Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Randy Parrish Date

TIGAR-V2 with nonparametric Bayesian eQTL weights estimated from GTEx V8 & Leveraging multiple reference panels to improve TWAS power by ensemble machine learning

By

Randy Parrish
Master of Science in Public Health

Biostatistics and Bioinformatics

Jingjing Yang, PhD
Thesis Committee Chair

Michael P. Epstein, PhD
Committee Member

TIGAR-V2 with nonparametric Bayesian eQTL weights estimated from GTEx V8 & Leveraging multiple reference panels to improve TWAS power by ensemble machine learning

By

Randy Parrish
B.A., B.S.
University of Louisville
2019

Thesis Committee Chair: Jingjing Yang, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2021

Abstract

TIGAR-V2 with nonparametric Bayesian eQTL weights estimated from GTEx V8 & Leveraging multiple reference panels to improve TWAS power by ensemble machine learning
By Randy Parrish

Background: Transcriptome-wide association study (TWAS) is a popular technique for integrating reference transcriptomic data with data from genome-wide association studies (GWAS) to conduct gene-based association studies. The standard two-stage TWAS methods train gene expression prediction models on reference data, and then test the association between the predicted genetically regulated gene expression (GReX) and phenotype of interest for test data. Limitations of existing TWAS tools make it difficult for users to train GReX prediction models using their own data and no methods currently exists for leveraging multiple reference panels to improve TWAS power.

Methods: In part one, we develop a new version of the Transcriptome-Integrated Genetic Association Resource (TIGAR-V2), train nonparametric Bayesian DPR gene expression prediction models for 49 tissues from the Genotype-Tissue Expression (GTEx) project V8 reference panel, and validate the TIGAR-V2 method using application TWAS of breast and ovarian cancer. In part two, we develop a novel Stacked Regression based TWAS (SR-TWAS) method for leveraging multiple reference panels using ensemble machine learning and validated our method using simulation studies and real TWAS leveraging two reference panels of brain frontal cortex tissue.

Results: TIGAR-V2 identified 88 TWAS risk genes for breast cancer, most of which are known or near previously identified GWAS (84; 95%) or TWAS (35; 40%) risk genes. TIGAR-V2 identified 37 TWAS risk genes of ovarian cancer, most of which are known or near previously identified GWAS (35; 95%) or TWAS (13; 35%) risk genes. TIGAR-V2 identified 1 novel independent risk gene of breast cancer with known biological functions involved in carcinogenesis and 2 novel independent risk genes of both breast and ovarian cancer which are near such genes. SR-TWAS models had higher gene expression prediction accuracy and TWAS power than the models trained on single cohorts in all simulation scenarios and outperformed both single cohort models in the real data application GReX prediction.

Conclusions: We believe our improved TIGAR-V2 and SR-TWAS tools will provide a useful resource for mapping risk genes of complex diseases by TWAS.

TIGAR-V2 with nonparametric Bayesian eQTL weights estimated from GTEx V8 & Leveraging multiple reference panels to improve TWAS power by ensemble machine learning

By

Randy Parrish

B.A., B.S.
University of Louisville
2019

Thesis Committee Chair: Jingjing Yang, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2021

Acknowledgments

First I would like to thank my parents for their constant, unwavering financial support that allowed me to focus on my degree, my grandmother for her encouragement, and my sibling for the emotional support and consistenty hilarious, often terrible, "advice."

I am forever grateful to Dr. Jingjing Yang for providing me the opportunity to conduct this thesis and the direction I needed to see it through to the end. I would like to thank my thesis committee, Dr. Jingjing Yang and Dr. Michael Epstein, for their helpful advice and valuable insight.

I want to thank the faculty and staff of Department of Biostatistics and Bioinformatics for the incredible assistance, knowledge, and community they have shared with me, with special thanks to Dr. David Benkeser, Dr. John Hanfelt, Dr. Reneé Moore, and Dr. Lance Waller.

Though I wish more of it could have been spent in-person, I am thankful for the time I have had as a student here. The incredible enthusiasm, talent, and character of the people in this community has simultaneously made the distance more difficult to endure and easier to bear.

I am grateful for the friendships I have made with the members of my biostatistics cohort and the RSPH class of 2021. To my friends at Emory and elsewhere, thank you for the encouragement, for the solidarity, and for enriching my life with your presence. To Kathleen Keh–thank you for being there.

Finally, I want to express my deepest gratitude to Dr. Vicki Ragsdell for the guidance I needed to make it. This journey would not have been possible without her.

Contents

List of Figures

List of Tables

1 Introduction

The majority of genetic risk loci successfully identified by genome-wide association studies (GWAS) lie within noncoding regions of the genome, and the underlying mechanisms through which such variants affect complex traits remain mostly undetermined [1–6]. Recent studies revealed that GWAS signals were enriched with expression quantitative trait loci (eQTL) [7–12]–regulatory variants which explain a fraction of the variance in transcript abundance for a target gene [7–12]. Multiple techniques have been proposed to integrate transcriptomic data, including eQTL summary statistics, with GWAS data in order to improve the power for identifying GWAS risk loci and to illustrate the underlying biological mechanism of GWAS loci [7, 13–18].

Transcriptome-wide association study (TWAS) [13, 16, 18–20] is a popular, widely used technique for integrating reference transcriptomic data with GWAS data to conduct gene-based association studies. The standard two-stage TWAS methods [13, 16, 18] first fit gene expression prediction models using reference transcriptomic and genetic data profiled for the same samples, and then test the association between the predicted genetically regulated gene expression (GReX) and phenotype of interest for the test GWAS cohort. Recent application studies show that TWAS is capable of identifying risk genes whose genetic effects are potentially mediated through gene expression [21–23]. TWAS has the advantages of using publicly available reference transcriptomic data such as the Genotype-Tissue Expression (GTEx) project [10, 24] and summary-level GWAS data and has successfully identified novel candidate risk genes for age-related macular degeneration [23], rheumatoid arthritis [25], schizophrenia [26], pancreatic cancer [27], and broad types of complex traits [17].

However, most of the existing tools [13, 16, 20] require specific genotype dosage data

format per gene, miss the goodness-of-fitting evaluation for trained gene expression prediction models, and fail to implement parallel computing within the tool. The use of non-standard formats for genotype input files necessitates additional data preparation. Failing to evaluate goodness-of-fit can result in invalid follow-up tests based on models for genes that lack sufficient information for prediction. These limitations bring difficulties for users who need to train gene expression prediction models by using their own reference transcriptomic and genetic data.

In part one, we develop a new version of the Transcriptome-Integrated Genetic Association Resource (referred to as TIGAR-V2) that takes genotype data of the Variant Call Format (VCF) as input, conducts cross-validation [13, 16, 28] to evaluate trained gene expression prediction models, and enables parallel computation to make use of high performance computing clusters. Additionally, TIGAR-V2 implements both general linear regression with Elastic-Net penalty [13] and nonparametric Bayesian Dirichlet process (DPR) regression methods [29–31] for training gene expression prediction models, and tests gene-based association by both Burden type [13, 16, 21] and Variance-Component statistics [32].

To make the TIGAR-V2 a convenient resource for the public, we trained nonparametric Bayesian DPR gene expression prediction models for 49 tissues from the GTEx V8 reference panel [24]. These tissue-specific eQTL weights are provided along with this TIGAR-V2 tool, which can be conveniently used for follow-up gene-based association studies using both individual-level and summary-level GWAS data. In our example application studies, we used eQTL weights obtained from transcriptomic data of breast mammary tissue and ovary tissue from the GTEX V8 reference panel along with publicly available GWAS summary statistics [33, 34] to conduct TWAS for studying breast cancer and ovarian cancer.

We identified 88 significant TWAS risk genes for breast cancer, 84 (95%) of which have either been previously identified by GWAS or are within 1MB of a known GWAS

locus [33, 35–42], and 35 (40%) of which were identified by previous TWAS [43–48]. Of the 37 significant TWAS risk genes for ovarian cancer, 35 (95%) have either been previously identified by GWAS or are within 1MB of a known GWAS locus [34, 49, 50], and 13 (35%) have been identified by previous TWAS [22, 47, 51]. Additionally, we identified 3 novel independent significant risk genes (*KLHL25* of breast cancer; *UBE2MP1* and *FRG1EP* of both breast and ovarian cancer). Gene *KLHL25* has known biological functions involved in carcinogenesis, while genes *UBE2MP1* and *FRG1EP* are near such a gene [52–56].

The use of a single reference panel for GReX prediction model training is a major limitation of existing TWAS tools, including TIGAR-V2. There are multiple studies that generate both transcriptomic and genetic data of the same samples for the same tissue type. For example, the Religious Orders Study (ROS) [57, 58], Rush Memory and Aging Project (MAP) [58, 59], Mount Sinai Brain Bank (MSBB) [60], Mayo Clinic Brain Bank (MCBB) [61], and GTEx [24] studies all profile transcriptomic data of prefrontal cortex tissue. Leveraging multiple reference panels will increase the effective training sample size, thus leading to improved accuracy of predicting genetically regulated gene expression (GReX) and increased power of the follow-up TWAS. The idea is analogous to the meta-analysis of multiple GWAS cohorts [62, 63], but relaxes the assumption of same effect-size distribution per genetic variant across multiple cohorts made by single variant tests.

In part two, we develop a novel TWAS method to leverage multiple reference panels by ensemble machine learning [64]. We employ the ensemble machine learning technique of stacked regression [65, 66] to combine gene expression prediction models trained from multiple reference panels of the same tissue type, in order to improve the prediction accuracy of GReX and the power of TWAS. We refer this novel TWAS method as Stacked Regression based TWAS (i.e., SR-TWAS). We validated our SR-TWAS method by simulation studies using real whole genome sequencing (WGS)

genotype data from GTEx V8 [24] and ROS/MAP [57–59], as well as real TWAS leveraging reference panels of the GTEx V8 and ROS/MAP of brain frontal cortex tissue.

Under all simulation scenarios, SR-TWAS models had higher gene expression prediction accuracy and TWAS power than either of the models trained on single cohorts, which performed similarly. SR-TWAS achieves the greatest gains in power over single cohort models under scenarios in which a gene has a high proportion of true causal eQTLs with relatively small effect sizes. In the real data application to GTEx V8 and ROS/MAP, SR-TWAS performance was similar to the ROS model, but outperformed both single cohort models.

In Chapter 2, we first outline the TIGAR-V2 framework in Section 2.1.1. In the following sections, we describe the GTEx V8 dataset (Section 2.1.2) and the application of TIGAR-V2 to train gene expression prediction models with the GTEx V8 reference data (Section 2.1.3). In Section 2.1.1.2 we conduct an application TWAS of breast cancer and ovarian cancer from the trained GReX prediction models. Results of GReX prediction model training and application TWAS are described in Sections 2.2.1 and 2.2.2. We end Chapter 2 with a discussion in Section 2.3.

In Chapter 3, we discuss the stacked regression model (Section 3.1.1) and the SR-TWAS tool framework and implementation of the (Section 3.1.2). In Section 3.1.4 we describe the design of a simulation study used to asses the SR-TWAS method. Real data application studies using GTEx V8 and ROS/MAP reference panels are described in Section 3.1.5. Simulation study results are described in Section 3.2.1. Results of the application Studies leveraging GTEx V8 and ROS/MAP datasets are described in Section 3.2.2. The chapter ends with a discussion (Section 3.3). We provide our final conclusions regarding the work presented in this thesis in Chapter 4.

2 TIGAR-V2 with nonparametric Bayesian eQTL weights estimated from GTEx V8

In this chapter, we develop TIGAR-V2, train nonparametric Bayesian DPR GReX prediction models for 49 tissues from the GTEx V8 reference panel [24], and conduct an example application TWAS of breast and ovarian cancer.

2.1 Methods

2.1.1 TIGAR-V2

The standard two-stage TWAS [13, 16, 18] first fits gene expression prediction models by taking genotype data ($\mathbf{G}$) of *cis*-SNPs (within $\pm$1MB of the target gene g) as predictors, assuming the following additive genetic model for expression quantitative traits ($\mathbf{E}_g$) with respect to the target gene g,

$$\mathbf{E}_g = \mathbf{G}\mathbf{w} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 \mathbf{I}). \tag{2.1}$$

The *cis*-eQTL effect size vector $\mathbf{w}$ can be estimated by different regression methods from the reference training data. For example, PrediXcan estimates $\mathbf{w}$ by a general linear regression model with Elastic-Net penalty [13, 67]; FUSION estimates $\mathbf{w}$ by the Bayesian Sparse Linear Mixed Model (BSLMM) [16, 68]; and the initial TIGAR tool estimates $\mathbf{w}$ by a nonparametric Bayesian DPR model [18].

TIGAR-V2 implements both general linear regression with Elastic-Net penalty as used by PrediXcan [13] and nonparametric Bayesian DPR methods for estimating $\mathbf{w}$, where the nonparametric Bayesian DPR method includes both Elastic-Net and BSLMM as used by FUSION [16] as special cases [18, 31]. Additionally, TIGAR-V2 runs 5-fold cross validation with the reference data by default to provide an average

prediction R^2 per gene across 5 folds of validation data (referred to as 5-fold CV R^2). The 5-fold CV R^2 can be used to evaluate if the trained gene expression prediction model contains enough information for follow-up TWAS (e.g., using the threshold of 5-fold CV $R^2 > 0.005$).

2.1.1.1 Nonparametric Bayesian DPR Model. Given reference genotype matrix $\mathbf{G}_0$ of SNPs within g and within $\pm$1Mb of g and expression data $\mathbf{E}_g$, both centered at 0, the nonparametric Bayesian Dirichlet process regression model assumes

$$\mathbf{E}_g = \mathbf{G}_0\mathbf{w} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2_{\boldsymbol{\epsilon}}\mathbf{I}), \qquad \sigma^2_{\boldsymbol{\epsilon}} \sim \mathrm{IG}(a_{\boldsymbol{\epsilon}}, b_{\boldsymbol{\epsilon}}) \tag{2.2}$$

where $\mathbf{w}$ denotes the *cis*-eQTL effect-size vector for gene g. Genetically regulated gene expression of g for test genotype matrix $\mathbf{G}_t$ are then imputed by

$$\widehat{\mathbf{GReX}}_g = \mathbf{G}_t\widehat{\mathbf{w}} \tag{2.3}$$

Further, the effect-size of each *cis*-eQTL (w_{ig}; $i = 1, \ldots, p$) in gene g is assumed to follow a normal prior $\mathrm{N}(0, \sigma^2_w)$ with an effect-size variance σ^2_w that is assumed to follow a Dirichlet process (DP) prior D with an inverse gamma (IG) base distribution and concentration parameter ξ.

$$w_i \sim \mathrm{N}(0, \sigma^2_w), \qquad \sigma^2_w \sim D, \qquad D \sim \mathrm{DP}(\mathrm{IG}(a, b), \xi) \tag{2.4}$$

Effect-size variance σ^2_w can be treated as a latent variable that can be integrated out in order to derive an equivalent non-parametric prior distribution for w_{ig}

$$w_i \sim \sum_{k=1}^{\infty} \pi_k \mathrm{N}(0, \sigma^2_k), \quad \sigma^2_k \sim \mathrm{IG}(a_k, b_k), \quad \pi_k = v_k \prod_{l=1}^{k-1}(1 - v_l), \quad v_k \sim \mathrm{Beta}(1, \xi) \tag{2.5}$$

The resulting mixture normal prior is the weighted sum of an infinite number

of normal distributions $\mathrm{N}(0,\sigma_k^2)$ with weights π_k determined by v_l. The number of components with non-zero weights in the mixture normal prior is determined by concentration parameter ξ, which is assumed to follow a $\mathrm{Gamma}(a_\xi, b_\xi)$ hyper prior.

Non-informative priors for σ_k^2, σ_ϵ^2, ξ are induced by setting hyperparameters a_k, b_k, a_ϵ, b_ϵ, $b_\xi = 0.1$ and $a_\xi = 1$. A data-driven variational Bayesian algorithm, an approximation for the MCMC with greater computational efficiency, is then used to adaptively estimate the parameters σ_k^2, σ_ϵ^2, ξ and obtain the Bayesian posterior estimate for $\mathbf{w}$.

The DP normal mixture prior [29–31] used by DPR both covers the Elastic-Net [67] and BSLMM [68] models as special cases and lacks their shared limitation of assuming a parametric prior. The DPR model is close to the infinitesimal model, which assumes that a large number of variants each with small effect size contribute to phenotype [69, 70], and is preferred for modeling genes with many eQTLs of relatively small effect sizes [18]. The more flexible DPR model can robustly model the underlying complex genetic architecture of transcriptomes and improve GReX prediction accuracy [18].

2.1.1.2 TWAS. With the estimates of *cis*-eQTL effect sizes $\widehat{\mathbf{w}}$ and individual-level GWAS data of test samples, the standard two-stage TWAS would test the association between predicted GReX values given by $\widehat{\mathbf{GReX}}_g = \mathbf{G_t}\widehat{\mathbf{w}}$ and the phenotype of interest. TIGAR-V2 predicts GReX values by taking estimates of *cis*-eQTL effect sizes (outputs from training gene expression prediction model per gene with reference training data) and genotype data (VCF files) of test samples as inputs. TIGAR-V2 tests the association between $\widehat{\mathbf{GReX}}_g$ and the phenotype of interest (pedigree (PED) format as used by PLINK [71] and GATK [72]) based on the general linear regression model, with the phenotype as response variable and predicted GReX as an explanatory test variable.

