Powerful Variance-Component TWAS method identifies novel and known risk genes for clinical and pathologic Alzheimer's dementia phenotypes

Shizhen Tang[1,2], Aron S. Buchman[3], Philip L. De Jager[4], David A. Bennett[3], Michael P. Epstein[1], Jingjing Yang[1*]

1. Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, 30322, USA
2. Department of Biostatistics and Bioinformatics, Emory University School of Public Health, Atlanta, GA, 30322, USA
3. Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, 60612, USA
4. Center for Translational and Computational Neuroimmunology, Department of Neurology and Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Irving Medical Center, New York, NY10032, USA

* Correspondence author
J.Y., jingjing.yang@emory.edu

Keywords: Transcriptome-wide association study; Variance component test; cis-eQTL; Non-parametric Bayesian Dirichlet process regression; GWAS; Alzheimer's dementia; Brain pathology

Abstract

Transcriptome-wide association studies (TWAS) have been widely used to integrate transcriptomic and genetic data to study complex human diseases. Within a test dataset lacking transcriptomic data, existing TWAS methods impute gene expression by creating a weighted sum that aggregates SNPs with their corresponding cis-eQTL effects on transcriptome estimated from reference datasets. Existing TWAS methods then apply a linear regression model to assess the association between imputed gene expression and test phenotype, thereby assuming the effect of a cis-eQTL SNP on test phenotype is a linear function of the eQTL's estimated effect on reference transcriptome. Thus, existing TWAS methods make a strong assumption that cis-eQTL effect sizes on reference transcriptome are reflective of their corresponding SNP effect sizes on test phenotype. To increase TWAS robustness to this assumption, we propose a Variance-Component TWAS procedure (VC-TWAS) that assumes the effects of cis-eQTL SNPs on phenotype are random (with variance proportional to corresponding cis-eQTL effects in reference dataset) rather than fixed. By doing so, we show VC-TWAS is more powerful than traditional TWAS when cis-eQTL SNP effects on test phenotype truly differ from their eQTL effects within reference dataset. We further applied VC-TWAS using cis-eQTL effect sizes estimated by a nonparametric Bayesian method to study Alzheimer's dementia (AD) related phenotypes and detected 13 genes significantly associated with AD, including 6 known GWAS risk loci. All significant loci are proximal to the major known risk loci *APOE* for AD. Further, we add this VC-TWAS function into our previously developed tool TIGAR for public use.

Introduction

Genome-wide association studies (GWAS) have succeeded in identifying thousands of genetic loci associated with complex traits and diseases[1-3]. However, for the most part the molecular mechanisms linking these genes with these complex traits and diseases remain

unexplained[4]. Studies have shown that gene expression plays a key role in the phenotypic manifestation of human diseases [5]. Many common genetic variants associated with the phenotypes manifested by human diseases are highly likely to be expression quantitative trait loci (eQTL) [6; 7]. Therefore, integrating gene expression data together with genetic data from GWAS is expected to help identify novel risk genes as well as elucidate the mechanisms underlying the associations of genetic loci with complex traits and disease phenotypes.

The transcriptome-wide association study (TWAS) is an innovative strategy to leverage the enhanced power of integrating gene expression of the transcriptome together with genetic data from GWAS [8] for gene-based association studies of complex traits. Using a reference transcriptomic panel like Genotype-Tissues Expression (GTEx) [9], TWAS first trains a regression model that treats gene expression as the outcome and cis-SNP genotype data as predictors, which can be viewed as a gene expression imputation model. Existing TWAS tools such as PrediXcan[10], FUSION[11], and TIGAR[8] employ different regression methods to fit such gene expression imputation models, where corresponding cis-eQTL effect sizes are derived. For example, PrediXcan method[10] implements the Elastic-Net penalized regression method[12] while TIGAR[8] implements the nonparametric Bayesian Dirichlet process regression (DPR)[13] method. Regardless of technique, these existing TWAS methods produce a set of estimates of cis-eQTL effects from the reference expression data. Within a GWAS lacking transcriptomic data, existing TWAS tools then imputes genetically-regulated gene expression (GReX) by weighted summing cis-eQTL SNP data from the GWAS with corresponding estimates of cis-eQTL effects on expression from the reference dataset. Once created, existing TWAS tools then test for association between the imputed gene expression GReX and traits of interest within the GWAS based on a linear regression model. TWAS methods have been deployed widely since their inception and have had substantial success in improving our understanding the genetic regulation of various complex traits[14].

