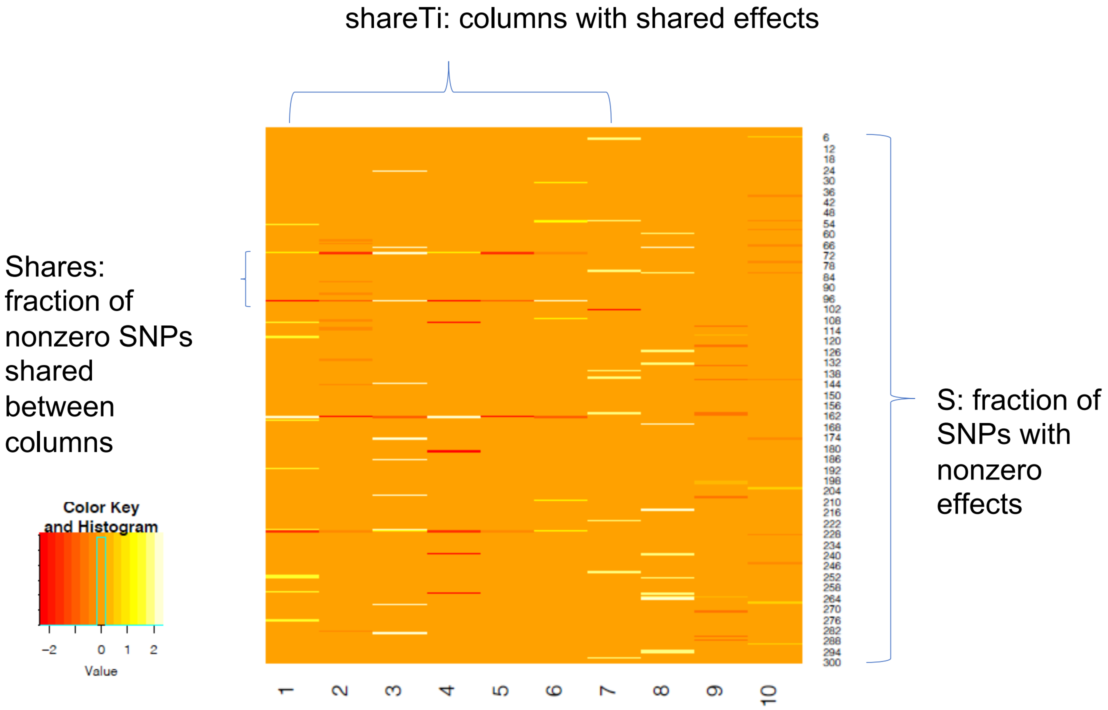


Figure 2.10: Design of weight matrix W in simulation

2.6.2 Effect of reference panel sample size on type I error for MSG

Table 2.4: Comparison of type I error for MSG using individual GWAS (MSG_{IND}), MSG with GWAS summary statistics and reference genome of 400 individuals (MSG_{REF400}), MSG with GWAS summary statistics and reference genome of 5000 individuals ($MSG_{REF5000}$) in simulation with different number of effect-sharing isoforms $(2,4,8)$, sparsity $(0.01,0.05,0.1)$ and splicing heritability $(0.01,0.05,0.1)$. The number of replication is 20000 for each scenario.

shareTi	s	h_c	mcca.ind	MSG (REF400)	MSG (REF5000)
2	0.01	0.01	0.048	0.059	0.049
	0.01	0.05	0.050	0.059	0.050
	0.01	0.10	0.047	0.058	0.049
	0.05	0.01	0.051	0.061	0.051
	0.05	0.05	0.049	0.059	0.051
	0.05	0.10	0.050	0.060	0.052
	0.10	0.01	0.012	0.015	0.013
	0.10	0.05	0.013	0.016	0.013
	0.10	0.10	0.012	0.015	0.013
4	0.01	0.01	0.050	0.060	0.050
	0.01	0.05	0.052	0.061	0.051
	0.01	0.10	0.046	0.057	0.048
	0.05	0.01	0.050	0.059	0.051
	0.05	0.05	0.048	0.057	0.047
	0.05	0.10	0.047	0.060	0.051
	0.10	0.01	0.011	0.015	0.012
	0.10	0.05	0.013	0.015	0.013
	0.10	0.10	0.013	0.016	0.013
8	0.01	0.01	0.051	0.060	0.051
	0.01	0.05	0.051	0.059	0.050
	0.01	0.10	0.050	0.060	0.049
	0.05	0.01	0.046	0.055	0.047
	0.05	0.05	0.049	0.060	0.050
	0.05	0.10	0.050	0.060	0.050
	0.10	0.01	0.012	0.014	0.012
	0.10	0.05	0.012	0.015	0.012
	0.10	0.10	0.012	0.015	0.013

2.6.3 Compilation of well-known trait associated gene lists

2.6.3.1 Compilation of AD-associated genes

We obtained AD-associated genes from published literature (Lin et al., 2021). The authors performed intensive hand-curation to identify confident AD-associated genes (positives) from various disease gene resources, including AlzGene, AlzBase, OMIM, DisGenet, DistiLD, and UniProt, Open Targets, GWAS Catalog, differentially expressed genes (DEGs) in ROSMAP9, and published literature. The genes include: APOE, SORL1, GAB2, CR1, PICALM, CLU, CD33, ABCA7, ADAM10, CD2AP, BIN1, APOC1, TOMM40, INPP5D, PSEN2, EPHA1, APP, MTHFD1L, CNTNAP2, HLA-DRB1, CASS4, BCAM, ABCA1, PTK2B, MS4A6A, FRMD4A, BCL3, SLC24A4, GLIS3, FERMT2, PSEN1, TREM2, ZCWPW1, EXOC3L2, MS4A4A, ACE, APOC4, BZW2, SUCLG2, APOB, SCIMP, SCARB1, RELB, CRY2, PVRL2, CLASRP, ADAMTS4, MMP3, UBE2L3, PPP1R37, ECHDC3, TCF7L2, IL6R, MS4A2, LIPG, MAN2A1, MAPT, ALDH1A2, ABI3, LILRA5, CELF1, PLCG2, HMGCR, OARD1, APH1B, APOC2, OR4S1, STAT4, MS4A4E, PVR, MT-ND2, HS3ST1, CCR2, VASP, CYP8B1, BLOC1S3, PPP1R13L, NFIC, NKPD1, INSR, CNTNAP5, BCAS3, BCHE, BCL2, NME8, CLPTM1, CLNK, UBQLN1, CLMN, IL1B, TRAPPC6A, VSNL1, SORCS1, PPARG, IGSF23, CRH, PSMA1, CHRNB2, FBXL7, CHRNA7, SPON1, MYO16, CHRNA2, VLDLR, KIR3DL2, KIT, HLA-DRB5, BACE1, HLADRA, DSG2, CALHM1, RBFOX1, HFE, PILRA, LRP4, HARBI1, TFCP2, CBLC, DPP10, SYNJ1, CDC25B, ACP2, ACHE, PACSIN3, MADD, ZNF652, GSK3B, PFDN1, RIN3, MARK4, GRIN2A, PDGFRB, MAPK8IP1, GRIN3B, CCRL2, ECE1, SCN1A, HBEGF, CACNA1G, CEACAM16, MMP13, ESR1, ALDH5A1, PLAU, SCN8A, CACNA2D1, MMP12

2.6.3.2 Compilation of LDL-associated genes

We obtained well-known lipid metabolism genes from (Zhou et al., 2020). The authors collected genes from literature and KEGG pathways. The genes include: VAPB, APOE, SOAT1, LRPAP1, ADH1B, NPC1, PPARG, ANGPTL3, PCSK9, CYP27A1, KPNB1, CETP, LPIN3, NCEH1, TNKS, FADS3, LDLRAP1, OSBPL5, FADS1, VDAC1, PLTP, APOC2, LIPA, LPA, LIPC, SORT1, MYLIP, SCARB1, ABCB11, VDAC3, LRP2, APOB, APOH, TSPO, VAPA, LIPG, APOA4, APOC3, ALDH2, APOA1, NPC2, LRP1, LDLR, APOC1, STARD3, FADS2, CD36, ABCG5, ABCG8, STAR, APOA2, ABCA1, VDAC2, ANGPTL4, SOAT2, CYP7A1, IRF2BP2, LPL, LCAT

2.6.3.3 Compilation of SCZ-associated genes

Since there are lots of genes previously reported to be associated with SCZ, We focused on genes reported to be associated with SCZ via splicing from literature. The genes are: IRAK4,CYC1,CHI3L1, FLJ46321,ATXN3,DENND1A, S100A12, ARAF, BICD2, DLG3, NRG3, DISC1, KCNH2, GRM3, ZNF804A, ERBB4, DRD2, AS3MT, SNX19, ARL6IP4, APOPT1, CYP2D6.

CHAPTER 3

Leveraging gene-level prediction as informative covariate in hypothesis weighting improves power for gene-based rare variant association studies

3.1 Introduction

Rare variant association studies (RVASs) enable the identification of disease-associated genes with clear functional support (Liu et al., 2014). In RVASs, a large number of hypothesis tests are usually generated from scanning the human genome and one needs corrections to limit false positives while maximizing power. False discovery rate (FDR) (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003) control has become a popular approach for detecting weak effects by limiting the expected false discovery proportion (FDP). Of the FDR control procedures, the Benjamini Hochberg (BH) (Benjamini and Hochberg, 1995) procedure is one of the most commonly used. While BH is nearly optimal when all hypotheses are equally likely to be null (Zhang et al., 2019), it suffers from suboptimal power when tests are heterogeneous (Genovese et al., 2006), which is often the case in modern applications like RVASs.

Different from the BH procedure, hypothesis-weighting FDR control procedures have been proposed to incorporate prior information to up-weight and down-weight hypotheses (Roeder and Wasserman, 2009). The idea is that more FDR budget can be allocated to hypotheses with greater prior probability of being non-null, hence has the potential to increase detection power (Li et al., 2013; Zhang et al., 2019).

We reason that hypothesis weighting might help improve detection power in RVASs. Previous research has shown the effectiveness of hypothesis weighting in multiple genetic applications, like differential expression gene discovery (Ignatiadis et al., 2016); GWAS (Kichaev et al., 2019; Li et al., 2013; Andreassen et al., 2013; Yurko et al., 2020; Fortney et al., 2015), eQTL discovery (Ignatiadis et al., 2016; Zhang et al., 2019), and trait-associated epigenetic marks discovery (Huang et al., 2020). Recently, many hypothesis weighting methods have been developed, and a detailed comparison of these methods were reviewed elsewhere (Korthauer et al., 2019; Ignatiadis and Huber, 2017). Among these methods, independent hypothesis weighting (IHW) (Ignatiadis et al., 2016) have been recommended due to its powerful, robust, and computational efficient nature (Huang et al., 2020; Korthauer et al., 2019). In addition, in IHW framework, the relationship between covariates and p-values is assumed to be not fully known and can be learned as a function of the covariates from data without overfitting. This enables us to harness prior information that does not precisely specify but are informative of the likelihood of hypotheses being non-null. Leveraging these desired properties, we hypothesize that there might be opportunities to derive gene-level scores reflecting the prior likelihood of genes' association with traits and using them as covariates in IHW framework to increase detection power in RVASs.

Genome-wide association studies (GWASs) provide opportunities for us to derive gene-level scores to facilitate RVASs discovery, as recent findings suggest the convergence of trait-associated genes from common and rare variants (Singh et al., 2020). To map SNP-level GWAS signals to gene-level probability of disease association. we first obtain a starting set of genes from methods like iRIGS (Wang et al., 2019) that can classify genes near GWAS hits to high-confidence genes (HRG) and local background genes (LBG). However, iRIGS and other similar methods do not provide genome-wide prediction of risk genes. We tackle this by leveraging the rich collection of gene-level annotations to identify patterns that are shared among the starting set of genes inferred by iRIGS (i.e. HRG, LBG), and leverage those patterns to assign scores to all genes based on the similarity of profiles to these genes.

Here, we propose a three-stage pipeline to improve the power to identify risk genes from RVASs. First, we identify a small set of training genes by applying iRIGS to classify genes near GWAS significant loci to HRGs and LBGs. Second, we derive genome-wide probability of disease association from training set genes and relevant biological features using machine-learning models. Specifically, we frame this as a classification problem (i.e., classify genes into risk genes versus non-risk genes) to be solved by supervised machine-learning algorithms like random forest (Breiman, 2001). We validate the prediction scores by testing for increased burden of SNP-heritability and enrichment with gene lists repeatedly implicated in disease among top predicted genes. Finally, we use prediction scores as covariates to weight published *p*-values from RVASs through the IHW framework. To demonstrate this pipeline, we propose to detect genes associated with SCZ from *p*-values in a recent RVASs (Singh et al., 2020) using predictions informed from a recent GWAS (Ripke et al., 2014). As there is significant overlap of rare variant risk between SCZ and autism spectrum disorders (ASD) (Singh et al., 2020), we also propose to use the same predictions as covariates to adjust recent published ASD RVAS *p*-values (Satterstrom et al., 2020).

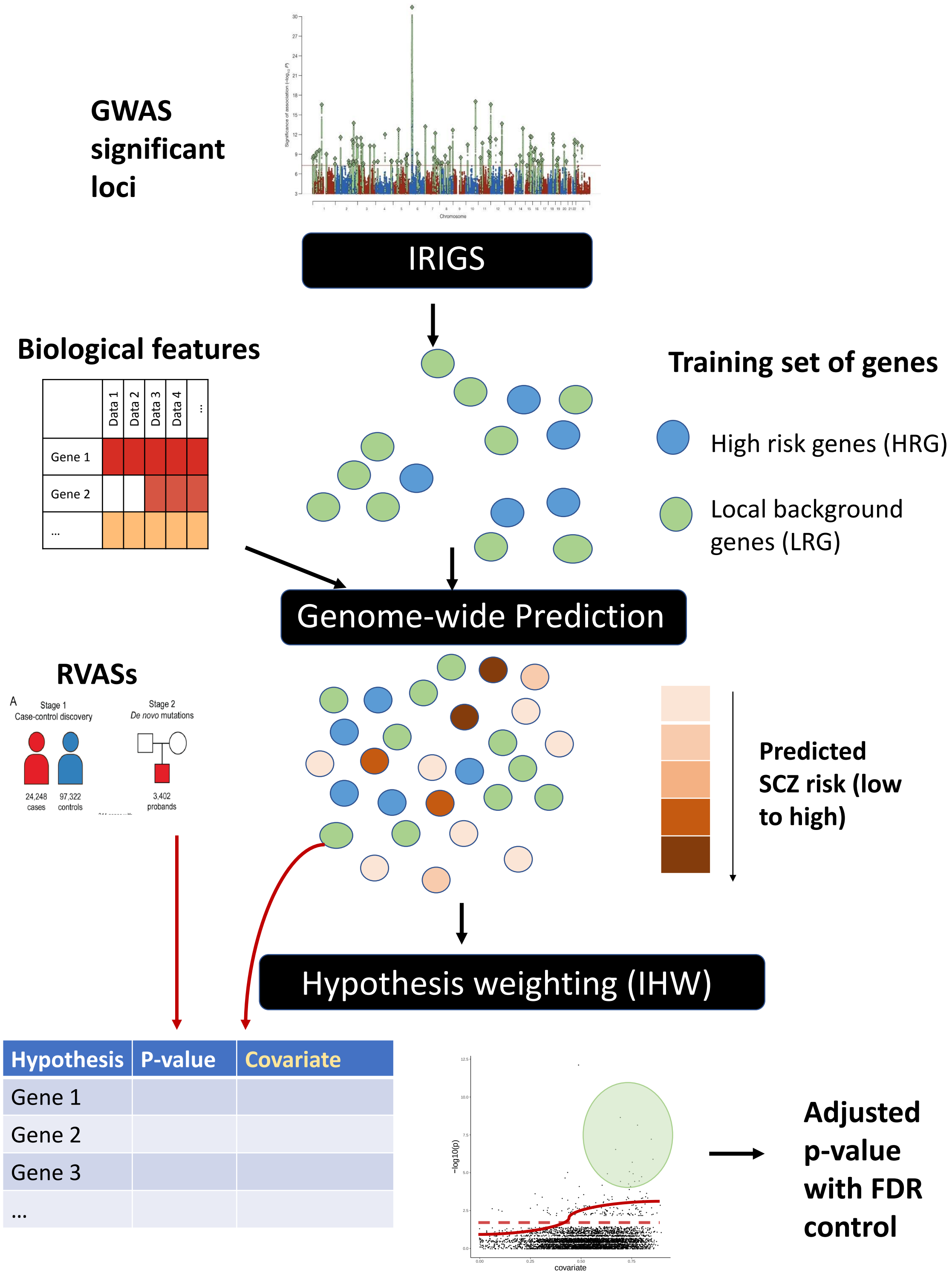
3.2 Methods

3.2.1 Method overview

As shown in Fig 3.1, our approach involves three stages. First, we obtained HRGs and LBGs near significant SCZ GWAS loci through probabilistically ranked genes based on their strength of genomic evidence and closeness in the network space via iRIGS. These two sets of genes served as positive and negative instances for the subsequent training. Second, we constructed features using selected biological annotations and predicted the SCZ association for all genes using the labeled training set of genes and selected features. Third, we use the prediction as the informative covariate in a published FDR-based method IHW to improve gene discovery from RVASs.

3.2.2 Obtain the training set of genes

To facilitate the supervised training, a training set of both ”SCZ-genes” and ”non SCZ-genes” are needed. We used iRIGS to obtain these genes. All genes within a 2 Mb region



Hypothesis	P-value	Covariate
Gene 1		
Gene 2		
Gene 3		
...		

Figure 3.1: Overview of the workflow

centered at the SCZ GWAS significant index SNPs are considered to be candidates (Ripke et al., 2014), the genes with highest iRIGS posterior probability (HRG) from all GWAS hits are used as ”positive” instances and the genes with posterior probability less than the median of all candidate genes (LBGs) are used as ”negative” instances.

3.2.3 Feature pre-processing

BRAINSPAN (Miller et al., 2014) is a dataset with RNA sequencing profiling of different cortical and subcortical structures across the full course of human brain development. The dataset includes 524 samples with developmental time points ranging from 5 post-conceptional weeks (pcw) to over 40 years of age from 26 brain structures. We used the genes in the dataset as instances (i.e., rows of the feature matrix), and their expression values measured in RPKM (reads per kilobase of exon model per million mapped reads) for the different developmental time points and brain structure as columns of the feature matrix for the training dataset.