With summary-level GWAS data (i.e., Z-score statistic values from single variant GWAS tests) of test samples, TIGAR-V2 tests the gene-based association by using both

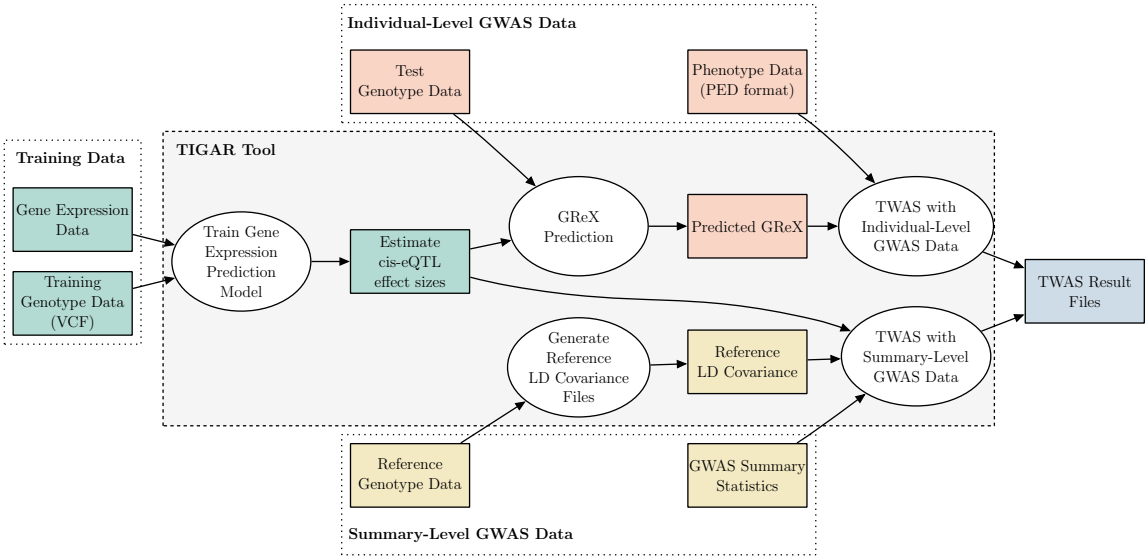
Burden [73, 74] and Variance-Component [32] test statistics, where *cis*-eQTL effect size estimates $\widehat{\mathbf{w}}$ are taken as variant weights. The required linkage disequilibrium (LD) covariance matrix among test *cis*-SNPs for Burden and Variance-Component tests can be obtained by TIGAR-V2 from reference genotype data ($\mathbf{G}_0$) or genotype data from other reference panels such as 1000 Genome [75]. For Burden test, FUSION Z-score statistic [16] as given by Equation 2.6 will be used if $\widehat{\mathbf{w}}$ are estimated using standardized training gene expression and genotype data, and S-PrediXcan test statistic [21] as given by Equation 2.7 will be used if $\widehat{\mathbf{w}}$ are estimated using only centered training gene expression and genotype data.

$$\widetilde{Z}_{g,\text{FUSION}} = \sum_{i \in \text{Model}_g} \frac{\widehat{w_i} Z_i}{\sqrt{\widehat{\mathbf{w}}' \mathbf{V} \widehat{\mathbf{w}}}}, \qquad \mathbf{V} = \text{Corr}(\mathbf{G}_0) \tag{2.6}$$

$$\widetilde{Z}_{g,\text{PrediXcan}} = \sum_{i \in \text{Model}_g} \frac{(\widehat{w_i}\widehat{\sigma_i}) Z_i}{\sqrt{\widehat{\mathbf{w}}' \mathbf{V} \widehat{\mathbf{w}}}}, \qquad \widehat{\sigma_i^2} = \text{Var}(x_i), \quad \mathbf{V} = \text{Cov}(\mathbf{G}_0) \tag{2.7}$$

2.1.1.3 Tool framework. The tool framework of TIGAR-V2 is shown in Figure 2.1, where all TWAS steps in TIGAR-V2 are enabled using Python and Bash scripts. Python libraries "pandas" [76, 77], "numpy" [76, 78], "scipy" [79], "sklearn" [80, 81], and "statsmodels" [82] are used to develop TIGAR-V2. Genotype data in VCF saved as one file per chromosome can be taken as input files for TIGAR-V2. TABIX tool [83] is used to extract genotype data per target gene efficiently from VCF genotype files. Parallel computation is enabled by using the "multiprocessing" Python library, allowing users to train gene expression prediction models and test gene-based association of multiple genes in parallel.

Figure 2.1: TIGAR-V2 framework including TWAS steps of training gene expression prediction models from reference data, predicting GReX with individual-level GWAS data, and testing gene-based association.



2.1.2 GTEx V8 Data

The Genotype-Tissue Expression (GTEx) project was created to further scientific study the relationship between genetic variation and gene expression by establishing a comprehensive reference catalog of genotype data linked to genome-wide gene expression patterns across multiple human tissues [10, 84, 85].

GTEx collects normal biospecimens from postmortem donors identified by partner organ procurement organizations and low-PMI (post-mortem-interval) autopsy progams [10, 86]. Donors between the ages of 21-70, who do not meet medical exclusion criteria are eligible only if biospecimin collection can begin within 24 hours of death [10, 86].

The version 8 release provides comprehensive profiling of whole genome sequencing (WGS) genotype data and RNA-sequencing (RNA-seq) transcriptomic data (15,253 normal samples) across 54 tissue types of 838 donors [10, 24, 84, 85]. The GTEx V8 data set provides ideal reference data for training tissue-specific gene expression prediction models for a diverse of tissue types on human bodies. Both PrediXcan

and FUSION tools use GTEx V8 data as the reference data, and provide estimated *cis*-eQTL effect sizes per gene with respect to 49 tissue types that have >70 samples with profiled WGS genotype and RNA-seq transcriptomic data (Figure 2.2A) as a public resource for TWAS.

WGS data of GTEx V8 donors were obtained through dbGaP with accession phs000424.v8.p2. Gene expression data of Transcripts Per Million (TPM) per sample per tissue were downloaded from the GTEx portal (`www.gtexportal.org`).

2.1.3 Training gene expression prediction models with GTEx V8

In TIGAR-V2, we fitted nonparametric Bayesian DPR models per gene with respect to the 49 tissue types described in Section 2.1.2. The *cis*-eQTL effect size estimates by nonparametric Bayesian DPR method can be used to test TWAS association and are shared with the public along with the tool.

Genotype data was filtered such that only variants with missing rate $< 20\%$, minor allele frequency > 0.01, and Hardy-Weinberg equilibrium p-value $> 10^{-5}$ were considered for fitting the gene expression prediction models. Genes with > 0.1 TPM in ≥ 10 samples were considered. Raw gene expression data (TPM) were then adjusted for age, body mass index (BMI), top five genotype principal components, and top probabilistic estimation of expression residuals (PEER) factors [87]. The raw gene expression data of Breast mammary tissue were further adjusted for *ESR1* expression following previous TWAS of breast cancer [47]. WGS genotype data of *cis*-SNPs within $\pm$1MB around gene transcription start sites (TSS) of the target gene were used as predictors. Five-fold cross validation was conducted by default to obtain 5-fold CV R^2 (the average GReX prediction R^2 of the 5-folds) per gene per tissue. Only significant gene expression prediction models with 5-fold CV $R^2 > 0.005$ were retained in the output files.

2.1.4 Application TWAS of Breast and Ovarian Cancer

We conducted application TWAS of breast cancer and ovarian cancer by using *cis*-eQTL effect sizes estimated from GTEx V8 [24] of breast mammary tissue and ovary tissue and summary-level GWAS data [33, 34]. The GWAS summary data of breast and ovarian cancer were obtained from the Breast Cancer Association Consortium (BCAC) with 122,977 cases and 105,974 controls of European ancestry [33] and the Ovarian Cancer Association Consortium (OCAC) with 22,406 cases and 40,941 controls of European ancestry [34], respectively.

We further compared TWAS results by using the S-PrediXcan [21] Burden-type test statistic (Equation 2.7) with variant weights given by *cis*-eQTL effect sizes estimated by DPR (i.e., TIGAR-V2) and Elastic-Net (i.e., PrediXcan) methods.

2.2 Results

2.2.1 Model Training Results

From the application studies with GTEx V8 reference data as described in Section 2.1.3, a total of 1,104,305 significant gene expression prediction models with 5-fold CV $R^2 > 0.005$ were successfully trained by TIGAR-V2 (using the nonparametric Bayesian DPR method) for transcripts on the autosomal chromosomes of 49 tissue types. On average, $\sim$22K gene expression prediction models were obtained per tissue type. The median of 5-fold CV R^2 and the median of training R^2 per tissue type were respectively presented in Figure 2.2(B, C). As expected, nonparametric Bayesian DPR method over-fitted gene expression prediction models with training data [18], resulting in inflated training R^2 as shown in Figure 2.2C. Whereas, 5-fold CV R^2 as shown in Figure 2.2B demonstrates the gene expression prediction performance with independent test data.

As shown in Figure 2.2(B, C), top median 5-fold CV R^2 across genome-wide transcripts were obtained for kidney cortex tissue (cyan bar), various brain tissues

(yellow bars), and uterus tissue (hot pink bar), which all have sample sizes ~ 100. Whereas, muscle skeletal, skin, and whole blood tissues that have relatively large sample sizes $400 \sim 600$ have median 5-fold CV $R^2 \approx 0.02$. Multiple factors might contribute to this: i) overfitting; ii) increased sample size may result in more successfully trained models for relatively low-heritability genes; iii) tissue types with small sample sizes are generally of limited accessibility in human donors which might have an overall higher heritability for gene expression quantitative traits. We trained both PrediXcan and TIGAR models on breast tissue data downsampled to $N = 140$ in order to assess the influence of sample size and cross validation threshold on model training results. Model training results for the downsampled breast data compared to model training results with breast ($N = 337$) and ovary ($N = 140$) tissue data (Figures A.1 and A.2) suggest that increased sample size mainly influences results for genes with relatively low heritability and shows potential overfitting issues for tissues with smaller sample sizes. The DPR method gives a higher prediction R^2 for low heritability genes but is likely subject to more overfitting.

The training computation costs in CPU hours per chromosome per tissue type with GTEx V8 reference data by TIGAR-V2 (using the nonparametric Bayesian DPR method) were shown in Figure 2.3. The computation cost per chromosome ranged from 5 CPU hours to over 474, with a median of 50.6 and mean of 69.1, which is mainly due to various numbers of transcripts (or genes) per chromosome and various sample sizes per tissue type. That is, with sample size ~ 300, the average computation time for training a nonparametric Bayesian DPR gene expression prediction model per transcript (or gene) with 5-fold cross-validation is ~ 4 minutes by TIGAR-V2. The computation complexity is linear with respect to training sample size.

Figure 2.2: Results of training gene expression prediction models with GTEx V8 reference data of 49 tissue types, by TIGAR-V2 using the nonparametric Bayesian DPR method. A) Number of subjects (i.e., sample size), B) Median 5-Fold CV R^2, and C) Median Training R^2, per tissue type.

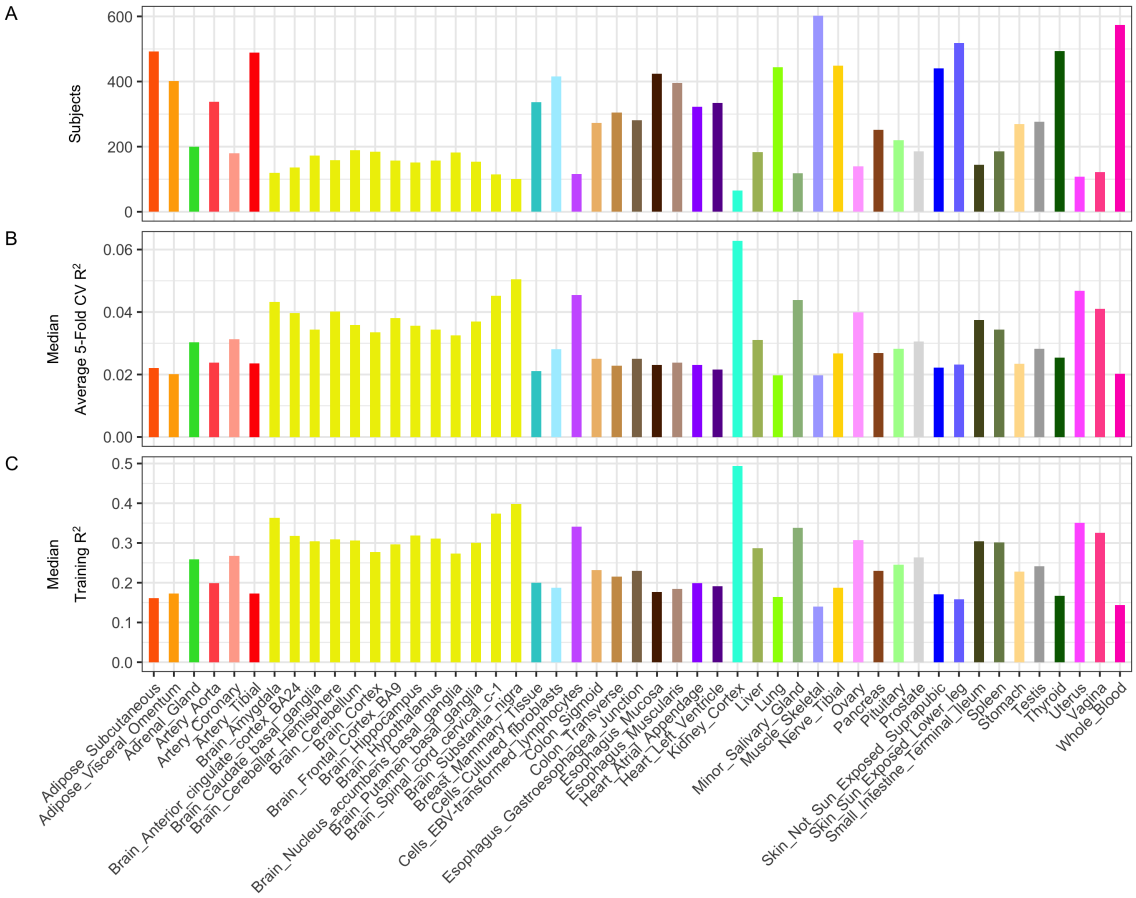
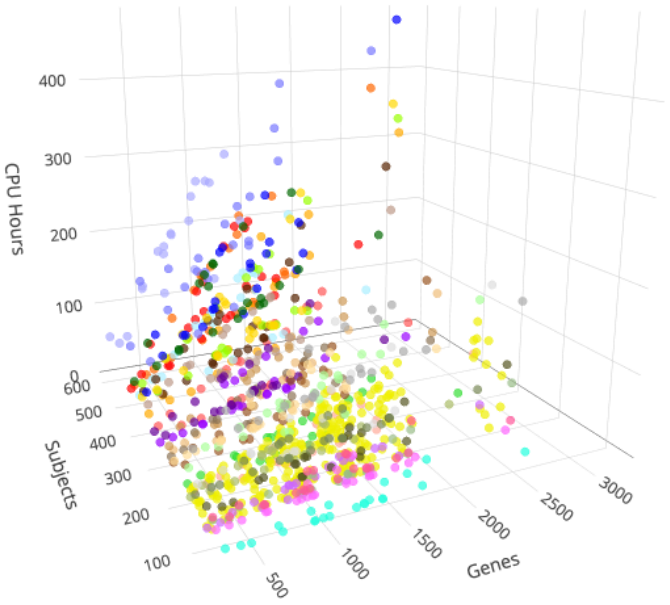


Figure 2.3: Computation costs in CPU Hours per chromosome per tissue type for training gene expression prediction models by TIGAR-V2 (using the nonparametric Bayesian DPR method). The same color codes with respect to different tissue types as used in Figure 2.2 were used here.



2.2.2 Application TWAS Results

From the model training results in the above Section 2.2.1 by TIGAR-V2, we obtained 22,781 and 22,823 significant gene expression prediction models by using the nonparametric Bayesian DPR method for breast ($N_{\text{training}} = 337$) and ovarian ($N_{\text{training}} = 140$) tissue types, respectively. Using GWAS summary statistics of breast cancer and ovarian cancer [33, 34] and *cis*-eQTL effect sizes estimated with respect to the corresponding tissue type, TIGAR-V2 detected 88 significant TWAS genes (p-values $< 2.5 \times 10^{-6}$) for breast cancer and 37 significant TWAS genes (p-values $< 2.5 \times 10^{-6}$) for ovarian cancer (Figure 2.4).

Out of these 88 significant TWAS genes for breast cancer by TIGAR-V2, 20 are known GWAS risk genes of breast cancer [33, 35–42] and 64 are located within 1MB region of a previously identified GWAS risk gene of breast cancer [33, 35–42]. Furthermore, 35 of these TWAS significant genes have been identified by previous

TWAS using PrediXcan and FUSION [43–48]. Similarly, out of these 37 significant TWAS genes for ovarian cancer by TIGAR-V2, 35 are located on chromosome 17 including two known GWAS risk genes (*NSF* and *PLEKHM1*) [34, 49, 50] and 33 genes within 1MB of these two known GWAS risk genes. Of these significant TWAS risk genes of ovarian cancer by TIGAR-V2, 13 (including *NSF* [51]) have been identified by previous TWAS using PrediXcan [22, 47, 51].

Since TWAS is conducted using genotype data within ±1MB region of the test gene (i.e., test region), genes with overlapped test regions often have highly correlated GReX values. Thus, these nearby significant TWAS genes are often not representing independent associations. In Tables 2.1 and 2.2, we listed the most significant genes among genes that have shared test regions, which represent the number of significant and potentially independent TWAS risk genes for breast cancer and ovarian cancer. All independent TWAS risk genes of breast cancer except *KLHL25*, *UBE2MP1*, and *FRG1EP* (31 significant genes; Table 2.1) were either identified by previous GWAS/TWAS or are within 1MB region of previously identified risk genes of breast cancer. For example, TIGAR-V2 identified *L3MBTL3* (previously identified by GWAS [33] and TWAS [44–46, 48]) and an additional 6 significant genes within the 1MB region of *L3MBTL3*. Of the independent TWAS genes of breast cancer, 17 (54%) have been identified by previous TWAS using PrediXcan and FUSION [43–48].

Similarly, as shown in Tables 2.2 and B.1, TIGAR-V2 identified 4 independent significant TWAS genes for ovarian cancer. In particular, TWAS risk gene *RP11-798G7.8* on chromosome 17 was identified by previous TWAS [47] and lies within 1MB of known GWAS risk gene *PLEKHM1* [34, 49]. Moreover, all independent TWAS risk genes of ovarian cancer by TIGAR-V2 (*PRC1-AS1*, *UBE2MP1*, *RP11-798G7.8*, and *FRG1EP*) are also TWAS risk genes of breast cancer [33, 47, 88], which demonstrates likely pleiotropy effect for these TWAS risk genes.

TIGAR-V2 identified 3 novel independent TWAS risk genes (*KLHL25*, *UBE2MP1*,

and *FRG1EP*) for breast cancer. Interestingly, genes *UBE2MP1* and *FRG1EP* were also identified for ovarian cancer by TIGAR-V2 (Table B.7), and all three genes are involved with biological functions in carcinogenesis, either directly or indirectly. The protein encoded by *KLHL25* was reported acting as an adaptor protein for a suspected lung cancer tumor-suppressing protein CUL3 to form an enzyme complex that targets ACLY, a protein often over-expressed in cancers, for degradation [54]. Pseudogene *UBE2MP1* was found to have a significant expression-methylation-correlation difference between normal and cancerous breast tissue [52]. *UBE2MP1* was also found to be amplified in gastric cancers with amplified copy number variations in the 16p11.2 region, a mutation found to be associated with shorter overall survival [56], and was predicted to be a driver of lung adenocarcinoma [55]. The test region of *FRG1EP* overlaps with the test region of pseudogene *ANKRD20A21P*, another TWAS risk gene identified by TIGAR-V2, which has been implicated as a potentially important lncRNA regulator of endometrial carcinogenesis [53].