Existing TWAS tools derive GReX as the summation of cis-eQTL SNP genotypes weighted by their corresponding effects on expression derived from the reference dataset. Thus, existing TWAS methods that test the association between GReX and outcome are essentially weighted burden tests[8] (referred to as Burden-TWAS in this paper). Burden-TWAS methods assume the effects of cis-eQTL SNPs on phenotype is a linear function of their corresponding estimated effects on expression based on the reference data (see Methods). This assumption may not be true in practice. For example, the effect of cis-eQTL SNP on phenotype may be a non-linear, rather than a linear, function of the cis-eQTL effect on transcriptome in the reference sample. Moreover, the assumed cis-eQTL effects on transcriptome in the test dataset could be mis-specified due to ancestral differences between the test and reference datasets. For example, recent studies show that gene-expression prediction models trained in one population perform poorly when applied to a different population[15]. The non-linearity assumption about the effects of cis-eQTL SNPs on the phenotype of interest does not affect the validity of TWAS but it can reduce the power of Burden-TWAS methods.

This study derives a novel Variance-Component TWAS (VC-TWAS) method which relaxes the main assumption of Burden-TWAS. This modification makes this method more robust to misspecification of cis-eQTL effect size estimates derived from the reference transcriptome improving its power as compared to the Burden-TWAS methods. VC-TWAS is a TWAS analogue of the Sequence Kernel Association Test (SKAT)[16-19] used for gene-based association studies. 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 (which can be derived using DPR methods in TIGAR or Elastic-Net penalized regression methods implemented in PrediXcan). 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 relaxes the main assumption of Burden-TWAS and is robust to misspecification of cis-eQTL effect size estimates derived from the reference transcriptome; both in terms of the direction and magnitude of the effect.

This manuscript is as organized as follows. First, we provide an overview of standard Burden-TWAS as well as existing techniques for estimating cis-eQTL SNP effects from reference transcriptome data. We then derive our novel VC-TWAS approach and compare its performance with Burden-TWAS using simulated data generated using different eQTL models. We then apply 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)[20-23] and Mayo Clinic Late-Onsite Alzheimer's disease (LOAD)[24; 25] 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*. Finally, we provide a brief discussion summarizing our findings and describing implementation of our method into our previously developed software tool TIGAR[8] for public use.

Methods

Standard 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 $\boldsymbol{g}$. Cis-eQTL effect size estimates $\widehat{\boldsymbol{w}}$ from reference data, ***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[8; 10; 11] that test for association between $\widehat{\boldsymbol{GReX}}$ and the phenotype of interest can be written as:

$$\boldsymbol{E}[\boldsymbol{g}(\boldsymbol{Y}|\boldsymbol{G_{new}})] = \beta\widehat{\boldsymbol{GReX}} + \boldsymbol{\alpha}'\boldsymbol{Z} = \beta(\boldsymbol{G_{new}}\widehat{\boldsymbol{w}}) + \boldsymbol{\alpha}'\boldsymbol{Z}, \qquad \text{(Equation 3)}$$

where $\boldsymbol{Y}$ denotes the phenotype of interest, $g(.)$ denotes a link function, $\widehat{\boldsymbol{GReX}}$ is imputed gene expression levels, and $\boldsymbol{\alpha}'$ 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 to be a linear function of $\widehat{\boldsymbol{w}}$[8; 10; 11]. As noted above the assumption that there is a linear relationship between cis-eQTL and phenotypes may not be true when analyzing human data and limits the power of the existing Burden-TWAS methods.