DEPICT (Pers et al., 2015) provides a dataset with 14,461 ”reconstituted” gene sets with a membership probability for each gene in each gene set based on co-regulation of gene expression and previously annotated gene sets representing a wide spectrum of biological annotations. We used the genes in the dataset as instances (i.e., rows of the feature matrix), and their membership probability across the 14,461 reconstituted gene sets as columns as a feature matrix for the training dataset.

FANTOM5 (Andersson et al., 2014) project used cap analysis of gene expression (CAGE) technique to measure promoter utilization across 975 human samples. We used the genes in the dataset as instances, and their CAGE expression TPM values as columns of the feature matrix for the training dataset.

LAKE (Lake et al., 2018) includes nuclear transcriptomic data for more than 60,000 single cells from human adult visual cortex, frontal cortex, and cerebellum from six different individuals. It is a unique resource that enables us to observe expression signatures of different cell types and states to resolve the heterogeneity within tissues. To reduce the dimension of this dataset, we take an average of all expression in each cell type and state (with labels provided in the dataset), resulting in a matrix with 61 columns.

3.2.4 Model training and genome-wide prediction of SCZ risk

We seek to prioritize unlabelled genes with feature profiles similar to positive-labeled genes and different from negative-labeled genes using the random forest classifier. The input features for the random forest classifiers are: BRAINSPAN, DEPICT, FANTOM5, and LAKE. We obtained labeled genes as the training set from SCZ GWAS using iRIGS. As there are usually one HRG and multiple LBGs at each significant locus, the training set is highly imbalanced. To tackle this, we develop a workflow with 100 iterations, and we create a random balanced training set by down-sampling the negative-labeled genes in each iteration. In each iteration, we perform a 3-fold split on each balanced training set, and train on 2 folds, followed by prediction each time on the third fold (”test” fold). We evaluate the AUC of the test folds using predicted probabilities and true labels. We then make prediction of the whole dataset using the learned model and record the predicted score. The practice of learning from different random balanced training sets allow the prediction to be less prone to bias of a small set of genes and thus are more robust. Finally, after 100 iterations, we average the prediction probabilities assigned to each gene from all the iterations. We used the R package ”randomForest” (Liaw and Wiener, 2002) for the implementation and the only parameter we set is the number of trees (ntree=3000). We choose to not fine-tune other parameters to avoid over-fitting.

3.2.5 Application of IHW for hypothesis weighting

IHW is a general method with established type I error control and stability. Intuitively, individual tests may differ in their statistical properties and a covariate might provide information for such properties. For our case of gene-level rare variant association hypotheses, each gene may differ in their relevance to SCZ risk, and this risk can be indexed by gene-level covariates obtained by prediction in the previous step. Then, instead of using a flat p-value threshold in conventional methods, we can use an adaptive threshold informed by the covariate: allocate more FDR budget to the hypothesis with a certain covariate value.

To explain the methods, suppose we have m hypotheses to test based on p-values (p_1,...,p_m) with covariates X_1,...,X_m. Conventional BH-approach use this decision rule:

$$\text{Reject hypothesis i if } p_i \leq \hat{t}, \tag{3.1}$$

with cutoff $\hat{t}$ determined at a defined level using only p-values by a multiple testing procedure FWER control or FDR control, such as Bonferroni correction (Bonferroni, 1936) or BH (Benjamini and Hochberg, 1995) respectively, to protect against spurious discoveries.

Instead of using the conventional approach illustrated in Equation 3.1, we used "IHW" (Ignatiadis et al., 2016; Ignatiadis and Huber, 2017), a general and flexible hypothesis weighting approach unique in that it can learn weights from covariates and p-values without overfitting (i.e., losing type-I error control) using cross-weighting. In IHW, a decision rule is:

$$\text{Reject hypothesis i if } p_i \leq \hat{t}\widehat{W(X_i^{-l})} \text{ where } i \in I_l, \tag{3.2}$$

where I_l, $l = 1,..,k$ is a partition of the hypotheses into k folds to avoid overfitting. $\widehat{W(X_i^{-l})}$ are weight functions depending on covariates, with the weight function used for fold l being learned from p-values and covariates X fron the $k-1$ folds excluding the fold l. Compare equation 3.2 and equation 3.1, it is equivalent of using weighted p-values ($p_i/\widehat{W(X_i^{-l})}$) instead of p-values (p_i) in multiple hypothesis testing. The genes with large weights yield smaller weighted p-values, and the associated genes are more likely to be declared significant. Here, IHW splits the hypotheses into different strata (selected using the default mode "auto") based on increasing value of the predicted gene-level risk score. Within each stratum, IHW randomly split them into folds. IHW learns the weights for each stratum and fold combination to achieve the highest number of discoveries. Details of IHW can be found in published papers (Ignatiadis et al., 2016; Ignatiadis and Huber, 2017).

3.2.6 The SCZ RVAS data

Association *p*-values for SCZ RVAS were obtained from the Schizophrenia Exome Sequencing Meta-analysis (SCHEMA) consortium website https://schema.broadinstitute.org/results. The data contains the meta-analysis of whole-exomes from 24,248 cases and 97,322 controls from diverse global populations. Three classes of variants are included in the meta-analysis: PTVs (defined as stop-gained, frameshift, essential splice donor and acceptor variants), missense variants with MPC pathogenicity score ¿ 3, and MPC 2-3. PTVs and MPC ¿ 3 variants (class I) were analyzed by a burden test to generate gene-level *p*-values; genes with MPC 2-3 variants were aggregated and combined with class I *p*-values using a weighted Z-score method, please refer to Singh et al. (2020) for details. We extracted the meta-analysis *p*-values (column ”P_meta” in the online table) for the analysis.

3.2.7 The ASD RVAS data

Association test results for ASD RVAS were obtained from Table S2 of Satterstrom et al. (2020). FDR qvalues are transformed to *p*-values for analysis (code in https://github.com/yingji15/SCZIHW_public/). The data contains the largest exome sequencing study of ASD to date (n = 35,584 total samples, 11,986 with ASD). Two categories of rare variation, namely protein-truncating variants (PTVs; i.e., frameshift, stop gained, canonical splice site disruption) and “probably damaging” missense variants according to PolyPhen-2 (Mis3) (Adzhubei et al., 2010), in the context of three categories of inheritance pattern: de novo, inherited, and case-control are included.

3.3 Results

3.3.1 Evaluation of prediction scores

As shown in Fig 3.2, we first evaluated our prediction models using cross-validation and the model achieved an average area under the receiver–operator curve (AUC) of 0.74, 0.86, 0.87, and 0.89 from BRAINSPAN, DEPICT, LAKE, FANTOM5 respectively from the hold-out sets among all iterations. Among the selected features, DEPICT, LAKE and FANTOM5 showed comparable performance in terms of AUC, while BRAINSPAN based prediction showed lower AUC compared to the other three. The AUC values from all features are much higher than 0.5, suggesting they all contain informative signals about SCZ risk.

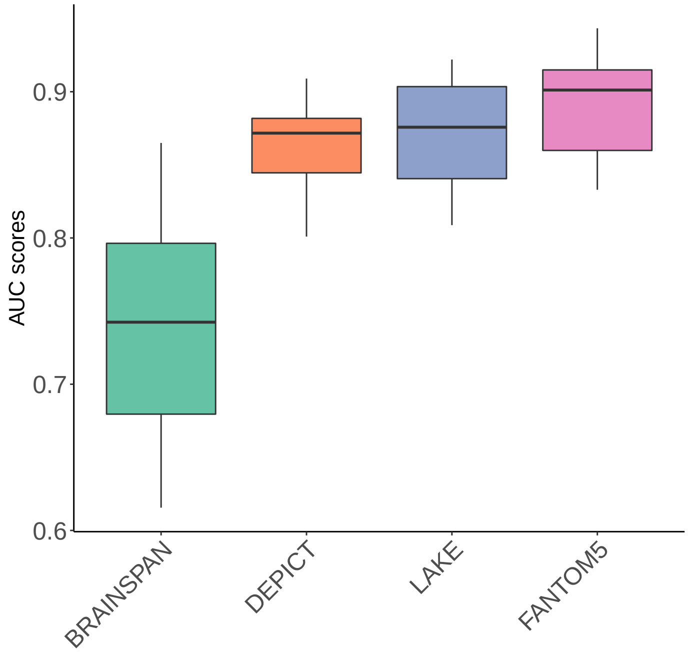


Figure 3.2: Distribution of AUC scores across the predictions from different features.

Since different features may characterize SCZ risk genes from different angles, we generated an ”ensemble score” as the final gene prediction by averaging the scaled predictions from all features. Then we performed a systematic empirical evaluation based on the enrichment of SNP-based heritability by stratified LD score regression (LDSC) (Bulik-Sullivan et al., 2015; Finucane et al., 2015) according to the rankings of the ensemble score. As shown in Fig 3.3, we found that the top ranked genes are significantly enriched for SNP-based heritability through applying LDSC on a most recent SCZ GWAS (Ripke et al., 2020). While there is a small ”bump” at around genes rank 10,000, the general trend of more pronounced heritability enrichment is observed for more prioritized genes.

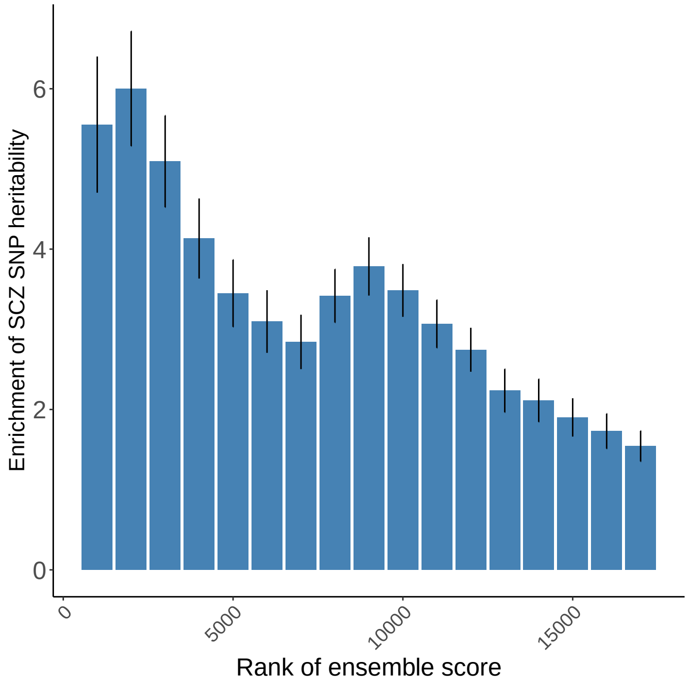


Figure 3.3: Enrichment of schizophrenia-SNP heritability with the number of ranked genes (calculated using LDSC). The most recent SCZ GWAS published in 2020 (Satterstrom et al., 2020) was used in the analysis.

We further evaluated the ensemble score based gene ranking using enrichment analyses with gene lists repeatedly implicated in SCZ (Wang et al., 2019). As shown in Table 3.1, we evaluated the top 1000 predicted genes by ensemble score versus the rest of the genome for enrichment using one-sided Fisher's exact test. We found strong enrichment in the target genes of FMRP ($p = 6.10 \times 10^{-249}$), which is an RNA-binding protein that regulates translation and needed at synapses for glutamate receptor signaling and neurogenesis (Purcell et al., 2014; Callan and Zarnescu, 2011). We also found top predicted genes to be significantly enriched in synaptic genes, including postsynaptic density (PSD, $p = 5.82 \times 10^{-126}$), protein cytoskeleton-associated scaffold protein (ARC, $p = 2.19 \times 10^{-8}$), NMDAR network ($p = 3.54 \times 10^{-24}$), mGluR5 ($p = 2 \times 10^{-5}$). We also observed significant enrichment in RFBOX1 ($p = 2.26 \times 10^{-140}$) and miR-137 targets ($p = 2.19 \times 10^{-22}$). A detailed description of the gene lists is in Table 3.4.

Table 3.1: Enrichment of top 1000 predicted genes in gene sets implicated in SCZ.

Gene set[a]	OR[b]	p-value[c]
FMRP-Darnel (832)	14.23	6.10×10^{-249}
RBFOX1 (556)	11.14	2.2×10^{-140}
PSD (1444)	5.15	5.82×10^{-126}
ECG (998)	5.38	2.28×10^{-99}
PRP (336)	5.06	4.53×10^{-34}
PRAZ (209)	5.87	6.90×10^{-27}
NMDAR (59)	18.17	3.54×10^{-24}
miR-137 targets (281)	4.32	2.19×10^{-22}
GABA (18)	46.06	1.07×10^{-11}
SYV (107)	4.32	7.38×10^{-09}
ARC (25)	13.82	2.19×10^{-08}
CRF (56)	5.54	6.81×10^{-07}
mGluR5 (37)	6.28	2.00×10^{-05}
CCS (73)	3.72	1.06×10^{-04}

[a] The numbers of genes in the corresponding gene sets are in parentheses. The source and short description of these gene sets are included in Table S1.

[b] Odds ratio from one-sided Fisher's exact test

[c] *p*-value from one-sided Fisher's exact test after Bonferroni correction.

3.3.2 Leverage prediction as covariates to identify SCZ risk genes

Having evaluated our predicted scores using different evidence, we sought to examine the utility of leveraging the predictions for the identification of risk genes from RVAS results. Here, we extracted the published association *p*-values from SCZ RVAS (Singh et al., 2020) and investigated the ensemble scores as covariates to conduct hypothesis weighting in IHW.

As an exploratory analysis, we first checked whether the ensemble score as a covariate is informative about power under the alternative. We started with SCZ RVAS results by partitioning all hypotheses into three equally sized groups: ”low score” group with the ensemble score less than its 33% quantile, ”medium score” group with the ensemble score between its 33% and 67% quantile, and ”high score” group with the ensemble score larger than its 67% quantile. As shown in Fig 3.4, we observe a successive increase of hypotheses with *p*-values near zero for increasing scores indicating that the proportion of non-null effects varies across different groups.

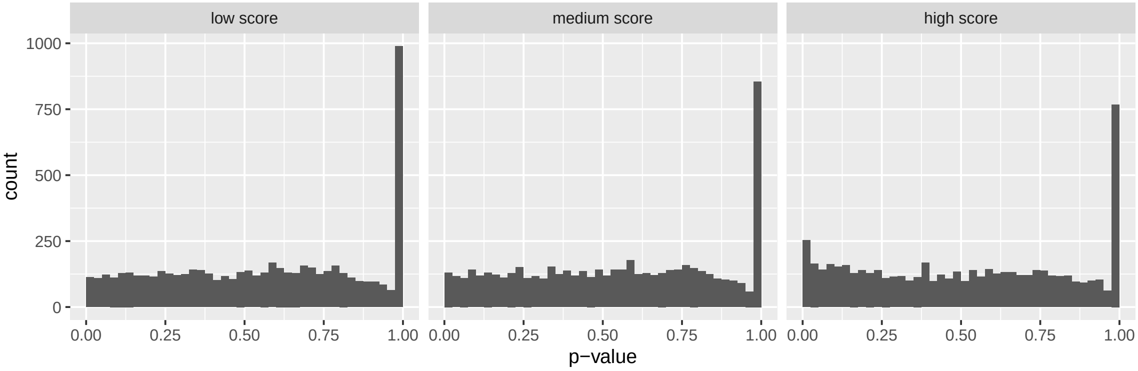


Figure 3.4: Histograms of SCHEMA *p*-values after splitting the hypotheses into three groups by the prediction score.

Since the ensemble scores are informative of the prior probability of each individual test, to maximize power for discovery, all gene-level tests should not be treated exchangeably. Thus, we used the ensemble scores as covariates to adjust RVAS gene-level *p*-values under different target FDR levels (α=0.05, 0.1, 0.2, 0.3) using IHW. The range of α is chosen to reflect the FDR control level commonly used in practice. We also included the prediction scores from individual feature sets (i.e., BRAINSPAN, DEPICT, LAKE, FAN-

TOM5) as covariates for comparison purpose. As shown in Table 3.2, when using the ensemble score as the covariate to adjust SCZ p-values, although we did not find an increase of significant genes when $\alpha = 0.05$ potentially due to insufficient power, we did observe 22%, 28% and 109% increase of significant genes for higher target FDR levels at $\alpha = 0.1, 0.2, 0.3$ respectively. For the single feature based scores, we observe more improvement from DEPICT and LAKE data, and less improvement from BRAINSPAN and FANTOM5.

One might doubt whether the power gain is just by chance. To check this, we randomly shuffled the ensemble score (i.e., ”IHW - shuf ensemble” in Table), and used the shuffled score as the covariate for adjustment. The number of rejections is similar to that using BH approach, providing evidence that a mis-specified covariate would not cause much power increase or decrease. This result is line with the previous findings that hypothesis weighting can lead to power improvements with informative weights and cause little power lost with uninformative weights (Roeder and Wasserman, 2009; Roeder et al., 2007; Andersson et al., 2014).

Table 3.2: Number of discoveries from SCZ dataset, by different methods and covariates[a]

method[b]	0.05[c]	0.1	0.2	0.3
IHW - BRAINSPAN	29	34	51	63
IHW - FANTOM5	30	37	54	64
IHW - DEPICT	30	36	57	93
IHW - LAKE	30	35	55	97
IHW - ensemble	30	38	59	134
IHW - shuf ensemble	31	31	48	57
BH	31	33	46	64

[a] at a range of target FDR levels α from 0.05 to 0.3.

[b] Methods used for p-value adjustment.

[c] $\alpha = 0.05$.

Using the expanded set of significant genes identified, we next sought biological insights. We applied gene ontology (GO) enrichment analysis to the 84 genes that are insignificant using BH-adjustment but significant after IHW-adjustment at FDR level $\alpha = 0.3$. As shown in Fig 3.5, we discovered an enrichment of biological processes like synapse assembly (OR=5.90), neuron projection guidance (OR=5.84) , consistent with previous knowledge on SCZ (Egbujo et al., 2016). We also evaluated the 64 genes that are significant under the conventional BH-adjustment at $\alpha = 0.3$ for GO term enrichment and observed none of the GO terms are significant at FDR=0.1 level. The lack of GO enrichment in BH-adjustment identified genes might be caused by a lack of power or specificity of the original results.

