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:

_____ _____

Shizhen Tang Date

Powerful variance-component method for TWAS identifies novel and known risk genes for Alzheimer's dementia

By

Shizhen Tang

Master of Science in Public Health

Biostatistics and Bioinformatics

Michael P. Epstein, PhD

(Thesis Advisor)

Jingjing Yang, PhD

(Reader)

Powerful variance-component method for TWAS identifies novel and known risk genes for Alzheimer's dementia

By

Shizhen Tang

B.S.

Xiamen University

2018

Thesis Committee Chair: Michael P. Epstein, 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
2020

Abstract

Powerful variance-component method for TWAS identifies novel and known risk genes for Alzheimer's dementia

By Shizhen Tang

Background: Existing method for detecting disease related genes of complex disease including Genome-wide association studies (GWAS) and Transcriptome-wide association studies (TWAS). Typically, GWAS focuses on detecting the association between common single nucleotide polymorphisms (SNPs) and traits. However, the biology mechanisms for the majority of GWAS signals remain to be determined. Existing TWAS methods such as PrediXcan, FUSION, and TIGAR employ different regression models to estimate cis-eQTL effect sizes from reference panels, but conduct gene-based association studies by Burden approach that models the variant effect size as a linear function of their corresponding cis-eQTL effect size estimate which may not be true for majority genes.

Methods and Materials: We proposed a novel TWAS method based on Sequential Kernel Association Test (SKAT) as VC-TWAS method, which takes cis-eQTL effect size estimates as variant weights but does not model the directions of variant effect sizes. In our studies, we applied PrediXcan and the nonparametric Bayesian Dirichlet process regression (DPR) model to estimate the cis-eQTL effect sizes. In simulation studies, we compared the performance of VC-TWAS and Burden-TWAS and simulated the data using the real genotype data from ROS/MAP dataset to simulation gene expression level and phenotype in two models. In real application, we applied VC-TWAS with the nonparametric Bayesian Dirichlet process regression (DPR) model to study Alzheimer's dementia related phenotypes.

Results: From simulation studies, Compared to Burden-TWAS, VC-TWAS with weights derived from DPR method was shown obtaining the highest power when phenotypes were simulated under the assumption of random effects. From Meta-analysis result, we detected 13 significant TWAS (FDR < 0.05) genes for AD diagnosis, including the well-known GWAS risk gene *TOMM40* with FDR $= 2.86 \times 10^{-9}$.Top novel risk Gene *ZNF234* with FDR $= 1.40 \times 10^{-12}$ and previously detected Gene *TRAPPC6A* by Burden type TWAS with FDR $= 1.52 \times 10^{-10}$ are identified by VC-TWAS. All significant loci are proximal to the major known risk loci *APOE* for Alzheimer's dementia.

Conclusion: Based on those result, our finding provided potential biological interpretations for the known AD risk genes that also had significant TWAS p-values, with respect to the mediated genetic effects through gene expression and the significant association with both AD diagnosis and AD pathology indices.

Powerful variance-component method for TWAS identifies novel and known risk genes for Alzheimer's dementia

By

Shizhen Tang

B.S.

Xiamen University

2018

Thesis Committee Chair: Michael P. Epstein, 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
2020

Contents

1. Introduction

Genome-wide association studies (GWAS) have succeeded in identifying thousands of genetic loci associated with complex traits and diseases. Typically, GWAS focuses on detecting the association between common single nucleotide polymorphisms (SNPs) and the trait of interest[1]. However, the genetic effect sizes on complex traits are often small and rely on large sample sizes to identify significant associations[2]. Single variant tests across all genome-wide SNPs by standard GWAS are also subject to the complication of multiple testing. Additionally, the molecular and biological mechanisms for majority GWAS signals remain to be determined[3]. Studies have shown that gene expression plays a key role in explaining the etiology of complex diseases[4]. It is shown that many common variants associated with diseases are highly likely to be expression quantitative trait loci (eQTL) [5; 6]. Therefore, integrating gene expression information in GWAS is expected to help identify novel risk genes as well as provide biological interpretation.

A novel gene-based test approach, the transcriptome-wide association study (TWAS) has been proposed to integrate transcriptomic and GWAS data[7]. Basically, TWAS first fits imputation models for expression quantitative traits by taking cis-SNP genotype data as predictors, where the broad sense of cis-eQTL effect sizes are estimated using reference panels such as Genotype-Tissues Expression (GTEx) [8]. Then TWAS tests for the association between the imputed genetically regulated gene expression (GReX) per gene and the trait of interest within additional GWAS samples, where GReX is obtained by using the corresponding gene expression imputation model fitted with reference panels. Existing TWAS tools such as PrediXcan[9], FUSION[10], and TIGAR[7]

employ different regression methods to fit gene expression imputation models. For example, PrediXcan method[9] implements the Elastic-Net penalized regression method and TIGAR[7] implements the nonparametric Bayesian Dirichlet process regression (DPR)[11] method. However, the essence of these existing TWAS[7] methods is a weighted burden test[7] (referred to as Burden-TWAS in this paper), which assumes SNP effect sizes on phenotype are of a linear function of their corresponding cis-eQTL effect sizes estimated from gene expression imputation models. This strong assumption does not hold for most genes and complex traits with unknown underlying genetic architectures, thus limiting the potential power of TWAS.

To relax this assumption for general studies, we derive a novel Variance-Component TWAS (VC-TWAS) method that is analogous to the previously proposed Sequence Kernel Association Test (SKAT)[12-14] for gene-based test. Unlike Burden-TWAS methods, our VC-TWAS aggregates genetic information across SNPs within the test gene region using a kernel similarity function that allows upweighting or downweighing of specific variants in the similarity score based on cis-eQTL effect size magnitudes. The test statistic can be thought of as a variance-component score statistic based on a linear mixed model where each variant in the gene has a random effect whose variance is a linear function of the squared values of corresponding cis-eQTL effect size. By modeling variants with random effects, the technique is robust to weight misspecification; both in terms of the direction and magnitude of the weight. That is, our VC-TWAS is expected to be robust to the direction and magnitude of cis-eQTL effect size estimates. In particular, our VC-TWAS uses variant weights that are cis-eQTL effect sizes estimated by either the nonparametric Bayesian Dirichlet process regression (DPR)

method implemented in TIGAR[7] or the Elastic-Net penalized regression method implemented in PrediXcan[9].

In this paper, we conducted in-depth simulation studies to validate the power performance as well as type I errors of our VC-TWAS method, under scenarios with various gene expression heritability and proportion of true causal eQTL and two phenotype models as assumed respectively by Burden-TWAS and VC-TWAS. Compared to Burden-TWAS, VC-TWAS with weights derived from DPR method was shown obtaining the highest power when phenotypes were simulated under the assumption of random effects. Then we applied our VC-TWAS method with weights derived from DPR and PrediXcan methods to GWAS data from Religious Order Study and Memory Aging Project (ROS/MAP)[15-18] and Mayo Clinic late-onsite Alzheimer's disease (LOAD)[19; 20] cohorts for studying Alzheimer's dementia (AD) related phenotypes. Our application studies demonstrated that VC-TWAS with weights derived from DPR method identified both novel and known risk genes for AD within 2MB of the well-known major risk gene *APOE* of AD, including the known risk gene *TOMM40*. Additionally, we integrate this novel VC-TWAS method into our previously developed software tool TIGAR[7] for public use.

In the following sections, we first provide descriptions about TWAS procedure, VC-TWAS, cis-eQTL effect size estimation, GWAS data of ROS/MAP and Mayo Clinic LOAD cohorts, as well as simulation study design in the Methods section. Second, we describe our results from simulation studies and real application studies of AD related phenotypes. Last, we end with a brief discussion about potential impact, computation tool, and current limitations of VC-TWAS.

2. Methods

2.1. TWAS Procedure

TWAS first fits gene expression imputation models by taking genotype data as predictors and assuming the following additive genetic model for expression quantitative traits,

$$\boldsymbol{E_g} = \boldsymbol{Gw} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma_\epsilon^2 \boldsymbol{I}). \qquad \text{(Equation 1)}$$

Here, $\boldsymbol{G}$ is the genotype matrix for all cis-genotypes (encoded as the number of minor alleles or genotype dosages of SNPs within 1MB of the target gene region), $\boldsymbol{w}$ is the cis-eQTL effect size vector, and $\boldsymbol{E_g}$ is the profiled gene expression levels for the target gene $\boldsymbol{g}$. With cis-eQTL effect size estimates $\widehat{\boldsymbol{w}}$ from reference data, $\boldsymbol{GReX}$ will be imputed by the following equation

$$\widehat{\boldsymbol{GReX}} = \boldsymbol{G_{new}}\widehat{\boldsymbol{w}}, \qquad \text{(Equation 2)}$$

where $\boldsymbol{G_{new}}$ is the genotype matrix for the test cohort.

The general test framework of Burden-TWAS[7; 9; 10] that test for association between $\widehat{\boldsymbol{GReX}}$ and the phenotype of interest can be written as:

$$\boldsymbol{Y} = \beta\widehat{\boldsymbol{GReX}} + \boldsymbol{\alpha'Z} + \boldsymbol{\varepsilon} = \beta(\boldsymbol{G_{new}}\widehat{\boldsymbol{w}}) + \boldsymbol{\alpha'Z} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}), \qquad \text{(Equation 3)}$$

where $\widehat{\boldsymbol{GReX}}$ is imputed gene expression levels, $\boldsymbol{Y}$ denotes the phenotype of interest, $\boldsymbol{\alpha'}$ denotes the coefficient vector for other non-genomic covariates $\boldsymbol{Z}$. Basically, Burden-TWAS tests the null hypothesis of $H_0: \beta = 0$, where cis-eQTL effect size estimates ($\widehat{\boldsymbol{w}}$) are taken as variant weights and SNP effect sizes on phenotype ($\beta\widehat{\boldsymbol{w}}$) are assumed of a linear function of $\widehat{\boldsymbol{w}}$[7; 9; 10]. However, this strong assumption of linear relationship is often not true in real studies, which limits the potential power of TWAS.

2.2. VC-TWAS Method

Here, we propose a powerful VC-TWAS method that is analogous to the previously proposed method SKAT for SNP-set based association test[12]. Similar to SKAT, the general test framework of VC-TWAS can be written as

$$\boldsymbol{Y} = \boldsymbol{\beta'G} + \boldsymbol{\alpha'Z} + \boldsymbol{\varepsilon}, \quad \beta_j' \sim N\left(0,\ w_j^2\tau\right), \quad \boldsymbol{\varepsilon} \sim N\ (0,\ \sigma_\epsilon^2), \qquad \text{(Equation 4)}$$

for continuous quantitative traits, and

$$logit\ p(Y_i = 1) = \boldsymbol{\beta'}G_i + \boldsymbol{\alpha'}Z_i,\ \beta_j' \sim N\ (0, \tau w_j^2), \qquad \text{(Equation 5)}$$

for dichotomous traits of sample i. Here, $\boldsymbol{\beta'}$ is the genetic effect size vector, $\boldsymbol{G}$ is the genotype matrix for all test SNPs within the test gene, $\boldsymbol{Z}$ is the non-genomic covariate matrix, and $\boldsymbol{\varepsilon}$ is the error term. VC-TWAS will test $H_0: \tau = 0$, which is equivalent to testing $H_0: \boldsymbol{\beta} = 0$. The variance-component score statistic used by VC-TWAS is given by

$$Q = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})'\boldsymbol{K}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}),\ \boldsymbol{K} = \boldsymbol{GWG'}, \qquad \text{(Equation 6)}$$

where $\widehat{\boldsymbol{\mu}}$ is the estimated phenotype mean under $\boldsymbol{H_0}$ and $\boldsymbol{W} = diag(w_j^2, \dots)$ with weight w_j for the jth variant.