<u>Cis-eQTL effect size estimation</u>

Different methods can be used to estimate cis-eQTL effect sizes $\boldsymbol{w}$ from Equation 1. In this study, we applied PrediXcan (Elastic-Net penalized regression) and TIGAR (nonparametric Bayesian DPR methods)[8; 10] to estimate $\boldsymbol{w}$. We briefly describe both methods here. PrediXcan TWAS method[10] employs Elastic-Net penalized regression method[12] to estimate cis-eQTL effect sizes $\boldsymbol{w}$ from Equation 1. Basically, the Elastic-Net method assumes a combined LASSO (L_1)[26] and Ridge (L_2)[27] penalty and estimate $\boldsymbol{w}$ by the following equation

$$\widehat{\boldsymbol{w}} = argmin_{\boldsymbol{w}}(\|\boldsymbol{E_g} - \boldsymbol{G}\boldsymbol{w}\|_2^2 + \lambda(\alpha||\boldsymbol{w}||_1 + \frac{1}{2}(1-\alpha)||\boldsymbol{w}||_2^2)) \qquad \text{(Equation 4)}$$

Where $\|\cdot\|_1$ denotes L_1 norm, $\|\cdot\|_2$ denotes L_2 norm. Particularly, $\boldsymbol{\alpha}$ is taken as 0.5 by PrediXcan method[10] and penalty parameter $\boldsymbol{\lambda}$ can be tuned by a 5-fold cross validation.

TIGAR[8] provides a more flexible approach to nonparametrically estimate cis-eQTL effect sizes by a Bayesian DPR method[13]. The DPR method assumes a normal prior distribution $N(0, \sigma_w^2)$ for cis-eQTL effect sizes and a Dirichlet process prior[28] for effect-size variance $\boldsymbol{\sigma_w^2}$ as follows:

$$\mathrm{w_i} \sim \mathrm{N}(0, \sigma_\mathrm{w}^2),\ \ \sigma_\mathrm{w}^2 \sim \mathrm{D},\ \ \mathrm{D} \sim \mathrm{DP}(\mathrm{IG}(\mathrm{a}, \mathrm{b}), \xi). \qquad \text{(Equation 5)}$$

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[29; 30] is implemented to efficiently obtain posterior estimates $\widehat{\boldsymbol{w}}$.

<u>VC-TWAS Method</u>

Here, we propose a powerful VC-TWAS method that is analogous to the previously proposed SKAT method for SNP-set based association studies[16]. Similar to SKAT, the general test framework of VC-TWAS can be written as

$$\boldsymbol{Y} = \boldsymbol{\beta}'\boldsymbol{G} + \boldsymbol{\alpha}'\boldsymbol{Z} + \boldsymbol{\varepsilon},\ \ \beta_j' \sim N\left(0,\ w_j^2\tau\right),\ \ \ \boldsymbol{\varepsilon} \sim N\ (0,\ \sigma_\epsilon^2), \qquad \text{(Equation 6)}$$

for continuous quantitative traits, and

$$E[logit\ \left(Prob(Y_i = 1)\right)] = \boldsymbol{\beta}' G_i + \boldsymbol{\alpha}' Z_i,\ \beta_j' \sim N\ (0, w_j^2\tau), \qquad \text{(Equation 7)}$$

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

$$Q = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})'\boldsymbol{K}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}),\ \ \boldsymbol{K} = \boldsymbol{GWG}', \qquad \text{(Equation 8)}$$

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

In contrast to SKAT methods, 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 to be 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 8) follows a mixture of chi-square distributions under the null hypothesis[31; 32], p-value can be conveniently obtained from several approximation and exact methods like the Davies exact method[33].

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 ***K*** in Equation 8 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[8]), 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. As we show using simulated data, 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.

Compared to Burden-TWAS, VC-TWAS is based on a random effect model and test for if the variation of random effects differs from 0. That is, the variance component score statistic (Equation 8) used by VC-TWAS models no directions of cis-eQTL effect sizes (w_j) and is robust for mis-specifications of cis-eQTL effect sizes.