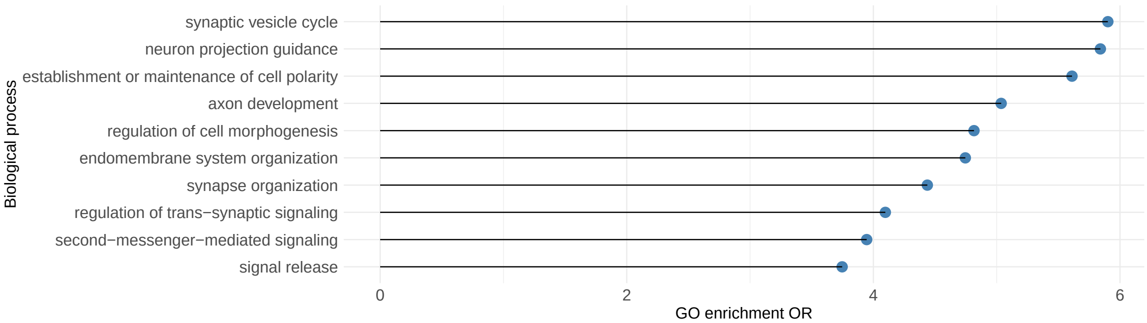


Figure 3.5: SCZ gene ontology(GO) enrichment analysis results for top biological processes with FDR≤ 0.1.

Then we further investigated genes not significant using BH-adjustment but ”boosted” to significance after adjustment using the ensemble score in IHW (referred to as IHW-adjustment). The FDR level $\alpha = 0.1$ is chosen since that’s the more stringent level at which the adjustment leads to improvements. Since the RVAS study we used is comprehensive and included the most available RVAS studies of moderate size, we were not able to conduct replication studies. Instead, we looked for literature support for genes ”boosted”. CACNA2D1 is one example, not significant under traditional BH-adjustment ($p = 0.23$) but significant after IHW-adjustment ($p = 0.065$). A deletion in CACNA2D1 have been observed in one Japanese SCZ patient from a independent study (Malhotra and Sebat, 2012). There are also support for CACNA2D1 in other psychiatric disorders that are correlated with SCZ like epilepsy and intellectual disability (Vergult et al., 2015), it has been identified as a potential drug target in MDD from GWAS (Gaspar et al., 2019). Another example is FABP7, not significant under traditional BH-adjustment ($p = 0.21$) but significant after IHW-adjustment ($p = 0.065$) respectively. There were non-synonymous polymorphisms identified from SCZ and ASD in FABP7 (Shimamoto et al., 2014). FYN not significant under traditional BH-adjustment ($p = 0.17$) but significant after IHW-adjustment ($p = 0.045$). Previous study has identified an excess of disruptive and damaging variants in FYN among SCZ patients (Tsavou and Curtis, 2019). For all these genes, while previous studies found the numbers of variants involved are too small to draw firm conclusions, adding an ensemble score as covariates provides extra confidence that these genes might increase SCZ risk.

3.3.3 Leverage prediction as covariates to identify ASD risk genes

As another application, we sought to evaluate whether our ensemble score can serve as covariates for the detection of risk genes from RVAS in ASD. While the prediction scores are for SCZ, multiple lines of evidence have suggested that SCZ and ASD partially share underlying genetic mechanisms: SCZ and ASD are genetically correlated (Lee et al., 2013); up to 30% of individuals diagnosed with ASD during childhood will develop SCZ in adulthood (Burbach and van der Zwaag, 2009); CNVs and rare alleles show overlap between ASD and SCZ in synaptic related genes (Walsh et al., 2008; Szatmari et al., 2007). Thus, we used the same SCZ-risk scores as covariates to adjust p-values from the largest RVAS of ASD (Satterstrom et al., 2020) using IHW. Exploratory plots in Fig 3.6 suggested that our ensemble score is also informative in stratifying ASD test results.

As shown in Table 3.3, when using the ensemble score as covariate, we observe 47%, 77%, 125% and 230% increase of significant genes for $\alpha = 0.05, 0.1, 0.2, 0.3$ respectively, showing a similar trend of increased association detection after adjustment as in SCZ. For the single feature based scores, we observe more improvement from BRAINSPAN, DEPICT and LAKE data, and less improvement from FANTOM5. These findings are consistent with the overlapping genetic basis in SCZ and ASD.

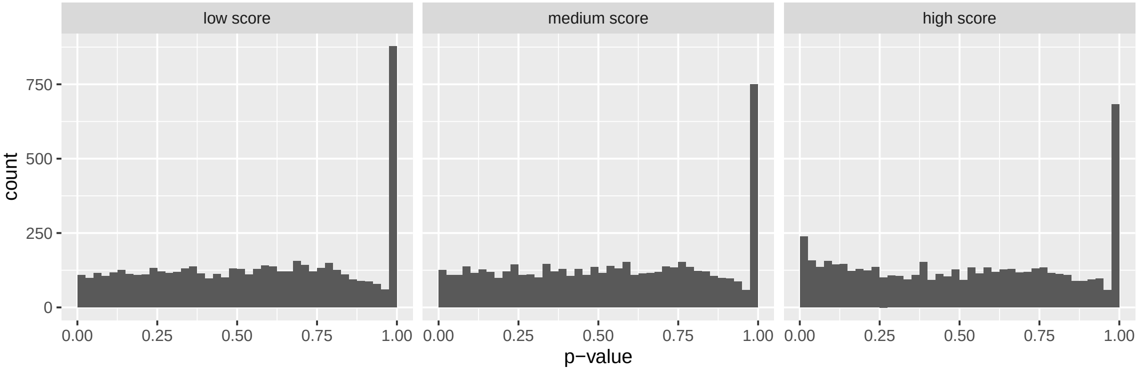


Figure 3.6: Histograms of ASD p-values after splitting the hypotheses into three groups by the prediction score.

Table 3.3: Number of discoveries from ASD dataset, by different methods and covariates[a]

method[b]	0.05[c]	0.1	0.2	0.3
IHW - BRAINSPAN	89	142	294	477
IHW - FANTOM5	98	119	207	363
IHW - DEPICT	107	158	329	495
IHW - single_LAKE	105	149	287	439
IHW - avg scaled	112	176	323	658
IHW - shuffled avg	79	95	151	201
BH	76	99	143	199

[a] At a range of target FDR levels α from 0.05to 0.3.
[b] Methods used forp-value adjustment.
[c] $\alpha = 0.05$.

Similar to previous analysis on SCZ, using the expanded set of significant genes identified, we next sought biological insights. We applied gene ontology enrichment analysis to the 488 genes that are insignificant using BH-adjustment but significant after IHW-adjustment at FDR level $\alpha = 0.3$. As shown in Fig 3.7, we discovered the enrichment of biological processes like cell part morphogenesis (OR=3.56), neuron projection development (OR=3.29), neuron differentiation (OR=2.86), consistent with previous knowledge on ASD (Gilbert and Man, 2017).

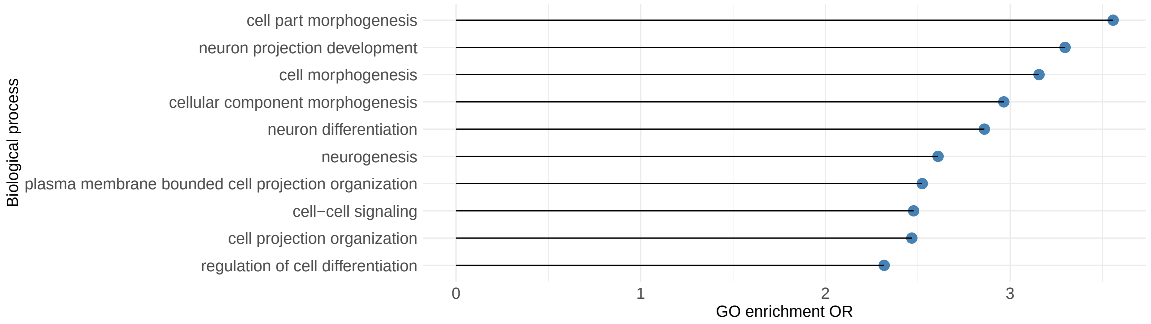


Figure 3.7: ASD gene ontology(GO) enrichement analysis results for top 10 biological processes with FDR≤ 0.05.

Then we further investigated genes not significant using BH-adjustment but ”boosted” to significance after adjustment using the ensemble score in IHW (referred to as IHW-adjustment). The FDR level $\alpha = 0.1$ is chosen and 84 genes are ”boosted” by IHW-adjustment. Since the RVAS study we used is comprehensive and included the most available RVAS studies of moderate size, we were not able to conduct replication studies. Instead, we looked for literature support for genes ”boosted”. COBL is not significant under traditional BH-adjustment (p= 0.37) but significant after IHW-adjustment (p=0.095). Previous studies have shown deletions of COBL cause defects in neuronal cytoskeleton morphogenesis in model vertebrates (Ahuja et al., 2007). It has also been supported by case-unique CNVs in autism case-control studies (Griswold et al., 2012). GABRA1 is not significant under traditional BH-adjustment (p= 0.34) but significant after IHW-adjustment (p=0.086). Previous studies have found significant reductions of GABRA1 expression in several brain regions of subjects with ASD (Fatemi et al., 2009).

3.4 Discussion

In this study, we explored the use of IHW in analyzing RVAS results with gene-level predicted scores as covariates, and investigated the biology of SCZ and ASD in the process. The covariates were the predicted gene-level susceptibility to SCZ obtained through supervised learning using biological features BRAINSPAN, FANTOM5, DEPICT and LAKE as inputs. An ensemble score which is the average of all single-feature-based predictions is also derived to capture support from all features. Applications to SCZ and ASD gene-level RVASs p-values using the ensemble score lead to more significant genes than using any single feature, suggesting the benefits of integrating diverse biological evidence. This is consistent with previous findings that integrating multiomics covariates improves power in identifying SNPs from GWAS analysis and eGenes from eQTL analysis (Yurko et al., 2020). When using the ensemble score as covariate, we observed 22%, 28% and 109% increase of significant genes for target FDR levels at $\alpha = 0.1, 0.2, 0.3$ respectively for SCZ RVAS analysis; 47%, 77%, 125% and 230% increase of significant genes for $\alpha = 0.05, 0.1, 0.2, 0.3$ respectively for ASD RVAS analysis.

Previous studies have shown the hypothesis weighting adjustment mostly has an impact on the genes with ”borderline significance”. Genes with very small p-values already have high power, genes with very large p-values have extremely low power and benefit little by weighting. Therefore, the weighting approach is most useful for genes with a marginal effect (Roeder and Wasserman, 2009). Here for SCZ, we observe more improvement when FDR > 0.1, the reason of which might be that there are more genes at the borderline when FDR> 0.1 yet very few genes are at the borderline when FDR=0.05. On the other hand, for ASD, we observe improvements across different FDR levels, suggesting there are more borderline genes at each FDR level. This might come from the larger power from rare-variant gene-level tests in ASD.

There are a few limitations to our pipeline. First, the training set in the prediction scoring process is from genes inferred by iRIGS near GWAS hits, there might be false positives and false negatives in this set. Therefore, the candidate genes we identified still require thorough experimentation. Second, as with most supervised learning methods, our pipeline depends on existing patterns of labelled genes and are less powerful to identify disease genes with unexpected characteristics.

Opportunities for future expansion of this strategy include exploring more features to include and applying better approaches to integrate signals from multiple features. Currently, we chose features from gene expression and biological processes in the prediction. There are other data resources that could potentially be included as features, such as proteomics, epigenomics. As IHW takes a single covariate, we took an average of the single feature based predictions to derive an ensemble score for hypothesis weighting. We explored other methods that could include multiple dimensions of covariates like AdaFDR (Zhang et al., 2019) and AdaPT (Lei and Fithian, 2016). However, our application of AdaFDR did not provide improvements in the genes identified and tend to be less stable; AdaPT takes many iterations of optimization and is computationally expensive as it uses a p-value masking procedure. Therefore, we chose IHW in this analysis. There might be room for further improvement in the way of integrating multiple covariates, which are worthy of future explorations.

3.5 Conclusions

In this paper, we present a three-stage pipeline to identify risk genes from both GWASs and RVASs: we first obtain training genes close to GWAS significant loci via iRIGS, then build machine-learning prediction models to predict each gene's probability to associate with SCZ using training genes and biological features; finally we use the prediction scores as informative covariates for hypothesis weighting to improve gene detection power from RVASs using IHW. We applied the pipeline to SCZ and ASD RVASs and observed sizeable improvements on the number of genes discovered. As an increasing volume of contextual information are being generated, we believe that our approach that leverages prediction as covariates in hypothesis weighting provides a valuable contribution to boost statistical significance in RVASs. This approach requires little investment and can be easily applied to the analysis of existing and future studies beyond RVASs.

3.6 Availability of data and materials

The original data used is documented on https://github.com/yingji15/SCZIHW_public The processed datasets during the current study will be available in Zenodo. The code for performing these analysis is freely available at https://github.com/yingji15/SCZIHW_public

3.7 Abbreviations

RVAS: rare variant association study

GWAS: genome-wide associations study

SCZ: schizophrenia

ASD: autism spectrum disorder

FDR: false discovery rate

AUC: area under the receiver-operating characteristic (ROC) curve

3.8 Supplementary Materials

Table 3.4: Gene sets implicated in SCZ

Gene set	Short description
FMRP-Darnel (Darnell et al., 2011)	Fragile X mental retardation (FMRP) protein targets
RBFOX1 (Weyn-Vanhentenryck et al., 2014)	targets of RNA binding protein, fox-1 homolog 1
PSD (Bayés et al., 2011)	post synaptic genes
ECG (Samocha et al., 2014)	evolutionary constrained genes
PRP (Pirooznia et al., 2012)	genes related to presynaptic proteins
PRAZ (Pirooznia et al., 2012)	genes in the presynaptic active zone
NMDAR (Purcell et al., 2014)	components of the N-methyl-D-aspartate (NMDA) network
miR-137 targets (Ripke et al., 2011)	miRNA-137 targets
GABA (Pocklington et al., 2015)	components of the GABA receptor complex
SYV (Pirooznia et al., 2012)	synaptic vesicles
ARC	neuronal activity-regulated cytoskeleton-associated proteins
CRF (Shipra et al., 2006)	chromatin remodeling factors
mGluR5 (Walsh et al., 2008)	components of the metabotropic glutamate receptor 5 complex
CCS (Müller et al., 2010)	calcium channel and signaling genes

CHAPTER 4

Incorporating European GWAS findings improves polygenic risk prediction accuracy of breast cancer among East Asians [1]

4.1 Introduction

Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer death in females worldwide (Torre et al., 2015). Although the rate of getting breast cancer has stabilized in some high-income countries, it continues to rise in most Asian and other low and middle-income countries, stressing the need for establishing early risk prediction and management strategies (Denny et al., 2017). Genetic risk factors play an important role in breast cancer predisposition (Nathanson et al., 2001). Large scale genome-wide association studies (GWASs) have identified more than 200 loci to be associated with risk of breast cancer (Cai et al., 2014; Michailidou et al., 2015, 2017; Shu et al., 2020; Zheng et al., 2009, 2013). Polygenic risk score (PRS), a weighted aggregation of risk allele counts derived from GWASs, is emerging as a useful tool for breast cancer risk stratification in multiple populations, including Europeans (EURs) and East Asians (EASs) (Khera et al., 2018; Mavaddat et al., 2019; Wen et al., 2016).

The small sample size in GWASs of non-EUR samples and the differences of genetic architecture between EUR and other populations make it challenging to develop powerful and well-calibrated PRSs across diverse populations. To date, large-scale breast cancer GWASs were conducted by the Breast Cancer Association Consortium (BCAC), with summary statistics publicly available for 123,000 (Michailidou et al., 2015) and 220,000 EUR samples (Michailidou et al., 2017). The largest non-EUR GWAS was conducted by the Asia Breast Cancer Consortium (ABCC), which included more than 40,000 EAS samples (Cai et al., 2014; Shu et al., 2020; Zheng et al., 2009, 2013). From study of 17 anthropometric and blood-panel traits, applying PRSs derived from EUR GWASs directly to non-EUR

[1] This chapter has been previously published in Genetic Epidemiology (Ji et al., 2021)

samples showed poor transferability in general, with approximately 37%, 50%, and 78% lower prediction R2 in South Asians, EASs, and Africans, respectively, compared to that in EUR populations (Martin et al., 2019). Therefore, developing polygenic risk prediction models for diverse populations is imperative to translate the GWAS findings to clinical use. This calls for efforts to improve the performance of PRS in non-EUR samples to mitigate the racial disparity.

Given that non-EUR GWASs are usually of insufficient sample size, and that there is extensive genetic sharing across populations (Consortium et al., 2015), there are several recent studies that sought to incorporate a large EUR GWAS and a smaller non-EUR GWAS to improve risk prediction in non-EUR populations. Specifically, Coram et al. (2015) proposed a cross-population best linear unbiased prediction method based on multi-component linear mixed models, where SNPs were placed in classes defined by GWAS evidence from different ancestries and allelic effects computed in a population-specific fashion (Coram et al., 2015, 2017); this method requires individual-level training data in the target population. When only GWAS summary statistics are available, Márquez-Luna et al. (2017) constructed a trans-ethnic PRS from a weighted linear combination of PRSs from two populations (Márquez-Luna et al., 2017), which could improve the prediction accuracy for type II diabetes in Hispanic/Latinos and South Asians. Grinde et al. (2019) found that this approach did not perform well for several anthropometric, blood count, and blood pressure traits in their Hispanic/Latino cohorts (Grinde et al., 2019). Instead, they proposed to construct PRSs by selecting single-nucleotide polymorphisms (SNPs) and their corresponding weights based on different combinations of EUR GWASs, Hispanic/Latino GWASs, or meta-analyses of both. They found that PRSs using an EUR GWAS for SNP selection and a Hispanic/Latino GWAS or meta-analysis for SNP weights performed the best in their empirical studies. These findings suggested that PRS performances could differ by population- and disease-specific genetic architectures. For breast cancer, EUR-based PRSs were reported to perform equally well in Hispanic/Latinos as in EURs but poorly in

African Americans (Allman et al., 2015). The performance of EUR- and EAS-based PRSs in EASs remains unclear. This motivates us to evaluate breast cancer PRS predictions in EASs and develop new strategies to construct PRSs targeting non-EUR populations.