Different from SKAT, VC-TWAS takes cis-eQTL effect size estimates from Equation 1 as variant weights (w_j). That is, the variances (τw_j^2) of SNP effect sizes on phenotype are assumed of a linear function of cis-eQTL effect size estimates, which is robust to both direction and magnitude of cis-eQTL effect size estimates. Since the variance-component score statistic Q (Equation 6) follows a mixture of chi-square distributions under the null hypothesis[21; 22], p-value can be conveniently obtained from several approximation and exact methods like the Davies exact method[23].

2.3. Cis-eQTL effect size estimation

Different methods can be used to estimate cis-eQTL effect sizes $\boldsymbol{w}$ from Equation 1. In this study, we applied PrediXcan and nonparametric Bayesian DPR methods[7; 9] to estimate $\boldsymbol{w}$ and compared the performance of VC-TWAS using cis-eQTL effect sizes estimated by these two methods. Here, we briefly describe the PrediXcan and nonparametric Bayesian DPR methods for estimating $\boldsymbol{w}$.

PrediXcan TWAS method[9] employs Elastic-Net penalized regression method[24] to estimate cis-eQTL effect sizes $\boldsymbol{w}$ from Equation 1. Basically, the Elastic-Net method assumes a combined LASSO (L_1)[25] and Ridge (L_2)[26] penalty and estimate $\boldsymbol{w}$ by the following equation

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{argmin}\left(\left\|\boldsymbol{E_g} - \boldsymbol{Gw}\right\|_2^2 + \lambda(\alpha||\boldsymbol{w}||_1 + \frac{1}{2}(1-\alpha)||\boldsymbol{w}||_2^2)\right) \quad \text{(Equation 7)}$$

Where $\|\cdot\|_1$ denotes L_1 norm, $\|\cdot\|_2$ denotes L_2 norm. Particularly, $\boldsymbol{\alpha}$ is taken as 0.5 by PrediXcan method[9] and penalty parameter $\boldsymbol{\lambda}$ can be tuned by a 5-fold cross validation.

The nonparametric Bayesian DPR method[11] provides a more flexible approach to nonparametrically estimate cis-eQTL effect sizes. The DPR method assumes a normal prior distribution $N(0, \sigma_w^2)$ for cis-eQTL effect sizes and a Dirichlet process prior[27] for effect-size variance $\boldsymbol{\sigma_w^2}$ as follows:

$$\mathrm{w_i} \sim \mathrm{N}(0, \sigma_\mathrm{w}^2),\ \sigma_\mathrm{w}^2 \sim \mathrm{D},\ \mathrm{D} \sim \mathrm{DP}(\mathrm{IG}(\mathrm{a}, \mathrm{b}), \xi). \quad \text{(Equation 8)}$$

That is, the prior distribution D of effect-size variance deviates from a Dirichlet Process (DP) with an inverse gamma (IG) distribution and concentration parameter $\boldsymbol{\xi}$.As proposed by previous studies, variational Bayesian algorithm[28; 29] is implemented to obtain posterior estimates $\widehat{\boldsymbol{w}}$.

2.4. Computational considerations of VC-TWAS

We note that our VC-TWAS method is computationally more complex than standard Burden-TWAS given the need to perform eigen-decomposition of the kernel matrix $\boldsymbol{K}$ in Equation 6 to obtain an analytic p-value. Such eigen-decomposition has computational complexity $O(m^3)$ for considering m SNPs with non-zero cis-eQTL effect sizes. As DPR method produces non-zero cis-eQTL effect size estimates for almost all SNPs within a test gene region (with most cis-eQTL effect size estimates being close to zero[7]), we explored an alternate VC-TWAS that considered a reduced set of SNPs by filtering out those with cis-eQTL effect size estimates smaller than the median cis-eQTL effect size estimate. By doing so, we can reduce up to 80% computation time while having negligible impact on performance relative to using all SNPs with non-zero cis-eQTL effect size estimates. These are validated by our following simulation studies.

2.5. ROS/MAP data

ROS/MAP data are generated from the Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) [15-18], which are ongoing prospective cohort studies of studying aging and dementia. Participants are senior adults showed no signs of dementia when enrolled, who underwent annual clinical evaluation. Brain autopsy was done at the time of death for each participant. All participants signed an informed consent and Anatomic Gift Act, and the studies are approved by an Institutional Review Board of Rush University Medical Center, Chicago, IL. All participants in this study also signed a repository consent to allow their data to be re-purposed. Currently, we have microarray genotype data generated for 2,093 European-decent subjects from ROS/MAP[15-18], which

are further imputed to the 1000 Genome Project Phase 3[30] in our analysis[31]. Post-mortem brain samples (gray matter of the dorsolateral prefrontal cortex) from ~30% these ROS/MAP participants with assayed genotype data are profiled for transcriptomic data by next-generation RNA-sequencing[32], which are used as reference data to train GReX prediction models.

Using ROS/MAP data, we conducted TWAS for clinical diagnosis of LOAD as well as pathology indices of AD quantified with β-antibody specific immunostains, including PHFtau tangle density, β-amyloid load, and a global measure of AD pathology (a combination of neuritic and diffuse plaques and neurofibrillary tangles)[15; 16; 18]. The tangle density quantifies the average PHFtau tangle density within two or more 20μm sections from eight brain regions — hippocampus, entorhinal cortex, midfrontal cortex, inferior temporal, angular gyrus, calcarine cortex, anterior cingulate cortex, and superior frontal cortex. The β-amyloid load quantifies the average percent area of cortex occupied by β-amyloid protein in adjacent sections from the same eight brain regions. These two are based on immunohistochemistry. The global measure of AD pathology is based on counts of neuritic and diffuse plaques and neurofibrillary tangles (15 counts) on 6μm sections stained with modified Bielschowsky[15; 16; 18].

2.6. Mayo Clinic LOAD GWAS data

Mayo Clinic LOAD GWAS data contain samples from two clinical AD Case-Control series: Mayo Clinic Jacksonville (MCJ: 353 AD cases and 331 Controls), Mayo Clinic Rochester (MCR: 291 AD cases and 787 Controls) and a neuropathological series of autopsy-confirmed subjects from the Mayo Clinic Brain Bank (MCBB: 298 AD cases

and 223 non-AD Controls)[19; 20]. In total, we have 844 cases with LOAD and 1,255 controls without a dementia diagnosis. Mayo Clinic LOAD GWAS data have microarray genotype data profiled for 2,099 European-decent samples that are further imputed to the 1000 Genome Project Phase 3[30] in our analysis[31]. This cohort only profiles the phenotype of clinical diagnosis of AD.

2.7. Simulation Study Design

The purpose of this simulation study is to compare the performance of Burden-TWAS and VC-TWAS with variant weights estimated by PrediXcan and DPR methods. We used the real genotype data from ROS/MAP[33] participants to simulate quantitative gene expression and phenotype traits, where the genotype data were of 2,799 cis-SNPs (with $MAF > 5\%$ and Hardy Weinberg p-value $> 10^{-5}$) of the arbitrarily chosen gene *ABCA7*. Specifically, quantitative gene expression traits are generated by the following equation

$$\boldsymbol{E_g} = \boldsymbol{Gw} + \boldsymbol{\varepsilon_E}, \qquad \text{(Equation 9)}$$

where $\boldsymbol{G}$ denotes the genotype matrix of randomly selected true causal eQTL based on a target proportion of causal eQTL (p_{causal}) within the test gene, $\boldsymbol{w}$ denotes cis-eQTL effect sizes generated from $N(0, \sigma_w^2 \boldsymbol{I})$ with variance σ_w^2 chosen to ensure a target gene expression heritability (h_e^2), and $\boldsymbol{\varepsilon_E}$ is the error term generated from $N(0, (1 - h_e^2)\boldsymbol{I})$. Phenotype data are generated based on two models to mimic two different genetic architectures of complex traits in practice.

Model I: The genetic effects on the trait of study are completely driven by genetically regulated gene expression (GReX), where SNP effect sizes are of a linear

function of their corresponding cis-eQTL effect sizes as assumed by Burden-TWAS methods[7; 9; 10]. Phenotype data are generated from the following equation

$$\boldsymbol{Y} = \phi \boldsymbol{E_g} + \boldsymbol{\varepsilon_Y} \ = \phi(\boldsymbol{Gw} + \boldsymbol{\varepsilon_E}) + \boldsymbol{\varepsilon_Y}\ ,\ \ \boldsymbol{\varepsilon_Y} \sim N(0, (1 - h_p^2)\boldsymbol{I}\ ), \qquad \text{(Equation 10)}$$

where $\boldsymbol{E_g}$ is the gene expression generated from Equation 9 and $\phi = \sqrt{h_p^2 / Var\ (\boldsymbol{E_g})}$ is a scalar chosen to ensure a target phenotype heritability (h_p^2).

Model II: The magnitudes of SNP effect sizes on phenotype are driven by their corresponding cis-eQTL effect sizes, while the directions of SNP effect sizes are not restricted. Specifically, variances of SNP effect sizes on phenotype are taken as a linear function of squared values of cis-eQTL effect sizes, as assumed by VC-TWAS. Phenotype data are generated from the following equation

$$\boldsymbol{Y} = \boldsymbol{G\beta} + \boldsymbol{\varepsilon_Y}\ , \boldsymbol{\varepsilon_Y} \sim N\ (0,\ (1 - h_p^2)\boldsymbol{I}\ ), \qquad \text{(Equation 11)}$$

where $\boldsymbol{G}$ denotes the genotype matrix of randomly selected true causal SNPs that are also true causal cis-eQTL as in Equation 9, respective SNP effect sizes are generated from $\beta_i \sim N(0, \varphi {w_i}^2)$ with corresponding cis-eQTL effect size w_i as used in Equation 9 and φ chosen to ensure a phenotype heritability h_p^2.

We considered scenarios with various proportions of causal cis-eQTL/SNPs $p_{causal} = (0.001, 0.01, 0.1, 0.2)$ for the test gene, and various combinations of expression heritability (h_e^2) and phenotype variance/heritability (h_p^2) that were chosen to ensure TWAS power falling within the range of $(25\%, 90\%)$. The values of h_e^2 and h_p^2 were taken as $\left(h_e^2, h_p^2\right) = ((0.001,0.2), (0.01,0.3), (0.05,0.4), (0.1,0.4), (0.2,0.5))$ for simulating phenotypes from Model I, while taken as $\left(h_e^2, h_p^2\right) =$

$((0.001,0.1), (0.01,0.1), (0.05,0.15), (0.1,0.15), (0.2,0.15))$ for simulating phenotypes from Model II.