ROS/MAP data

In this study we used clinical and postmortem data from older adults participating in one of two ongoing community-based cohorts studies the Religious Orders Study (ROS) and Rush

Memory and Aging Project (MAP) [20-23] which document risk factors and chronic conditions of aging including dementia. ROS recruits nuns, priests, and brothers across the United States. MAP recruits participants living in private homes, subsidized housings, and retirement facilities across the greater Chicago metropolitan area.

Participants are senior adults without known dementia, who agree to annual clinical evaluation and brain autopsy at the time of death. 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. ROS/MAP data can be requested at www.radc.rush.edu.

Both studies employ harmonized data collection methods performed by the same staff for annual testing during life and for structured autopsy and collection of genomic data from blood and brain biospecimens. Harmonized data collection facilitates joint analyses of the studies' data. Details of the studies are described elsewhere.

Currently, we have microarray genotype data generated for 2,093 European-decent subjects from ROS/MAP[20-23], which are further imputed to the 1000 Genome Project Phase 3[34] in our analysis[35]. Post-mortem brain samples (gray matter of the dorsolateral prefrontal cortex) from ~30% these ROS/MAP participants with assayed genotype data are also profiled for transcriptomic data by next-generation RNA-sequencing[36], which are used as reference data to train GReX prediction models in our application studies.

Using ROS/MAP data, we conducted TWAS for clinical diagnosis of LOAD as well as pathology indices of AD quantified with β-antibody and PHFtau specific immunostains. We quantified β-amyloid load and PHFtau tangle density; we also generated a summary measure of the burden of AD pathology (a combination of neuritic and diffuse plaques and neurofibrillary tangles based on modified Bielschowsky silver stain) [20; 21; 23].

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[20; 21; 23].

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)[24; 25]. 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[34] in our analysis[35]. This cohort only profiles the phenotype of clinical diagnosis of AD.

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[37] 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 that may be encountered with human data.

Model I: The genetic effects on the phenotype of interest 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[8; 10; 11]. Phenotype data are generated from the following equation

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

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 the phenotype are no longer a linear function of corresponding cis-eQTL effect sizes as in Model I, but rather are derived randomly from a distribution whose variance is a linear function of such effects. By doing this, the SNP effect sizes on phenotype are a function of cis-eQTL effect sizes but the assumption of a linear relationship between the two (assumed by Burden-TWAS) is relaxed. 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 $(h_e^2, h_p^2) = ((0.001,0.2), (0.01,0.3), (0.05,0.4), (0.1,0.4), (0.2,0.5))$ for simulating phenotypes from Model I, while taken as $(h_e^2, h_p^2) = ((0.001,0.1), (0.01,0.1), (0.05,0.15), (0.1,0.15), (0.2,0.15))$ for simulating phenotypes from Model II.

In our simulation studies, we used 499 ROS/MAP samples as training data that were also used as training data in our application studies, and randomly selected 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 computationally-efficient form that considered only the filtered set of variants with cis-eQTL effect size estimates greater than the median effect size value.

Results

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, $(h_e^2, h_p^2) = ((0.001,0.2), (0.01,0.3), (0.05,0.4), (0.1,0.4), (0.2,0.5))$, 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 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 compared to using DPR weights. Whereas, when $\mathrm{p_{causal}} > 0.01$, TWAS methods using DPR weights achieved higher power compared to using PrediXcan weights. These results are consistent with previous Burden-TWAS findings[8]. 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 had similar performance 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 $(h_e^2, h_p^2) = ((0.001,0.1), (0.01,0.1), (0.05,0.15), (0.1,0.15), (0.2,0.15))$. When $\mathrm{p_{causal}} \geq 0.01$, the power of VC-TWAS method using DPR weights was double that 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 had similar performance to that using complete DPR.