Differential linkage disequilibrium (LD) and minor allele frequency (MAF) are major contributors to the poor transferability of PRSs among populations (Wang et al., 2020b). In a PRS of the form $\sum w_j X_j$, where X_j and w_j stand for the standardized genotype and weight for SNP j, respectively, the ideal weight satisfies $w_j^2 = h_j^2$, where h_j^2 is the disease heritability directly contributed by SNP j (Speed and Balding, 2019). Because the disease heritability contributed by a SNP varies according to local LD (Gazal et al., 2017; Speed et al., 2017), explicit incorporation of EUR LD information can improve prediction accuracy for EURs (Hu et al., 2017; Marquez-Luna et al., 2020; Vilhjálmsson et al., 2015). In addition, SNPs can serve as better proxies for the underlying “true effects” in populations in which they have high LD scores compared to populations in which they have low LD scores, where the LD score of a SNP is the sum of LD r^2 measured between this SNP and all other SNPs (Bulik-Sullivan et al., 2015). MAF has been used as an indication of the strength of natural selection, thus the differences in strengths of selection between ancestries might have an impact on PRSs (Wang et al., 2020b). This motivates us to examine whether modeling the LD and MAF differences between EURs and EASs could improve effect size estimation and genetic risk prediction in EAS populations.

In this paper, we are primarily interested in 1) evaluating the transferability of EUR GWAS data to breast cancer risk prediction in EASs, 2) improving risk predication for breast cancer in EASs, and 3) exploring the effects of LD and MAF differences between EUR and EAS ethnicities in PRS construction. We propose a rescaled meta-analysis framework that upweights EAS signals over EUR signals, yielding effect size estimates closer to the true effect sizes in EASs while taking advantage of the large sample sizes of EUR GWASs. We constructed PRSs using summary statistics from the rescaled meta-analysis of EUR and EAS GWAS data and then evaluated their performances in an independent EAS validation dataset. Our PRS outperforms PRSs derived from the EUR or EAS GWAS alone as well as the conventional meta-analysis of EAS and EUR GWASs. The EUR and EAS GWASs used in the analysis are from the BCAC and ABCC, respectively.

4.2 Materials and methods

4.2.1 GWAS training data in samples of European ancestry

Two large, publicly available summary statistics datasets based on European ancestry were used in this study. The data were from the Breast Cancer Association Consortium (BCAC) (details see S1 Table). EUR_2015 (Michailidou et al., 2015) includes two subsets, GWAS (N = 32,498) and COGS (N = 89,677). EUR_2017 (Michailidou et al., 2017) is the largest available GWAS study of breast cancer in European ancestry population. This study consisted of three subsets, GWAS, COGS, and OncoArray (N = 106,776).

4.2.2 GWAS training data in samples of Asian ancestry

The GWASs in samples of East Asian ancestry were from the Asian Breast Cancer Consortium (ABCC), which includes 14,958 breast cancer cases and 15,843 controls of Asian ancestry (Cai et al., 2014; Zheng et al., 2009, 2013). Samples were from studies conducted in mainland China, South Korea, Japan, Thailand, Malaysia, Singapore, Canada, U.S., Hong Kong, Taiwan and other countries and regions. Details are in S1 Table. We used the meta-analyzed summary statistics data from the study.

4.2.3 Validation data of East Asian ancestry

The validation set of EAS ancestry is from the Shanghai breast cancer genetic study, including 1794 cases and 2059 controls. Samples were all genotyped on MEGA chip (Illumina), with 80k custom Asian content selected to improve the coverage of low-frequency SNPs in Asian populations. Data were imputed using the Phase 3 release of the 1000 Genomes Project as reference.

4.2.4 Meta-analysis of EUR and EAS

Let $\beta_{EAS,j}$ and $V_{EAS,j}$ be the expectation and its variance of SNP j in EAS GWAS. The corresponding estimates in EUR are $\beta_{EUR,j}$ and $V_{EUR,j}$. In conventional meta-analysis, we have

$$\hat{\beta} = \frac{w_{inv\_EAS,j}\beta_{EAS,j} + w_{inv\_EUR,j}\beta_{EUR,j}}{w_{inv\_EAS,j} + w_{inv\_EUR,j}}$$

where $w_{inv\_EAS,j} = \frac{1}{V_{EAS,j}}$ and $w_{inv\_EUR,j} = \frac{1}{V_{EUR,j}}$ are the inverse-variance weights for EAS and EUR, respectively. As our goal is to obtain weights to construct PRS for EAS, we multiply a scaling factor (denoted as α, $\alpha > 1$) for EAS to the inverse variance weight to obtain a rescaled estimate

$$\hat{\beta} = \frac{\alpha \times w_{inv\_EAS,j}\beta_{EAS,j} + w_{inv\_EUR,j}\beta_{EUR,j}}{\alpha \times w_{inv\_EAS,j} + w_{inv\_EUR,j}}$$

with variance

$$\hat{V} = [\frac{\alpha \times w_{inv\_EAS,j}}{\alpha \times w_{inv\_EAS,j} + w_{inv\_EUR,j}}]^2 V_{EAS,j} + [\frac{w_{inv\_EUR,j}}{\alpha \times w_{inv\_EAS,j} + w_{inv\_EUR,j}}]^2 V_{EUR,j}$$

To select the scaling factor, we tried a grid of $\alpha = 1,2,3,4,5$ to derive the resulting summary statistics. Then we selected the scale factor to use along with p-value threshold that achieved a high prediction accuracy by cross-validation in the EAS validation set.

4.2.5 "P+T"

The P+T method refers to the calculation of PRS using informed LD-pruning and p-value thresholding. In this study, we use the implementation of the P+T method in the software package PRSice-2 (Euesden et al., 2015) with the default threshold of $r^2 = 0.2$ for clumping correlated markers and clumping windows of 250 kb. We examined varying strengths of LD among SNPs by repeating the procedures and changing threshold for clumping correlated markers ($r^2 = 0.1, 0.2, 0.4, 0.6, 0.8$) and the sizes of clumping windows (250 kb, 500 kb), and found results to be similar. For any pair of SNPs that have a physical distance smaller than the clumping window or r2 greater than the selected threshold, the less significant SNP is removed. PRS is computed by summing risk alleles weighted by effect sizes derived from input summary statistics. The p-value threshold are selected using validation data to optimize prediction accuracy. We constructed PRS for EAS using selected SNPs and effect size estimates $PRS = \sum_{j=1}^{J} w_{EAS,j} x_j$, which $w_{EAS,j}$ is the weight for the jth SNP.

4.2.6 LDpred

LDpred is a method that uses the GWAS summary statistics and LD information from the external LD reference sample to infer the posterior mean effect size of each SNP, conditioning on the SNP effect estimates of other correlated SNPs. This method assumes a point-normal prior on the distribution of SNP effects such that only a fraction of SNPs have non-zero estimated effects. These fractions of causal SNPs (denoted as f) were used in the validation set: 1 (i.e., all SNPs), 0.3, 0.1, 0.03, 0.01, 0.003, and 0.001, with an LD radius of 400 (i.e., $M/3000$, M is the number of SNPs, around 1.2 million Hapmap SNPs is included in the current analysis) to obtain local LD information, as suggested by the authors.

4.2.7 Assessment of PRSs

Our analysis used genotypes and phenotypes in independent validation data of East Asian ancestry from training GWAS. We reported area under the ROC curve (AUC) in a logistic regression model using the disease as outcome. When using a model with only PRS as the predictor, we report the in-sample fit using all individuals in the validation set. When using models with PRS, age and first 2 genotype PCs, we use the 10-fold cross-validation procedure. To compare AUCs from different training GWAS data, we conducted one-sided Delong's test for paired AUC curves using "roc.test" implemented in R package pROC.

We also include average Nagelkerke's pseudo R2 liability-scale pseudo R2 for the models (Lee et al., 2012) and the likelihood ratio test p-value. Nested models are considered to provide performance estimates of PRSs: the full model (PRS + covariates including age and first two PCs of genotype) and the reduced model (covariates only). Nagelkelke's pseudo R2 was calculated comparing the full model with the reduced model with the covariates alone, thus yielding an estimate of how well the variable (PRS here) explains the data. R packages "rcompanion" (see Web Resources) was used in the analysis. Since Nagelkelke's R2 suffers from bias when case/control proportion is different, we included liability-scale R2 that accounted for an ascertained case/control ratio (Lee et al., 2012).

To assess the relationship of PRS with breast cancer case/control status, individuals in the validation set were binned into 10 deciles according to the PRS, and the percentage of cases within each bin was determined. We calculated the odds ratio (OR) comparing top 10% of the individuals with the remaining 90% of the samples as the reference group, as well as OR comparing top 10% of individuals with individuals in the 40th-60th percentiles.

4.2.8 Data Availability Statement

Access to the ABCC data could be requested by submission of an inquiry to Dr. Wei Zheng (wei.zheng@vanderbilt.edu). Request of access to the BCAC data could be submitted directly to BCAC (http://bcac.ccge.medschl.cam.ac.uk/).

4.2.9 Web Resources

LD score: https://data.broadinstitute.org/alkesgroup/LDSCORE/

PRSice2 package (for P+T): https://www.prsice.info/

Rcompanion package (for pseudo-R2):

https://cran.rproject.org/web/packages/rcompanion/index.html

2015 EUR and 2017 EUR GWAS summary statistics: http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/gwas-icogs-and-oncoarray-summary-results/

Liability-adjusted R2 implementation: adapted from https://github.com/armartin/pgc_scz_asia/blob/master/eur_eas_prs.R

4.3 Results

4.3.1 Trade-off between training GWAS sample size and matched genetic ancestry in PRS prediction in EASs

We first constructed PRSs using SNP-level effect sizes from single EUR or EAS GWASs and evaluated their performance in an independent EAS validation dataset. To assess the impact of sample size on PRS performance, we used EUR GWAS from the BCAC published in 2015 (62,533 cases and 60,976 controls) (Michailidou et al., 2015) and EAS GWAS from the ABCC (14,958 cases and 15,843 controls) (Cai et al., 2014; Zheng et al., 2009, 2013) for PRS construction; we used an earlier, smaller EUR GWAS rather than the most recent, much larger GWAS (Michailidou et al., 2017) from the BCAC to illustrate our strategy of combing EAS and EUR data to avoid the EAS data being overwhelmed by the EUR data. For each training GWAS dataset, we derived predictors based on the "P+T" method (Materials and methods) and chose parameters that maximize prediction accuracy through cross-validation in an independent EAS validation dataset of 3,853 subjects (1,794 cases and 2,059 controls) (MEGA Shanghai; see Materials and methods and Table 4.1).

Table 4.1: Studies contributing to the current analysis (Shu et al., 2020)

Study	Cases	Controls	Sub-total	Genotyping platform
SBCGS (EAS)	1563	2396	3959	Illumina HumanExome-12v1_A Beadchip
SBCGS (EAS)	2511	2135	4646	Affymetrix GenomeWide Human SNP Array 6.0
HCES-Br (EAS)	274	273	547	Illumina Multi-Ethnic Genotyping Array
KPOP (EAS)	963	921	1884	Illumina Multi-Ethnic Genotyping Array
BBJ1 (EAS)	2642	2099	4741	Illumina OmniExpress BeadChip
SeBCS (EAS)	2246	2052	4298	Affymetrix Genome-Wide Human SNP Array 6.0
BCAC iCOGS (EAS)	4759	5967	10716	Illumina iSelect Genotyping Array
BCAC GWAS (EUR)	14910	17588	32498	Illumina 370K/550K/610K/670K/1.2M, Affymetrix 5.0/6.0
BCAC iCOGS (EUR)	46785	42892	89677	Illumina iSelect Genotyping Array
SBCGS (EAS validation)	1794	2059	3853	Illumina Multi-Ethnic Genotyping Array

We reported area under the receiver operating characteristics curve (AUC) of the PRSs along with their p-value threshold and the number of SNPs included in Fig 4.1 and Table 4.2. We found that a single individual sample in the EAS training dataset was substantially more informative about predicting breast cancer risk in the EAS validation dataset compared to that in the EUR training datasets. The best PRS derived from the EAS GWAS yielded an AUC of 0.5782. The best PRS derived from the EUR GWAS, which is ~ 4 times the sample size of the EAS GWAS, yielded a comparable AUC of 0.5809 in the EAS validation set. These findings are consistent across the different *p*-value thresholds used and the numbers of SNPs included in the PRSs (Fig 4.1), demonstrating the trade-off between higher prediction accuracy conferred by the larger EUR sample size and the matched EAS ancestry. Similar findings for type II diabetes in Latinos have been reported before (Márquez-Luna et al., 2017).

Table 4.2: Prediction AUC, Nagelkerke's R^2 and liability adjusted R^2 in the EAS validation set [a]

Model[b]	p[c]	N_{SNPs}[d]	AUC[e]	$R^2_{Nagelkelke}$[f]	R^2_{adj}[g]	p over reduced[h]
EAS	5×10^{-6}	44	0.5782	0.05	0.02	9.64×10^{-29}
EUR	5×10^{-8}	107	0.5809	0.05	0.02	1.99×10^{-28}
META_FE	1×10^{-5}	257	0.6008	0.06	0.03	1.33×10^{-38}
META_2EAS	5×10^{-6}	190	0.6049	0.07	0.03	1.30×10^{-40}
META_3EAS	1×10^{-5}	193	0.6059	0.07	0.03	1.08×10^{-40}
ADD3	NA	265	0.6096	0.07	0.03	7.70×10^{-43}

[a] We reported AUC, Nagelkerke's R2 and liability adjusted R2 for each of the PRSs in the EAS validation dataset (adjusted for age and first 2 principal components of genotype);
[b] Models that PRSs are based on: EAS: EAS GWAS derived PRS; EUR: EUR GWAS derived PRS; META_FE: conventional fixed-effect meta-analysis; META_2EAS: rescaled meta-analysis that up-weights the EAS summary statistics by a factor of two; META_3EAS: rescaled meta-analysis that up-weights the EAS summary statistics by a factor of three; ADD3: summation of the three best PRSs within each LD category;
[c] *p*-value cutoff for including SNPs in model;
[d] Number of SNPs in model;
[e] Area under ROC curve;
[f] Nagelkerke's R2 of a full model;
[g] Liability adjusted Nagelkerke's R2 that accounted for case/control ratio;
[h] *p*-value from likelihood ratio test compare full model (PRS, covariates including age and first 2 principal components of genotype) with a reduced model (a model with covariates age and first 2 principal components of genotype only);

To explore whether incorporating GWASs from both the EUR and EAS populations can boost prediction performance, we performed a meta-analysis of the EUR and EAS

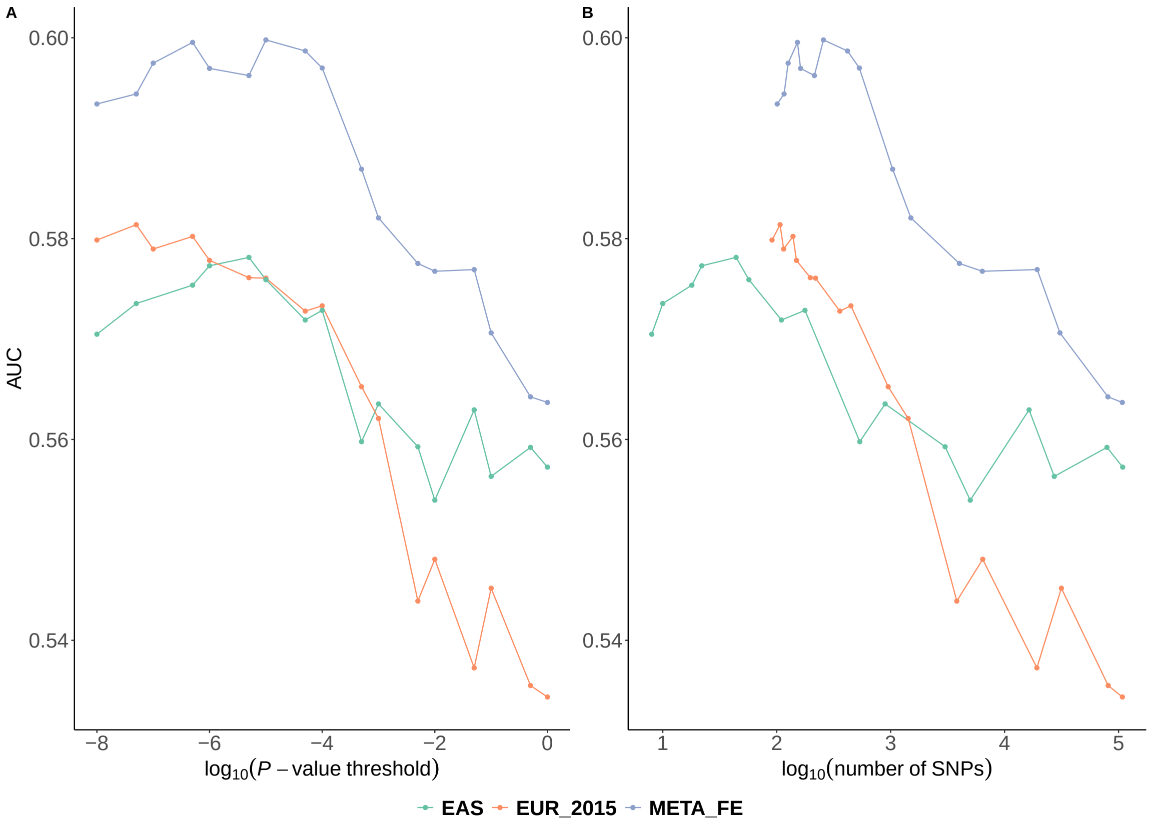


Figure 4.1: The AUC of PRSs derived from the EASEUR GWASs alone and meta-analysis of both. PRSs are derived from EAS, EUR GWASs and fixed effect meta-analysis of the EAS and EUR (denoted as `META_FE`). PRSs are evaluated in an independent EAS validation dataset for predicting breast cancer risk. Each PRS was plotted against the (A) *p*-value threshold and (B) number of SNPs included. The corresponding numerical results were reported in Table 4.3.

GWASs and used the resulting summary statistics to derive PRSs. We obtained an AUC of 0.6008 in the validation EAS dataset, which was higher than any PRS derived from the EAS or EUR GWASs alone. These results demonstrated that combining information from a higher-powered EUR GWAS and ancestry matched EAS GWAS helped improve breast cancer risk predictions in EASs, which was consistent with previous reports for other traits in EASs (Lam et al., 2019a) and Hispanic/Latinos (Grinde et al., 2019).