In our simulation studies, we randomly selected 499 ROS/MAP samples as training data and 1,200 ROS/MAP samples as test data. We estimated cis-eQTL effect sizes from training data by using PrediXcan and DPR methods and then conducted Burden-TWAS and VC-TWAS with test data. For each scenario, we repeated simulations for 1,000 times and obtained the power as the proportion of simulations that had test p-value $< 2.5 \times 10^{-6}$ (genome-wide significance threshold for gene-based test). Additionally, we simulated phenotype under the null hypothesis $Y \sim N(0, 1)$ for 10^6 times and evaluated type I errors of Burden-TWAS and VC-TWAS, using variant weights derived from PrediXcan and DPR methods. For each VC-TWAS, we considered both our original form of the test as well as the alternate form that considered only the filtered set of variants with cis-eQTL effect size estimates greater than the median effect size value to improve computational efficiency.

3. Results

3.1. Simulation results

We compared the performance of VC-TWAS and Burden-TWAS using PrediXcan weights (cis-eQTL effect size estimates by Elastic-Net penalized regression) and DPR weights (cis-eQTL effect size estimates by DPR) under various scenarios. We also evaluated the performance of VC-TWAS and Burden-TWAS using filtered DPR weights as described in Methods.

First, we compared TWAS power for studying phenotypes simulated from Model I that assumed SNP effect sizes on phenotypes were of a linear function of their corresponding cis-eQTL effect sizes. As shown in Figure 1A, for scenarios with various proportions of true causal cis-eQTL/SNPs within the test gene region, $p_{causal} = (0.001, 0.01, 0.1, 0.2)$, and various combinations of expression heritability h_e^2 and phenotype heritability h_p^2, $\left(h_e^2, h_p^2\right) = ((0.001,0.2),\ (0.01,0.3),\ (0.05,0.4),\ (0.1,0.4),\ (0.2,0.5))$, we observed that Burden-TWAS had comparable power with VC-TWAS when $\mathrm{p_{causal}} = (0.001, 0.01)$ with sparse true causal signals, and slightly higher power when $\mathrm{p_{causal}} = (0.1, 0.2)$. In particular, when $\mathrm{p_{causal}} = (0.001, 0.01)$, TWAS methods using PrediXcan weights achieved higher power than using DPR weights. Whereas, when $\mathrm{p_{causal}} > 0.01$, TWAS methods using DPR weights achieved higher power than using PrediXcan weights. These results are consistent with previous Burden-TWAS findings[7]. This is because DPR method is preferred for modeling quantitative gene expression traits when a gene harbors a considerable proportion of true cis-eQTL with relatively smaller effect sizes, e.g., scenarios with $\mathrm{p_{causal}} > 0.01$ in our simulation studies. Across all

considered scenarios, TWAS using filtered DPR weights performed as powerful as using complete DPR.

Second, we compared TWAS power for studying phenotypes simulated from Model II that assumes variances of SNP effect sizes on phenotype were of a linear function of the squared values of their corresponding cis-eQTL effect sizes. As shown in Figure 1B, we found that the VC-TWAS obtained higher power than Burden-TWAS across all scenarios, with $p_{causal} = (0.001, 0.01, 0.1, 0.2)$ and $\left(h_e^2, h_p^2\right) = ((0.001,0.1), (0.01,0.1), (0.05,0.15), (0.1,0.15), (0.2,0.15))$. Especially, when $\mathrm{p_{causal}} \geq 0.01$, the power of VC-TWAS method using DPR weights was twice higher than using PrediXcan weights on average (88.99% vs. 38.45%), while the power of burden-TWAS using DPR weights was comparable with using PrediXcan weights except when $\mathrm{p_{causal}} = 0.001$. When $\mathrm{p_{causal}} = 0.001$ and $h_e^2 \in (0.1,0.2)$, both TWAS approaches using PrediXcan weights performed better than using DPR weights. Again, across all considered scenarios, TWAS using filtered DPR weights performed as powerful as using complete DPR.

In addition, to evaluate type I errors of both TWAS approaches, we conducted 10^6 times simulations under the null hypothesis where phenotypes were not associated with genetic data of the test gene. Without loss of generality, we used gene expression data simulated with $\mathrm{p_{causal}} = 0.2$, $h_e^2 = 0.1$ and generated phenotypes randomly from a $N(0,1)$ distribution. We evaluated type I errors (Table 1) with multiple significant levels $(10^{-2}, 10^{-4}, 2.5 \times 10^{-6})$, demonstrating that both TWAS approaches had type I errors well controlled with all considered significance levels. We also presented the quantile-quantile (QQ) plots of p-values by all methods in Supplementary Figure 1.

To summarize, VC-TWAS performed similarly to Burden-TWAS for studying phenotypes simulated from Model I, while outperformed Burden-TWAS for studying phenotypes simulated from model II. This is because the genetic architecture assumed under Model I is the one assumed by Burden-TWAS with linear relationship between SNP effect sizes on phenotype and cis-eQTL effect sizes. Whereas, Model II assumes a genetic architecture that is more general in practice, where the variances of SNP effect sizes on phenotype are of a linear function with squared values of cis-eQTL effect sizes. In particular, TWAS methods using DPR weights achieved higher power than using PrediXcan weights when $p_{causal} \geq 0.01$, which is consistent with previous studies[7]. Additionally, TWAS using filtered DPR weights achieved similar power as using complete DPR weights, while saving up to 80% computation time.

3.2. Application results

We applied VC-TWAS to the GWAS data of ROS/MAP and Mayo Clinic LOAD cohorts, using SNP weights (i.e., cis-eQTL effect sizes) generated by PrediXcan and filtered DPR methods with 499 ROS/MAP training samples that had both transcriptomic and genetic data profiled[7]. As suggested by previous studies[9; 34], our TWAS results included genes with 5-fold cross validation (CV) $R^2 > 0.005$ for predicting quantitative gene expression traits by either PrediXcan or DPR. We obtained VC-TWAS p-values for 5,710 genes using PrediXcan weights and 12,650 genes using filtered DPR weights. Filtered DPR weights were used in our VC-TWAS such that on average the variance component test considered ~50% SNPs from the test gene region and costs ~3 CPU munities (Supplementary Figure 2). Specifically, the median number of SNPs considered

by VC-TWAS per gene is 2,872 for using filtered DPR weights and 6,632 for using complete DPR weights.

With ROS/MAP GWAS data, we conducted VC-TWAS for four AD related phenotypes including dichotomous phenotype of AD clinical diagnosis, quantitative pathology indices of AD including β-amyloid load, PHFtau tangle density (tangles), and a global measure of AD pathology (gpath). For the dichotomous phenotype of AD clinical diagnosis (N=1,436), we took patients diagnosed with Alzheimer's dementia as cases (N=609), and patients either without cognitive impairment or diagnosed with mild cognitive impairment (MCI) as controls (N=827). For continuous AD pathology phenotypes, among all samples with profiled genetic data, we had 1,294 participants with profiled β-Amyloid, 1,303 participants with profiled tangles, and 1,329 participants with gpath values. In the VC-TWAS of all four phenotypes, we adjusted for covariates of age, smoking status, sex, study group (ROS or MAP), education, and the top three principal components of ancestry.

With Mayo Clinic cohort, we conducted VC-TWAS for AD clinical diagnosis with 844 cases diagnosed with LOAD and 1,255 controls showed no signal of dementia, which adjusted for covariates age, sex, and top three principal components of ancestry. Since only the phenotype of AD clinical diagnosis was profiled by both ROS/MAP and Mayo Clinic cohorts (under different diagnosis criteria) and different sets of covariates were adjusted in VC-TWAS, we conducted meta-analysis with VC-TWAS summary statistics obtained per study by using Fisher's method (meta VC-TWAS) to leverage the power of considering a larger sample size[35].

By meta VC-TWAS, we detected 13 significant risk genes with FDR < 0.05 that were located within ~2MB region around the well-known AD risk gene *APOE* on chromosome 19 (Figure 2A; Table 2). Seven of those significant genes were known risk genes by previous GWAS (*CLASRP, TOMM40, MARK4, CLPTM1, CEACAM19, RELB*)[36] and Burden-TWAS (*TRAPPC6A*)[7]. To investigate whether these significant genes were also involved in the mechanisms of AD pathologies, we investigated the VC-TWAS p-values of these significant genes with respect to AD pathology phenotypes (β-amyloid, tangles and gpath) (Table 3; Figure 2B; Supplementary Figure 3). Interestingly, 11 out of these 13 genes had at least one VC-TWAS p-value <0.05 with respect to one of the AD pathology phenotypes. Recall that gpath is a global measure of AD pathology, which includes weighted quantitative measures of β-amyloid and tangles. Six of these significant genes (*ZNF234, CLASRP, TRAPPC6A, CLPTM1, CEACAM19* and *GIPR*) have VC-TWAS p-value <0.05 for all three AD pathology phenotypes, which are likely to be involved in the biological mechanisms of both β-amyloid and tangles. The other five genes (*TOMM40, MARK4, PPP1R13L, EML2, FBXO46*) have VC-TWAS p-value <0.05 for β-amyloid and gpath, which are likely to be involved only in the biological mechanism of β-amyloid.

For example, the top significant gene *ZNF234* (with FDR $= 1.40 \times 10^{-12}$) by meta VC-TWAS of AD clinical diagnosis is also the top significant gene (p-value $= 2.10 \times 10^{-4}$) by VC-TWAS of β-amyloid, the second most significant gene (p-value $= 6.39 \times 10^{-5}$) by VC-TWAS of gpath, and has p-value $= 1.06 \times 10^{-3}$ by VC-TWAS of tangles. These results showed that the genetic factor of gene *ZNF234* on AD is potentially mediated through its gene expression, and the expression of this gene is also potentially

involved in the mechanisms of both AD pathology indices of β-amyloid and tangles. Besides AD, gene *ZNF234* is also a known risk gene for lipid traits[37]. The genetically regulated gene expression of this gene might also affect lipid traits, thus leading to a pleiotropy phenomenon of AD and lipid traits. Additionally, *ZNF234* is known to be involved in the super pathway of gene expression and is annotated with the Gene Ontology term of nucleic acid binding and DNA-binding transcription factor activity[38].

Another gene of interest is *TOMM40*, which has FDR $= 2.86 \times 10^{-9}$ by meta VC-TWAS for AD clinical diagnosis and VC-TWAS p-values $= (4.44 \times 10^{-4}, 6.95 \times 10^{-2}, 1.91 \times 10^{-4})$ for β-amyloid, tangles, and gpath, respectively. These findings suggest that the genetic effect of this well-known AD risk gene *TOMM40*[39] could be mediated through its gene expression, and the disease mechanism caused by this gene is likely to be only involved with the AD pathology of β-amyloid.

For all SNPs considered by meta VC-TWAS for genes *ZNF234* and *TOMM40*, we colocalized meta GWAS results for AD clinical diagnosis with ROS/MAP and Mayo Clinic cohorts and their corresponding cis-eQTL effect sizes estimated by DPR with ROS/MAP training data. Interestingly, we found that the VC-TWAS association of these two genes were likely to be driven by SNPs around *APOE/TOMM40* loci that also possessed major cis-eQTL effect sizes (Figure 3).

In addition, our VC-TWAS identified a significant gene *HSPBAP1* (FDR = 0.058) for tangles (Supplementary Figure 3B). As shown by previous studies, mRNA of gene *HSPBAP1* was abnormally expressed in the anterior temporal neocortex of patients with intractable epilepsy[40]. Based on our VC-TWAS results, gene *HSPBAP1* might not have a significant genetic effect on AD, but have a significant effect on the global measurement

of the brain pathology (p-value = 4.57×10^{-6}). This showed that gene *HSPBAP1* could be involved in the mechanism of brain pathology tangles and other neurological diseases such as intractable epilepsy[40].