Additionally, we compared our TWAS results by TIGAR-V2 with the ones obtained by PrediXcan that used *cis*-eQTL effect sizes estimated by Elastic-Net method (see Figures B.1, B.2 and Tables B.2-B.6). Quantile-quantile (QQ) plots of TWAS p-value results (Figure B.3) show similar inflation for TIGAR and PrediXcan models, which may be due to correlations with strongly associated genes, as many significant genes share test regions and are not independently significant. We showed that 37.5% independent significant TWAS genes by PrediXcan were also identified by TIGAR-V2. Whereas, only TIGAR-V2 identified TWAS genes *UBE2MP1* on chromosome 16 and *FRG1EP* on chromosome 20 of ovarian cancer, and known GWAS risk genes *FGF10* [33, 89] and *TOX3* [37, 39] of breast cancer. Other exclusive independent TWAS genes identified by TIGAR-V2 include lncRNA *RP11-758M4.4* which was shown to be a potential biomarker of breast cancer [90], *RPS23* which was found to be overexpressed in advanced colorectal adenocarcinomas [91], and *ZNF404* whose

dysregulation was linked to breast cancer pathogenesis by eQTL analyses [92, 93].

Interestingly, TWAS risk genes *PRC1-AS1* and *LRRC37A4P* were identified by both PrediXcan and TIGAR-V2 for both breast cancer and ovarian cancer. Gene *PRC1-AS1* on chromosome 15 is a long non-coding RNA (lncRNA) gene previously identified as being associated with breast carcinoma [33, 88]. Regulation of *PRC1-AS1* is known to differ with respect to different types of breast cancers [94] and increased expression of *PRC1-AS1* lncRNA is associated with hepatocellular carcinoma [95]. Pseudogene *LRRC37A4P* on chromosome 17 lies within 1MB downstream from the known risk gene *PLEKHM1* of breast cancer and ovarian cancer [34, 49].

Overall, these TWAS results not only validated our TIGAR-V2 tool with findings consistent with previous GWAS and TWAS of breast cancer and ovarian cancer, but also identified novel risk genes that were shown to be possibly involved in the biological mechanisms of oncogenesis.

Figure 2.4: Manhattan plots of TWAS results by TIGAR-V2 for studying breast cancer with 88 significant risk genes (A) and ovarian cancer with 37 significant risk genes (B). Significant gene *FCGR1B* of breast cancer (p-value: 4.12e-63) was removed from (A) to reduce the upper limit of the y-axis.

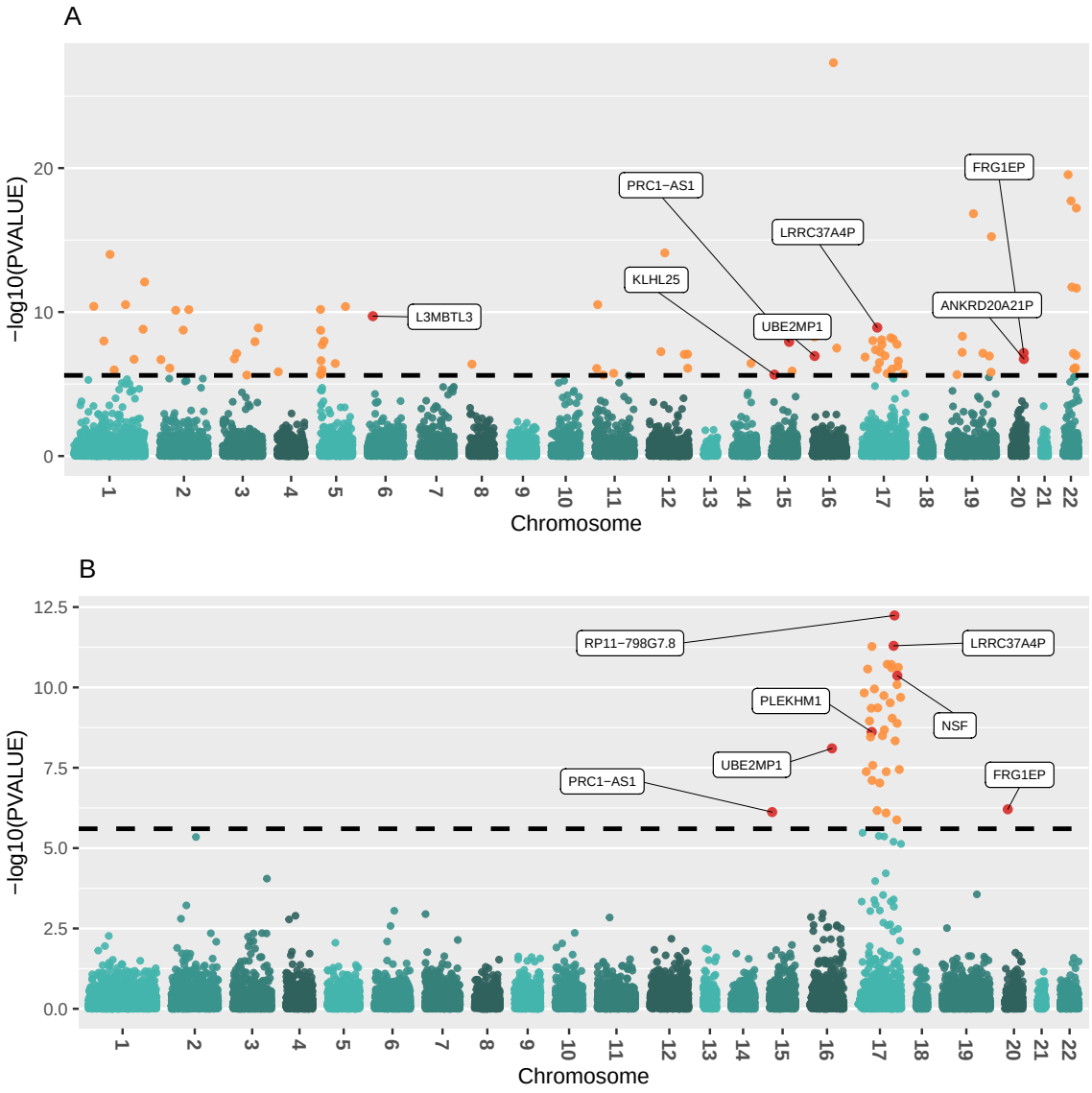


Table 2.1: Independent TWAS risk genes of breast cancer identified by TIGAR-V2.

Gene	Chrom	Start	End	Zscore	Pvalue
FCGR1B[b]	1	121087345	121096310	-16.77	4.12e-63
KLHDC7A[a]	1	18480982	18486126	-6.04	1.56e-09
MTX1P1[b]	1	155230975	155234325	5.21	1.92e-07
AC010136.2[b]	2	217978707	217992615	-6.52	6.80e-11
CASP8[a]	2	201233443	201287711	-6.51	7.56e-11
EOMES[b]	3	27715949	27722711	6.07	1.28e-09
PSMD6-AS2[b]	3	64004022	64012148	-5.38	7.50e-08
FAM114A1[b]	4	38867677	38945739	-4.82	1.41e-06
FGF10[a]	5	44303544	44389706	6.60	4.13e-11
SLC22A5[b]	5	132369752	132395614	6.53	6.63e-11
ANKRD55[b]	5	56099678	56233359	-5.63	1.85e-08
RPS23[b]	5	82273358	82278577	4.77	1.86e-06
L3MBTL3[a]	6	130013699	130141451	6.37	1.93e-10
RP11-758M4.4[b]	8	74798784	74866939	5.06	4.17e-07
PIDD1[a]	11	799191	809646	-6.64	3.04e-11
CCDC91[a]	12	28133249	28581511	-7.77	7.76e-15
RP11-116D17.4[b]	12	115318657	115320405	-5.36	8.40e-08
CTD-2325P2.4[b]	14	68627166	68628445	-5.09	3.65e-07
RCCD1[a]	15	90955796	90963125	-6.29	3.26e-10
MAN2C1[b]	15	75358201	75368154	-4.85	1.25e-06
KLHL25	15	85759323	85795030	-4.73	2.22e-06
TOX3[a]	16	52438005	52547802	10.98	4.82e-28
UBE2MP1	16	35169692	35170241	-5.31	1.13e-07
LRRC37A4P[b]	17	45506741	45550335	6.08	1.20e-09
CBX8[a]	17	79792132	79801683	5.76	8.46e-09
TOM1L1[b]	17	54899387	54960627	4.77	1.84e-06
SSBP4[a]	19	18418864	18434387	8.53	1.47e-17
ZNF404[b]	19	43872363	43884051	5.41	6.31e-08
FRG1EP	20	29480147	29497179	5.39	6.95e-08
DNAJB7[b]	22	40859549	40861617	-9.22	2.89e-20
TMEM184B[b]	22	38219291	38273034	4.92	8.72e-07

a known GWAS risk genes of breast cancer.

b genes within 1MB of known GWAS risk genes of breast cancer.

Table 2.2: Independent TWAS risk genes of ovarian cancer identified by TIGAR-V2.

Gene	Chrom	Start	End	Zscore	Pvalue
PRC1-AS1[a]	15	90972860	90988624	4.95	7.56e-07
UBE2MP1	16	35169692	35170241	5.77	7.88e-09
RP11-798G7.8[a]	17	45531577	45533838	-7.21	5.77e-13
FRG1EP	20	29480147	29497179	-4.99	6.19e-07

a genes within 1MB of known GWAS risk genes of ovarian cancer.

2.3 Discussion

In this chapter, we develop a new version of the TIGAR tool with improved computation efficiency [18], referred to as TIGAR-V2. This new version uses fewer Python library dependencies for easier set up, speeds up computation by using functions from the "numpy" Python library, reduces required memory usage by loading genotype data from VCF files in small block increments with pre-specified genotype data format, and adds the function to conduct the recently published Variance-Component gene-based association test [32]. For example, for DPR model training using the example data provided with the tool (129 samples, 4 genes, 1800-1891 SNPs per gene) and a single core, the computation time is reduced by up to 90% and memory usage by up to 50%, compared to the original TIGAR version. Gene expression prediction model training for the GTEx V8 data require less than 8GB of memory per transcript/gene.

TIGAR-V2 can efficiently train gene expression prediction models by using both Elastic-Net and nonparametric Bayesian DPR methods, as well as test gene-based association by using both Burden (FUSION [16] and S-PrediXcan [21] Z-score test statistics) and Variance-Component TWAS test statistics with both individual-level and summary-level GWAS data.

In addition, we trained gene expression prediction models with the reference GTEx V8 data by using the nonparametric Bayesian DPR method. We demonstrated the usefulness of these trained models along with estimated *cis*-eQTL effect sizes in an

application TWAS of breast and ovarian cancer by using GWAS summary statistics.

Our TWAS application studies by TIGAR-V2 identified 88 significant risk genes for breast cancer and 37 significant risk genes (four independent risk genes) for ovarian cancer, where the majority significant TWAS genes are either known GWAS risk genes or within 1MB region of known GWAS risk genes. In particular, these four independent risk genes of ovarian cancer were also identified as risk genes of breast cancer. These findings demonstrate potential pleiotropy effects shared with breast cancer is likely exist for the risk genes of ovarian cancer. Moreover, three novel risk genes were identified by TIGAR-V2 for breast cancer and two of these were also identified as novel risk genes for ovarian cancer.

3 Leveraging multiple reference panels to improve TWAS power by ensemble machine learning

In this chapter, we develop a novel TWAS method that can leverage multiple reference panels and thus improve TWAS power, by using the ensemble machine learning technique of stacked regression [65, 66] (referred to as SR-TWAS). The stacked regression learns linear combinations of gene expression predictors trained from different reference panels of the same tissue type to improve GReX prediction accuracy. We evaluated the SR-TWAS methods by both simulation and real studies.

3.1 Methods

3.1.1 Stacked Regression

Stacked regression is a machine learning method for forming linear combinations of different predictors to improve prediction accuracy [66]. The theoretical background for combining predictors rather than selecting a single best predictor is well-established and has been developed since the 1970s [66, 96, 97]. The "stacking" method of combining predictors originated in a 1991 paper by Wolpert, who described the concept as any scheme for feeding information from a set of cross-validated models to another before forming the final prediction in order to reduce prediction error [65]. Breiman further expanded the idea with stacked regression, a specific framework for combining the initial predictors by weighted average with coefficient constraints to control for multicolinearity [66].

Here, the eQTL results obtained per cohort can each be viewed as a trained predictor. Let $\widehat{\mathbf{w}}_k$ be the kth set of estimated eQTL effect sizes of gene g (e.g., posterior Bayesian estimates) from K different multi-omics datasets ($k = 1, \ldots, K$). Consider an independent multi-omics validation dataset with profiled gene expression

$\mathbf{E}_{vg}$ of gene g and genotype matrix $\mathbf{G}_v$. Then the predicted GReX of the validation samples is given by $\mathbf{G}_v\widehat{\mathbf{w}}_k$. We will solve for a set of optimal weights $\zeta_1,\ldots,\zeta_K$ such that the R^2 between the profiled gene expression $\mathbf{E}_{vg}$ and the weighted average of multiple trained predictors is maximized (ie $1-R^2$ is minimized). Then the loss function can be written

$$\underset{(\zeta_k;k=1,\ldots,K)}{\text{minimize}} \frac{\left\|\mathbf{E}_{vg}-\sum_{k=1}^{K}\zeta_k\mathbf{G}_v\widehat{\mathbf{w}}_k\right\|^2}{\left\|\mathbf{E}_{vg}-\bar{E}_{vg}\right\|^2}, \qquad \text{s.t. } \sum_{k=1}^{K}\zeta_k=1, \quad \zeta_k\in[0,1] \tag{3.1}$$

As a result, we will obtain a set of weights ζ_k for $k=1,\ldots,K$ and a final GReX prediction model $\widetilde{\mathbf{w}}$ for gene g given by the weighted average of the eQTL effect sizes of K trained models

$$\widetilde{\mathbf{w}}=\sum_{k=1}^{K}\zeta_k\widehat{\mathbf{w}}_k \tag{3.2}$$

Then the final predicted GReX for test genotype data $\mathbf{G}_t$ is given by

$$\widehat{\mathbf{GReX}}_g=\mathbf{G}_t\widetilde{\mathbf{w}} \tag{3.3}$$

This stacked regression technique has been shown to almost always obtain better prediction than a single prediction model [66].

3.1.2 SR-TWAS Tool Framework

SR-TWAS was designed to be compatible with the TIGAR-V2 tool framework as described in Section 2.1.1.3; it accepts models trained by TIGAR-V2 as input, imports utility functions from TIGAR-V2, and outputs model files which can be used as input for TIGAR-V2 GReX prediction and summary-level TWAS. Much of the structure of the SR-TWAS code was derived from existing TIGAR-V2 scriptsand it shares dependencies on TABIX [83] and the Python numpy [76, 78], pandas [76], scipy [79], and statsmodels [82] libraries. However, it was written in Python 3.6 (rather than

Python 3.5) in order to employ features offered in later releases of scikit-learn [80, 81], previously used by TIGAR-V2 only for cross validation and elastic-net training.

The SR-TWAS script utilizes scikit-learn's consistent, extensible interfaces for defining estimators and predictors and for initializing objects [81]. The script trains a stacked regression model using a modified version of scikit-learn's StackingRegressor class, which trains a final estimator from cross-validated predictions from base estimators fitted on the full design matrix. The script defines two custom classes to be used as input for the stacking regressor object: a base estimator class (WeightEstimator) which converts trained GReX prediction models into scikit-learn-compatible estimator objects and a final estimator class (ZetaEstimator) which obtains the values of $\zeta_1, \ldots, \zeta_K$ that minimize the loss function (Equation 3.1) under the constraints $\zeta_k \geq 0$ and $\sum_{k=1}^{K} \zeta_k = 1$ [66].

During the stacked regression, SNP minor allele frequencies and effect sizes for the specified target are first read from each of the K user-specified weight files. The SNPs are then matched to SNPs in the validation genotype data and filtered to exclude effect sizes of SNPs for which the difference between the MAF of the genotype data and the MAF from the corresponding weight file exceeds a user-specified MAF difference threshold. The effect sizes from each weight file are used to initialize K separate instances of the WeightEstimator class. These K WeightEstimator objects are used as base estimators and fit on genotype and expression data from the validation data.

Only SR-TWAS models trained from $K = 2$ base models are presented in the following sections. The code was designed to accept any $K \geq 2$, and while the stacked regression script has been primarily tested using $K = 2$ base models, preliminary testing with dummy weight files confirms it can train stacked regression models from $K = 3, 4, 5$ base models.

3.1.3 ROS/MAP Data

The Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) are two ongoing longitudinal, epidemiologic clinical-pathologic cohort studies of aging and Alzheimer's disease [57–59] collectively referred to as ROS/MAP [58]. ROS enrolls Catholic nuns, priests, and brothers from religious groups across the United States, primarily from communal living settings [57]. While the similar adult lifestyle of participants allows for more control of potential confounders such as education and socioeconomic status, it simultaneously limits the ability to study such variables [57, 59].

MAP was designed to complement and extend studies like ROS by including subjects from a wider range of life experiences, socioeconomic status, and educational attainment [57, 98] and recruits participants primarily from retirement communities in the Chicago area, but also subsidized housing, retirement homes, and through organizations serving minorities and low-income elderly [57, 98]. All participants in both studies are without known dementia and agree to annual clinical evaluations and brain donation upon death [57, 59, 98]. Similarity in study design and data collection procedures allows the ROS and MAP datasets to be merged for use in joint analyses [57, 99].

Quality-controlled ROS/MAP genotype data for European subjects [100] was used for both the real data application and simulation studies. The real data application used imputed dosage genotype data while the simulation study used WGS data of ROS/MAP samples to avoid possible heterogeneity due to different genotype assay technology. ROS/MAP data for both the application and simulation studies was lifted from human reference genome GRCh37 to GRCh38 in order to be compatible with GTEx V8 data.

3.1.4 Simulation Study Design

In this section, we describe in depth simulation studies under four different scenarios to assess the performance of weighted eQTL effect-sizes obtained from model ensemble compared to the constituent eQTL effect-sizes trained from single cohorts. Performance was assessed with respect to prediction imputation R^2 in the test data and the power of TWASs.

Simulations were conducted using real WGS genotype data for gene $ABCA7$ on chromosome 19 from 1665 (465 training, 400 validation, 800 testing) ROS/MAP participants and 465 GTEx participants. Genotype data for each cohort was filtered to include 2,202 SNPs which had minor allele frequency (MAF) $> 5\%$ and Hardy-Weinberg p-value $> 10^{-5}$ for both cohorts.

Under each scenario, the proportion of true causal SNPs is varied among values $p_{\text{causal}} = (0.001, 0.01, 0.05, 0.1)$ while the expression heritability (ie the proportion of expression variation attributable to genetic variation [69]) is held constant at $h_e^2 = 0.2$ and the phenotype heritability (ie the proportion of phenotype variation attributable to genetic variation [70]) h_p^2 is varied to ensure the follow-up TWAS power within a range of 25% to 90%.