Table 4.3: Odds ratio (OR) comparing those with high PRS (10%) with the middle (40%-60%) of the population [a]

Model[b]	OR[c]	OR (95% CI)[d]	p-value[e]
EAS	2.23	1.82 - 2.98	1.90×10^{-12}
EUR	1.68	1.32 - 2.14	1.30×10^{-5}
META_FE	2.41	1.88 - 3.09	3.40×10^{-13}
META_2EAS	2.36	1.84 - 3.02	8.36×10^{-13}
META_3EAS	2.52	1.97 - 3.24	1.42×10^{-14}
ADD3	2.60	2.03 - 3.34	2.10×10^{-15}

[a] We reported ORs of breast cancer for the top 10% PRS relative to 40%-60% of population (adjusted for age and first 2 principal components of genotype) ;
[b] Models that PRSs are based on: EAS: EAS GWAS derived PRS; EUR: EUR GWAS derived PRS; META_FE: conventional fixed-effect meta-analysis; META_2EAS: rescaled meta-analysis that up-weights the EAS summary statistics by a factor of two; META_3EAS: rescaled meta-analysis that up-weights the EAS summary statistics by a factor of three; ADD3: summation of the three best PRSs within each LD category;
[c] OR: odds ratio;
[d] OR (95% CI): 95% confidence internal for OR;
[e] p-value: logistic regression association test p-values.

4.3.2 Upweighting the EAS effect size estimates in meta-analysis improves PRS prediction in EASs

In conventional fixed-effect meta-analysis, the effect size of each SNP is calculated by an inverse variance weighted sum of the effect size estimates from the participating GWASs. When the true effect sizes are equal across the GWASs, this formula is optimal and entails no efficiency loss compared to a joint analysis of the GWASs using individual-level data (Lin and Zeng, 2010). However, this formula may not be optimal when the goal is to provide effect size estimate from EAS and EUR GWASs for constructing PRS in EASs. Instead, we propose to tip the trans-ethnic meta-analysis towards the EAS population by up-scaling EAS effect size estimates by a factor of α in addition to the inverse variance weighting scheme (Materials and methods). This strategy enables us to shrink the meta effect size estimates towards the estimates in the EAS GWAS and increase the power to detect EAS-specific signals. To find a good up-scaling factor α, we conducted a grid search (i.e., $\alpha = 1,2,3,4,5$) and then evaluated the AUC of the resulting PRSs using cross-validation (S1 Fig). The conventional fixed-effect meta-analysis is a special case with $\alpha = 1$.

We evaluated these PRSs on the EAS validation dataset and found that up-scaling the EAS GWAS with a factor of two or three in meta-analysis could result in increased predicting accuracy (Fig 4.2). For example, we obtained an AUC of 0.6059 when using $\alpha = 3$, compared to an AUC of 0.6008 in conventional fixed-effect meta-analysis (i.e., $\alpha = 1$). In general, an up-scaling factor of two or three resulted in better prediction performance than conventional fixed-effect meta-analysis across a range of *p*-value thresholds used and numbers of SNPs included in the PRS in our analysis (Fig 4.2). We conducted one-sided Delong's test and did not observe significant difference between the AUCs of PRSs derived from the rescaled meta-analysis ($\alpha = 3$) and conventional fixed-effect meta-analysis (*p*-value = 0.14). This is somewhat expected as recent literature on PRS evaluation also showed small AUC differences between PRSs constructed using different methods using the same training dataset (Khera et al., 2018). We note that the difference between the AUCs of PRSs derived from the rescaled meta-analysis ($\alpha = 3$) and EAS GWAS was statistically significant (*p*-value = $4.6 \times 10^{(} - 4)$), so was the difference between the AUCs of PRSs derived from the rescaled meta-analysis ($\alpha = 3$) and EUR GWAS (*p*-value = $1.9 \times 10^{(} - 4)$).

4.3.3 GWAS effect size heterogeneity is related to LD score differences between ancestries

Wojcik et al. (2019) observed inconsistent effect size estimates between populations, which could contribute to reduced transferability of PRSs between populations. The extent of effect size differences between populations differs across the genome. As LD score measures

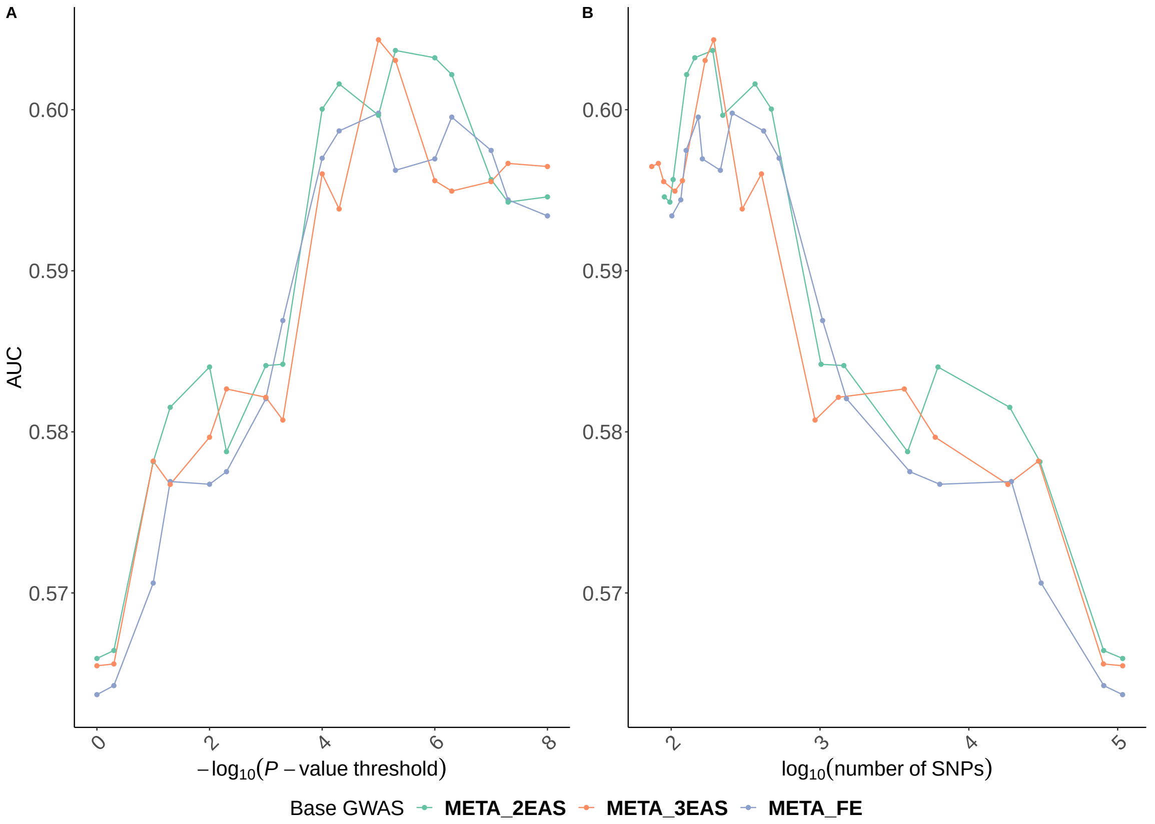


Figure 4.2: The AUC of PRSs derived from rescaled meta-analysis of the EAS and EUR_2015. PRSs are evaluated in predicting breast cancer risk in the EAS validation dataset. META_FE denotes the conventional fixed-effect meta-analysis; META_2EAS denotes the rescaled meta-analysis that upweights the EAS effect size estimates by a factor of two; and META_3EAS denotes the rescaled meta-analysis that upweights the EAS effect size estimates by a factor of three. Each PRS was plotted against the (A) *p*-value threshold and (B) number of SNPs. The corresponding numerical results were provided in Table 4.3

the tagging capacity of a SNP, a natural topic of investigation is whether the extent of LD differences is related to the extent of effect size differences between EURs and EASs.

To examine this, we used the difference in EAS and EUR LD score (Bulik-Sullivan et al., 2015): $l_{diff,j} = l_{EAS,j} - l_{EUR,j}$ as an indication of a SNP's tagging capacity divergence between populations, where $l_{EAS,j}$ and $l_{EUR,j}$ are the ancestry-specific LD scores in EASs and EURs, respectively, estimated from the corresponding populations in the 1000 Genomes Project (Consortium et al., 2015). We partitioned all available SNPs into three equally sized groups: "low EAS/high EUR LD score" group with l_{diff} less than its 33%

quantile, “similar EAS/EUR LD score” group with l_{diff} between its 33% and 67% quantiles, and “high EAS/low EUR LD score” group with l_{diff} larger than its 67% quantile. To account for the impact of differential GWAS sample sizes on SNPs’ effect size estimates, we calculated “standardized” z-scores by dividing the original z-scores by the square root of GWAS sample size (Wojcik et al., 2019). Then, we compared the standardized z-scores in the EUR and EAS GWASs (denoted as z_{EUR} and z_{EAS}, respectively) for SNPs with p-value $< 5 \times 10^{-8}$ in either the EUR or EAS GWAS (Fig 4.3). In general, we observed a reduction of standardized Z scores in EASs compared to EURs regardless of l_{diff} categories, with an overall slope of 0.64 ($z_{EAS} = 0.64 \times z_{EUR}$, 95% confidence interval: 0.61 - 0.67). In addition, we observed that standardized z-scores tended to be even lower in EASs for SNPs with “low EAS/high EUR LD score” ($z_{EAS} = 0.52 \times z_{EUR}$, 95% confidence interval: 0.48 - 0.55), and higher in EASs for SNPs with “high EAS/low EUR LD score” ($z_{EAS} = 0.83 \times z_{EUR}$, 95% confidence interval: 0.75 - 0.90), suggesting that LD differences are related to observed effect size differences between populations.

4.3.4 Effects of LD differences on genetic risk prediction

We further investigated the impact of differential tagging capacity due to differential LD levels between populations on PRS performance in EASs. After classifying SNPs into three groups based on l_{diff}, we constructed group specific PRSs based on both conventional and rescaled meta-analyses and applied the PRSs to the validation EAS dataset. We observed that the performance of PRS in the low EAS/high EUR LD score group is noticeably lower than that in the other two groups. This is true for both the conventional and rescaled meta-analyses (Fig 4.4). For example, the AUC of the PRS derived from the conventional meta-analysis using SNPs in the low EAS/high EUR LD score group is 0.5677, while the AUC of the PRSs derived using SNPs in the similar EAS/EUR LD score and high EAS/low EUR LD score groups are 0.5837 and 0.5890, respectively. Comparing the rescaled versus conventional meta-analysis, we found that upweighting the EAS effect size estimates with

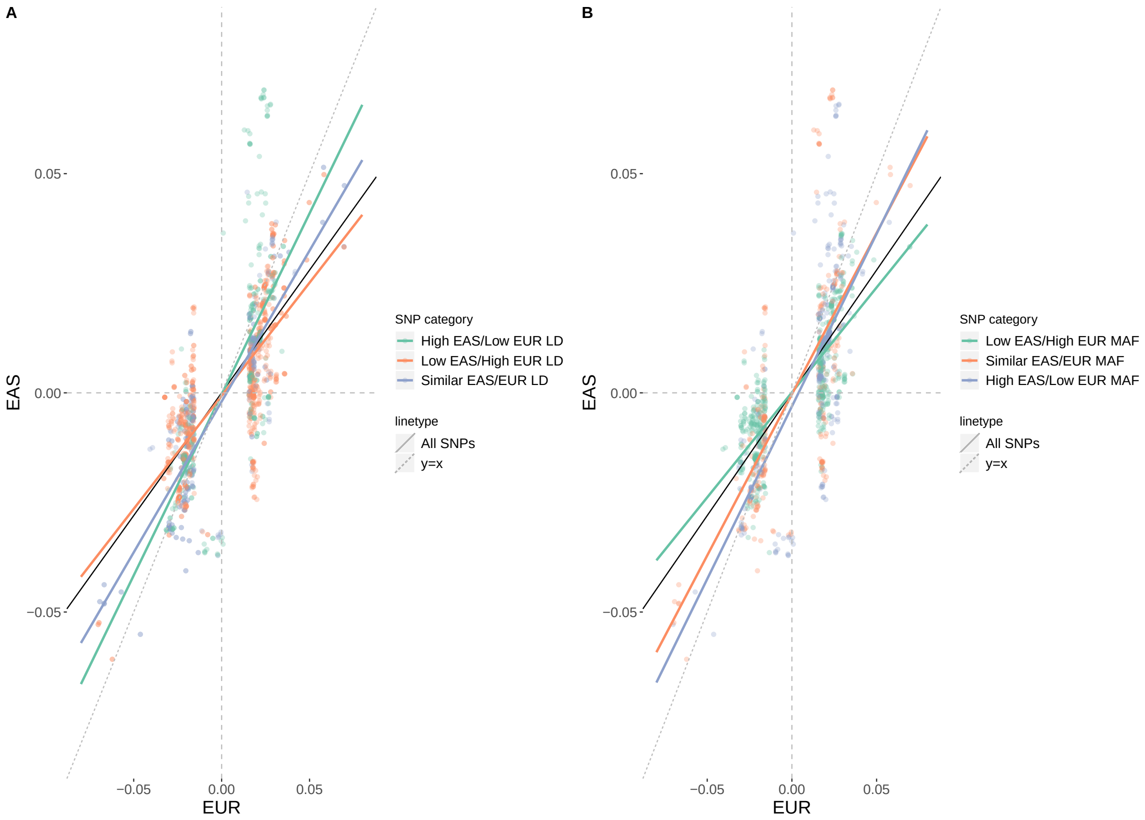


Figure 4.3: Standardized Z-scores of SNPs from EUR and EAS GWAS classified into different LD and MAF groups. SNPs with p-values $< 5 \times 10^{-8}$ in either `EUR_2015` or the EAS GWAS were included. (A) Low EAS/High EUR LD: SNPs with l_{diff} less than its 33% quantile; Similar EAS/EUR LD: SNPs with l_{diff} between its 33% and 67% quantiles; High EAS/Low EUR LD: SNPs with l_{diff} larger than its 67% quantile. (B) Low EAS/high EUR MAF: SNPs with $MAF_{diff,j}$ less than its 33% quantile, similar EAS/EUR MAF: SNPs with $MAF_{diff,j}$ between its 33% and 67% quantiles, and high EAS/low EUR MAF: SNPs with $MAF_{diff,j}$ larger than its 67% quantile. The black dashed line is the slope of the fitted line of standardized Z score in EAS over standardized Z score in EUR with all SNPs included; the red/blue/green lines are fitted lines of standardized Z score in EAS over standardized Z score in EUR for SNPs in corresponding LD group or MAF group.

a factor of two or three resulted in a more dramatic increase in prediction accuracy for SNPs in the low EAS/high EUR LD score and similar EAS/EUR LD score groups, while the performance gain in the high EAS/low EUR LD score group appeared to be minimal.

As the LD difference is likely a key factor contributing to the poor transferability of EUR-derived PRS to non-EUR populations, we also explored whether taking LD differences among population into account has potential to increase PRS accuracy. We con-

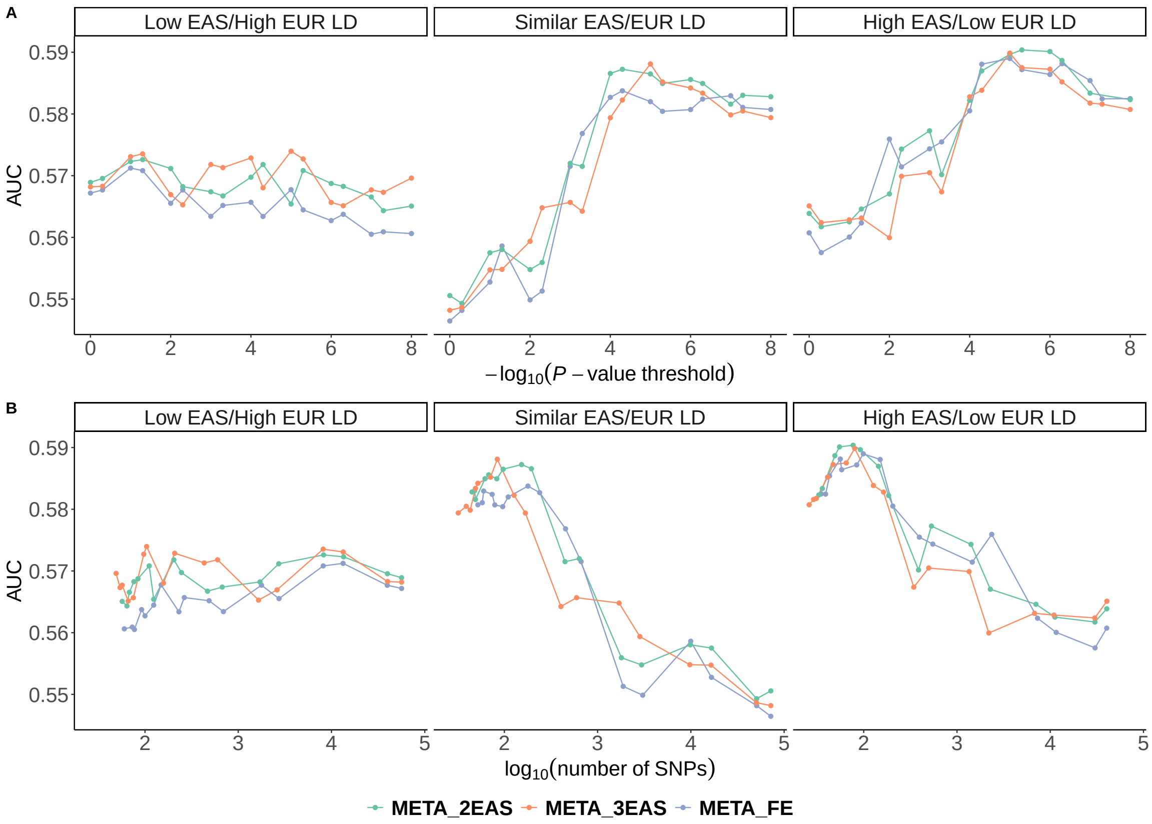


Figure 4.4: The AUC of PRSs constructed with SNPs in different l_{diff} groups. All SNPs were classified into three groups: low EAS/high EUR LD score group with l_{diff} less than its 33% quantile, similar EASEUR LD score group with l_{diff} between its 33% and 67% quantiles, and high EASlow LD score group with l_{diff} larger than its 67% quantile. META_FE denotes the conventional fixed-effect meta-analysis; META_2EAS denotes the rescaled meta-analysis that upweights the EAS effect size estimates by a factor of two; and META_3EAS denotes the rescaled meta-analysis that upweights the EAS effect size estimates by a factor of three. Each PRS was plotted against the (A) p-value threshold and (B) number of SNPs.

structed an l_{diff}-stratified PRS by the summation of the three best PRSs within each l_{diff} group (referred to as "ADD3"). We observed marginal significant improvement of AUC from ADD3 (AUC = 0.6096) over conventional fixed-effect meta-analysis PRS (AUC = 0.6008, one-sided Delong's test p-value = 0.05).