In conclusion, our VC-TWAS method with filtered DPR weights identified both novel and known risk genes for AD clinical diagnosis that are proximal to the major known risk gene *APOE* (Supplementary Figure 4). These results provided potential biological interpretations for the known AD risk genes that also had significant VC-TWAS p-values, with respect to the mediated genetic effects through gene expression and the significant association with both AD clinical diagnosis and AD pathology indices. However, compared with the results of VC-TWAS with PrediXcan weights, no significant genes were identified with FDR <0.05 (Supplementary Figures 5-7), and the genes with smallest p-values were not proximal to *APOE*. This further showed that our VC-TWAS method with filtered DPR weights provided more insight into the genetic origins of AD in these datasets.

4. Discussion

In this paper, we propose a novel variance-component TWAS method which takes cis-eQTL effect sizes as variant weights without the strong assumption of a linear relationship between SNP effect sizes on phenotypes and cis-eQTL effect sizes. By implementing this VC-TWAS with cis-eQTL effect sizes estimated by DPR method[7; 11], we created a powerful test statistic that had good performance in simulation studies and obtained biologically meaningful TWAS results for AD related phenotypes. In particular, we detected 13 significant TWAS genes for AD clinical diagnosis, including the well-known GWAS risk gene *TOMM40* and previously identified TWAS gene *TRAPPC6A*[7]. Moreover, 6 out of these 13 significant genes were identified by previous GWAS[36]. The pleiotropy effects of 11 of these genes with respect to AD clinical diagnosis and pathology indices demonstrated the possible biological mechanisms involved with the AD pathologies of β-amyloid and tangles.

To help users to conduct our VC-TWAS method conveniently and efficiently, we added this function into our previously developed tool — Transcriptome Integrated Genetic Association Resource (TIGAR)[7]. We enabled the choices of using either cis-eQTL effect sizes estimated by PrediXcan method (i.e., Elastic-Net)[24] or nonparametric Bayesian DPR method[12]. Further, we also enabled this new VC-TWAS method by using individual-level for continuous and dichotomous phenotypes. Since the variance component test statistic used by VC-TWAS involves calculating and performing an eigen-decomposition of a genotypic kernel matrix, efficient computation is required (even when filtering variants to include only those variants with larger cis-acting eQTL estimates) for obtaining the corresponding p-values for genome-wide genes. Our TIGAR

packages implements multi-threaded computation to take advantage of high-performance cloud computing clusters and enable practical computation for testing genome-wide genes.

Of course, current TWAS methods including our VC-TWAS still have their limitations. First, because of genetic and transcriptomic heterogeneities across different ethnicities, one may not be able to translate cis-eQTL effect size estimates across cohorts with different ethnicities. That is, applying TWAS methods to GWAS data of a different ethnicity from the reference samples that were used to estimate cis-eQTL effect sizes is likely to fail. Reference panels with diverse ethnicities and multiple tissue types will be needed to expand TWAS to study complex diseases for ethnicities besides Caucasian. As shown in our application studies, since both ROS/MAP and Mayo Clinic cohorts are consisted with European samples, TWAS on Mayo Clinic GWAS data with cis-eQTL effect sizes derived from ROS/MAP training data resulted in interesting results. Second, current TWAS methods fail to account for the uncertainty of cis-eQTL effect-size estimates and trans-eQTL information. Advanced statistical approaches are needed to address such gaps, which may further improve the power of existing TWAS methods. Third, our current VC-TWAS method requires individual-level GWAS data. By deriving the variance component test statistic using GWAS summary statistics will make our method and tool more practically useful for the field. We will be addressing these limitations in our continuing research.

In conclusion, compared with Burden-TWAS methods, our VC-TWAS method assumes a linear relationship between variances of SNP effect sizes on phenotypes and squared values of cis-eQTL effect sizes, which is preferred in practical studies as shown

by our application studies for AD related phenotypes. Compared with SKAT method for gene-based association studies that generally uses variant weights derived from the corresponding minor allele frequency (MAF), our VC-TWAS method integrated transcriptomic data with GWAS data by taking the cis-eQTL effect size estimates as variant weights. That is, our proposed method in this paper provides a powerful TWAS method based on variance component test, which not only employs transcriptomic data in gene-based association studies but also flexibly accounts for the unknown genetic architectures underlying the test genes. As a result, our VC-TWAS method provides the public a useful tool for illustrating the genetic etiology of complex diseases by providing a list of risk genes whose effects on phenotypes might be mediated through transcriptomes.

5. Appendix

5.1. Tables

Significance Level	Burden-TWAS			VC-TWAS		
	DPR	Filtered DPR	PrediXcan	DPR	Filtered DPR	PrediXcan
1.00×10^{-2}	9.82×10^{-3}	9.86×10^{-3}	9.27×10^{-3}	9.43×10^{-3}	9.46×10^{-3}	9.23×10^{-3}
1.00×10^{-4}	8.64×10^{-5}	8.44×10^{-5}	9.95×10^{-5}	8.64×10^{-5}	9.05×10^{-5}	8.24×10^{-5}
2.50×10^{-6}	2.00×10^{-6}	2.00×10^{-6}	2.00×10^{-6}	2.00×10^{-6}	1.00×10^{-6}	6.00×10^{-6}

Table 1. Type I errors under null simulation studies with $p_{causal} = 0.2$, $h_e^2 = 0.1$ for Burden-TWAS and VC-TWAS with DPR weights, filtered DPR weights, and PrediXcan weights, at significant levels $(10^{-2}, 10^{-4}, 2.5 \times 10^{-6})$.

Gene name	Chr	Start	End	P-value	FDR
ZNF234	19	44,645,710	44,664,462	1.11×10^{-16}	1.40×10^{-12}
CLASRP	19	45,542,298	45,574,214	4.44×10^{-16}	2.81×10^{-12}
TRAPPC6A	19	45,666,187	45,681,485	3.60×10^{-14}	1.52×10^{-10}
TOMM40	19	45,394,477	45,406,935	9.05×10^{-13}	2.86×10^{-9}
MARK4	19	45,754,550	45,808,541	4.62×10^{-16}	1.17×10^{-8}
PPP1R13L	19	45,882,892	45,909,607	1.82×10^{-10}	3.84×10^{-7}
CLPTM1	19	45,457,848	45,496,598	5.71×10^{-8}	1.03×10^{-4}
EML2	19	46,112,660	46,148,726	1.88×10^{-7}	2.97×10^{-4}
FBXO46	19	46,213,887	46,234,151	4.13×10^{-7}	5.80×10^{-4}
CEACAM19	19	45,174,724	45,187,631	3.93×10^{-6}	4.68×10^{-3}
GIPR	19	46,171,502	46,185,704	4.07×10^{-6}	4.68×10^{-3}
RELB	19	45,504,695	45,541,452	6.63×10^{-6}	6.99×10^{-3}
ZNF225	19	44,617,548	44,637,255	2.59×10^{-5}	2.51×10^{-2}

Table 2. Significant genes with FDR < 0.05 for phenotype AD clinical diagnosis by meta VC-TWAS with filtered DPR weights, using samples of ROS/MAP and Mayo Clinic cohorts. Known AD risk genes by previous GWAS are shaded in grey.

Gene name	Chr	β-Amyloid	Tangles	Gpath	Phenotypes with p-value <0.05
ZNF234	19	2.10×10^{-4}	1.06×10^{-3}	6.39×10^{-5}	β-amyloid, tangles, gpath
CLASRP	19	1.39×10^{-3}	8.69×10^{-3}	3.76×10^{-4}	β-amyloid, tangles, gpath
TRAPPC6A	19	4.44×10^{-4}	3.74×10^{-3}	1.91×10^{-4}	β-amyloid, tangles, gpath
TOMM40	19	9.55×10^{-4}	6.95×10^{-2}	2.08×10^{-4}	β-amyloid, gpath
MARK4	19	1.73×10^{-2}	2.08×10^{-1}	2.62×10^{-2}	β-amyloid, gpath
PPP1R13L	19	2.64×10^{-2}	1.57×10^{-1}	2.98×10^{-2}	β-amyloid, gpath
CLPTM1	19	1.18×10^{-2}	1.59×10^{-2}	9.24×10^{-3}	β-amyloid, tangles, gpath
EML2	19	8.47×10^{-3}	5.97×10^{-2}	2.16×10^{-2}	β-amyloid, gpath
FBXO46	19	2.02×10^{-3}	8.90×10^{-2}	2.68×10^{-3}	β-amyloid, gpath
CEACAM19	19	1.03×10^{-3}	1.21×10^{-2}	3.19×10^{-5}	β-amyloid, tangles, gpath
GIPR	19	6.37×10^{-3}	3.22×10^{-2}	3.25×10^{-2}	β-amyloid, tangles, gpath

Table 3. VC-TWAS p-values with respect to phenotypes of β-amyloid, tangles, and gpath for genes with FDR <0.05 from meta VC-TWAS with filtered DPR weights on the phenotype of AD clinical diagnosis. Known AD risk genes by previous GWAS are shaded in grey.