For each scenario, the expression of $ABCA7$ is simulated 1,000 using a genotype matrix $\mathbf{G}^*$ of N_{causal} randomly chosen causal SNPs constructed from genotype data for all samplesand a matrix of effect sizes for those causal SNPs $\boldsymbol{\beta}_{N_{\text{causal}} \times 1000}$ generated such that each column $\boldsymbol{\beta}_{\bullet i} \sim N(0, 1)$. The gene expression $\mathbf{E}_i$ for the ith simulation is given by

$$\mathbf{E}_i = \gamma_i \mathbf{G}^* \boldsymbol{\beta}_{\bullet i} + \boldsymbol{\epsilon}_i, \qquad \gamma_i = \sqrt{\frac{h_e^2}{\text{Var}(\mathbf{G}^* \boldsymbol{\beta}_{\bullet i})}}, \qquad \boldsymbol{\epsilon}_i \sim \text{N}\left(0, \sqrt{1 - h_e^2}\right) \tag{3.4}$$

where γ_i is a scale factor to ensure the targeted h_e^2 value.

Per-cohort gene expression prediction models are then trained using the non-

parametric Bayesian model described in Section 2.1.1.1 and the genotype and simulated expression data of 465 ROS/MAP and 465 GTEx subjects. The SR-TWAS model for $K = 2$ predictors was then applied to the trained GTEx and ROS/MAP models using a validation cohort of 400 ROS/MAP subjects.

Once trained, each of the GTEx, ROS/MAP, and SR-TWAS models were used to obtain predicted gene expression values $\widehat{\mathbf{GReX}}_i$ from the genotype data of 800 ROS/MAP test samples. The true expression data for these samples was used to simulate phenotype values for the follow up TWAS. The phenotype vector $\mathbf{Y}_i$ for the ith target was simulated using the true simulated expression $\mathbf{E}_i$ of the using

$$\mathbf{Y}_i = \varphi_i \mathbf{E}_i + \boldsymbol{\epsilon}_i, \qquad \varphi_i = \sqrt{\frac{h_p^2}{\mathrm{Var}(\mathbf{E}_i)}}, \qquad \boldsymbol{\epsilon}_i \sim \mathrm{N}\Big(0, \sqrt{1 - h_p^2}\Big) \tag{3.5}$$

where φ_i is a scale factor to ensure the targeted h_p^2 value.

Finally, the GTEx, ROS/MAP, and SR-TWAS models were evaluated by expression prediction accuracy and TWAS power. The predicted gene expression $\widehat{\mathbf{GReX}}_i$ from each model GTEx, ROS/MAP, and SR-TWAS, was used to calculate expression prediction R^2 and phenotype prediction R^2 for the ith target

$$R^2_{E_i} = \mathrm{Cor}\Big(\mathbf{E}_i, \widehat{\mathbf{GReX}}_i\Big)^2 \tag{3.6}$$

Phenotype prediction R^2, and phenotype prediction p-value (ie TWAS power) was obtained from a simple linear regression of $\mathbf{Y}_i$ onto $\widehat{\mathbf{GReX}}_i$.

3.1.5 Application Studies Leveraging GTEx V8 and ROS/MAP Reference Panels

From the model training results described in above Section 2.2.1, we obtained 21,901 significant gene transcript expression prediction models by using the nonparametric Bayesian DPR method for brain frontal cortex BA9 ($N_{\text{training}} = 157$) tissue type.

Transcriptomic data for ROS samples ($N_{\text{training}} = 256$) were similarly used to train 14,957 significant nonparametric Bayesian gene expression prediction models from variants with minor allele frequency > 0.01 and Hardy-Weinberg equilibrium p-value $> 10^{-5}$.

The stacked regression model for $K = 2$ predictors (denoted SR-TWAS) was then applied to the trained GTEx and ROS models using a validation cohort of 121 MAP subjects. Trained models for each cohort GTEx, ROS, and SR-TWAS were then used to predict gene expression for 122 MAP test samples.

3.2 Results

3.2.1 Simulation Study Results

Density plots of the ROS/MAP ζ weight used to obtain the final SR-TWAS model from the the GTEx and ROS/MAP models are shown in Figure 3.2. Under a sparse cis-eQTL causality model with $p_{\text{causal}} = 0.001$, the majority of SR-TWAS models were derived from only one of the underlying base models and there was a slight preference for ROS/MAP-only SR-TWAS models. As the causal proportion increased, the less extreme weights were chosen and for the majority of SR-TWAS models, the contribution of each base model to the SR-TWAS model was equal or approximately equal.

As shown in the plot of average expression prediction R^2 by proportion of causal SNPs (Figure 3.3), the SR-TWAS models out-performed both of the single cohort models in predicting genetically regulated gene expression. Prediction accuracy for all models decreased as p_{causal} was increased. However, the magnitude of the difference in performance between the SR-TWAS and single cohort models was much greater in scenarios with $p_{\text{causal}} > 0.001$.

TWAS power results for each scenario are shown in Figure 3.4. In every scenario, the TWAS with SR-TWAS models had achieved higher TWAS power than either of

the single cohort models, which performed similarly. The difference in performance between the SR-TWAS and single cohort models was minimal under a sparse cis-eQTL causality model with $p_{\text{causal}} = 0.001$. The power improvement of the SR-TWAS models over the single cohort models is greatest when the proportion of causal SNPs is high and the phenotype heritability is low. The TWAS power results complement the prediction R^2 results and are consistent with previous findings that better imputation R^2 results in higher TWAS power [18].

Effect-sizes were assumed to be constant across training cohorts. The performance of SR-TWAS when effect-sizes are heterogeneous between training cohorts is an anticipated area of follow up analysis. However, homogeneity in base models is expected to minimize improvement by stacked regression [66]. In the analyses presented in the original stacked regression paper, the largest gains in performance occurred when dissimilar base models were used [66]. Furthermore, stacking never resulted in a worse prediction performance than selecting the single best predictor [66]. Stacking minimizes prediction error on the validation data [65, 66], so we expect the method to be robust to effect-size heterogeneity when validation and test cohorts are similar.

Figure 3.1: Plots of average CV R^2 and training R^2 for simulations under four different scenarios with varying proportion of true causal SNPs $p_{\text{causal}} = (0.001, 0.01, 0.05, 0.1)$ and true expression heritability $h^2_e = 0.2$. Dotted lines denote the R^2 density of single cohort base models validation data during the SR-TWAS training.

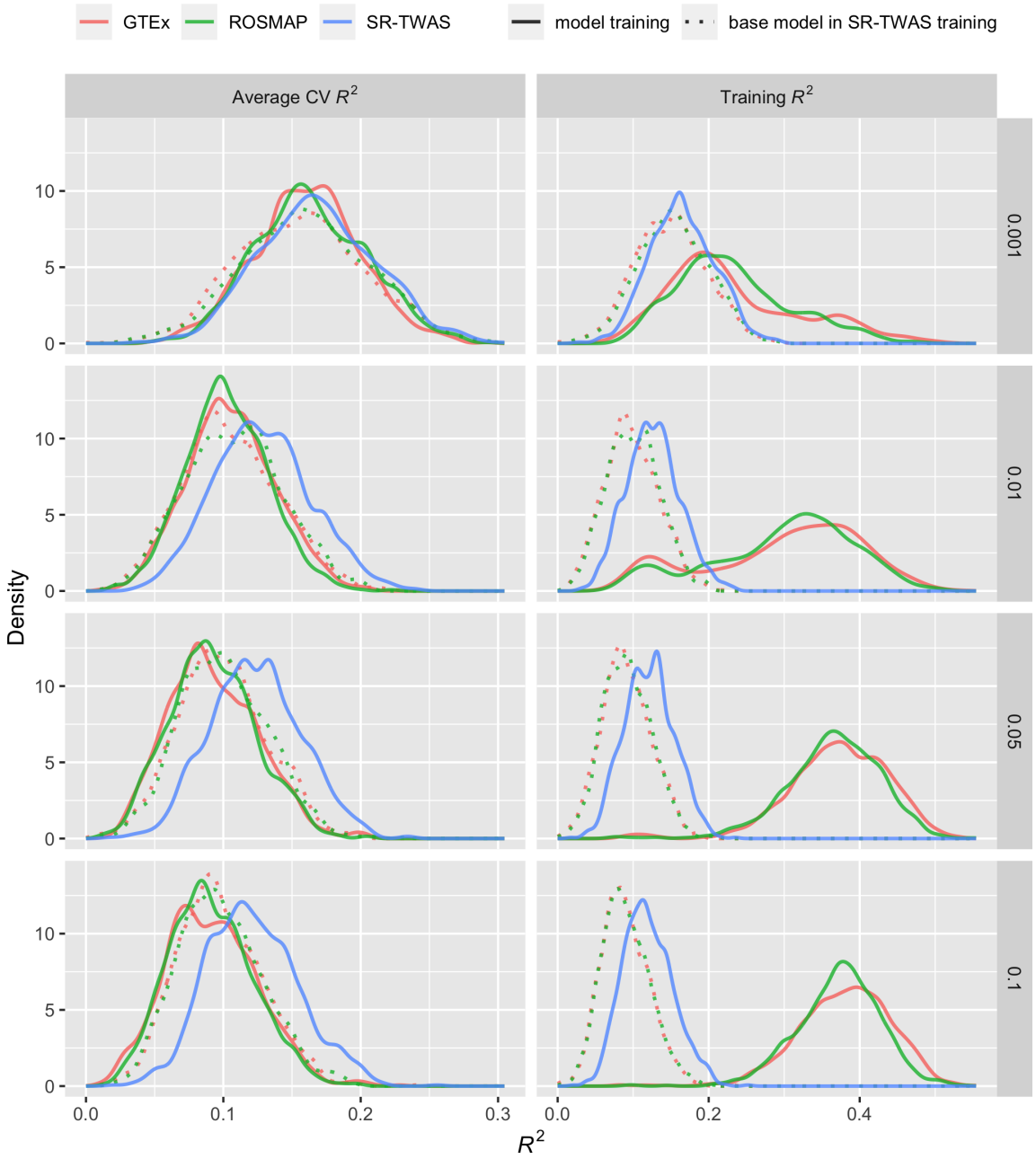


Figure 3.2: Plots of ROS/MAP ζ density for simulations under four different scenarios with varying proportion of true causal SNPs $p_{\text{causal}} = (0.001, 0.01, 0.05, 0.1)$ and true expression heritability $h_e^2 = 0.2$.

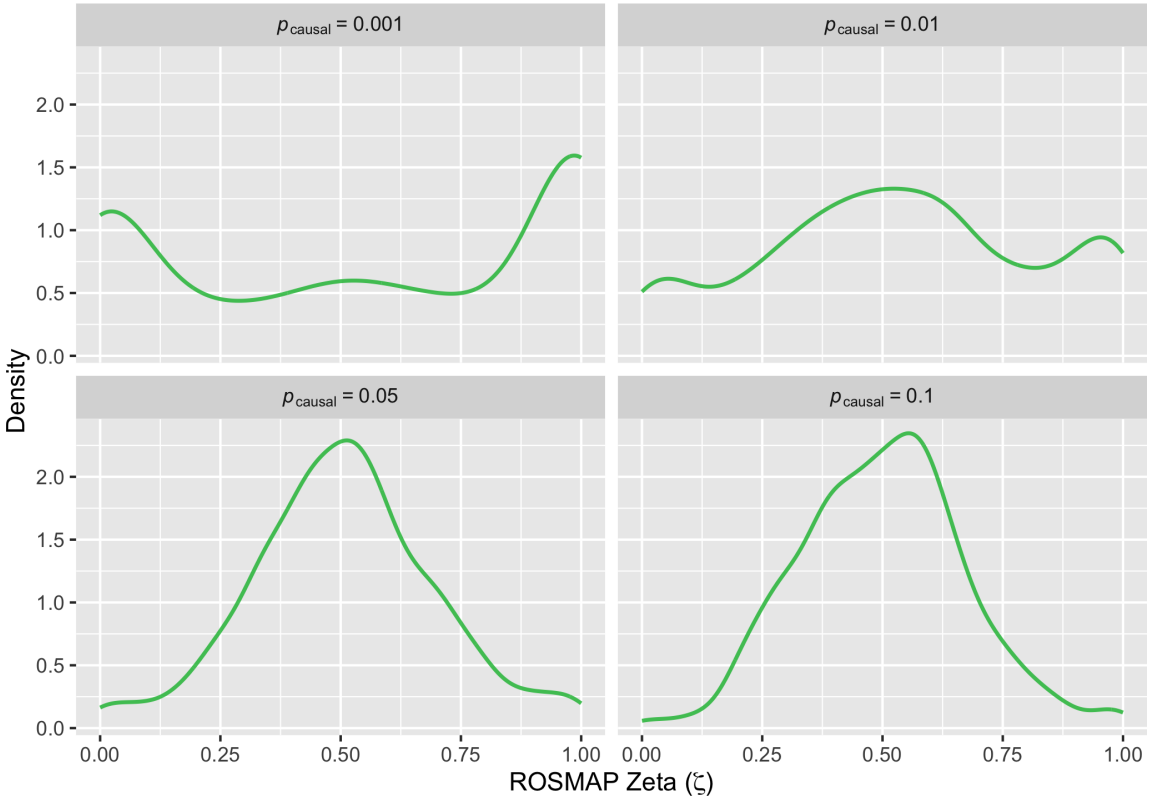


Figure 3.3: Plots of average expression prediction R^2 for simulations under four different scenarios with varying proportion of true causal SNPs $p_{\text{causal}} = (0.001, 0.01, 0.05, 0.1)$ and true expression heritability $h_e^2 = 0.2$.

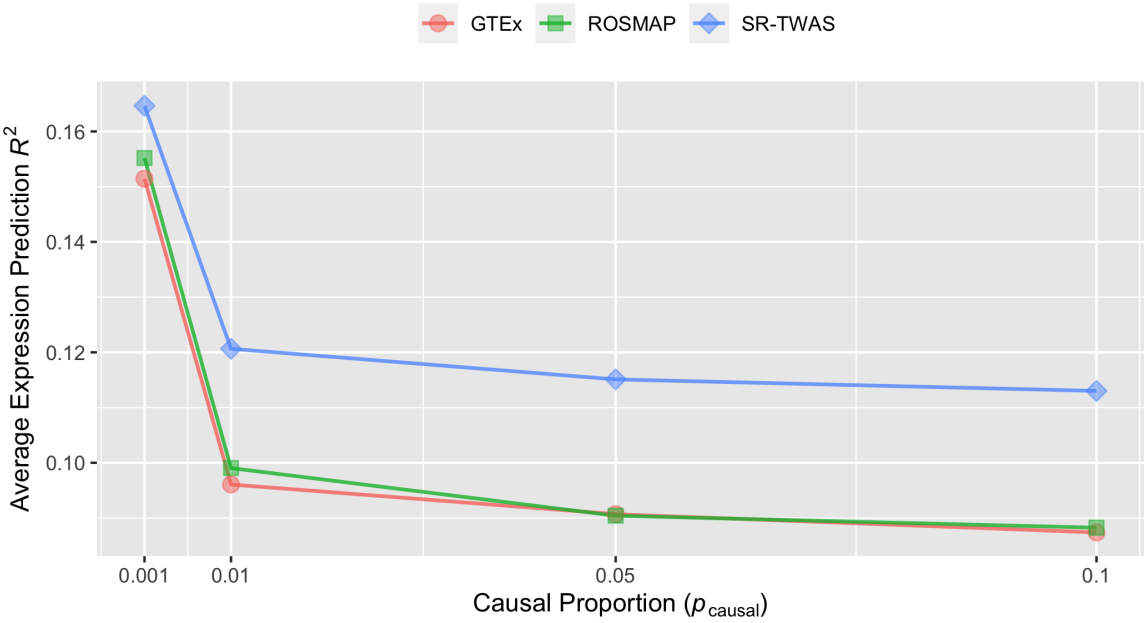
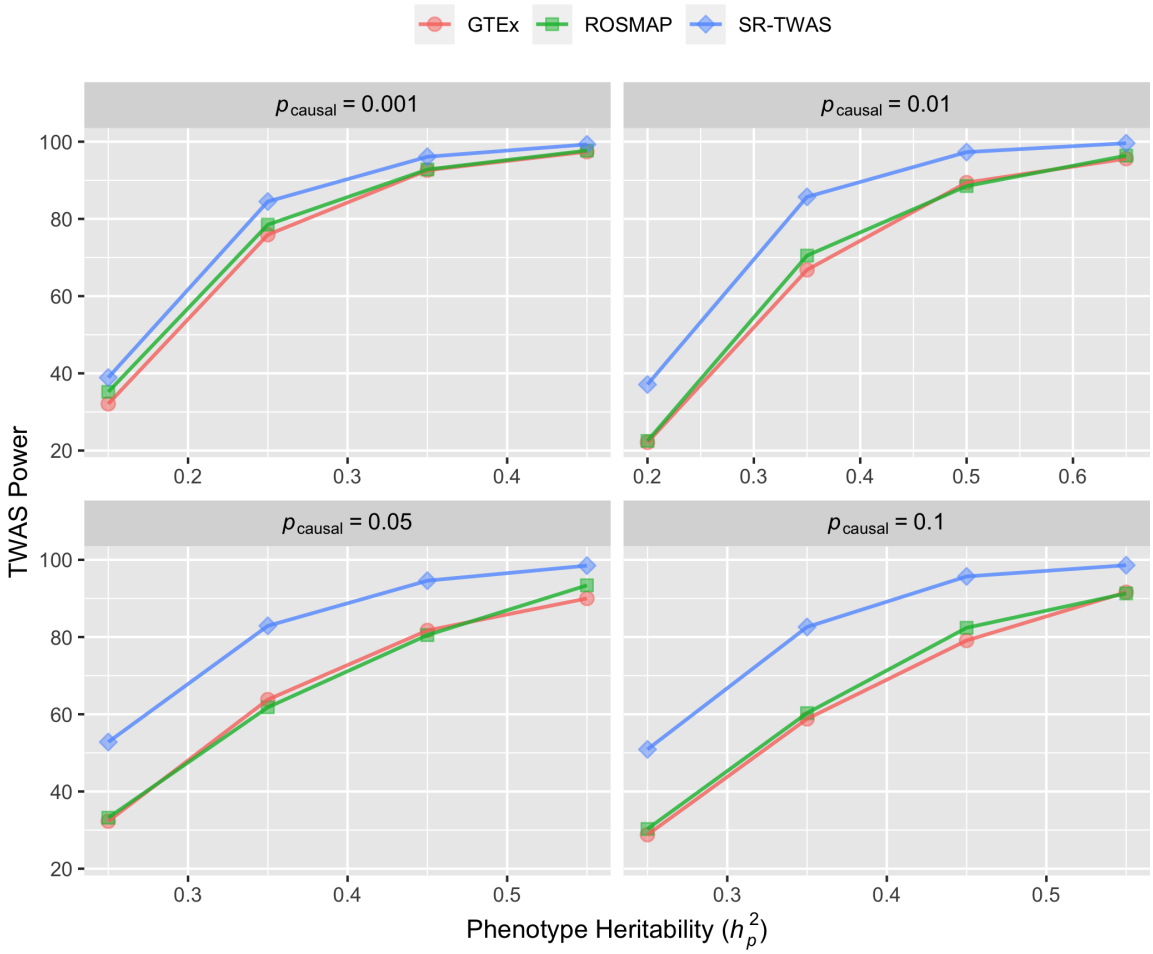


Figure 3.4: Plots of TWAS power for simulations under four different scenarios with varying proportion of true causal SNPs $p_{\text{causal}} = (0.001, 0.01, 0.05, 0.1)$, true expression heritability $h_e^2 = 0.2$, and phenotype heritability h_p^2 chosen for power in range of approximately 25% to 90%.