In addition, to evaluate type I error of Burden-TWAS and VC-TWAS, 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 across 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 in many situations, while outperforming 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 a function of cis-eQTL effect sizes but not explicitly a linear relationship (leading to more general models of effect). TWAS methods using DPR weights achieved higher power than using PrediXcan weights when $\mathrm{p_{causal}} \geq 0.01$, which is consistent with previous studies[8]. Additionally,

TWAS using filtered DPR weights achieved similar power as using complete DPR weights, while saving up to 80% of computation time.

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[8]. As suggested by previous studies[10; 38], 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.

Leveraging the clinical and postmortem AD data available in ROS/MAP, we were able to apply VC-TWAS to four clinical and pathologic AD phenotypes. We examined final cognitive status diagnosis of 1,436 decedents [AD Dementia (N=609) versus No Dementia (N=827)], as well as three postmortem AD phenotypes including: continuous outcomes of β-amyloid load (N=1,294), PHFtau tangle density (tangles, N=1,303), and a global AD pathology (N=1,329). In the VC-TWAS of all four AD 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, while adjusting 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 for each study by using Fisher's method (meta VC-TWAS)[39] to increase power with a larger sample size.

When we employed 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 1). Seven of those significant genes were known risk genes by previous GWAS (*CLASRP, TOMM40, MARK4, CLPTM1, CEACAM19, RELB*)[40; 41] and Burden-TWAS (*TRAPPC6A*)[8].

As clinical AD dementia can be associated with AD and related pathologies, in further analyses we investigated whether the genes which were found to be associated with AD dementia were also associated with pathologic AD phenotypes. We examined the VC-TWAS p-values of these significant genes with respect to AD pathology phenotypes (β-amyloid, tangles and global AD pathology) (Table 3; Figure 2B; Supplementary Figures 3-4). Interestingly, 11 out of these 13 genes had at least one VC-TWAS p-value <0.05 for one of the three pathologic AD phenotypes. The global AD pathology measure which we examined 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 global AD pathology, which are likely to be involved only in the biological mechanism of β-amyloid.

For example, the top significant gene *ZNF234* (with FDR = 1.40×10^{-12}) by meta VC-TWAS of AD clinical diagnosis is also the top significant gene (p-value = 2.10×10^{-4}) by VC-TWAS of β-amyloid, the second most significant gene (p-value = 6.39×10^{-5}) by VC-TWAS of global AD pathology, and has p-value = 1.06×10^{-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[42]. The genetically regulated gene expression of this gene might also affect lipid traits, thus leading to a pleiotropy phenomenon of AD and lipid traits[43]. 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[44].

Another gene of interest is *TOMM40*, which has FDR = 2.86×10^{-9} by meta VC-TWAS for AD dementia 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 global AD pathology, respectively. These findings suggest that the association of this well-known AD risk gene *TOMM40*[45] could be mediated through its gene expression via β-amyloid but not tangles.

For all SNPs considered by meta VC-TWAS for genes *ZNF234* and *TOMM40*, we colocalized meta GWAS results for AD dementia 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[46]. Based on our VC-TWAS results, gene *HSPBAP1* might not have a significant genetic effect on AD dementia, but might have a significant effect on tangle pathology (p-value = 4.57×10^{-6}) and may account for the increasing recognition of non-cognitive AD phenotypes[47].

This showed that gene *HSPBAP1* could be involved in the mechanism of brain pathology tangles and other neurological diseases such as intractable epilepsy[46; 48].

VC-TWAS with PrediXcan weights, no significant genes were identified with FDR <0.05 (Supplementary Figures 5-7), neither the genes with smallest p-values were proximal to APOE. In contrast, our results with VC-TWAS method using filtered DPR weights provided potential biological interpretations for several known AD risk genes via gene expression and for their associations with both clinical and pathologic AD phenotype.