5.2. Figures

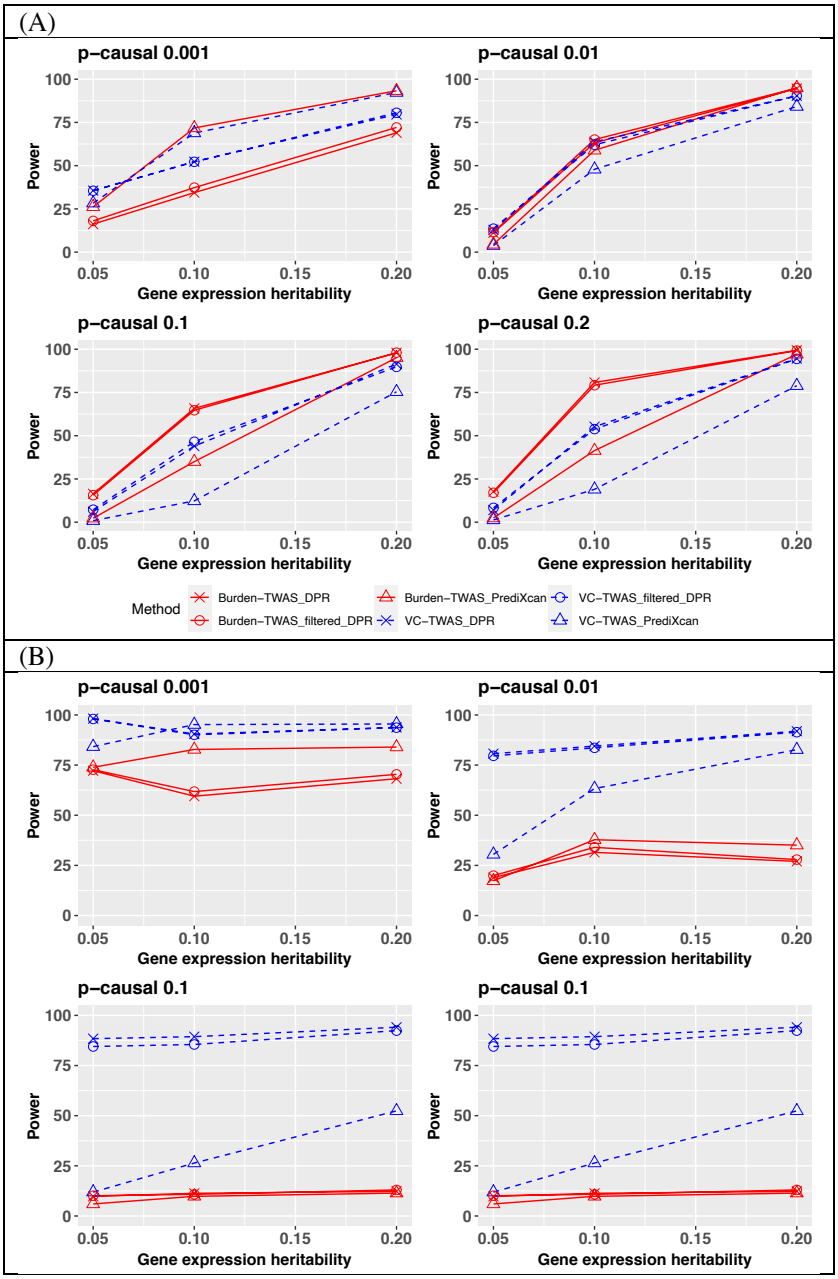


Figure 1. TWAS power comparison for VC-TWAS and Burden-TWAS with phenotypes simulated from Model I (A) and Model II (B). Various types of SNP weights were considered, including those derived from PrediXcan method, DPR method, and filtered DPR weights. Using DPR weights resulted in higher TWAS power than using PrediXcan weights across all scenarios with $p_{causal} \geq 0.01$. TWAS using filtered DPR weights had comparable performance as using complete DPR weights across all scenarios. For phenotypes simulated from Model I (panel A), Burden-TWAS had either comparable power with VC-TWAS when $p_{causal} = (0.001, 0.01)$ or slightly higher power $p_{causal} = (0.1, 0.2)$. For phenotypes simulated from Model II (panel B), VC-TWAS had higher power than Burden-TWAS under all scenarios. VC-TWAS with DPR weights resulted in the highest power.

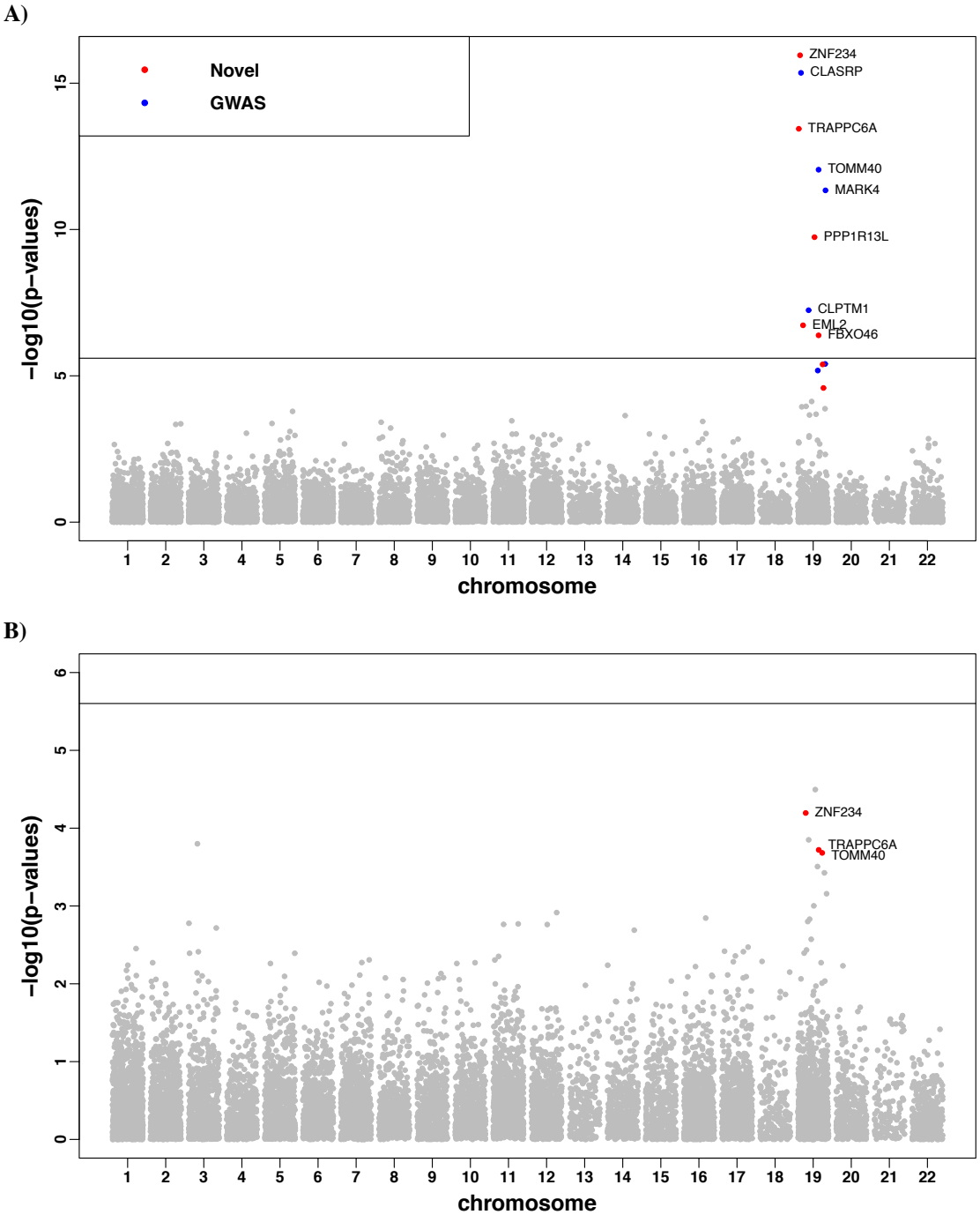


Figure 2. Manhattan plots of meta-analysis VC-TWAS for AD clinical diagnosis (A) and VC-TWAS of global AD pathology (B) with DPR weights. Genes with FDR < 0.05 are highlighted in (A), where red dots denote novel risk genes identified by meta VC-TWAS and blue dots denote known AD risk genes. Genes with FDR < 0.05 in meta VC-TWAS of AD clinical diagnosis and p-value < 0.05 in VC-TWAS of global AD pathology are highlighted in red in (B).

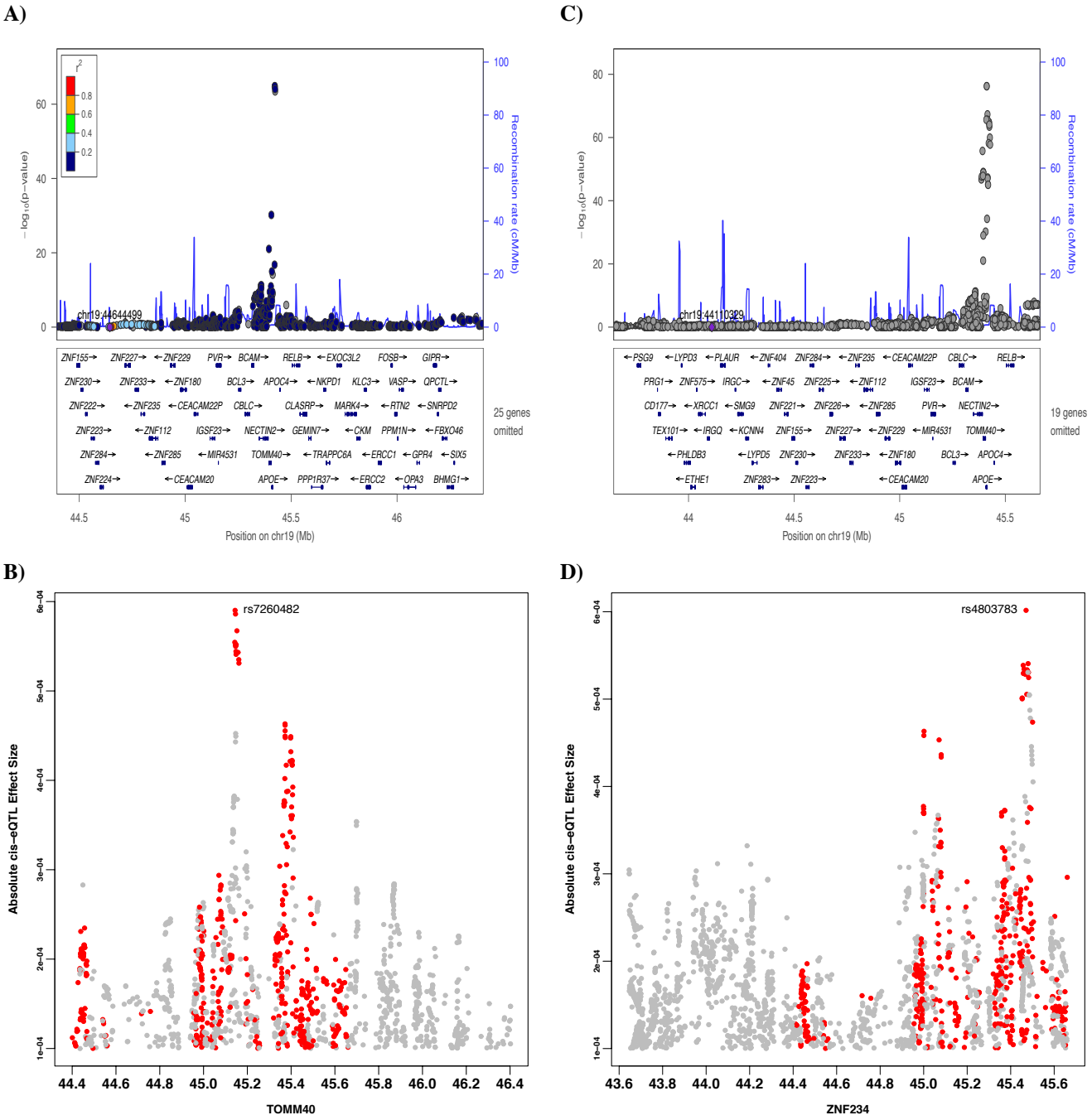
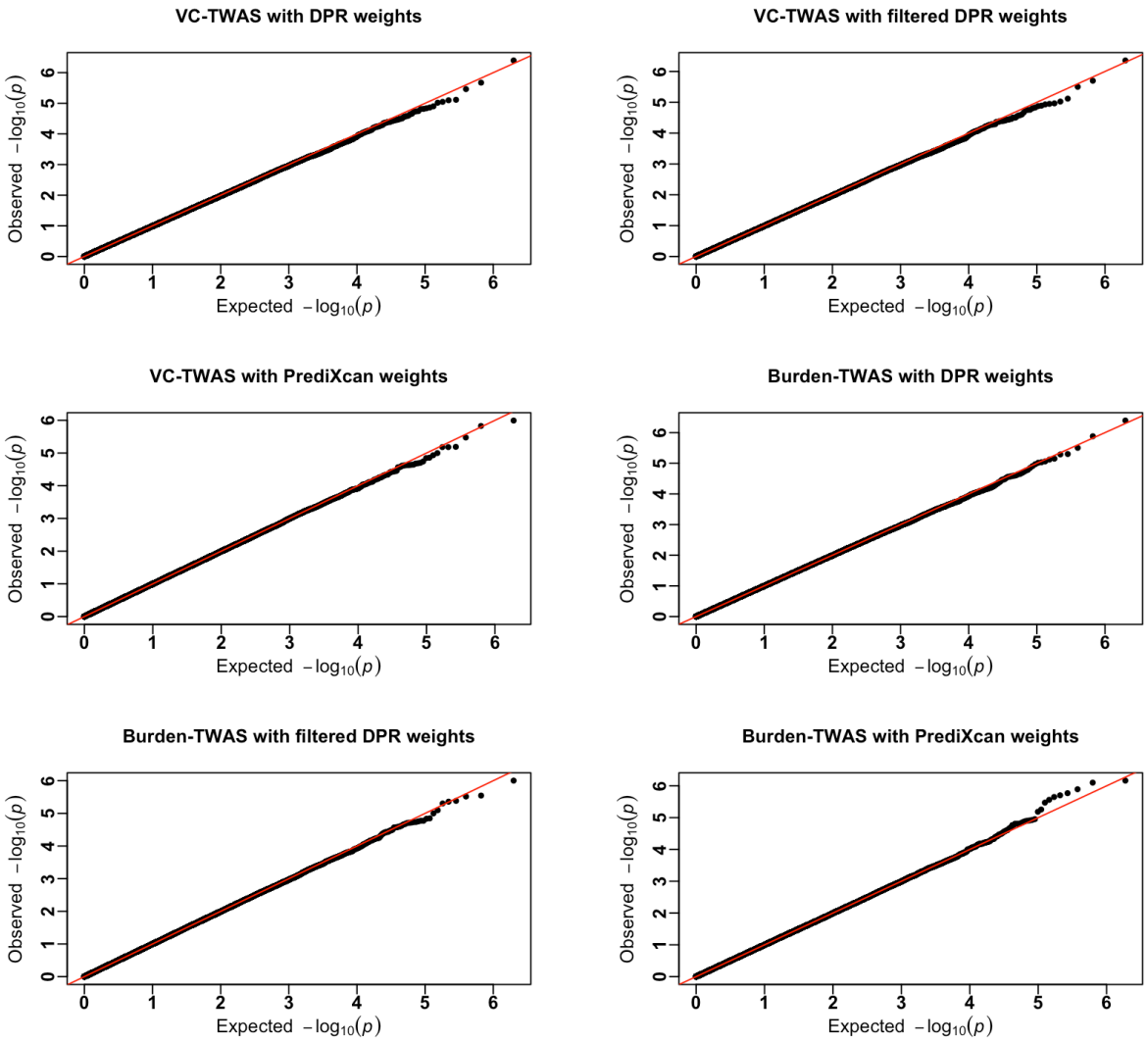