3.2.2 Application Studies Leveraging GTEx V8 and ROS/MAP Reference Panels

The training computation costs in CPU hours per chromosome with GTEx V8 and ROS base models by SR-TWAS were shown in Figure 3.5. The computation cost per chromosome ranged from 2 to 26.5 CPU hours, with a median of 8.9 and mean of 10.9, which is mainly due to various numbers of transcripts (or genes) per chromosome. With 121 validation samples, the average computation time for training a SR-TWAS gene expression prediction model per transcript (or gene) with 5-fold cross-validation is ~ 35 seconds.

Density plots of the ROS ζ weight used to obtain the final SR-TWAS model from the the GTEx and ROS models are shown in Figure 3.6. The majority of SR-TWAS models were derived from only one of the underlying base models, more frequently ROS than GTEx. The shape is similar to that of the ζ density curve for the sparse cis-eQTL causality model simulation scenario in the first panel of Figure 3.2.

As shown in Figure 3.7, SR-TWAS and ROS models show similar performance in gene expression prediction results and both outperformed the GTEx models. SR-TWAS performance is comparable, but slightly better than the ROS models. Results from Table 3.1 of median and mean prediction R^2 values are similar with SR-TWAS showing the highest (median 0.86%; mean 3.74%) R^2 values comparable to ROS results (median 0.78%; mean 3.49%), with both greatly outperforming the GTEx model (median 0.33%; mean 1.71%).

Comparisons of the SR-TWAS models with each of the two single cohort models gene expression prediction R^2 results are given in Figure 3.8 and Table 3.2. In Figure 3.8 prediction R^2 values for SR-TWAS are plotted against single cohort prediction R^2 for each gene. SR-TWAS performs noticeably better than GTEx, with the majority of genes plotted above the diagonal. SR-TWAS performance is again comparable with ROS model performance. Table 3.2 numerically describes the results plotted in Figure 3.8; when prediction $R^2 > 0.005$ for both models, SR-TWAS R^2 is greater than the single cohort model R^2 for the majority of genes. Pairwise comparisons of R^2 values for genes with $R^2 > 0.005$ for both models are shown in Table 3.3.

The similar performance of the SR-TWAS and ROS models compared to GTEx may be due to features of the training and prediction datasets. The GTEx models were trained from 157 brain frontal cortex BA9 samples while ROS models were trained from 256. Study demographics and tissue collection procedures are dissimilar for ROS/MAP compared to GTEx. The GTEx Project was established to study the effect of genetic variation on gene expression in normal human tissues [10, 86] while

ROS and MAP were designed to study aging and dementia [57–59]. The difference in objectives between the studies is reflected in the donor demographics. The mean age of ROS/MAP donors with RNA-Seq data is 86.7 [99], but GTEx does not enroll donors over the age of 70 [86]. ROS/MAP data may also be more heterogeneous than GTEx data. The ROS/MAP studies include data from specimens with neuropathology [99], while GTEx medically excludes some donors to avoid collecting diseased tissue [10, 86] and conducts pathology reviews of all collected biospecimens to ensure they are non-diseased [86].

Figure 3.5: Computation costs in CPU Hours per chromosome for training gene expression prediction models by SR-TWAS with ROS and GTEx V8 base models.

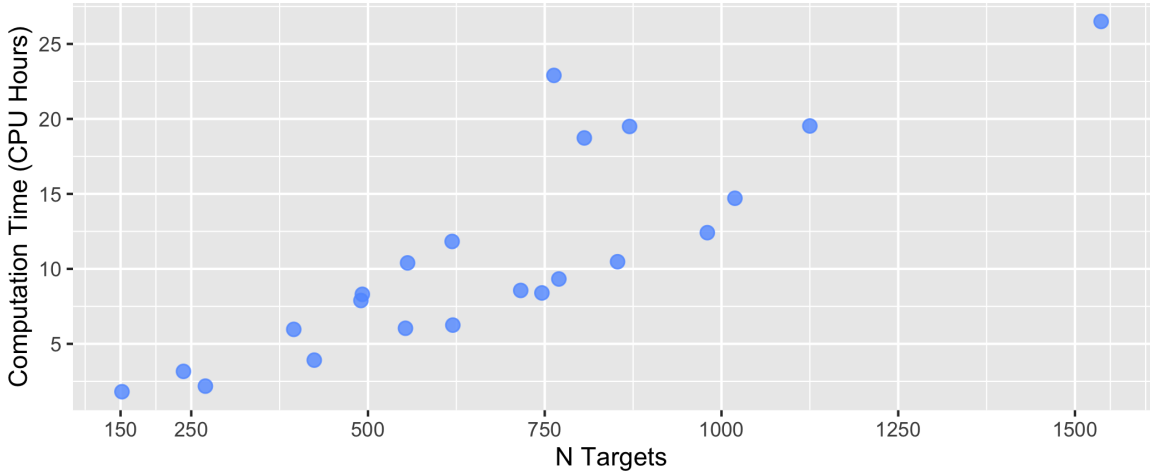


Figure 3.6: Plot of ROS ζ density for SR-TWAS of real data.

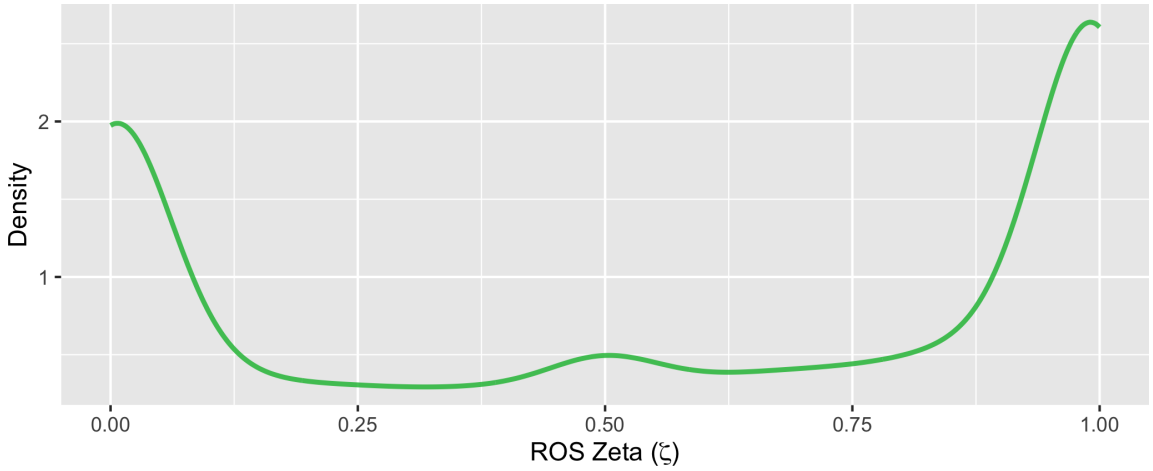


Figure 3.7: Plot of scaled prediction R^2 density for SR-TWAS of real data for 8337 genes where prediction $R^2 > 0.005$ for SR-TWAS and either single cohort model.

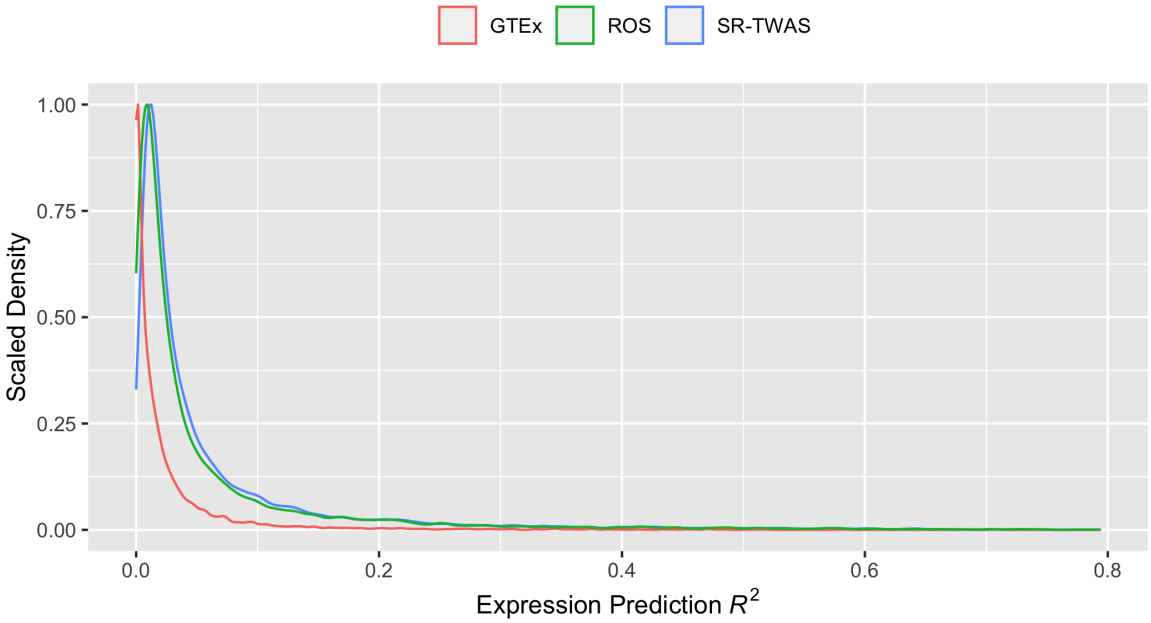


Table 3.1: Prediction R^2 results for SR-TWAS vs single cohort models.

	Median R^2	Mean R^2
GTEx	0.33%	1.71%
ROS	0.78%	3.49%
SR-TWAS	0.86%	3.73%

Figure 3.8: Gene expression prediction R^2 results for SR-TWAS vs single cohort models.

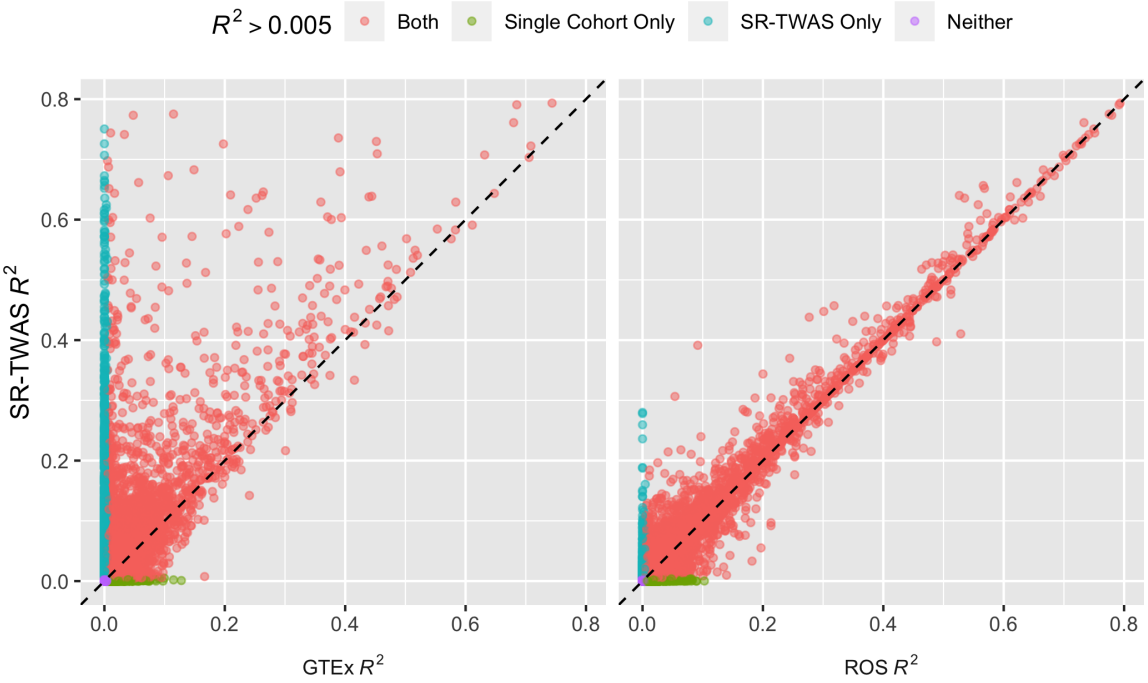


Table 3.2: Number of genes with prediction $R^2 > 0.005$ for both models that are above/below the diagonal in Figure 3.8.

	GTEx	ROS
above diagonal	3149 (64%)	4087 (56%)
below diagonal	1734 (36%)	3204 (44%)

Table 3.3: Pairwise comparison of model prediction R^2 by number of genes with $R^2_{\text{row}} \geq R^2_{\text{col}}$ and $R^2 > 0.005$ for both models. The diagonal shows the total number of genes with prediction $R^2 > 0.005$ for that model.

$R^2_{\text{row}} \geq R^2_{\text{col}}$	SR-TWAS	GTEx	ROS
SR-TWAS	8979	1734	3204
GTEx	3149	6579	2669
ROS	4087	1639	8722

3.3 Discussion

In this chapter, we present SR-TWAS–a novel tool for leveraging multiple transcriptomic reference panels of the same tissue type by ensemble machine learning

technique of stacked regression [64–66]. The power advantage of SR-TWAS model over single cohort models was demonstrated in both simulation studies and application to real data. SR-TWAS models had higher gene expression prediction accuracy and TWAS power under all simulations scenarios. In the application studies leveraging GTEx V8 and ROS/MAP reference panels, SR-TWAS achieved higher prediction R^2 than the underlying base models.

4 Conclusion

In this work we present and validate the TIGAR-V2 and SR-TWAS software tools for mapping TWAS risk genes of complex diseases.

TIGAR-V2 tool has its limitations such as using only *cis*-eQTL data, assuming a two-stage model for TWAS, and only testing a single phenotype. SR-TWAS shares the TIGAR-V2 limitations of using only *cis*-eQTL data in model training and assuming a two-stage model TWAS and requires an additional validation dataset independent of those used for base model training. Due to heterogeneity in genetic and transcriptomic data between populations of different ancestry, TIGAR-V2 and SR-TWAS model performance may be reduced when the test data is not derived from the same population used in model training or SR-TWAS validation [101].

There are many other useful TWAS tools available. For example, BGW-TWAS [19] uses both *cis*- and *trans*- genotype data to train gene expression prediction model of the target gene, CoMM [102] and PMR-Egger [20] assume a joint model with reference and test data that can achieve higher power when both data sets are homogeneous, and moPMR-Egger [103] tests the gene-based association with respect to multiple phenotypes. Overall, recent TWAS of complex diseases using these tools show promising results such as finding an increasing number of risk genes and revealing potential mediation effects through transcriptome and pleiotropy effects of these TWAS risk genes [23, 25, 46, 47].

TIGAR-V2 tool along with *cis*-eQTL effect sizes estimated by nonparametric Bayesian DPR methods with the GTEx V8 reference data are shared with the public on GitHub, `https://github.com/yanglab-emory/TIGAR`. We will also share LD reference covariance data files obtained from Europeans with WGS genotype data in the GTEx V8 data set, which are required for conducting TWAS using *cis*-eQTL effect

size estimates and GWAS summary-level data. The SR-TWAS tool is also publicly available on GitHub, `https://github.com/yanglab-emory/SR-TWAS`.

TIGAR-V2 and SR-TWAS tools implement user-friendly features, including accepting standard VCF-format genotype data as input, enabling parallel computation, and using efficient computation strategies to reduce time and memory usage. The tools are flexible; TIGAR-V2 provides users with options for using different training models and TWAS test statistics. While only SR-TWAS models with $K = 2$ base models are presented here, the SR-TWAS tool allows users to input additional trained base models. We believe our improved TIGAR-V2 and SR-TWAS tools will provide a useful resource for mapping risk genes of complex diseases by TWAS.

Appendix A: Downsampled Study Results

Figure A.1: Density plots of average 5-fold CV R^2 for TIGAR and PrediXcan model training for Ovary ($N = 140$), Breast ($N = 337$), and downsampled Breast ($N = 140$) tissues with varying CV R^2 threshold values: $(0.005, 0.01, 0.05, 0.1, 0.2)$.