4.3.5 Effect of MAF differences on GWAS effect size and genetic risk prediction

As MAF differences between populations might also contribute to GWAS heterogeneity and decreased transferability of PRS, we sought to investigate the effects of MAF empirically. We used a similar approach as in the interrogation of LD differences to partition all available SNPs into three equally sized groups by MAF differences. Specifically, we used the difference in EAS and EUR MAF: $MAF_{diff,j} = MAF_{EAS,j} - MAF_{EUR,j}$ as an indication of a SNP's MAF divergence between populations, where $MAF_{EAS,j}$ and $MAF_{EUR,j}$ are the ancestry-specific MAFs reported by the single population GWASs. We partitioned all available SNPs into three equally sized groups: "low EAS/high EUR MAF" group with $MAF_{diff,j}$ less than its 33% quantile, "similar EAS/EUR MAF" group with $MAF_{diff,j}$ between its 33% and 67% quantiles, and "high EAS/low EUR MAF" group with $MAF_{diff,j}$ larger than its 67% quantile. Then, we compared the standardized z-scores in the EUR and EAS GWASs for SNPs with p-value $< 5 \times 10^{-8}$ in either the EUR or EAS GWAS similar to what we did previously for LD differences (Fig 3B). We observed that standardized z-scores tended to be lower in EASs for SNPs with "low EAS/high EUR MAF" ($z_{EAS} = 0.48 \times z_{EUR}$, 95% confidence interval: 0.44 - 0.51), and higher in EASs for SNPs with "high EAS/low EUR MAF" ($z_{EAS} = 0.78 \times z_{EUR}$, 95% confidence interval: 0.64 - 0.83), suggesting that MAF differences are related to observed effect size differences between populations.

We constructed MAF group-specific PRSs based on both conventional and rescaled meta-analyses and applied the PRSs to the validation EAS dataset. We observed that the performance of PRS in the low EAS/high EUR MAF group is noticeably lower than that in the other two groups (Fig 4.5). This is true for both the conventional and rescaled meta-analyses. For example, the AUC of the PRS derived from the conventional meta-analysis using SNPs in the low EAS/high EUR MAF group is 0.5530, while the AUC of the PRSs derived using SNPs in the similar EAS/EUR MAF and high EAS/low EUR MAF groups are 0.5876 and 0.5993, respectively.

We explored whether integrating MAF and LD difference information would further

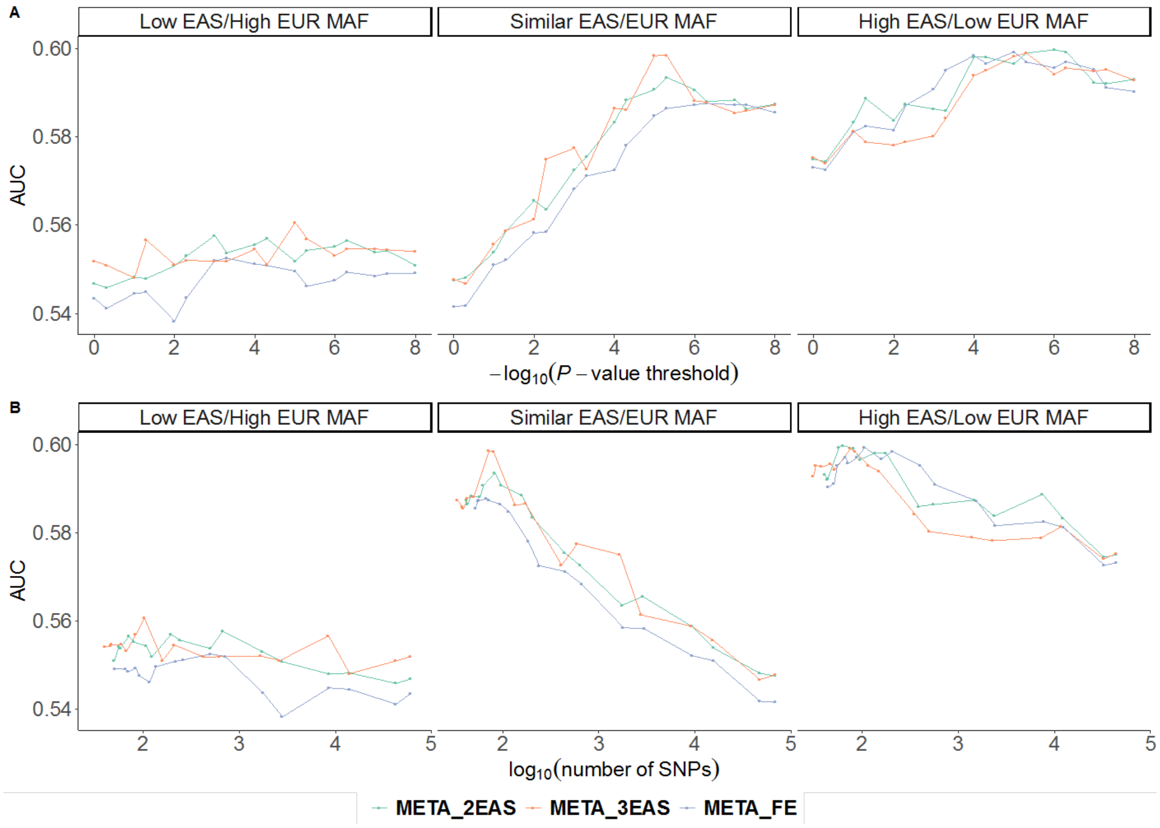


Figure 4.5: The AUC of PRSs constructed with SNPs in different MAF groups.

improve PRS. We stratified SNPs into nine groups cross-tabulated by the three LD score groups and three MAF groups (Fig 4.6). We observed that the prediction AUC of PRS derived from SNPs in "low EAS/high EUR MAF + high EAS/low EUR LD score" group was the lowest, while the AUC of PRS derived from SNPs in "high EAS/low EUR MAF + high EAS/low EUR LD score" group was the highest. Then, we evaluated the performance of the added score of the best PRSs from the nine groups (referred to as "ADD9"), similar to what we did with ADD3. We found no improvement on AUC for ADD9 (AUC = 0.6084) compared to ADD3 (AUC = 0.6096), indicating that the impact of MAF differences on PRS prediction might overlap with that of LD score differences (the correlation between MAF differences and LD score differences was 0.36).

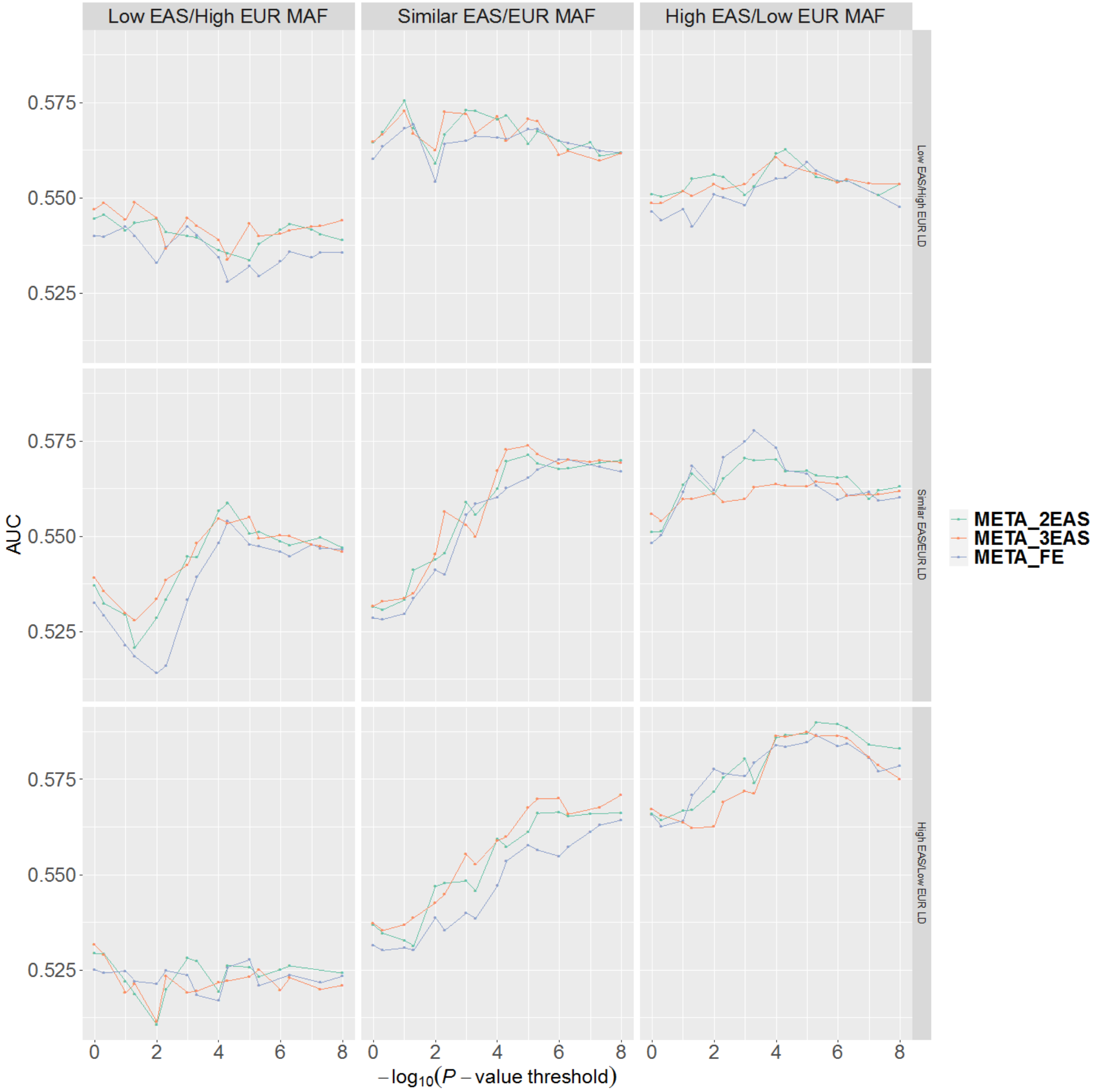


Figure 4.6: The AUC of PRSs constructed with SNPs in different MAF and LD groups.

4.3.6 Evaluation of the PRSs using the prediction Nagelkerke's pseudo R^2 and odds-ratio in EASs

We evaluated the candidate PRSs using the prediction Nagelkerke's pseudo R^2 and liability-adjusted R^2. We included the PRSs constructed from the EUR and EAS GWAS alone and the rescaled meta-analyses that upweights the EAS effect size estimates by a factor of two or three. We also included an l_{diff}-stratified PRS constructed by the summation of the three best PRSs within each l_{diff} group. We used logistic regression and included age and the first two principal components as covariates. The results were shown in Table 4.3. The PRS derived from the rescaled meta-analysis that upweights the EAS effect size estimates by a factor of three increased the prediction Nagelkerke's pseudo R2 (liability-adjusted R^2) by 40% (41%), 41% (42%), and 5% (6%), respectively, compared to PRSs derived from the EAS GWAS only, EUR GWAS only, and conventional fixed-effects meta-analysis of both. The l_{diff}-stratified PRS performed better than the other models, although the improvement was marginal (Table 1).

We assessed the odds ratio (OR) of developing breast cancer in the top 10% individuals with the highest PRSs versus the remaining 90% in the EAS validation dataset. We observed that PRSs derived from the meta-analyses of EAS and EUR GWASs resulted in improved ORs compared to PRSs derived from the EAS GWAS or EUR alone (Table 4.4). For example, we obtained ORs in the range of 2.31 - 2.50 for PRSs derived from the meta-analyses, while we obtained ORs of 2.23 and 1.73 for PRSs derived from the EAS GWAS or EUR, respectively. We also compared the top 10% individuals with the middle 40%-60% and observed similar results (Table 4.3).

4.4 Discussion

There has been tremendous progress in discovery of GWAS loci associated with breast cancer, making it feasible to construct PRS for individualized risk stratification. However, there is a lack of well-powered GWAS in non-EUR populations, a challenge that may exacerbate disparity in clinical use. The primary goal of this work is to explore strategies that can improve PRS of non-EUR populations, particularly in EAS. We found that incorporating information from well powered EUR GWAS and explicitly modeling LD and MAF differences are promising to improve PRS for breast cancer risk prediction in EAS. We proposed an approach to construct PRS from a rescaled meta-analysis of EUR and EAS GWAS which upweights the EAS component relative to the conventional inverse-variance based weightings. We observed improvement in PRS prediction accuracy using rescaled

Table 4.4: OR of developing breast cancer in individuals with the top 10% PRSs versus the remaining 90% in EAS validation data set[a]

Model[b]	OR[c]	OR (95% CI)[d]	p-value[e]
EAS	2.23	1.78 - 2.80	2.33×10^{-13}
EUR	1.73	1.39 - 2.16	3.93×10^{-7}
META_FE	2.50	1.99 - 3.14	8.40×10^{-17}
META_2EAS	2.31	1.85 - 2.91	1.86×10^{-14}
META_3EAS	2.43	1.94 - 3.06	5.34×10^{-16}
ADD3	2.50	1.99 - 3.14	8.40×10^{-17}

[a] We reported ORs of breast cancer for the top 10%PRS relative to remaining 90% of population (adjusted forage and first 2 principal components of genotype)

[b] META_FE denotes the conventional fixed-effect meta-analysis; META_2EAS denotes the rescaled meta-analysis that upweights the EAS effect size estimates by a factor of two; META_3EAS denotes the rescaled meta-analysis that upweights the EAS effect size estimates by a factor of three; ADD3 denotes the -stratified PRS constructed by the summation of the three best PRSs within each group

[c] OR, odds ratio;

[d] confidence interval of OR;

[e] logistic regression association test p-values

meta-analyses. As LD and MAF differences are likely key factors contributing to the poor transferability of EUR-derived PRS to non-EUR populations, we also explored whether taking them into account have potential to increase PRS accuracy. We observed marginal significant improvement of AUC from the l_{diff}-stratified PRS (“ADD3”) over conventional fixed-effect meta-analysis PRS (“META-FE”, p-value = 0.05) but no further improvement by stratifying on both LD and MAF differences (“ADD9”).

We further dissect why rescaled meta-analysis strategy is able to increase the PRS accuracy of the risk prediction in EAS in this study and in other non-EUR populations in general. We define the true model as $y = \sum \beta_{EAS,j} x_j$ for EAS and $y = \sum \beta_{EUR,j} x_j$ for EUR. Given the genetic differences between EAS and EUR, $\beta_{EAS,j}$ and $\beta_{EUR,j}$ are often unequal, and the extent of their differences depends on many factors, including LD, MAF, and environment factors. To construct a powerful PRS for EAS, i.e., $\sum w_{EAS,j}x_j$, the goal is to assign a weight for the jth SNP $w_{EAS,j}$ that is close to the true effect size $\beta_{EAS,j}$. When we use weights derived from EAS GWAS as $w_{EAS,j}$, the estimates are unbiased but of larger variance due to the smaller sample size. On the other hand, when the weights are derived from EUR GWAS, the estimates have smaller variance but are biased away from the true effect sizes in EAS, leading to poor transferability of EUR-derived PRS to EAS populations. To obtain estimates with a better bias and variance tradeoff, we proposed to combine EAS and EUR data to obtain estimates for use in PRS construction in EAS. Specifically, we proposed a rescaled meta-analysis strategy, with a rescale factor $\alpha > 1$ to "pull" the estimate towards the true effect size in EAS, i.e. $w_{EAS,j} = \alpha \times w_{inv\_EAS,j}\beta_{EAS,j} + w_{inv\_EUR,j}\beta_{EUR,j}$, where $w_{inv\_EAS,j}$ and $w_{inv\_EUR,j}$ are the inverse-variance weights in the conventional meta-analysis (Materials and methods). The magnitude of α controls for the extent of "pulling" towards EAS, with $\alpha = 0$ or $\alpha = inf$ corresponding to the extreme case where only EUR data ($\alpha = 0$) or only EAS data ($\alpha = inf$) are used. For $\alpha > 1$, it upweights EAS in a way that a sample in the EAS GWAS data contributes more than a sample in EUR GWAS to the resulting effect size estimate, thus achieving the effect of pulling toward EAS. The optimal magnitude of α depends on the relative sample size of EAS and EUR, and in general, α increases when EAS sample sizes increases relative to EUR due to a bias-variance tradeoff between the more accurate EAS results and more precise EUR results. When the EAS sample size is much smaller than the EUR sample size, the effect size estimates obtained from the EUR GWAS have much smaller standard errors. In this situation, upweighting the EAS effect size with a larger α may significantly increase the standard errors in the rescaled meta-analyzed compared to that in conventional fixed-effects meta-analysis, offsetting the potential benefit of reduced bias. For example, when we used the latest EUR GWAS data of 220,000 samples, the optimal α decreases to 1.3, achieving an AUC = 0.6195, compared to AUC = 0.6119 obtained in the traditional meta-analyses. To comprehensively study the relationship between optimal and relative EUR and EAS sample sizes, we would need access to individual-level training data so that we can conduct a series of GWASs at different sample sizes (e.g., 60k, 120k, 180k, and 220k for EUR and 10k, 20k, and 30k for EAS). For a variety of diseases, more non-EUR samples are being generated, e.g., PAGE (Wojcik et al., 2019) and TOPMed (Taliun et al., 2021), and the reweighting factor is expected to increase when meta-analyzing with EUR data, resulting in increased prediction accuracy for non-EUR samples. Note that we used only a single scaling factor to controls for the adjustment. Ideally, if we are able to assign a SNP-level scaling factor for every SNP, i.e. $w_{EAS,j} = \alpha_j \times w_{inv\_EAS,j}\beta_{EAS,j} + w_{inv\_EUR,j}\beta_{EUR,j}$ where α_j is the scaling factor for the jth SNP, the performance of PRS can be further improved. However, it is challenging to assign scaling factors for each individual risk allele, as it is unknown a priori which alleles have different effect sizes between populations, and to what extent of their differences are. It requires further efforts to model fine-scale genetic difference between populations to assign reasonable SNP-level scaling factors.