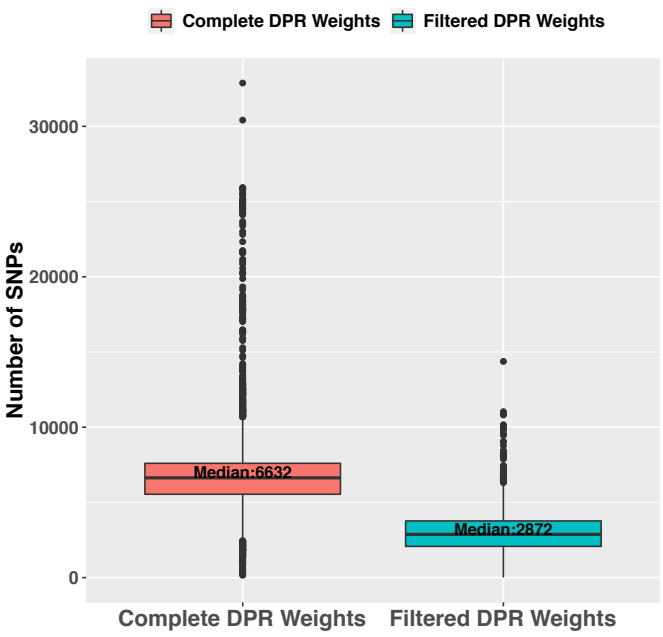
Figure 3. Locus zoom plots of GWAS results and the absolute values of cis-eQTL effect size estimates by DPR method for SNPs that were included in VC-TWAS of genes *TOMM40* (A, B) and *ZNF2334* (C, D). Only SNPs with the absolute values of estimated cis-eQTL effect sizes > 10^{-4} were included for VC-TWAS, same SNPs in *TOMM40* and *ZNF234* are labeled in red.

5.3. Supplemental Data

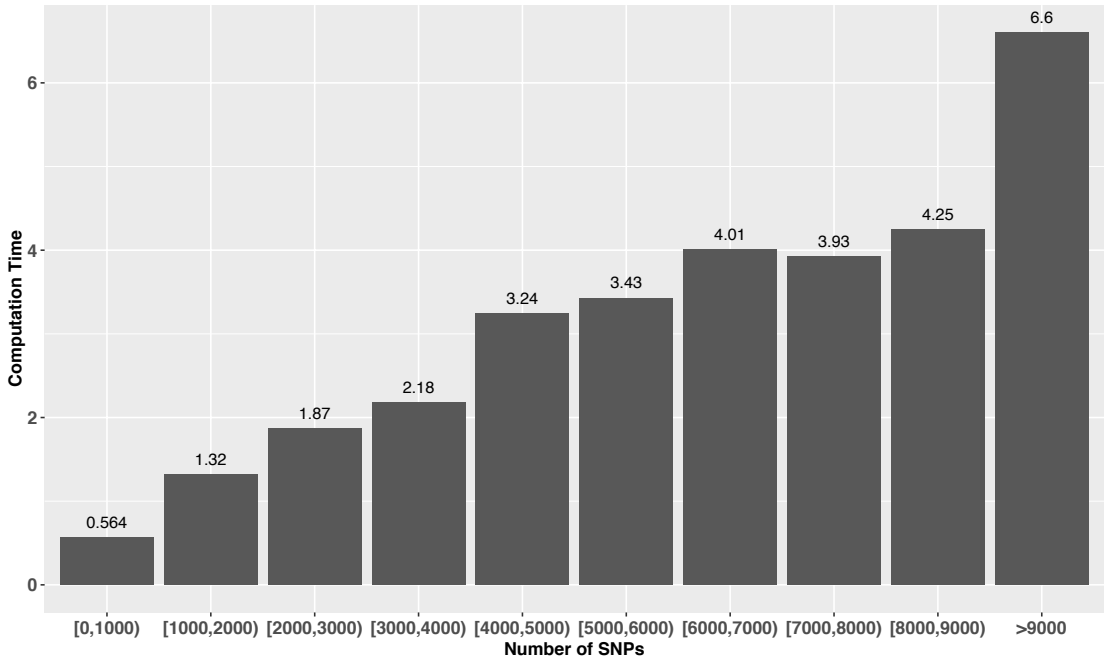


Supplementary Figure 1. Q-Q plots for VC-TWAS and Burden-TWAS with DPR weights, filtered DPR weights, and PrediXcan weights under null hypothesis, where quantitative gene expression traits were generated with $p_{causal} = 0.2$ and $h_e^2 = 0.1$.

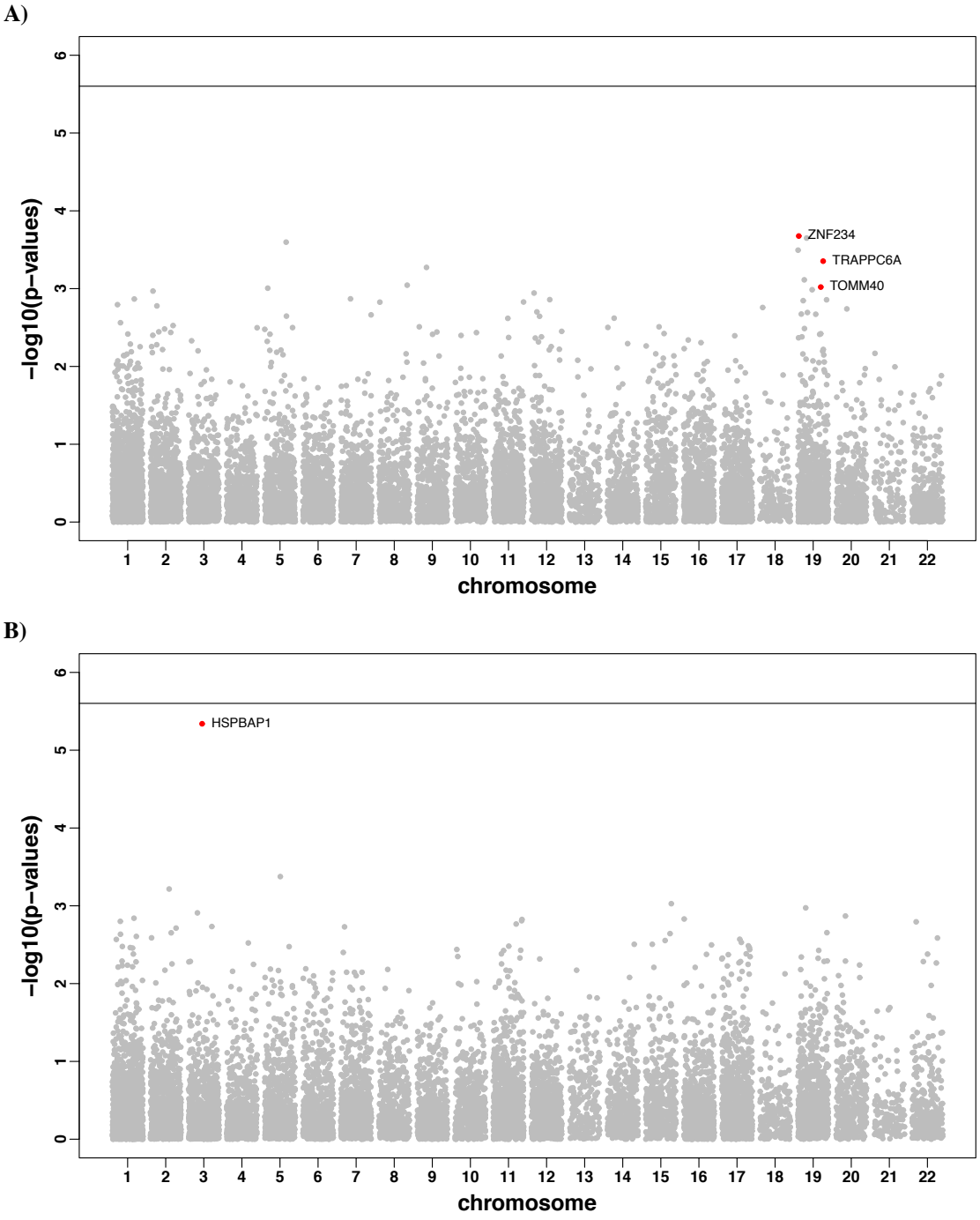
(A)