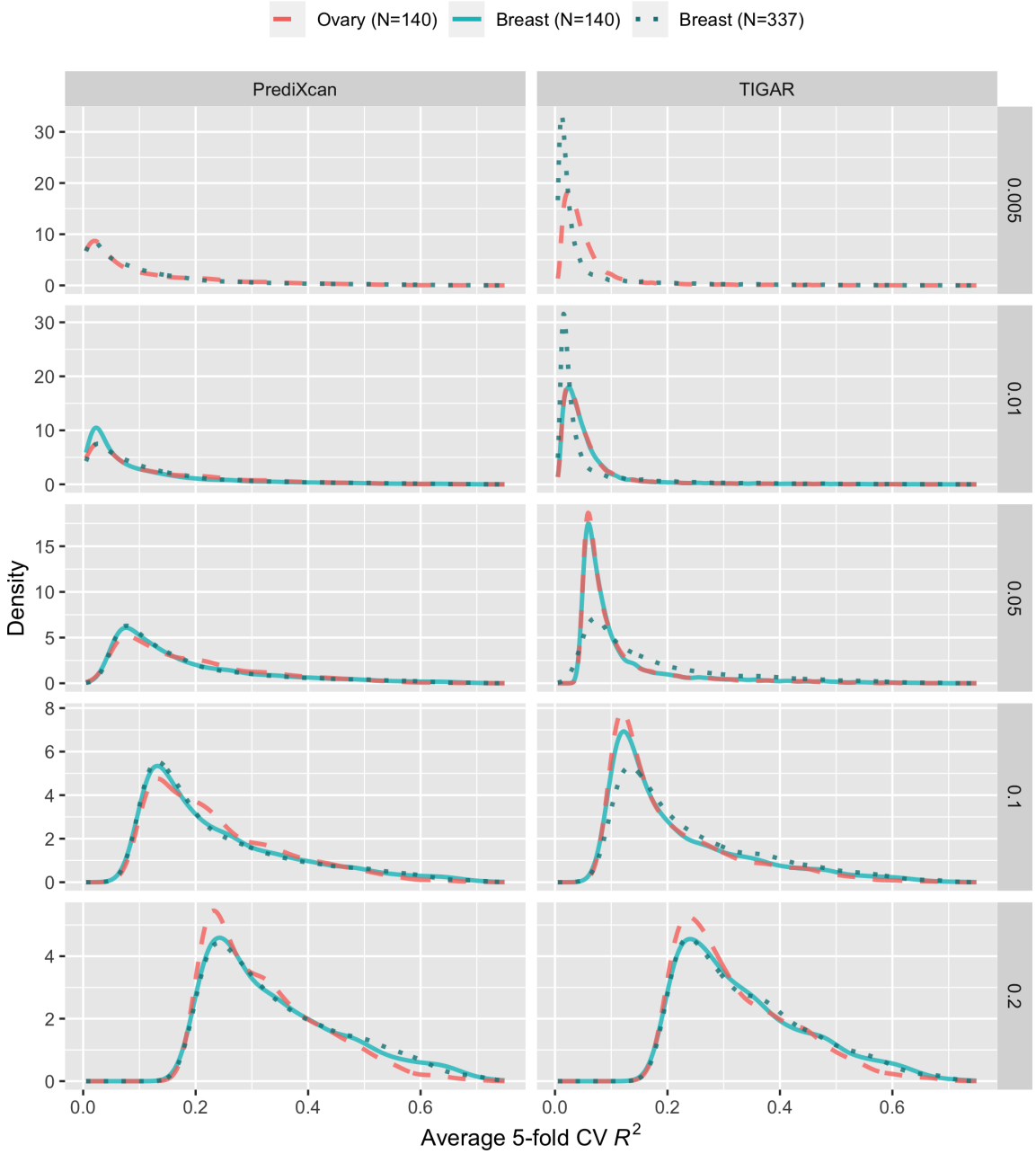
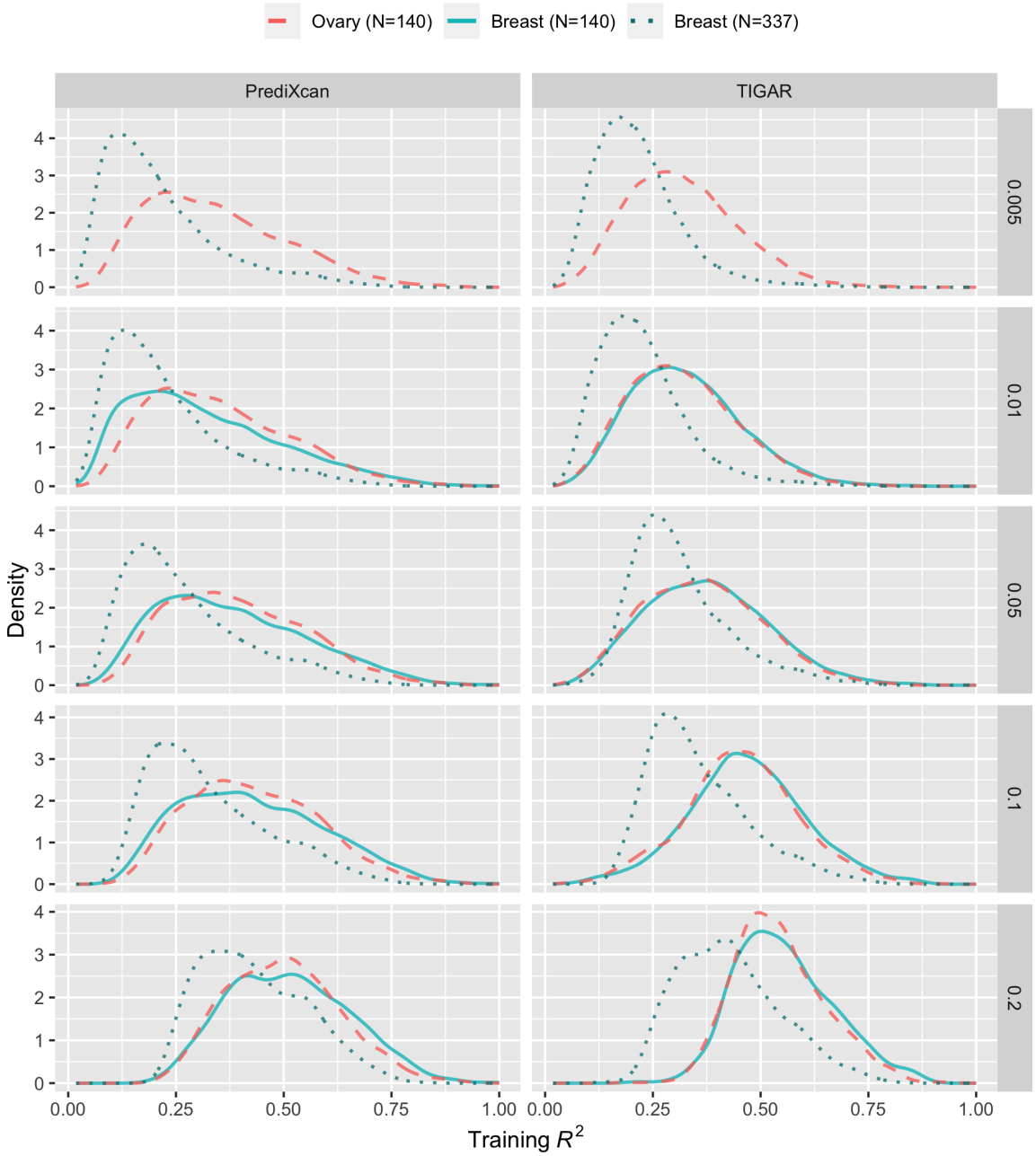


Figure A.2: Density plots of training R^2 for TIGAR and PrediXcan model training for Ovary ($N = 140$), Breast ($N = 337$), and downsampled Breast ($N = 140$) tissues with varying average 5-fold CV R^2 threshold values: $(0.005, 0.01, 0.05, 0.1, 0.2)$.



Appendix B: Application Breast and Ovarian Cancer TWAS Results

B.1 TIGAR Results

Table B.1: TWAS risk genes of both breast and ovarian cancer identified by TIGAR-V2.

				Breast		Ovary	
Gene	Chrom	Start	End	Zscore	Pvalue	Zscore	Pvalue
PRC1-AS1[a,c]	15	90972860	90988624	5.70	1.17e-08	4.95	7.56e-07
UBE2MP1	16	35169692	35170241	-5.31	1.13e-07	5.77	7.88e-09
ARHGAP27[a,c]	17	45393902	45434421	-5.27	1.33e-07	6.24	4.32e-10
AC091132.1[b,c]	17	45452844	45464065	-5.31	1.10e-07	-6.36	2.04e-10
LRRC37A4P[b,c]	17	45506741	45550335	6.08	1.20e-09	6.90	5.07e-12
DND1P1[b,c]	17	45585871	45586929	-5.81	6.11e-09	-6.90	5.31e-12
RP11-707O23.1[b,c]	17	45592621	45593369	-5.63	1.81e-08	-6.66	2.68e-11
MAPK8IP1P2[b,c]	17	45600869	45602340	-5.78	7.48e-09	-6.71	1.91e-11
LINC02210[b,c]	17	45620328	45655156	-5.48	4.23e-08	-6.50	8.23e-11
CRHR1[b,c]	17	45784280	45835828	-5.63	1.75e-08	-6.71	1.93e-11
MAPT[a,c]	17	45894382	46028334	-5.16	2.51e-07	6.24	4.45e-10
KANSL1-AS1[b,c]	17	46193576	46196723	-4.90	9.57e-07	-6.07	1.32e-09
RP11-259G18.3[b,c]	17	46259551	46260606	-4.99	6.03e-07	-6.09	1.11e-09
RP11-259G18.1[b,c]	17	46267037	46268694	-5.42	5.99e-08	-6.41	1.49e-10
LRRC37A2[b,c]	17	46511511	46553449	-5.11	3.24e-07	-6.12	9.08e-10
FAM215B[b,c]	17	46558830	46562795	-4.75	2.01e-06	-5.99	2.09e-09
FRG1EP	20	29480147	29497179	5.39	6.95e-08	-4.99	6.19e-07

a gene identified in previous breast cancer GWAS

b within 1MB of gene identified in previous breast cancer GWAS

c within 1MB of gene identified in previous ovarian cancer GWAS

B.2 PrediXcan Results

Figure B.1: Manhattan plots of TWAS results by PrediXcan for studying breast cancer with 56 significant genes (A) and ovarian cancer with 4 significant genes (B).

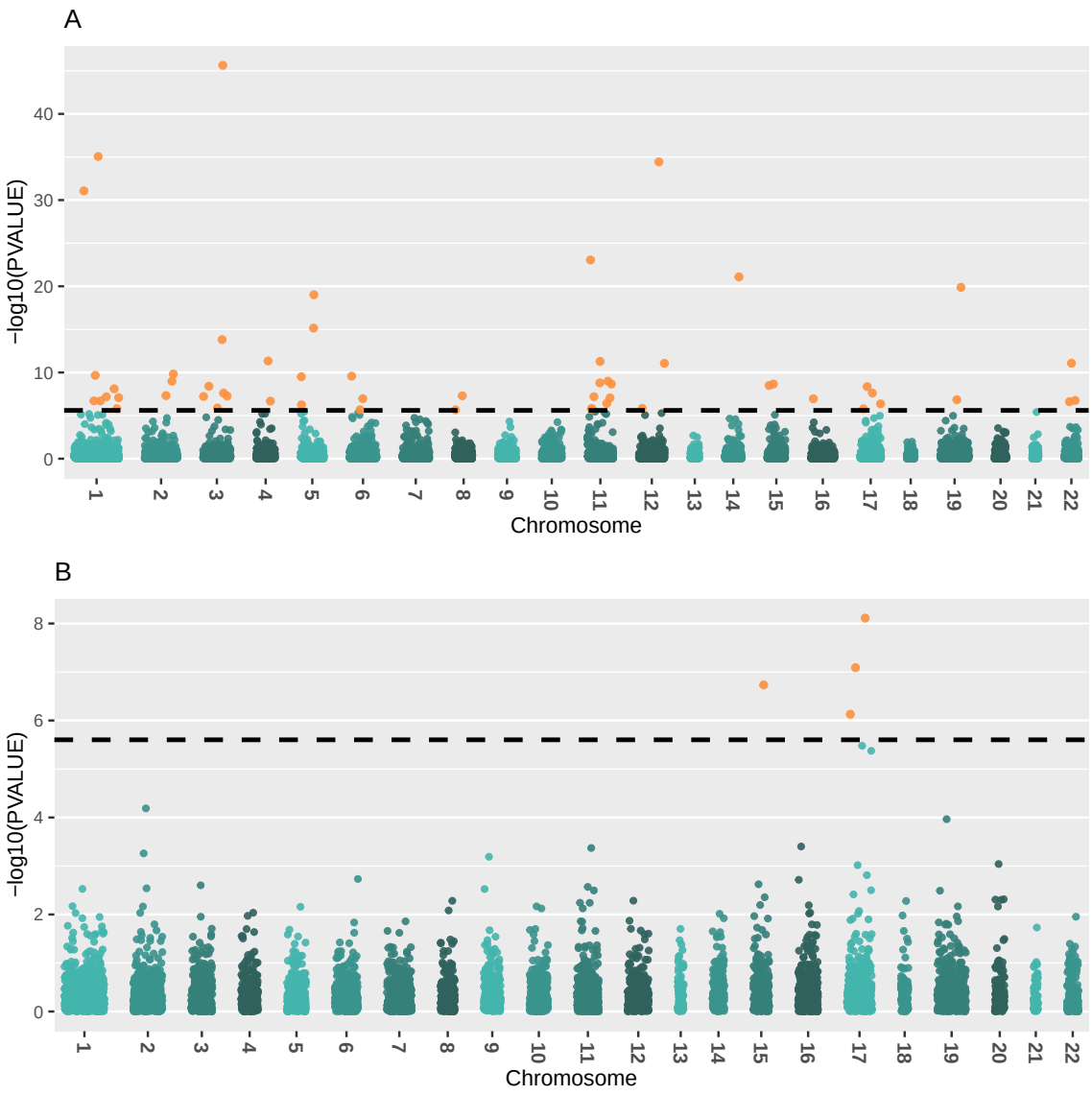


Table B.2: Independent TWAS risk genes of breast cancer identified by PrediXcan.

Gene	Chrom	Start	End	Zscore	Pvalue
CH17-437K3.1[b]	1	121396754	121463129	12.49	8.75e-36
ASH1L[a]	1	155335268	155562807	-6.35	2.16e-10
NSUN4[b]	1	46340177	46365152	5.77	7.81e-09
KLHDC7A[a]	1	18480982	18486126	-5.20	2.01e-07
ALS2CR12[b]	2	201288271	201357398	6.40	1.52e-10
RP11-337N6.3	2	177317715	177318471	5.46	4.82e-08
SLC4A7[a]	3	27372721	27484420	-14.30	2.34e-46
ZBTB38[a]	3	141324213	141449792	-7.68	1.54e-14
LINC00886[a]	3	156747346	156817062	5.89	3.96e-09
CMSS1[a]	3	99817834	100181732	5.44	5.31e-08
PSMD6-AS2[b]	3	64004022	64012148	-5.42	5.99e-08
EFCC1	3	129001629	129040742	4.84	1.32e-06
GLRA3[b]	4	174636914	174829314	6.92	4.66e-12
PPM1K[b]	4	88257620	88284769	-5.19	2.11e-07
SLC22A5[b]	5	132369752	132395614	9.09	9.64e-20
ANKRD55[b]	5	56099678	56233359	-6.29	3.09e-10
L3MBTL3[a]	6	130013699	130141451	6.32	2.67e-10
TOB2P1[b]	6	28217643	28218634	5.31	1.09e-07
ZNF703[b]	8	37695751	37700021	5.45	5.11e-08
PRR33[b]	11	1888577	1891772	-10.05	8.84e-24
EFEMP2[b]	11	65866441	65873592	6.03	1.60e-09
SPTY2D1[a]	11	18606401	18634791	4.80	1.57e-06
NTN4[b]	12	95657807	95791152	-12.37	3.75e-35
RP11-967K21.1[b]	12	28163298	28190738	6.83	8.68e-12
RPL12P7[b]	14	68693090	68693583	9.60	8.11e-22
RCCD1[a]	15	90955796	90963125	-5.98	2.18e-09
RP11-212I21.2[b]	16	55426797	55462297	5.31	1.11e-07
CBX8[a]	17	79792132	79801683	5.87	4.35e-09
LRRC37A4P[b]	17	45506741	45550335	5.58	2.43e-08
COX11[b]	17	54951902	54968764	4.79	1.63e-06
LRRC25[b]	19	18391144	18397617	9.31	1.32e-20
APOBEC3B[b]	22	38982347	38992804	6.83	8.25e-12

a gene identified in previous breast cancer GWAS

b within 1MB of gene identified in previous breast cancer GWAS

Table B.3: Independent TWAS risk genes of ovarian cancer identified by PrediXcan.

Gene	Chrom	Start	End	Zscore	Pvalue
PRC1-AS1[a]	15	90972860	90988624	5.21	1.85e-07
LINC02210[a]	17	45620328	45655156	-5.77	7.76e-09

a within 1MB of gene identified in previous ovarian cancer GWAS

Table B.4: TWAS risk genes of both breast and ovarian cancer identified by PrediXcan.

				Breast		Ovary	
Gene	Chrom	Start	End	Zscore	Pvalue	Zscore	Pvalue
PRC1-AS1[a,c]	15	90972860	90988624	5.92	3.19e-09	5.21	1.85e-07
LRRC37A4P[b,c]	17	45506741	45550335	5.58	2.43e-08	5.36	8.13e-08

a gene identified in previous breast cancer GWAS

b within 1MB of gene identified in previous breast cancer GWAS

c within 1MB of gene identified in previous ovarian cancer GWAS

B.3 Genes Significant in Multiple TWAS Results

Table B.5: Total number of TWAS risk genes identified by model and cancer type.

	TWAS	
Model	Breast	Ovary
TIGAR	88	37
PrediXcan	56	4

Table B.6: TWAS risk genes identified by multiple models or for multiple cancer types.

				TIGAR		PrediXcan	
Gene	Chrom	Start	End	Breast	Ovary	Breast	Ovary
PRC1-AS1[a,d]	15	90972860	90988624	X	X	X	X
LRRC37A4P[b,d]	17	45506741	45550335	X	X	X	X
DND1P1[b,d]	17	45585871	45586929	X	X	X	
LINC02210[b,d]	17	45620328	45655156	X	X		X
KLHDC7A[a]	1	18480982	18486126	X		X	
CH17-437K3.1[b]	1	121396754	121463129	X		X	
ASH1L[a,d]	1	155335268	155562807	X		X	
CASP8[a]	2	201233443	201287711	X		X	
ALS2CR12[b]	2	201288271	201357398	X		X	
SLC4A7[a,d]	3	27372721	27484420	X		X	
PSMD6-AS2[b]	3	64004022	64012148	X		X	
SLC22A5[b]	5	132369752	132395614	X		X	
PDLIM4[b]	5	132257671	132273454	X		X	
ANKRD55[b,d]	5	56099678	56233359	X		X	
C5orf56[b]	5	132410636	132488702	X		X	
L3MBTL3[a]	6	130013699	130141451	X		X	
PIDD1[a]	11	799191	809646	X		X	
AP006621.5[b]	11	777578	784297	X		X	
CCDC91[a]	12	28133249	28581511	X		X	
RCCD1[a,c]	15	90955796	90963125	X		X	
CBX8[a]	17	79792132	79801683	X		X	
LRRC25[b]	19	18391144	18397617	X		X	
UBE2MP1	16	35169692	35170241	X	X		
MAPK8IP1P2[b,d]	17	45600869	45602340	X	X		
CRHR1[b,d]	17	45784280	45835828	X	X		
RP11-707O23.1[b,d]	17	45592621	45593369	X	X		
RP11-259G18.1[b,d]	17	46267037	46268694	X	X		
AC091132.1[b,d]	17	45452844	45464065	X	X		
ARHGAP27[a,d]	17	45393902	45434421	X	X		
MAPT[a,d]	17	45894382	46028334	X	X		
LRRC37A2[b,d]	17	46511511	46553449	X	X		
RP11-259G18.3[b,d]	17	46259551	46260606	X	X		
KANSL1-AS1[b,d]	17	46193576	46196723	X	X		

				TIGAR		PrediXcan	
Gene	Chrom	Start	End	Breast	Ovary	Breast	Ovary
FAM215B[b,d]	17	46558830	46562795	X	X		
FRG1EP	20	29480147	29497179	X	X		
NSF[b,c]	17	46590669	46757464		X		X

a gene identified in previous breast cancer GWAS

b within 1MB of gene identified in previous breast cancer GWAS

c gene identified in previous ovarian cancer GWAS

d within 1MB of gene identified in previous ovarian cancer GWAS

Table B.7: TWAS risk genes not previously identified in breast or ovarian cancer GWAS.