We also examined the predictive performance on our validation set using Grinde et al. (2019)'s approach that performed well in several anthropometric and blood count traits in Hispanic Americans: select SNPs based on European GWASs and use meta-analysis weight estimates to construct PRS. We find resulting PRS perform worse (AUC = 0.5859) than using conventional meta-analysis GWAS for both SNP selection and weight estimates (AUC = 0.6008). This agrees with previous findings that the best performing PRS for a trait likely depends on the genetic architecture, differences in sample size between populations, and other factors.

We are aware that there are other PRS construction methods besides "P+T", such as LDpred (Vilhjálmsson et al., 2015) and SBayesR (Lloyd-Jones et al., 2019). Specifically, we applied LDpred using the EAS LD reference panel. The findings were similar compared to those obtained by "P+T": PRS derived from conventional fixed-effects meta-analysis performed better than those derived from single population GWASs; PRSs derived from rescaled meta-analysis and the l_{diff}-stratified PRS performed better than PRS derived from conventional fixed-effect meta-analysis. We observed comparable performance between PRSs constructed using LDpred and "P+T", which is consistent with existing literature on breast cancer PRS (Khera et al., 2018).

We studied the impact of LD and MAF PRS constructed by taking into account both LD score differences and MAF differences between EAS and EUR did not outperform PRS constructed by taking into account LD score differences alone in EAS. It would be worthy of further exploration on how to better leverage MAF and LD in PRS construction. In addition, we did not explore the impact of functional genomic annotations on trans-ethnic PRS prediction. As previous studies have shown that the use of epigenetic and functional annotations improves heritability estimation and PRS prediction in a single population (Hu et al., 2017; Lloyd-Jones et al., 2019), an interesting topic of investigation is to incorporate those annotations when constructing trans-ethnic PRS to further boosting prediction accuracy.

Our work is based on a target population of EAS, while there are potential opportunities to extend the strategies explored in this study to other ethnicities. For example, the explicit modeling of genetic difference between EUR and African has potential to improve PRS in African and in African Americans. Although our approach has been effective in a relatively homogenous population like EAS, its application remains challenging in admixed populations with complex LD patterns and demographic history like African Americans or Hispanic/Latinos. Since the genomes of admixed individuals are a mosaic of segments with different ancestral origins, a first step would be to get ancestry specific effect size estimates and p-values from training GWASs, which is often not available from publicly available summary statistics. If individual-level training GWAS data is available, recently developed methods like Tractor (Atkinson et al., 2020) could be applied to obtain ancestry specific summary statistics by generating ancestry dosage at each site from local ancestry inference calls and running a local ancestry-aware regression. Similarly, for the validation data, local ancestry haplotype dosage for each person at each variant need to be estimated and weighted by the ancestry specific effect size estimated in the previous step to allow the generation of “ancestry-specific” PRSs. After that, we can experiment with our strategy of globally upweighting the “more informative” ancestry-level PRS. However, local ancestry estimation in both training and validation sets might introduce bias and the anticipated large sample size discrepancies between EUR and African Americans GWAS studies might further complicate the application. We think this question is worthy of further exploration and we believe that the rapid expansion of genomic resources in admixed populations will be critical to improve PRS predictions. Besides genetic factors (e.g., LD and MAF), environmental factors also influence effect size differences among ancestries. We argue that, for admixed populations, it is critical for PRS to be ancestry-aware, especially for clinical use, since each individual admixed genome has unique local ancestry profiles, and without taking local ancestry into account it is hard to maintain desired sensitivity and specificity due to genetic differences among ancestries.

In summary, we proposed an approach to construct breast cancer PRS in EAS derived from a rescaled meta-analysis of EUR and EAS GWAS. Different from conventional inverse-variance based weighting framework, our approach upweights the EAS component over the EUR component. PRS derived from our rescaled meta-analysis outperforms PRS derived from single population GWAS or conventional meta-analysis. This strategy of integrating GWASs across ethnicities when building PRS prediction models could potentially be extended to other non-EUR populations.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

The emergence of GWAS over a decade ago has led to remarkable shifts in our ability to understand the genetic basis of complex human traits. The availability of multiple data sources, such as regulatory atlas, rare variants based findings, or GWAS from multiple ancestries, have eased the translation of GWAS findings into biological mechanisms. Advanced statistical analyses that leverage these data have been key in securing continued progress in gene discovery and disease risk prediction.

Despite the tremendous progress, there is still plenty of room for development of GWAS follow-up analyses. While great progress has been made in identifying genes that influence traits via integrative analysis of GWAS and expression QTLs, genes linked to traits via regulatory effects other than expression remain understudied. In Chapter 2 of this dissertation, we attempted to fill the gap by leveraging splicing for gene discovery. Moreover, common variants identified by GWASs only contribute to part of the genetic basis of complex traits. The complementary roles of rare and common variants in disease biology have long been recognized, yet few published methods integrate knowledge from both for gene discovery. To bridge this gap, in Chapter 3, we proposed an approach to leverage GWAS signals to improve power for rare-variant based risk gene discovery. Another popular application of GWAS data is the use of PRS in disease risk prediction. However, the lack of representation of diverse populations limits the transferability of GWAS results across populations. To address this, we developed a rescaled meta-analysis based framework in Chapter 4 to improve genetic risk prediction accuracy for minority populations.

In Chapter 2, using our Multidimensional Splicing Gene (MSG) discovery approach, we implicated novel risk genes through integrative modeling of GWAS summary statistics and multidimensional splicing data from GTEx. Overall, we identified 2.15 times and

3.23 times significant genes from MSG than from current state-of-the-art approaches (i.e., S-MultiXcan or UTMOST). There are some exciting future avenues to gain deeper understanding of splicing regulation through analyzing a growing abundance of splicing data with GWAS using MSG. For example, recent developments in cell-type-specific splicing signal detection (Benegas et al., 2021) and long-read transcriptomics (Amarasinghe et al., 2020) offer remarkable potential in understanding how genes influence complex traits via splicing with higher accuracy and resolution.

In Chapter 3, we presented a pipeline to improve gene detection power from RVASs using GWAS signals via hypothesis weighting. We applied the pipeline to SCZ and ASD RVASs and observed sizeable improvements on the number of genes discovered. Different from most previous studies that use hypothesis weighting, we used prediction scores as covariates, which enabled us to harness data from various sources that are not readily available for each hypothesis. This approach requires little investment and can be easily applied to the analysis of existing and future studies beyond RVASs. Multiple large scale study design like GWAS, QTL discovery, can all be reanalyzed using this strategy with appropriate prediction scores as covariates.

Lastly, in Chapter 4, we proposed an approach to construct breast cancer PRS in east asians (EAS) derived from a rescaled meta-analysis of European (EUR) and EAS GWAS. PRS derived from our rescaled meta-analysis outperforms PRS derived from single population GWAS or conventional meta-analysis. This strategy of integrating GWASs across ethnicities when building PRS prediction models could potentially be extended to other non-EUR populations like African Americans and Hispanics. With the rapid expansion of genomic resources for non-EUR populations coupled with the active development of analytic methods, we believe there are remarkable potentials to the deployment of PRS in clinical settings.

With the decrease of sequencing cost, GWAS using whole genome sequencing (WGS) in large samples across ethnic groups will become increasingly realistic. This will lead to the discovery of additional GWAS loci, rare variants, and population-specific variants. These anticipated developments have the potential to improve PRS prediction, and identify genes through the integration of rare and common variants. In addition, with the development of high-throughput technologies, our ability to map sites of regulatory impact will increase, which will accelerate the translation of GWAS findings into genes that influence traits via regulatory effects. In the future, the availability of large-scale data, along with continuing development of analytical approaches, will keep driving gene discovery and reduce health disparities from GWAS.

References

(2017). Meta-analysis of gwas of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Molecular autism*, 8:1–17.

Aberg, K. A., Liu, Y., Bukszár, J., McClay, J. L., Khachane, A. N., Andreassen, O. A., Blackwood, D., Corvin, A., Djurovic, S., Gurling, H., et al. (2013). A comprehensive family-based replication study of schizophrenia genes. *JAMA psychiatry*, 70(6):573–581.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249.

Ahuja, R., Pinyol, R., Reichenbach, N., Custer, L., Klingensmith, J., Kessels, M. M., and Qualmann, B. (2007). Cordon-bleu is an actin nucleation factor and controls neuronal morphology. *Cell*, 131(2):337–350.

Akula, N., Marenco, S., Johnson, K., Feng, N., Zhu, K., Schulmann, A., Corona, W., Jiang, X., Cross, J., England, B., et al. (2021). Deep transcriptome sequencing of subgenual anterior cingulate cortex reveals cross-diagnostic and diagnosis-specific rna expression changes in major psychiatric disorders. *Neuropsychopharmacology*, pages 1–9.

Allman, R., Dite, G. S., Hopper, J. L., Gordon, O., Starlard-Davenport, A., Chlebowski, R., and Kooperberg, C. (2015). Snps and breast cancer risk prediction for african american and hispanic women. *Breast cancer research and treatment*, 154(3):583–589.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, 21(1):1–16.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., et al. (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet*, 9(4):e1003455.

Aschard, H., Vilhjálmsson, B. J., Greliche, N., Morange, P.-E., Trégouët, D.-A., and Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics*, 94(5):662–676.

Atkinson, E. G., Maihofer, A. X., Kanai, M., Martin, A. R., Karczewski, K. J., Santoro, M. L., Ulirsch, J. C., Kamatani, Y., Okada, Y., Finucane, H. K., et al. (2020). Tractor: A framework allowing for improved inclusion of admixed individuals in large-scale association studies. *bioRxiv*.

Barbeira, A. N., Bonazzola, R., Gamazon, E. R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., et al. (2020). Exploiting the gtex resources to decipher the mechanisms at gwas loci. *BioRxiv*, page 814350.

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature communications*, 9(1):1–20.

Barbeira, A. N., Pividori, M., Zheng, J., Wheeler, H. E., Nicolae, D. L., and Im, H. K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS genetics*, 15(1):e1007889.

Bayés, À., Van De Lagemaat, L. N., Collins, M. O., Croning, M. D., Whittle, I. R., Choudhary, J. S., and Grant, S. G. (2011). Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature neuroscience*, 14(1):19–21.

Benegas, G., Fischer, J., and Song, Y. S. (2021). Robust and annotation-free analysis of isoform variation using short-read scrna-seq data. *bioRxiv*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012.

Burbach, J. P. H. and van der Zwaag, B. (2009). Contact in the genetics of autism and schizophrenia. *Trends in neurosciences*, 32(2):69–72.

Burkhardt, R., Kenny, E. E., Lowe, J. K., Birkeland, A., Josowitz, R., Noel, M., Salit, J., Maller, J. B., Pe'er, I., Daly, M. J., et al. (2008). Common snps in hmgcr in micronesians and whites associated with ldl-cholesterol levels affect alternative splicing of exon13. *Arteriosclerosis, thrombosis, and vascular biology*, 28(11):2078–2084.

Cai, Q., Zhang, B., Sung, H., Low, S.-K., Kweon, S.-S., Lu, W., Shi, J., Long, J., Wen, W., Choi, J.-Y., et al. (2014). Genome-wide association analysis in east asians identifies breast cancer susceptibility loci at 1q32. 1, 5q14. 3 and 15q26. 1. *Nature genetics*, 46(8):886–890.

Cai, X., Yang, Z.-H., Li, H.-J., Xiao, X., Li, M., and Chang, H. (2021). A human-specific schizophrenia risk tandem repeat affects alternative splicing of a human-unique isoform as3mt d2d3 and mushroom dendritic spine density. *Schizophrenia Bulletin*, 47(1):219–227.

Callan, M. A. and Zarnescu, D. C. (2011). Heads-up: New roles for the fragile x mental retardation protein in neural stem and progenitor cells. *Genesis*, 49(6):424–440.

Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

Consortium, G. (2015). The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Consortium, G. et al. (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330.

Consortium, I. S. G., 2, W. T. C. C. C., et al. (2012). Genome-wide association study implicates hla-c* 01: 02 as a risk factor at the major histocompatibility complex locus in schizophrenia. *Biological psychiatry*, 72(8):620–628.

Coram, M. A., Candille, S. I., Duan, Q., Chan, K. H. K., Li, Y., Kooperberg, C., Reiner, A. P., and Tang, H. (2015). Leveraging multi-ethnic evidence for mapping complex traits in minority populations: an empirical bayes approach. *The American Journal of Human Genetics*, 96(5):740–752.

Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L., and Tang, H. (2017). Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *The American Journal of Human Genetics*, 101(2):218–226.

Darnell, J. C., Van Driesche, S. J., Zhang, C., Hung, K. Y. S., Mele, A., Fraser, C. E., Stone, E. F., Chen, C., Fak, J. J., Chi, S. W., et al. (2011). Fmrp stalls ribosomal translocation on mrnas linked to synaptic function and autism. *Cell*, 146(2):247–261.

Denny, L., De Sanjose, S., Mutebi, M., Anderson, B. O., Kim, J., Jeronimo, J., Herrero, R., Yeates, K., Ginsburg, O., and Sankaranarayanan, R. (2017). Interventions to close the divide for women with breast and cervical cancer between low-income and middle-income countries and high-income countries. *The Lancet*, 389(10071):861–870.

Dichgans, M., Malik, R., König, I. R., Rosand, J., Clarke, R., Gretarsdottir, S., Thorleifsson, G., Mitchell, B. D., Assimes, T. L., Levi, C., et al. (2014). Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke*, 45(1):24–36.

Egbujo, C. N., Sinclair, D., and Hahn, C.-G. (2016). Dysregulations of synaptic vesicle trafficking in schizophrenia. *Current psychiatry reports*, 18(8):1–10.

Eicher, J. D., Landowski, C., Stackhouse, B., Sloan, A., Chen, W., Jensen, N., Lien, J.-P., Leslie, R., and Johnson, A. D. (2015). Grasp v2. 0: an update on the genome-wide repository of associations between snps and phenotypes. *Nucleic acids research*, 43(D1):D799–D804.

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2015). Prsice: polygenic risk score software. *Bioinformatics*, 31(9):1466–1468.

Fatemi, S. H., Reutiman, T. J., Folsom, T. D., and Thuras, P. D. (2009). Gaba a receptor downregulation in brains of subjects with autism. *Journal of autism and developmental disorders*, 39(2):223.

Feng, H., Mancuso, N., Gusev, A., Majumdar, A., Major, M., Pasaniuc, B., and Kraft, P. (2021). Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. *PLoS genetics*, 17(4):e1008973.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228.

Fortney, K., Dobriban, E., Garagnani, P., Pirazzini, C., Monti, D., Mari, D., Atzmon, G., Barzilai, N., Franceschi, C., Owen, A. B., et al. (2015). Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS genetics*, 11(12):e1005728.

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E. M., Aguet, F., et al. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature genetics*, 50(7):956–967.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091.

Gaspar, H. A., Gerring, Z., Hübel, C., Middeldorp, C. M., Derks, E. M., and Breen, G. (2019). Using genetic drug-target networks to develop new drug hypotheses for major depressive disorder. *Translational psychiatry*, 9(1):1–9.

Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B. M., Gusev, A., et al. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10):1421.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

Gilbert, J. and Man, H.-Y. (2017). Fundamental elements in autism: from neurogenesis and neurite growth to synaptic plasticity. *Frontiers in cellular neuroscience*, 11:359.

Glatt, S., Chandler, S., Bousman, C., Chana, G., Lucero, G., Tatro, E., May, T., Lohr, J., Kremen, W., Everall, I., et al. (2009). Alternatively spliced genes as biomarkers for schizophrenia, bipolar disorder and psychosis: A blood-based spliceome-profiling exploratory study (supplementry table). *Current Pharmacogenomics and Personalized Medicine (Formerly Current Pharmacogenomics)*, 7(3):164–188.

Gligorijević, V. and Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571.

Goes, F. S., McGrath, J., Avramopoulos, D., Wolyniec, P., Pirooznia, M., Ruczinski, I., Nestadt, G., Kenny, E. E., Vacic, V., Peters, I., et al. (2015). Genome-wide association study of schizophrenia in ashkenazi jews. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(8):649–659.

Grinde, K. E., Qi, Q., Thornton, T. A., Liu, S., Shadyab, A. H., Chan, K. H. K., Reiner, A. P., and Sofer, T. (2019). Generalizing polygenic risk scores from europeans to hispanics/latinos. *Genetic epidemiology*, 43(1):50–62.

Griswold, A. J., Ma, D., Cukier, H. N., Nations, L. D., Schmidt, M. A., Chung, R.-H., Jaworski, J. M., Salyakina, D., Konidari, I., Whitehead, P. L., et al. (2012). Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. *Human molecular genetics*, 21(15):3513–3523.

Guo, X., Wang, X., Wang, Y., Zhang, C., Quan, X., Zhang, Y., Jia, S., Ma, W., Fan, Y., and Wang, C. (2017). Variants in the smarca4 gene was associated with coronary heart disease susceptibility in chinese han population. *Oncotarget*, 8(5):7350.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252.

Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B. M., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature genetics*, 50(4):538–548.

Herold, C., Hooli, B. V., Mullin, K., Liu, T., Roehr, J. T., Mattheisen, M., Parrado, A. R., Bertram, L., Lange, C., and Tanzi, R. E. (2016). Family-based association analyses of imputed genotypes reveal genome-wide significant association of alzheimer's disease with osbpl6, ptprg, and pdcl3. *Molecular psychiatry*, 21(11):1608–1612.

Hoffmann, T. J., Theusch, E., Haldar, T., Ranatunga, D. K., Jorgenson, E., Medina, M. W., Kvale, M. N., Kwok, P.-Y., Schaefer, C., Krauss, R. M., et al. (2018). A large electronic-health-record-based genome-wide study of serum lipids. *Nature genetics*, 50(3):401–413.

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., Yu, Z., Li, B., Gu, J., Muchnik, S., et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature genetics*, 51(3):568–576.

Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS computational biology*, 13(6):e1005589.

Huang, J., Bai, L., Cui, B., Wu, L., Wang, L., An, Z., Ruan, S., Yu, Y., Zhang, X., and Chen, J. (2020). Leveraging biological and statistical covariates improves the detection power in epigenome-wide association testing. *Genome biology*, 21:1–19.

Ignatiadis, N. and Huber, W. (2017). Covariate powered cross-weighted multiple testing. *arXiv preprint arXiv:1701.05179*.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577–580.

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nature genetics*, 51(3):404–413.

Ji, Y., Long, J., Kweon, S.-S., Kang, D., Kubo, M., Park, B., Shu, X.-O., Zheng, W., Tao, R., and Li, B. (2021). Incorporating european gwas findings improve polygenic risk prediction accuracy of breast cancer among east asians. *Genetic Epidemiology*.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, 50(9):1219–1224.

Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M. K., Schoech, A., Pasaniuc, B., and Price, A. L. (2019). Leveraging polygenic functional enrichment to improve gwas power. *The American Journal of Human Genetics*, 104(1):65–75.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.

Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature reviews Molecular cell biology*, 14(3):153–165.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome biology*, 20(1):1–21.

Lake, B. B., Chen, S., Sos, B. C., Fan, J., Kaeser, G. E., Yung, Y. C., Duong, T. E., Gao, D., Chun, J., Kharchenko, P. V., et al. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature biotechnology*, 36(1):70.

Lam, M., Chen, C.-Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B. C., et al. (2019a). Comparative genetic architectures of schizophrenia in east asian and european populations. *Nature genetics*, 51(12):1670–1678.

Lam, M., Hill, W. D., Trampush, J. W., Yu, J., Knowles, E., Davies, G., Stahl, E., Huckins, L., Liewald, D. C., Djurovic, S., et al. (2019b). Pleiotropic meta-analysis of cognition, education, and schizophrenia differentiates roles of early neurodevelopmental and adult synaptic pathways. *The American Journal of Human Genetics*, 105(2):334–350.

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452–1458.

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2012). A better coefficient of determination for genetic profile analysis. *Genetic epidemiology*, 36(3):214–224.

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., Mowry, B. J., Thapar, A., Goddard, M. E., Witte, J. S., et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature genetics*, 45(9):984.

Lei, L. and Fithian, W. (2016). Adapt: an interactive procedure for multiple testing with side information. *arXiv preprint arXiv:1609.06035*.

Lewis, C. M. and Hagenaars, S. P. (2019). Progressing Polygenic Medicine in Psychiatry Through Electronic Health Records. *JAMA Psychiatry*, 76(5):470–472.

Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., Lin, X., and Liang, L. (2013). Using eqtl weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in genetics*, 4:103.

Li, M., Jaffe, A. E., Straub, R. E., Tao, R., Shin, J. H., Wang, Y., Chen, Q., Li, C., Jia, Y., Ohi, K., et al. (2016a). A human-specific as3mt isoform and borcs7 are molecular risk factors in the 10q24. 32 schizophrenia-associated locus. *Nature medicine*, 22(6):649.

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., and Pritchard, J. K. (2018). Annotation-free quantification of rna splicing using leafcutter. *Nature genetics*, 50(1):151–158.

Li, Y. I., Van De Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., and Pritchard, J. K. (2016b). Rna splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604.

Li, Y. I., Wong, G., Humphrey, J., and Raj, T. (2019). Prioritizing parkinson's disease genes using population-scale transcriptomic data. *Nature communications*, 10(1):1–10.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lin, C.-X., Li, H.-D., Deng, C., Liu, W., Erhardt, S., Wu, F.-X., Zhao, X.-M., Wang, J., Wang, D., Hu, B., et al. (2021). Genome-wide prediction and integrative functional characterization of alzheimer's disease-associated genes. *bioRxiv*.

Lin, D. and Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 34(1):60–66.

Liu, D. J., Peloso, G. M., Yu, H., Butterworth, A. S., Wang, X., Mahajan, A., Saleheen, D., Emdin, C., Alam, D., Alves, A. C., et al. (2017a). Exome-wide association study of plasma lipids in¿ 300,000 individuals. *Nature genetics*, 49(12):1758–1766.

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P. L., Goel, A., Zhang, H., et al. (2014). Meta-analysis of gene-level tests for rare variant association. *Nature genetics*, 46(2):200.

Liu, J. Z., Erlich, Y., and Pickrell, J. K. (2017b). Case–control association mapping by proxy using family history of disease. *Nature genetics*, 49(3):325.

Liu, L., Zeng, P., Xue, F., Yuan, Z., and Zhou, X. (2021). Multi-trait transcriptome-wide association studies with probabilistic mendelian randomization. *The American Journal of Human Genetics*, 108(2):240–256.

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). 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.

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature communications*, 10(1):1–11.

Loos, R. J. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, 11(1):1–3.

Love, J. E., Hayden, E. J., and Rohn, T. T. (2015). Alternative splicing in alzheimer's disease. *Journal of Parkinson's disease and Alzheimer's disease*, 2(2).

Luningham, J. M., Chen, J., Tang, S., De Jager, P. L., Bennett, D. A., Buchman, A. S., and Yang, J. (2020). Bayesian genome-wide twas method to leverage both cis-and trans-eqtl information through summary statistics. *The American Journal of Human Genetics*, 107(4):714–726.

Ma, L., Semick, S. A., Chen, Q., Li, C., Tao, R., Price, A. J., Shin, J. H., Jia, Y., Brandon, N. J., Cross, A. J., et al. (2020a). Schizophrenia risk variants influence multiple classes of transcripts of sorting nexin 19 (snx19). *Molecular psychiatry*, 25(4):831–843.

Ma, L., Shcherbina, A., and Chetty, S. (2020b). Variations and expression features of cyp2d6 contribute to schizophrenia risk. *Molecular psychiatry*, pages 1–11.

Malhotra, D. and Sebat, J. (2012). Cnvs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*, 148(6):1223–1241.

Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., and Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nature genetics*, 51(4):675–682.

Marioni, R. E., Harris, S. E., Zhang, Q., McRae, A. F., Hagenaars, S. P., Hill, W. D., Davies, G., Ritchie, C. W., Gale, C. R., Starr, J. M., et al. (2018). Gwas on family history of alzheimer's disease. *Translational psychiatry*, 8(1):1–7.

Marquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S. S., Furlotte, N., Auton, A., Price, A. L., 23andMe Research Team, et al. (2020). Ldpred-funct: incorporating functional priors improves polygenic prediction accuracy in uk biobank and 23andme data sets. *bioRxiv*, page 375337.

Márquez-Luna, C., Loh, P.-R., Consortium, S. A. T . D. S., Consortium, S. T . D., and Price, A. L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic epidemiology*, 41(8):811–823.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591.

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J. P., Chen, T.-H., Wang, Q., Bolla, M. K., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics*, 104(1):21–34.

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., et al. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature genetics*, 47(4):373–380.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94.

Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., Ebbert, A., Riley, Z. L., Royall, J. J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. *Nature*, 508(7495):199–206.

Mills, M. C. and Rahal, C. (2020). The gwas diversity monitor tracks diversity by disease in real time. *Nature genetics*, 52(3):242–243.

Müller, C. S., Haupt, A., Bildl, W., Schindler, J., Knaus, H.-G., Meissner, M., Rammner, B., Striessnig, J., Flockerzi, V., Fakler, B., et al. (2010). Quantitative proteomics of the cav2 channel nano-environments in the mammalian brain. *Proceedings of the National Academy of Sciences*, 107(34):14950–14957.

Musliner, K. L., Mortensen, P. B., McGrath, J. J., Suppli, N. P., Hougaard, D. M., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Andreassen, O., Pedersen, C. B., Pedersen, M. G., Mors, O., Nordentoft, M., Børglum, A. D., Werge, T., Agerbo, E., and for the Bipolar Disorder Working Group of the Psychiatric Genomics Consortium (2019). Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and Schizophrenia With Risk for Depression in the Danish Population. *JAMA Psychiatry*, 76(5):516–525.

Nathanson, K. N., Wooster, R., and Weber, B. L. (2001). Breast cancer genetics: what we know and what we need. *Nature medicine*, 7(5):552–556.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS Genet*, 6(4):e1000888.

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., et al. (2015). A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, 47(10):1121.

Oba, T., Saito, T., Asada, A., Shimizu, S., Iijima, K. M., and Ando, K. (2020). Microtubule affinity–regulating kinase 4 with an alzheimer's disease-related mutation promotes tau accumulation and exacerbates neurodegeneration. *Journal of Biological Chemistry*, 295(50):17138–17147.

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., et al. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics*, 50(3):381–389.

Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., Lui, J. C., Vedantam, S., Gustafsson, S., Esko, T., et al. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications*, 6(1):1–9.

Pirooznia, M., Wang, T., Avramopoulos, D., Valle, D., Thomas, G., Huganir, R. L., Goes, F. S., Potash, J. B., and Zandi, P. P. (2012). Synaptomedb: an ontology-based knowledgebase for synaptic genes. *Bioinformatics*, 28(6):897–899.

Pocklington, A. J., Rees, E., Walters, J. T., Han, J., Kavanagh, D. H., Chambert, K. D., Holmans, P., Moran, J. L., McCarroll, S. A., Kirov, G., et al. (2015). Novel findings from cnvs implicate inhibitory and excitatory signaling complexes in schizophrenia. *Neuron*, 86(5):1203–1214.

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'dushlaine, C., Chambert, K., Bergen, S. E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–190.

Raj, T., Li, Y. I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.-C., Ng, B., Gupta, I., Haroutunian, V., et al. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in alzheimer's disease susceptibility. *Nature genetics*, 50(11):1584–1592.

Reeskamp, L. F., Hartgers, M. L., Peter, J., Dallinga-Thie, G. M., Zuurbier, L., Defesche, J. C., Grefhorst, A., and Hovingh, G. K. (2018). A deep intronic variant in ldlr in familial hypercholesterolemia: Time to widen the scope? *Circulation: Genomic and Precision Medicine*, 11(12):e002385.

Richardson, T. G., Sanderson, E., Palmer, T. M., Ala-Korpela, M., Ference, B. A., Davey Smith, G., and Holmes, M. V. (2020). Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable mendelian randomisation analysis. *PLoS medicine*, 17(3):e1003062.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K.-H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., Bergen, S. E., Collins, A. L., Crowley, J. J., Fromer, M., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics*, 45(10):1150.

Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A., Lin, D.-Y., Duan, J., Ophoff, R. A., Andreassen, O. A., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature genetics*, 43(10):969.

Ripke, S., Walters, J. T., O'Donovan, M. C., of the Psychiatric Genomics Consortium, S. W. G., et al. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *MedRxiv*.

Rodosthenous, T., Shahrezaei, V., and Evangelou, M. (2019). Integrating multi-omics data through sparse canonical correlation analysis for predicting complex traits: A comparative study. *bioRxiv*, page 843524.

Roeder, K., Devlin, B., and Wasserman, L. (2007). Improving power in genome-wide association studies: Weights tip the scale. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 31(7):741–747.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 24(4):398.

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature genetics*, 46(9):944–950.

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, 180(3):568–584.

Schwartzentruber, J., Cooper, S., Liu, J. Z., Barrio-Hernandez, I., Bello, E., Kumasaka, N., Young, A. M., Franklin, R. J., Johnson, T., Estrada, K., et al. (2021). Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new alzheimer's disease risk genes. *Nature genetics*, pages 1–11.

Scotti, M. M. and Swanson, M. S. (2016). Rna mis-splicing in disease. *Nature Reviews Genetics*, 17(1):19.

Shi, X., Chai, X., Yang, Y., Cheng, Q., Jiao, Y., Chen, H., Huang, J., Yang, C., and Liu, J. (2020). A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. *Nucleic acids research*, 48(19):e109–e109.

Shimamoto, C., Ohnishi, T., Maekawa, M., Watanabe, A., Ohba, H., Arai, R., Iwayama, Y., Hisano, Y., Toyota, T., Toyoshima, M., et al. (2014). Functional characterization of fabp3, 5 and 7 gene variants identified in schizophrenia and autism spectrum disorder and mouse behavioral studies. *Human molecular genetics*, 23(24):6495–6511.

Shipra, A., Chetan, K., and Rao, M. (2006). Cremofac—a database of chromatin remodeling factors. *Bioinformatics*, 22(23):2940–2944.

Shu, X., Long, J., Cai, Q., Kweon, S.-S., Choi, J.-Y., Kubo, M., Park, S. K., Bolla, M. K., Dennis, J., Wang, Q., et al. (2020). Identification of novel breast cancer susceptibility loci in meta-analyses conducted among asian and european descendants. *Nature communications*, 11(1):1–9.

Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., Bass, N., Bigdeli, T. B., Breen, G., Bromet, E. J., et al. (2020). Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. *medRxiv*.

Speed, D. and Balding, D. J. (2019). Sumher better estimates the snp heritability of complex traits from summary statistics. *Nature genetics*, 51(2):277–284.

Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., and Balding, D. J. (2017). Reevaluation of snp heritability in complex human traits. *Nature genetics*, 49(7):986–992.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Sun, R., Hui, S., Bader, G. D., Lin, X., and Kraft, P. (2019). Powerful gene set analysis in gwas with the generalized berk-jones statistic. *PLoS genetics*, 15(3):e1007530.

Szatmari, P., Paterson, A. D., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X.-Q., Vincent, J. B., Skaug, J. L., Thompson, A. P., Senman, L., et al. (2007). Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature genetics*, 39(3):319.

Takata, A., Matsumoto, N., and Kato, T. (2017). Genome-wide identification of splicing qtls in the human brain and their enrichment among schizophrenia-associated loci. *Nature communications*, 8(1):1–11.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., et al. (2021). Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2):87–108.

Tsavou, A. and Curtis, D. (2019). In-silico investigation of coding variants potentially affecting the functioning of the glutamatergic n-methyl-d-aspartate receptor in schizophrenia. *Psychiatric genetics*, 29(2):44–50.

Vergult, S., Dheedene, A., Meurs, A., Faes, F., Isidor, B., Janssens, S., Gautier, A., Le Caignec, C., and Menten, B. (2015). Genomic aberrations of the cacna2d1 gene in three patients with epilepsy and intellectual disability. *European Journal of Human Genetics*, 23(5):628–632.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*, 97(4):576–592.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., , Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., Psaty, B. M., Kooperberg, C., Roden, D., Darbar, D., Arnett, D. K., Regan, E. A., Boerwinkle, E., Rotter, J. I., Allison, M. A., McDonald, M.-L. N., Chung, M. K., Smith, N. L., Ellinor, P. T., Vasan, R. S., Mathias, R. A., Rich, S. S., Heckbert, S. R., Redline, S., Guo, X., Chen, Y.-D. I., Liu, C.-T., de Andrade, M., Yanek, L. R., Albert, C. M., Hernandez, R. D., McGarvey, S. T., North, K. E., Lange, L. A., Weir, B. S., Laurie, C. C., Yang, J., and Visscher, P. M. (2019). Recovery of trait heritability from whole genome sequence data. *bioRxiv*.

Walker, R. L., Ramaswami, G., Hartl, C., Mancuso, N., Gandal, M. J., De La Torre-Ubieta, L., Pasaniuc, B., Stein, J. L., and Geschwind, D. H. (2019). Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell*, 179(3):750–771.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, A. S., Kusenda, M., Malhotra, D., Bhandari, A., et al. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *science*, 320(5875):539–543.

Wang, G.-S. and Cooper, T. A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):749–761.

Wang, H., Yang, J., Schneider, J. A., De Jager, P. L., Bennett, D. A., and Zhang, H.-Y. (2020a). Genome-wide interaction analysis of pathological hallmarks in alzheimer's disease. *Neurobiology of aging*, 93:61–68.

Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., Zhong, X., Tao, R., Wen, Z., Sutcliffe, J. S., et al. (2019). A bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia gwas data. *Nature neuroscience*, 22(5):691–699.

Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P. M., and Yengo, L. (2020b). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature communications*, 11(1):1–9.

Wen, W., Shu, X.-o., Guo, X., Cai, Q., Long, J., Bolla, M. K., Michailidou, K., Dennis, J., Wang, Q., Gao, Y.-T., et al. (2016). Prediction of breast cancer risk based on common genetic variants in women of east asian ancestry. *Breast Cancer Research*, 18(1):1–8.

Weyn-Vanhentenryck, S. M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P. A., Zhang, M. Q., et al. (2014). Hits-clip and integrative modeling define the rbfox splicing-regulatory network linked to brain development and autism. *Cell reports*, 6(6):1139–1152.

Wheeler, H. E., Shah, K. P., Brenner, J., Garcia, T., Aquino-Michaels, K., Consortium, G., Cox, N. J., Nicolae, D. L., and Im, H. K. (2016). Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS genetics*, 12(11):e1006423.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274.

Witten, D. and Tibshirani, R. (2020). *PMA: Penalized Multivariate Analysis*. R package version 1.2.1.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1).

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518.

Wu, Y., Cao, H., Baranova, A., Huang, H., Li, S., Cai, L., Rao, S., Dai, M., Xie, M., Dou, Y., et al. (2020). Multi-trait analysis for genome-wide association study of five psychiatric disorders. *Translational psychiatry*, 10(1):1–11.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., et al. (2012). Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375.

Yang, Y., Shi, X., Jiao, Y., Huang, J., Chen, M., Zhou, X., Sun, L., Lin, X., Yang, C., and Liu, J. (2020). Comm-s2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. *Bioinformatics*, 36(7):2009–2016.

Yu, C.-E., Seltman, H., Peskind, E. R., Galloway, N., Zhou, P. X., Rosenthal, E., Wijsman, E. M., Tsuang, D. W., Devlin, B., and Schellenberg, G. D. (2007). Comprehensive analysis of apoe and selected proximate markers for late-onset alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. *Genomics*, 89(6):655–665.

Yurko, R., G'Sell, M., Roeder, K., and Devlin, B. (2020). A selective inference approach for false discovery rate control using multiomics covariates yields insights into disease risk. *Proceedings of the National Academy of Sciences*, 117(26):15028–15035.

Zhang, M. J., Xia, F., and Zou, J. (2019). Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature communications*, 10(1):1–11.

Zheng, W., Long, J., Gao, Y.-T., Li, C., Zheng, Y., Xiang, Y.-B., Wen, W., Levy, S., Deming, S. L., Haines, J. L., et al. (2009). Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25. 1. *Nature genetics*, 41(3):324–328.

Zheng, W., Zhang, B., Cai, Q., Sung, H., Michailidou, K., Shi, J., Choi, J.-Y., Long, J., Dennis, J., Humphreys, M. K., et al. (2013). Common genetic determinants of breast-cancer risk in east asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. *Human molecular genetics*, 22(12):2539–2550.

Zhou, D., Jiang, Y., Zhong, X., Cox, N. J., Liu, C., and Gamazon, E. R. (2020). A unified framework for joint-tissue transcriptome-wide association and mendelian randomization analysis. *Nature Genetics*, 52(11):1239–1246.

Zhu, Z., Lin, Y., Li, X., Driver, J. A., and Liang, L. (2019). Shared genetic architecture between metabolic traits and alzheimer's disease: a large-scale genome-wide cross-trait analysis. *Human genetics*, 138(3):271–285.