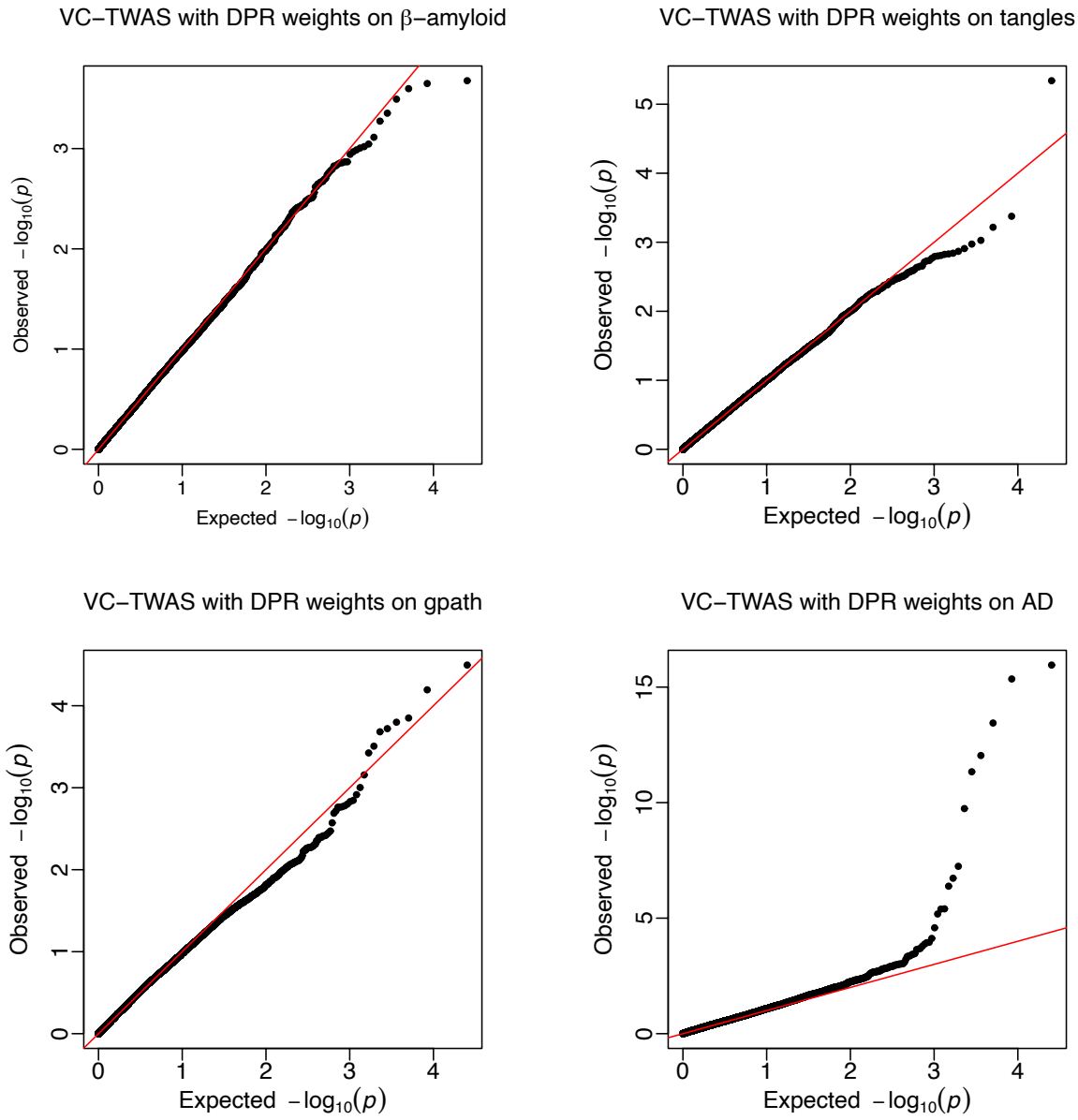
(B)



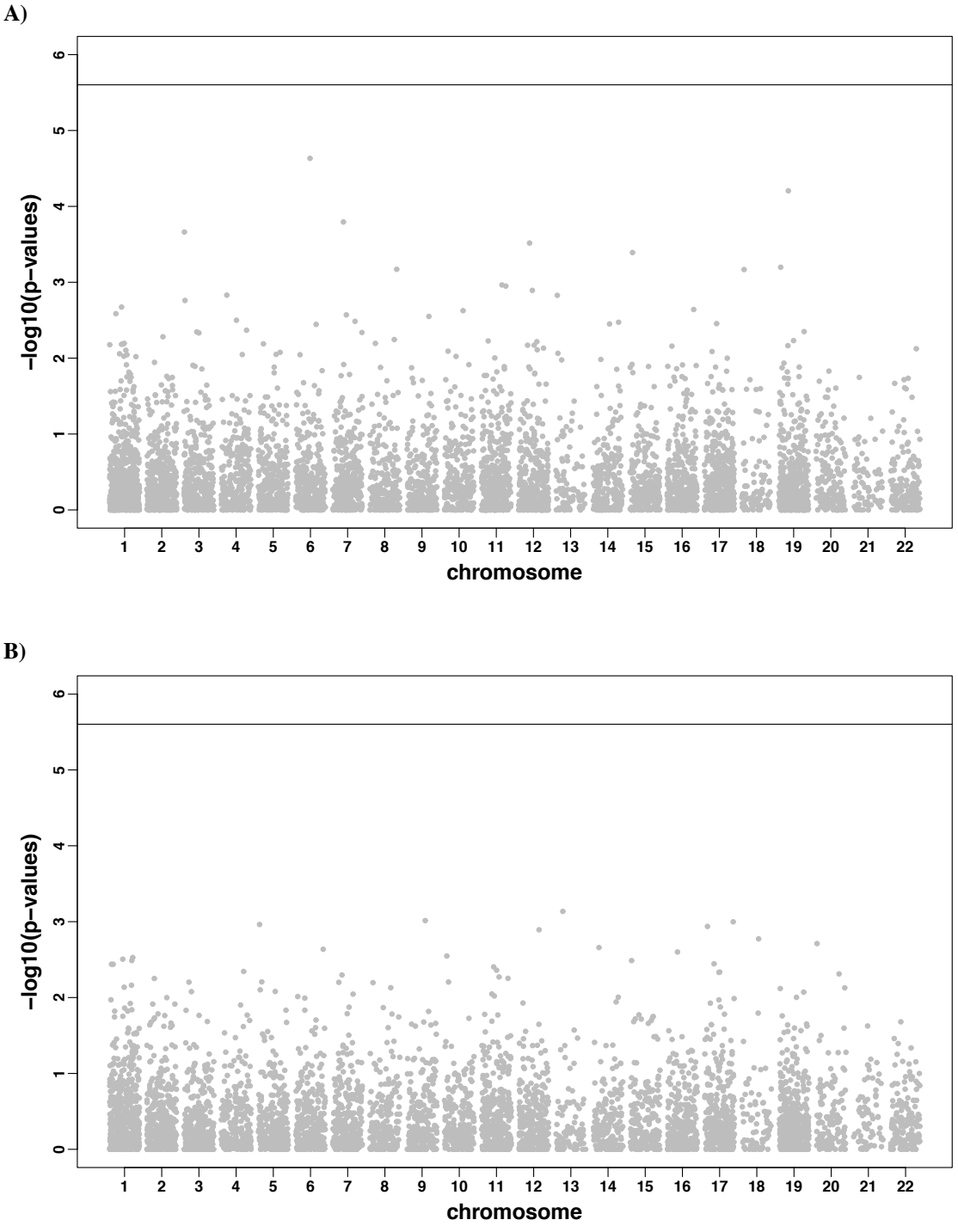
Supplementary Figure 2. (A) Box plot of the number of SNPs considered by VC-TWAS using complete DPR weights and filtered DPR weights derived from ROS/MAP training data. (B) Average computation time for VC-TWAS with respect to the number of SNPs considered in the test. SNPs with filtered DPR weights have cis-eQTL effect sizes estimates > 10^{-4} by DPR method.



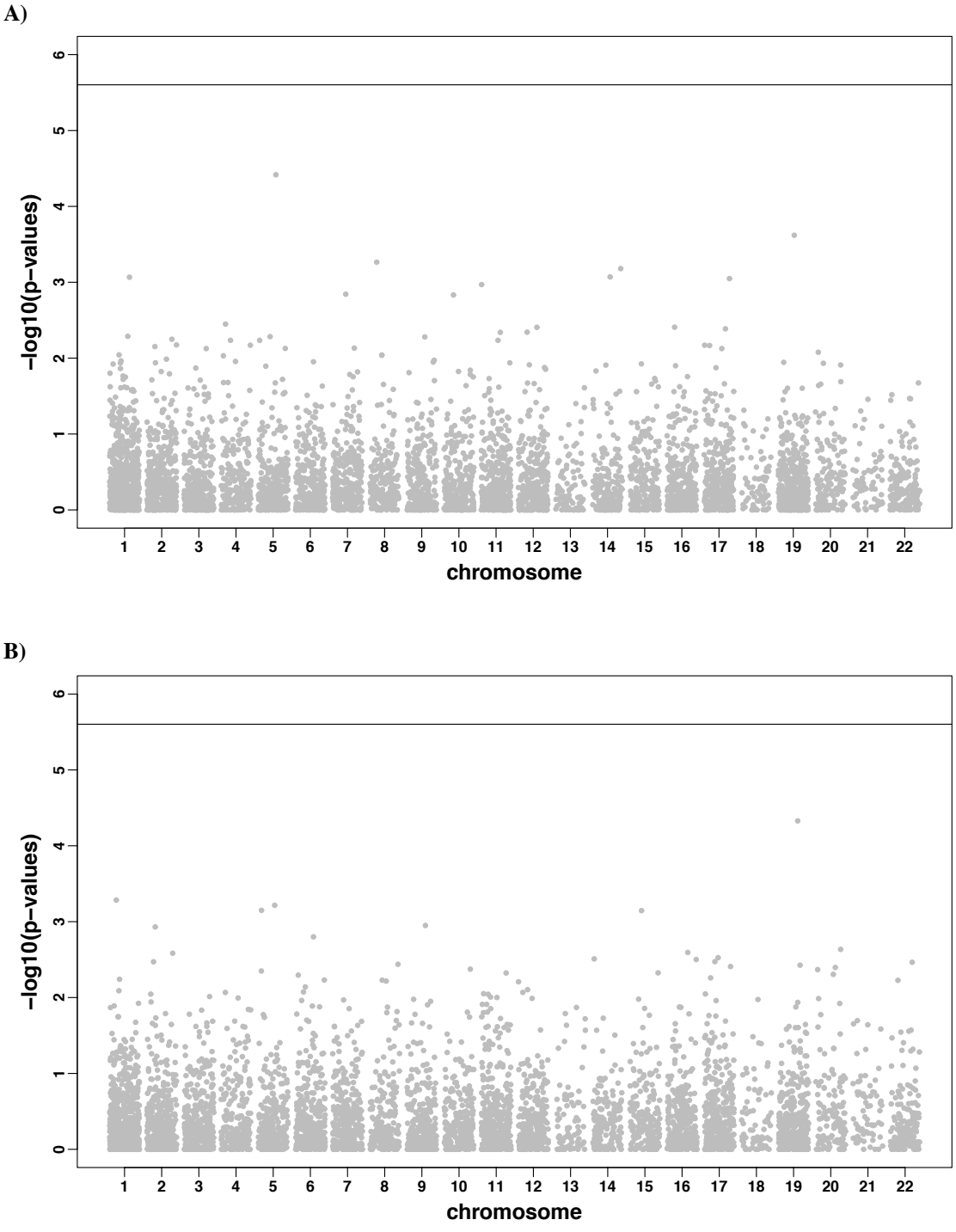
Supplementary Figure 3. Manhattan plots of VC-TWAS results with filtered DPR weights for studying quantitative AD pathology of β-Amyloid (A) and tangles (B). Genes with FDA < 0.05 by meta VC-TWAS of AD clinical diagnosis are colored in red in (A) and top significant gene by VC-TWAS with FDR = 0.058 of tangles is colored in red in (B).