Model	TWAS	Gene	Chrom	Start	End	Zscore	Pvalue
DPR	Breast	KLHL25	15	85759323	85795030	-4.73	2.22e-06
DPR	Breast	UBE2MP1	16	35169692	35170241	-5.31	1.13e-07
DPR	Ovary	UBE2MP1	16	35169692	35170241	5.77	7.88e-09
DPR	Breast	ANKRD20A21P	20	30656033	30723932	-5.22	1.83e-07
DPR	Breast	FRG1EP	20	29480147	29497179	5.39	6.95e-08
DPR	Ovary	FRG1EP	20	29480147	29497179	-4.99	6.19e-07
EN	Breast	RP11-337N6.3	2	177317715	177318471	5.46	4.82e-08
EN	Breast	EFCC1	3	129001629	129040742	4.84	1.32e-06

Figure B.2: Venn diagram of number of TWAS risk genes identified by model and cancer type.

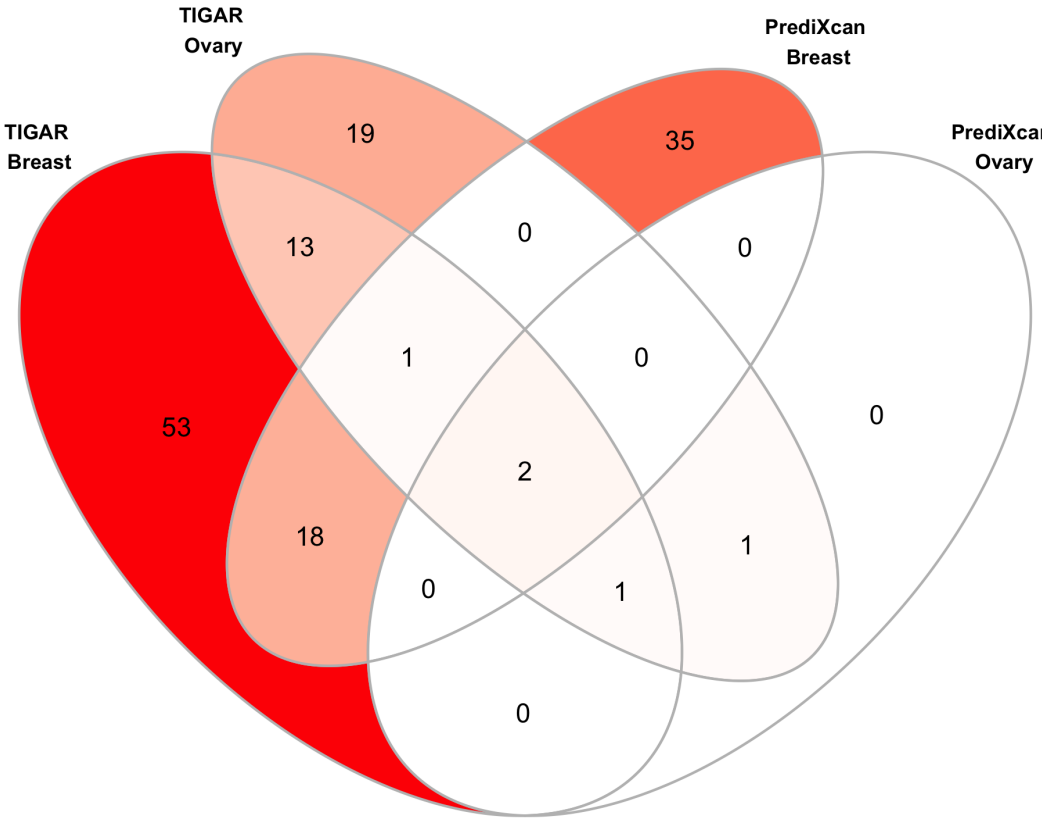
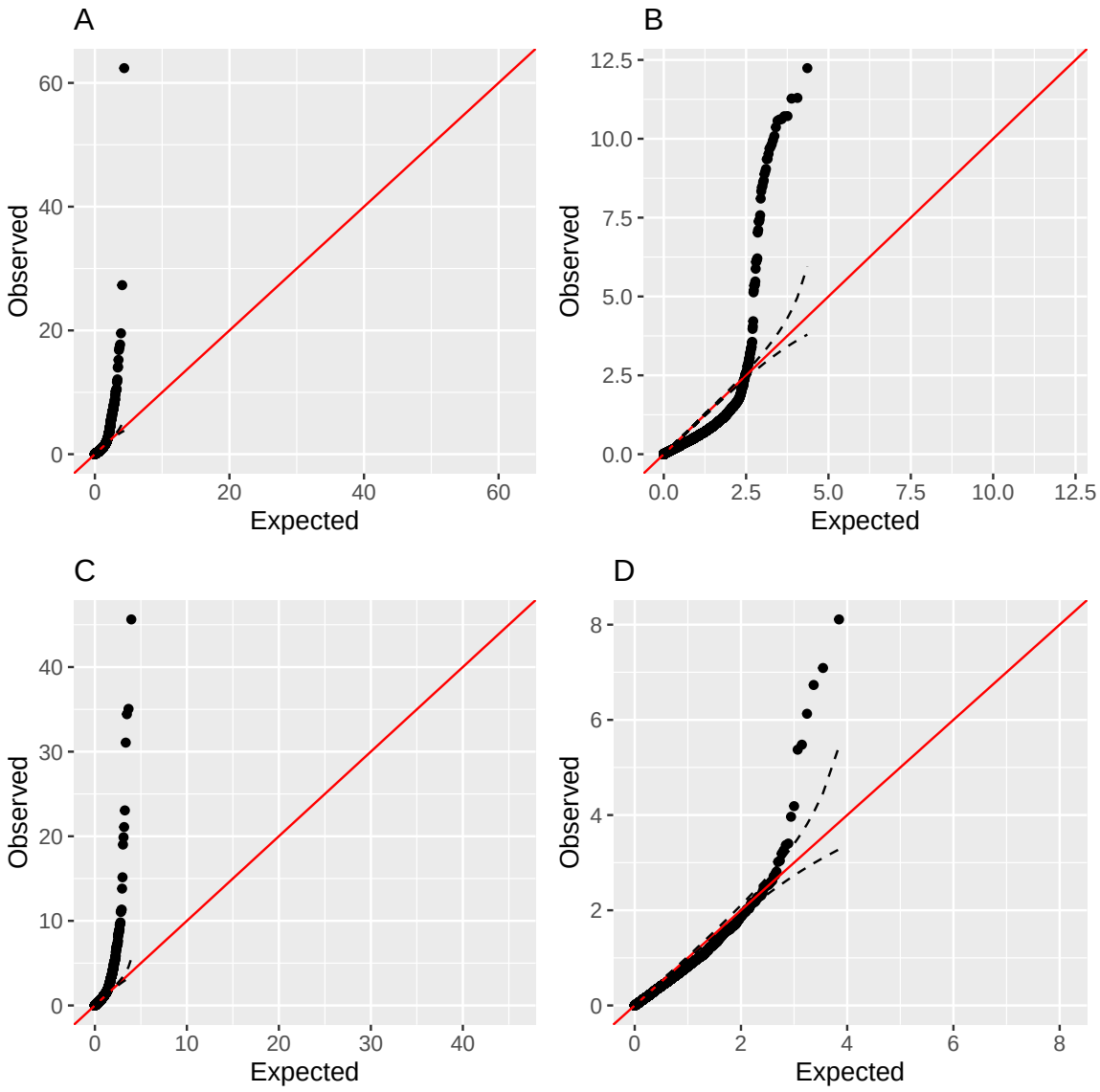


Figure B.3: QQ-Plots for TWAS results by TIGAR for studying breast cancer (A) and ovarian cancer (B) and TWAS results by PrediXcan for studying breast cancer (C) and ovarian cancer (D).



Appendix C: SR-TWAS Code

```
# return R^2
def get_r2(y, predY, pval=False):
  lm = sm.OLS(y, sm.add_constant(predY)).fit()
  if pval:
    return lm.rsquared, lm.f_pvalue
  return lm.rsquared

# flatten a nested list
def flatten(nested_list):
  return [j for i in nested_list for j in i]

# formats list [W0_N_SNP, W0_CVR2, W0_R2, W0_PVAL, ...,
#     ..., WK_N_SNP, WK_CVR2, WK_R2, WK_PVAL] for output even when
   one or more of the W_ks lacks data for a target
def format_final_est_out_vals(target_k_outvals, target_ks, K):
  target_k_ind = {target_ks[j]: j for j in range(len(target_ks))}
  out_list = [target_k_outvals[target_k_ind[k]] if k in target_ks
   else (0, 0, 0, 1) for k in range(K)]
  return flatten(out_list)

# estimator for individual trained models
class WeightEstimator(BaseEstimator):
  _estimator_type = 'regressor'

  def __init__(self, raw_weights):
    self.raw_weights = raw_weights

  def fit(self, X=None, y=None):
    self.coef_ = self.raw_weights.dropna()
    self.snpID = self.coef_.index.values
    self.n_features_in_ = self.coef_.size
    return self

  def predict(self, X):
    return np.dot(X[self.snpID], self.coef_)

  def score(self, X, y):
    return get_r2(y, self.predict(X))

  def r2_pval(self, X, y):
    return get_r2(y, self.predict(X), pval=True)

  def avg_r2_cv(self, X, y):
    return sum(cross_val_score(self, X, y)) / 5

  def est_out_vals(self, X, y):
```

```
    return [(self.n_features_in_, self.avg_r2_cv(X, y), *self.
   r2_pval(X, y))]

  def final_est_out_vals(self, X, y, ks, K):
    return format_final_est_out_vals((self.est_out_vals(X, y)), ks,
   K)

# final estimator for stacking regressor
class ZetasEstimator(BaseEstimator):
  _estimator_type = 'regressor'

  def __init__(self, min_method=None, tol=None):
    super().__init__()
    self.min_method = min_method
    self.tol = tol

  def _loss_func(self, zeta, X, y):
    if (len(zeta) == 1):
      Zeta = np.array([*zeta, 1 - np.sum(zeta)])
    else:
      Zeta = np.array(zeta)
    predY = np.dot(X, Zeta)
    R2 = get_r2(y, predY)
    return 1 - R2

  def fit(self, X, y, sample_weights=None):
    K = np.shape(X)[1]
    # if only one valid model set zeta = 1
    if (K == 1):
      self.coef_ = np.array([1])
      return self
    elif (K == 2):
      # initialize zeta list; all models weighted equally
      zeta_0 = np.full(K-1, 1/K)
      bnds = tuple([(0, 1)] * (K-1))
      # minimize loss function
      self.fit_res_ = minimize(
        self._loss_func,
        zeta_0,
        args=(X, y),
        bounds=bnds,
        tol=self.tol,
        method=self.min_method)
      zeta = self.fit_res_.x
      self.coef_ = np.array([*zeta, 1 - np.sum(zeta)])
    else:
      zeta_0 = np.full(K, 1/K)
      bnds = tuple([(0, 1)] * K)
      cons = ({'type': 'eq', 'fun': lambda x:  1 - sum(x)})
      # minimize loss function
      self.fit_res_ = minimize(
        self._loss_func,
        zeta_0,
        args=(X, y),
```

```
        bounds=bnds,
        tol=self.tol,
        method=self.min_method,
        constraints = cons)
      zeta = self.fit_res_.x
      self.coef_ = np.array(zeta)
    return self

  def predict(self, X):
    return np.dot(X, self.coef_)

  def score(self, X, y):
    return get_r2(y, self.predict(X))

# stacking regressor
class WeightStackingRegressor(StackingRegressor):
  def __init__(self, estimators, final_estimator=ZetasEstimator(),
    *, cv=None, n_jobs=None, passthrough=False, verbose=0):
    super().__init__(
      estimators=estimators,
      final_estimator=final_estimator,
      cv=cv,
      n_jobs=n_jobs,
      passthrough=passthrough,
      verbose=verbose)

  def fit(self, X, y, sample_weights=None):
    self = super().fit(X, y, sample_weights)
    self.zetas_ = self.final_estimator_.coef_
    return self

  def score(self, X, y):
    return get_r2(y, self.predict(X))

  def r2_pval(self, X, y):
    return get_r2(y, self.predict(X), pval=True)

  def avg_r2_cv(self, X, y):
    return sum(cross_val_score(self, X, y)) / 5

  def est_avg_cv_scores(self, X, y):
    return [est[1].avg_r2_cv(X, y) for est in self.estimators]

  def est_r2_pvals(self, X, y):
    return list(zip(*[est[1].fit(X,y).r2_pval(X,y) for est in self.
   estimators]))

  def est_out_vals(self, X, y):
    est_n_snps = [est[1].fit().n_features_in_ for est in self.
   estimators]
    est_avg_cv_score = self.est_avg_cv_scores(X, y)
    est_r2_pvals = self.est_r2_pvals(X, y)
    return list(zip(est_n_snps, est_avg_cv_score, *est_r2_pvals))
```

```
  def final_est_out_vals(self, X, y, ks, K):
    return format_final_est_out_vals(self.est_out_vals(X, y), ks, K)
```

References

[1] Mark I. McCarthy et al. “Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges”. In: *Nature Reviews Genetics* 9.5 (5 May 2008), pp. 356–369. DOI: 10.1038/nrg2344.

[2] Peter M. Visscher et al. “Five Years of GWAS Discovery”. In: *The American Journal of Human Genetics* 90.1 (Jan. 13, 2012), pp. 7–24. DOI: 10.1016/j.ajhg.2011.11.029.

[3] Qingyang Huang. “Genetic Study of Complex Diseases in the Post-GWAS Era”. In: *Journal of Genetics and Genomics* 42.3 (Mar. 20, 2015), pp. 87–98. DOI: 10.1016/j.jgg.2015.02.001.

[4] Peter M. Visscher et al. “10 Years of GWAS Discovery: Biology, Function, and Translation”. In: *American Journal of Human Genetics* 101.1 (July 6, 2017), pp. 5–22. DOI: 10.1016/j.ajhg.2017.06.005.

[5] Maren E. Cannon and Karen L. Mohlke. “Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci”. In: *American Journal of Human Genetics* 103.5 (Nov. 1, 2018), pp. 637–653. DOI: 10.1016/j.ajhg.2018.10.001.

[6] Yu-Ru Su et al. “A Mixed-Effects Model for Powerful Association Tests in Integrative Functional Genomics”. In: *The American Journal of Human Genetics* 102.5 (May 3, 2018), pp. 904–919. DOI: 10.1016/j.ajhg.2018.03.019.

[7] Alexandra C. Nica and Emmanouil T. Dermitzakis. “Using Gene Expression to Investigate the Genetic Basis of Complex Disorders”. In: *Human Molecular Genetics* 17.R2 (Oct. 15, 2008), R129–R134. DOI: 10.1093/hmg/ddn285.

[8] Dan L. Nicolae et al. “Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS”. In: *PLOS Genetics* 6.4 (Apr. 1, 2010), e1000888. DOI: 10.1371/journal.pgen.1000888.

[9] Jacek Majewski and Tomi Pastinen. “The Study of eQTL Variations by RNA-Seq: From SNPs to Phenotypes”. In: *Trends in Genetics* 27.2 (Feb. 1, 2011), pp. 72–79. DOI: 10.1016/j.tig.2010.10.006.

[10] GTEx Consortium. “The Genotype-Tissue Expression (GTEx) Project”. In: *Nature Genetics* 45.6 (6 June 2013), pp. 580–585. DOI: 10.1038/ng.2653.

[11] Alexandra C. Nica and Emmanouil T. Dermitzakis. “Expression Quantitative Trait Loci: Present and Future”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1620 (June 19, 2013). DOI: 10.1098/rstb.2012.0362.

[12] Damien C. Croteau-Chonka et al. “Expression Quantitative Trait Loci Information Improves Predictive Modeling of Disease Relevance of Non-Coding Genetic Variation”. In: *PLOS ONE* 10.10 (Oct. 16, 2015), e0140758. DOI: 10.1371/journal.pone.0140758.

[13] Eric R Gamazon et al. “A Gene-Based Association Method for Mapping Traits Using Reference Transcriptome Data”. In: *Nature Genetics* 47.9 (Sept. 2015), pp. 1091–1098. DOI: 10.1038/ng.3367.

[14] Zhihong Zhu et al. “Integration of Summary Data from GWAS and eQTL Studies Predicts Complex Trait Gene Targets”. In: *Nature Genetics* 48.5 (5 May 2016), pp. 481–487. DOI: 10.1038/ng.3538.

[15] Farhad Hormozdiari et al. “Colocalization of GWAS and eQTL Signals Detects Target Genes”. In: *American Journal of Human Genetics* 99.6 (Dec. 1, 2016), pp. 1245–1260. DOI: 10.1016/j.ajhg.2016.10.003.

[16] Alexander Gusev et al. “Integrative Approaches for Large-Scale Transcriptome-Wide Association Studies”. In: *Nature Genetics* 48.3 (3 Mar. 2016), pp. 245–252. DOI: 10.1038/ng.3506.

[17] Nicholas Mancuso et al. “Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits”. In: *The American Journal of Human Genetics* 100.3 (Mar. 2, 2017), pp. 473–487. DOI: 10.1016/j.ajhg.2017.01.031.

[18] Sini Nagpal et al. “TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits”. In: *The American Journal of Human Genetics* 105.2 (Aug. 2019), pp. 258–266. DOI: 10.1016/j.ajhg.2019.05.018.

[19] Justin M. Luningham et al. “Bayesian Genome-Wide TWAS Method to Leverage Both Cis- and Trans-eQTL Information through Summary Statistics”. In: *The American Journal of Human Genetics* 107.4 (Oct. 1, 2020), pp. 714–726. DOI: 10.1016/j.ajhg.2020.08.022.

[20] Zhongshang Yuan et al. “Testing and Controlling for Horizontal Pleiotropy with Probabilistic Mendelian Randomization in Transcriptome-Wide Association Studies”. In: *Nature Communications* 11.1 (1 July 31, 2020), p. 3861. DOI: 10.1038/s41467-020-17668-6.

[21] Alvaro N. Barbeira et al. “Exploring the Phenotypic Consequences of Tissue Specific Gene Expression Variation Inferred from GWAS Summary Statistics”. In: *Nature Communications* 9.1 (May 8, 2018), pp. 1–20. DOI: 10.1038/s41467-018-03621-1.

[22] Alexander Gusev et al. “A Transcriptome-Wide Association Study of High Grade Serous Epithelial Ovarian Cancer Identifies Novel Susceptibility Genes

and Splice Variants". In: *Nature Genetics* 51.5 (May 2019), pp. 815–823. DOI: 10.1038/s41588-019-0395-x.

[23] Tobias Strunz et al. "A Transcriptome-Wide Association Study Based on 27 Tissues Identifies 106 Genes Potentially Relevant for Disease Pathology in Age-Related Macular Degeneration". In: *Scientific Reports* 10.1 (1 Jan. 31, 2020), p. 1584. DOI: 10.1038/s41598-020-58510-9.

[24] GTEx Consortium. "The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues". In: *Science* 369.6509 (Sept. 11, 2020), pp. 1318–1330. DOI: 10.1126/science.aaz1776.

[25] Cuiyan Wu et al. "Transcriptome-Wide Association Study Identifies Susceptibility Genes for Rheumatoid Arthritis". In: *Arthritis Research & Therapy* 23 (2021). DOI: 10.1186/s13075-021-02419-9.