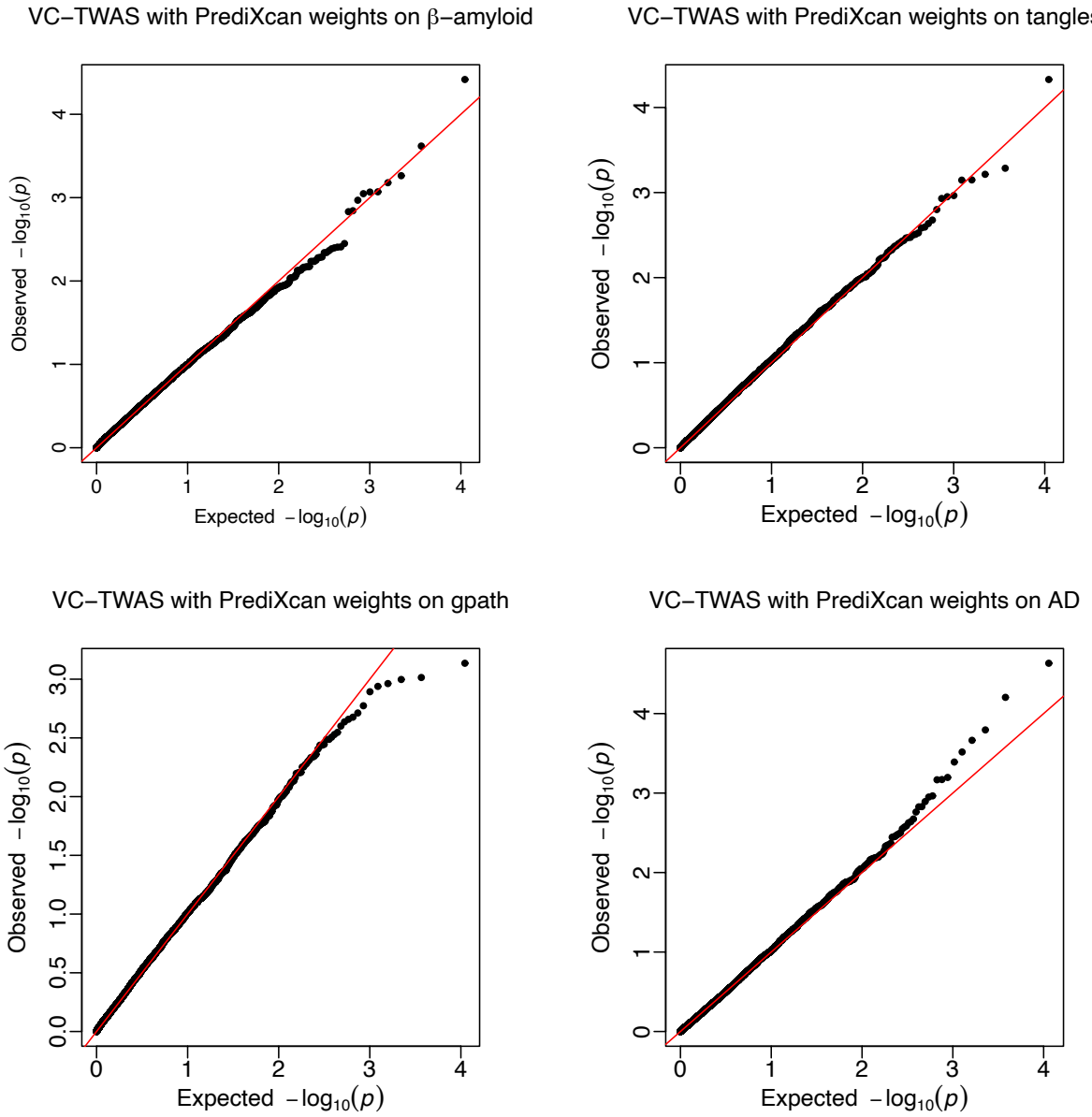
Supplementary Figure 4. Q-Q plots of VC-TWAS results with filtered DPR weights for studying β-amyloid, tangles, and gpath with ROS/MAO cohort, as well as meta VC-TWAS results with filtered DPR weights for studying AD clinical diagnosis with ROS/MAP and Mayo Clinic cohorts.



Supplementary Figure 5. Manhattan plots of VC-TWAS results with PrediXcan weights for studying AD clinical diagnosis (A) and global pathology (B).



Supplementary Figure 6. Manhattan plots of VC-TWAS results with PrediXcan weights for studying quantitative AD pathology of β-Amyloid (A) and tangles (B).



Supplementary Figure 7. Q-Q plots of VC-TWAS results with PrediXcan weights for studying β-amyloid, tangles, and gpath with ROS/MAO cohort, as well as meta VC-TWAS results with PrediXcan weights for studying AD clinical diagnosis with ROS/MAP and Mayo Clinic cohorts.

5.4. Web Resources

TIGAR, https://github.com/yanglab-emory/TIGAR
PrediXcan, https://github.com/hakyim/PrediXcan
RADC Research Resource Sharing Hub, http://www.radc.rush.edu/
ROS/MAP data, https://www.synapse.org/#!Synapse:syn3219045

6. References

1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. American journal of human genetics 90, 7-24.
2. Spencer, C.C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5, e1000477.
3. Cannon, M.E., and Mohlke, K.L. (2018). Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. Am J Hum Genet 103, 637-653.
4. He, X., Fuller, C.K., Song, Y., Meng, Q., Zhang, B., Yang, X., and Li, H. (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. American journal of human genetics 92, 667-680.
5. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS genetics 6, e1000888.
6. Gamazon, E.R., Huang, R.S., Cox, N.J., and Dolan, M.E. (2010). Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proc Natl Acad Sci U S A 107, 9287-9292.
7. Nagpal, S., Meng, X., Epstein, M.P., Tsoi, L.C., Patrick, M., Gibson, G., De Jager, P.L., Bennett, D.A., Wingo, A.P., Wingo, T.S., et al. (2019). TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. The American Journal of Human Genetics 105, 258-266.
8. Consortium, G.T., Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, et al. (2017). Genetic effects on gene expression across human tissues. Nature 550, 204-213.
9. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Consortium, G.T., Nicolae, D.L., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47, 1091-1098.
10. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics 48, 245-252.
11. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat Commun 8, 456.
12. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics 89, 82-93.
13. Kwee, L.C., Liu, D., Lin, X., Ghosh, D., and Epstein, M.P. (2008). A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet 82, 386-397.
14. Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 86, 929-942.

15. Bennett, D.A., Schneider, J.A., Arvanitakis, Z., and Wilson, R.S. (2012). Overview and findings from the religious orders study. Curr Alzheimer Res 9, 628-645.
16. Bennett, D.A., Schneider, J.A., Buchman, A.S., Barnes, L.L., Boyle, P.A., and Wilson, R.S. (2012). Overview and findings from the rush Memory and Aging Project. Curr Alzheimer Res 9, 646-663.
17. Ng, B., White, C.C., Klein, H.U., Sieberts, S.K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C., Bennett, D.A., et al. (2017). An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat Neurosci 20, 1418-1426.
18. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., and Schneider, J.A. (2018). Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis 64, S161-S189.
19. Carrasquillo, M.M., Zou, F., Pankratz, V.S., Wilcox, S.L., Ma, L., Walker, L.P., Younkin, S.G., Younkin, C.S., Younkin, L.H., Bisceglio, G.D., et al. (2009). Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nature genetics 41, 192-198.
20. Zou, F., Chai, H.S., Younkin, C.S., Allen, M., Crook, J., Pankratz, V.S., Carrasquillo, M.M., Rowley, C.N., Nair, A.A., Middha, S., et al. (2012). Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. PLoS genetics 8, e1002707.
21. Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 63, 1079-1088.
22. Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics 9, 292.
23. Moschopoulos, P.G., and Canada, W.B. (1984). The distribution function of a linear combination of chi-squares. Computers & Mathematics with Applications 10, 383-386.
24. Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Statistical Methodology) 67, 301-320.
25. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological) 58, 267-288.
26. Hoerl, A.E., and Kennard, R.W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 42, 80-86.
27. Muller, P., and Mitra, R. (2013). Bayesian Nonparametric Inference - Why and How. Bayesian Anal 8.
28. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational Inference: A Review for Statisticians. Journal of the American Statistical Association 112, 859-877.
29. Carbonetto, P., and Stephens, M. (2012). Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. Bayesian Anal 7, 73-108.

30. Buchanan, C.C., Torstenson, E.S., Bush, W.S., and Ritchie, M.D. (2012). A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. J Am Med Inform Assoc 19, 289-294.
31. Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nature genetics 48, 1284-1287.
32. De Jager, P.L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L.C., Yu, L., Eaton, M.L., Keenan, B.T., Ernst, J., McCabe, C., et al. (2014). Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci 17, 1156-1163.
33. De Jager, P.L., Shulman, J.M., Chibnik, L.B., Keenan, B.T., Raj, T., Wilson, R.S., Yu, L., Leurgans, S.E., Tran, D., Aubin, C., et al. (2012). A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol Aging 33, 1017 e1011-1015.
34. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.O., Lu, Y., Cai, Q., et al. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nature genetics 50, 968-978.
35. Rödel, E. (1971). Fisher, R. A.: Statistical Methods for Research Workers, 14. Aufl., Oliver & Boyd, Edinburgh, London 1970. XIII, 362 S., 12 Abb., 74 Tab., 40 s. Biometrische Zeitschrift 13, 429-430.
36. Jansen, I.E., Savage, J.E., Watanabe, K., Bryois, J., Williams, D.M., Steinberg, S., Sealock, J., Karlsson, I.K., Hagg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nature genetics 51, 404-413.
37. Andaleon, A., Mogil, L.S., and Wheeler, H.E. (2019). Genetically regulated gene expression underlies lipid traits in Hispanic cohorts. PLoS One 14, e0220827.
38. Tripathi, S., Christie, K.R., Balakrishnan, R., Huntley, R., Hill, D.P., Thommesen, L., Blake, J.A., Kuiper, M., and Laegreid, A. (2013). Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. Database (Oxford) 2013, bat062.
39. Chiba-Falek, O., Gottschalk, W.K., and Lutz, M.W. (2018). The effects of the TOMM40 poly-T alleles on Alzheimer's disease phenotypes. Alzheimers Dement 14, 692-698.
40. Xi, Z.Q., Sun, J.J., Wang, X.F., Li, M.W., Liu, X.Z., Wang, L.Y., Zhu, X., Xiao, F., Li, J.M., Gong, Y., et al. (2007). HSPBAP1 is found extensively in the anterior temporal neocortex of patients with intractable epilepsy. Synapse 61, 741-747.