[26] Alexander Gusev et al. "Transcriptome-Wide Association Study of Schizophrenia and Chromatin Activity Yields Mechanistic Disease Insights". In: *Nature Genetics* 50.4 (4 Apr. 2018), pp. 538–548. DOI: 10.1038/s41588-018-0092-1.

[27] Jun Zhong et al. "A Transcriptome-Wide Association Study Identifies Novel Candidate Susceptibility Genes for Pancreatic Cancer". In: *Journal of the National Cancer Institute* 112.10 (Oct. 1, 2020), pp. 1003–1012. DOI: 10.1093/jnci/djz246.

[28] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "7.10 Cross-Validation". In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, Feb. 2009, pp. 241–249.

[29] David M. Blei and Michael I. Jordan. "Variational Inference for Dirichlet Process Mixtures". In: *Bayesian Analysis* 1.1 (Mar. 2006), pp. 121–143. DOI: 10.1214/06-BA104.

[30] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (Apr. 3, 2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773.

[31] Ping Zeng and Xiang Zhou. "Non-Parametric Genetic Prediction of Complex Traits with Latent Dirichlet Process Regression Models". In: *Nature Communications* 8.1 (1 Sept. 6, 2017), pp. 1–11. DOI: 10.1038/s41467-017-00470-2.

[32] Shizhen Tang et al. "Novel Variance-Component TWAS Method for Studying Complex Human Diseases with Applications to Alzheimer's Dementia". In: *PLOS Genetics* 17.4 (Apr. 2, 2021), e1009482. DOI: 10.1371/journal.pgen.1009482.

[33] Kyriaki Michailidou et al. "Association Analysis Identifies 65 New Breast Cancer Risk Loci". In: *Nature* 551.7678 (Nov. 2, 2017), pp. 92–94. DOI: 10.1038/nature24284.

[34] Catherine M. Phelan et al. "Identification of Twelve New Susceptibility Loci for Different Histotypes of Epithelial Ovarian Cancer". In: *Nature Genetics* 49.5 (May 2017), pp. 680–691. DOI: 10.1038/ng.3826.

[35] Gilles Thomas et al. "A Multi-Stage Genome-Wide Association in Breast Cancer Identifies Two Novel Risk Alleles at 1p11.2 and 14q24.1 (RAD51L1)". In: *Nature Genetics* 41.5 (May 2009), pp. 579–584. DOI: 10.1038/ng.353.

[36] Kyriaki Michailidou et al. "Large-Scale Genotyping Identifies 41 New Loci Associated with Breast Cancer Risk". In: *Nature Genetics* 45.4 (Apr. 2013), 353–361e2. DOI: 10.1038/ng.2563.

[37] Habibul Ahsan et al. "A Genome-Wide Association Study of Early-Onset Breast Cancer Identifies PFKM as a Novel Breast Cancer Gene and Supports a Common Genetic Spectrum for Breast Cancer at Any Age". In: *Cancer*

Epidemiology, Biomarkers & Prevention 23.4 (Apr. 2014), pp. 658–669. DOI: 10.1158/1055-9965.EPI-13-0340.

[38] Kyriaki Michailidou et al. "Genome-Wide Association Analysis of More than 120,000 Individuals Identifies 15 New Susceptibility Loci for Breast Cancer". In: *Nature Genetics* 47.4 (Apr. 2015), pp. 373–380. DOI: 10.1038/ng.3242.

[39] Grazia Palomba et al. "Genome-Wide Association Study of Susceptibility Loci for Breast Cancer in Sardinian Population". In: *BMC Cancer* 15 (May 10, 2015). DOI: 10.1186/s12885-015-1392-9.

[40] Fergus J. Couch et al. "Identification of Four Novel Susceptibility Loci for Oestrogen Receptor Negative Breast Cancer". In: *Nature Communications* 7 (Apr. 27, 2016). DOI: 10.1038/ncomms11375.

[41] Roger L Milne et al. "Identification of Ten Variants Associated with Risk of Estrogen-Receptor-Negative Breast Cancer". In: *Nature Genetics* 49.12 (Dec. 2017), pp. 1767–1778. DOI: 10.1038/ng.3785.

[42] Sara R. Rashkin et al. "Pan-Cancer Study Detects Genetic Risk Variants and Shared Genetic Basis in Two Large Cohorts". In: *Nature Communications* 11 (Sept. 4, 2020). DOI: 10.1038/s41467-020-18246-6.

[43] Joshua D. Hoffman et al. "Cis-eQTL-Based Trans-Ethnic Meta-Analysis Reveals Novel Genes Associated with Breast Cancer Risk". In: *PLoS Genetics* 13.3 (Mar. 31, 2017). DOI: 10.1371/journal.pgen.1006690.

[44] Lang Wu et al. "A Transcriptome-Wide Association Study of 229,000 Women Identifies New Candidate Susceptibility Genes for Breast Cancer". In: *Nature Genetics* 50.7 (July 2018), pp. 968–978. DOI: 10.1038/s41588-018-0132-x.

[45] Manuel A. Ferreira et al. "Genome-Wide Association and Transcriptome Studies Identify Target Genes and Risk Loci for Breast Cancer". In: *Nature Communications* 10 (Apr. 15, 2019). DOI: 10.1038/s41467-018-08053-5.

[46] Helian Feng et al. “Transcriptome-Wide Association Study of Breast Cancer Risk by Estrogen-Receptor Status”. In: *Genetic Epidemiology* 44.5 (2020), pp. 442–468. DOI: 10.1002/gepi.22288.

[47] Siddhartha Kar et al. “Pleiotropy-Guided Transcriptome Imputation from Normal and Tumor Tissues Identifies New Candidate Susceptibility Genes for Breast and Ovarian Cancer”. In: *bioRxiv* (May 13, 2020). DOI: 10.1101/2020.04.23.043653.

[48] Xiang Shu et al. “Identification of Novel Breast Cancer Susceptibility Loci in Meta-Analyses Conducted among Asian and European Descendants”. In: *Nature Communications* 11 (Mar. 5, 2020). DOI: 10.1038/s41467-020-15046-w.

[49] Fergus J. Couch et al. “Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk”. In: *PLoS Genetics* 9.3 (Mar. 27, 2013). DOI: 10.1371/journal.pgen.1003212.

[50] Karoline B. Kuchenbaecker et al. “Identification of Six New Susceptibility Loci for Invasive Epithelial Ovarian Cancer”. In: *Nature Genetics* 47.2 (Feb. 2015), pp. 164–171. DOI: 10.1038/ng.3185.

[51] Yingchang Lu et al. “A Transcriptome-Wide Association Study among 97,898 Women to Identify Candidate Susceptibility Genes for Epithelial Ovarian Cancer Risk”. In: *Cancer Research* 78.18 (Sept. 15, 2018), pp. 5419–5430. DOI: 10.1158/0008-5472.CAN-18-0951.

[52] Adrián Mosquera Orgueira. “Hidden among the Crowd: Differential DNA Methylation-Expression Correlations in Cancer Occur at Important Oncogenic Pathways”. In: *Frontiers in Genetics* 6 (May 13, 2015). DOI: 10.3389/fgene.2015.00163.

[53] Juan Xu et al. “Distinct Expression Profile of lncRNA in Endometrial Carcinoma”. In: *Oncology Reports* 36.6 (Dec. 1, 2016), pp. 3405–3412. DOI: 10.3892/or.2016.5173.

[54] Cen Zhang et al. “Cullin3–KLHL25 Ubiquitin Ligase Targets ACLY for Degradation to Inhibit Lipid Synthesis and Tumor Progression”. In: *Genes & Development* 30.17 (Sept. 1, 2016), pp. 1956–1970. DOI: 10.1101/gad.283283.116.

[55] Ping Luo et al. “deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks”. In: *Frontiers in Genetics* 10 (Jan. 29, 2019). DOI: 10.3389/fgene.2019.00013.

[56] Zhenxin Zhu et al. “Whole-Exome Sequencing Identifies Prognostic Mutational Signatures in Gastric Cancer”. In: *Annals of Translational Medicine* 8.22 (Nov. 2020). DOI: 10.21037/atm-20-6620.

[57] David A. Bennett et al. “Overview and Findings from the Religious Orders Study”. In: *Current Alzheimer Research* 9.6 (July 1, 2012), pp. 628–645. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409291/.

[58] David A. Bennett et al. “Religious Orders Study and Rush Memory and Aging Project”. In: *Journal of Alzheimer’s Disease* 64 (Suppl 1 2018), S161–S189. DOI: 10.3233/JAD-179939.

[59] David A. Bennett et al. “Overview and Findings from the Rush Memory and Aging Project”. In: *Current Alzheimer Research* 9.6 (July 1, 2012), pp. 646–663. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439198/.

[60] Minghui Wang et al. “The Mount Sinai Cohort of Large-Scale Genomic, Transcriptomic and Proteomic Data in Alzheimer’s Disease”. In: *Scientific Data* 5 (Sept. 11, 2018). DOI: 10.1038/sdata.2018.185.

[61] Mariet Allen et al. “Human Whole Genome Genotype and Transcriptome Data for Alzheimer’s and Other Neurodegenerative Diseases”. In: *Scientific Data* 3 (Oct. 11, 2016). DOI: 10.1038/sdata.2016.89.

[62] Cristen J. Willer, Yun Li, and Gonçalo R. Abecasis. “METAL: Fast and Efficient Meta-Analysis of Genomewide Association Scans”. In: *Bioinformatics* 26.17 (Sept. 1, 2010), pp. 2190–2191. DOI: 10.1093/bioinformatics/btq340.

[63] D. Y. Lin and D. Zeng. “Meta-Analysis of Genome-Wide Association Studies: No Efficiency Gain in Using Individual Participant Data”. In: *Genetic Epidemiology* 34.1 (Jan. 2010). DOI: 10.1002/gepi.20435.

[64] Omer Sagi and Lior Rokach. “Ensemble Learning: A Survey”. In: *WIREs Data Mining and Knowledge Discovery* 8.4 (2018), e1249. DOI: 10.1002/widm.1249.

[65] David H. Wolpert. “Stacked Generalization”. In: *Neural Networks* 5.2 (Jan. 1, 1992), pp. 241–259. DOI: 10.1016/S0893-6080(05)80023-1.

[66] Leo Breiman. “Stacked Regressions”. In: *Machine Learning* 24.1 (July 1996), pp. 49–64. DOI: 10.1007/BF00117832.

[67] Hui Zou and Trevor Hastie. “Regularization and Variable Selection via the Elastic Net”. In: *Journal of the Royal Statistical Society*. Series B (Statistical Methodology) 67.2 (2005), pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

[68] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. “Polygenic Modeling with Bayesian Sparse Linear Mixed Models”. In: *PLOS Genetics* 9.2 (Feb. 7, 2013), e1003264. DOI: 10.1371/journal.pgen.1003264.

[69] Peter M. Visscher, William G. Hill, and Naomi R. Wray. “Heritability in the Genomics Era — Concepts and Misconceptions”. In: *Nature Reviews Genetics* 9.4 (4 Apr. 2008), pp. 255–266. DOI: 10.1038/nrg2322.

[70] Noah Zaitlen and Peter Kraft. “Heritability in the Genome-Wide Association Era”. In: *Human Genetics* 131.10 (Oct. 2012), pp. 1655–1664. DOI: 10.1007/s00439-012-1199-6.

[71] Shaun Purcell et al. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses”. In: *The American Journal of Human Genetics* 81.3 (Sept. 1, 2007), pp. 559–575. DOI: 10.1086/519795.

[72] Aaron McKenna et al. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data”. In: *Genome Research* 20.9 (Sept. 1, 2010), pp. 1297–1303. DOI: 10.1101/gr.107524.110.

[73] Bingshan Li and Suzanne M. Leal. “Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data”. In: *American Journal of Human Genetics* 83.3 (Sept. 12, 2008), pp. 311–321. DOI: 10.1016/j.ajhg.2008.06.024.

[74] Bingshan Li, Dajiang J. Liu, and Suzanne M. Leal. “Identifying Rare Variants Associated with Complex Traits via Sequencing”. In: *Current Protocols in Human Genetics* 78.1 (2013), pp. 1.26.1–1.26.22. DOI: 10.1002/0471142905.hg0126s78.

[75] Adam Auton et al. “A Global Reference for Human Genetic Variation”. In: *Nature* 526.7571 (7571 Oct. 2015), pp. 68–74. DOI: 10.1038/nature15393.

[76] Wes McKinney. “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference (SciPy 2010)*. Python in Science Conference. Austin, Texas, 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[77] Wes McKinney and PyData Development Team. *Pandas: Powerful Python Data Analysis Toolkit*. Version 0.23.4. Aug. 6, 2018. URL: https://pandas.pydata.org/pandas-docs/version/0.23.4/index.html.

[78] Charles R. Harris et al. "Array Programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[79] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17.3 (3 Mar. 2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[80] Fabian Pedregosa et al. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[81] Lars Buitinck et al. "API Design for Machine Learning Software: Experiences from the Scikit-Learn Project". In: *Proceedings of the European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECMPKDD'13)*. ECML PKDD Workshop: Languages for Data Mining and Machine Learning. 2013, pp. 108–122. URL: https://arxiv.org/abs/1309.0238.

[82] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and Statistical Modeling with Python". In: *Proceedings of the 9th Python in Science Conference (SciPy 2010)*. Python in Science Conference. Austin, Texas, 2010, pp. 92–96. DOI: 10.25080/Majora-92bf1922-011.

[83] Heng Li. "Tabix: Fast Retrieval of Sequence Features from Generic TAB-Delimited Files". In: *Bioinformatics* 27.5 (Mar. 1, 2011), pp. 718–719. DOI: 10.1093/bioinformatics/btq671.

[84] GTEx Consortium et al. “The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans”. In: *Science* 348.6235 (May 8, 2015), pp. 648–660. DOI: 10.1126/science.1262110.

[85] GTEx Consortium. “Genetic Effects on Gene Expression across Human Tissues”. In: *Nature* 550.7675 (7675 Oct. 11, 2017), pp. 204–213. DOI: 10.1038/nature24277.

[86] Latarsha J. Carithers et al. “A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project”. In: *Biopreservation and Biobanking* 13.5 (Oct. 1, 2015), pp. 311–319. DOI: 10.1089/bio.2015.0032.

[87] Oliver Stegle et al. “Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses”. In: *Nature Protocols* 7.3 (Feb. 16, 2012), pp. 500–507. DOI: 10.1038/nprot.2011.457.

[88] Qiuyin Cai et al. “Genome-Wide Association Analysis in East Asians Identifies Breast Cancer Susceptibility Loci at 1q32.1, 5q14.3 and 15q26.1”. In: *Nature Genetics* 46.8 (Aug. 2014), pp. 886–890. DOI: 10.1038/ng.3041.

[89] Haoyu Zhang et al. “Genome-Wide Association Study Identifies 32 Novel Breast Cancer Susceptibility Loci from Overall and Subtype-Specific Analyses”. In: *Nature Genetics* 52.6 (6 June 2020), pp. 572–581. DOI: 10.1038/s41588-020-0609-2.

[90] Nan Xu et al. “Clinical Significance of High Expression of Circulating Serum lncRNA RP11-445H22.4 in Breast Cancer Patients: A Chinese Population-Based Study”. In: *Tumor Biology* 36.10 (Oct. 1, 2015), pp. 7659–7665. DOI: 10.1007/s13277-015-3469-0.

[91] Tze Pheng Lau et al. “Pair-Wise Comparison Analysis of Differential Expression of mRNAs in Early and Advanced Stage Primary Colorectal Adenocarcinomas”. In: *BMJ Open* 4.8 (Aug. 8, 2014). DOI: 10.1136/bmjopen-2014-004930.

[92] Yunxian Liu et al. “Identification of Breast Cancer Associated Variants That Modulate Transcription Factor Binding”. In: *PLOS Genetics* 13.9 (Sept. 28, 2017), e1006761. DOI: 10.1371/journal.pgen.1006761.

[93] Tariq Ahmad Masoodi et al. “Computational Analysis of Breast Cancer GWAS Loci Identifies the Putative Deleterious Effect of STXBP4 and ZNF404 Gene Variants”. In: *Journal of Cellular Biochemistry* 118.12 (2017), pp. 4296–4307. DOI: 10.1002/jcb.26080.

[94] Zhengwei Du et al. “Identification of Long Non-Coding RNA-Mediated Transcriptional Dysregulation Triplets Reveals Global Patterns and Prognostic Biomarkers for ER+/PR+, HER2- and Triple Negative Breast Cancer”. In: *International Journal of Molecular Medicine* 44.3 (Sept. 2019), pp. 1015–1025. DOI: 10.3892/ijmm.2019.4261.

[95] Jufeng Xia et al. “Preliminary Investigation of Five Novel Long Non-Coding RNAs in Hepatocellular Carcinoma Cell Lines”. In: *BioScience Trends* 10.4 (2016), pp. 315–319. DOI: 10.5582/bst.2016.01140.

[96] J. N. K. Rao and Kathleen Subrahmaniam. “Combining Independent Estimators and Estimation in Linear Regression with Unequal Variances”. In: *Biometrics* 27.4 (1971), pp. 971–990. DOI: 10.2307/2528832.

[97] B. Efron and C. Morris. “Combining Possibly Related Estimation Problems”. In: *Journal of the Royal Statistical Society*. Series B (Statistical Methodology) 35.3 (1973), pp. 379–421.

[98] David A. Bennett et al. “The Rush Memory and Aging Project: Study Design and Baseline Characteristics of the Study Cohort”. In: *Neuroepidemiology* 25.4 (2005), pp. 163–175. DOI: 10.1159/000087446.

[99] Philip L. De Jager et al. “A Multi-Omic Atlas of the Human Frontal Cortex for Aging and Alzheimer’s Disease Research”. In: *Scientific Data* 5 (Aug. 7, 2018). DOI: 10.1038/sdata.2018.142.

[100] Philip L. De Jager et al. “A Genome-Wide Scan for Common Variants Affecting the Rate of Age-Related Cognitive Decline”. In: *Neurobiology of Aging* 33.5 (May 2012), 1017.e1–1017.e15. DOI: 10.1016/j.neurobiolaging.2011.09.033.

[101] Lauren S. Mogil et al. “Genetic Architecture of Gene Expression Traits across Diverse Populations”. In: *PLOS Genetics* 14.8 (Aug. 10, 2018), e1007586. DOI: 10.1371/journal.pgen.1007586.

[102] Can Yang et al. “CoMM: A Collaborative Mixed Model to Dissecting Genetic Contributions to Complex Traits by Leveraging Regulatory Information”. In: *Bioinformatics* 35.10 (May 15, 2019), pp. 1644–1652. DOI: 10.1093/bioinformatics/bty865.

[103] Lu Liu et al. “Multi-Trait Transcriptome-Wide Association Studies with Probabilistic Mendelian Randomization”. In: *The American Journal of Human Genetics* 108.2 (Feb. 4, 2021), pp. 240–256. DOI: 10.1016/j.ajhg.2020.12.006.