Discussion

In this paper, we propose a novel variance-component TWAS method that leverages cis-eQTL effect sizes from reference transcriptome but does not make the 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[8; 13], we created a powerful test statistic that had good performance in simulation studies and obtained biologically meaningful TWAS results for both clinical and pathologic AD phenotypes. In particular, we detected 13 TWAS genes for AD dementia, including the well-known GWAS risk gene *TOMM40* and previously identified TWAS gene *TRAPPC6A*[8]. Moreover, 6 out of these 13 genes were identified by previous GWAS[40]. The pleiotropy effects of 11 of these genes with respect to AD dementia and indices of AD pathology demonstrated the possible biological mechanisms linking AD risk genes via β-amyloid and tangles with AD dementia.

To help users conduct our VC-TWAS method conveniently and efficiently, we added this function into our previously developed tool — Transcriptome Integrated Genetic Association Resource (TIGAR)[8]. We enabled the choices of using either cis-eQTL effect sizes estimated by PrediXcan method (i.e., Elastic-Net)[12] 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. A typical VC-TWAS with ~20K genes can be accomplished with ~1,000 CPU hours.

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 have difficulty in translating cis-eQTL effect size estimates across cohorts with different ethnicities[49]. That being said, VC-TWAS is likely more robust to this phenomenon than Burden-TWAS given the former relaxes the assumption of the latter of a linear relationship between the SNP effect size and cis-eQTL effect in the transcriptome. Nevertheless, reference panels with diverse ethnicities and multiple tissue types are certainly needed to enhance TWAS to study complex diseases across different ancestral groups. We should note though that both ROS/MAP and Mayo Clinic cohorts that we considered in this work have similar (European) origins.

Second, current TWAS methods[50] 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 although the method can likely be expanded to include summary-level data following a similar framework to that of Lee et al[51]. We will explore this important enhancement in our continuing research.

Nevertheless, our novel technique allows a more flexible framework to account for the unknown genetic architectures underlying the relationships between SNPs and the phenotype of interest. Using simulated and real AD-related data, we show our method VC-TWAS 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.

Tables

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 $(10^{-2}, 10^{-4}, 2.5 \times 10^{-6})$.

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 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**	**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 3. VC-TWAS p-values with respect to phenotypes of β-amyloid, tangles, and global AD pathology 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. P-values <0.05 are superscripted with star.

Gene	Chr	β-Amyloid	Tangles	Global AD pathology
ZNF234	19	$2.10 \times 10^{-4^*}$	$1.06 \times 10^{-3^*}$	$6.39 \times 10^{-5^*}$
CLASRP	19	$1.39 \times 10^{-3^*}$	$8.69 \times 10^{-3^*}$	$3.76 \times 10^{-4^*}$
TRAPPC6A	19	$4.44 \times 10^{-4^*}$	$3.74 \times 10^{-3^*}$	$1.91 \times 10^{-4^*}$
TOMM40	19	$9.55 \times 10^{-4^*}$	6.95×10^{-2}	$2.08 \times 10^{-4^*}$
MARK4	19	$1.73 \times 10^{-2^*}$	2.08×10^{-1}	$2.62 \times 10^{-2^*}$
PPP1R13L	19	$2.64 \times 10^{-2^*}$	1.57×10^{-1}	$2.98 \times 10^{-2^*}$
CLPTM1	19	$1.18 \times 10^{-2^*}$	$1.59 \times 10^{-2^*}$	$9.24 \times 10^{-3^*}$
EML2	19	$8.47 \times 10^{-3^*}$	5.97×10^{-2}	$2.16 \times 10^{-2^*}$
FBXO46	19	$2.02 \times 10^{-3^*}$	8.90×10^{-2}	$2.68 \times 10^{-3^*}$
CEACAM19	19	$1.03 \times 10^{-3^*}$	$1.21 \times 10^{-2^*}$	$3.19 \times 10^{-5^*}$
GIPR	19	$6.37 \times 10^{-3^*}$	$3.22 \times 10^{-2^*}$	$3.25 \times 10^{-2^*}$

Figure Legends

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 $\mathrm{p_{causal}} \geq 0.01$. 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 $\mathrm{p_{causal}} = (0.001, 0.01)$ or slightly higher power $\mathrm{p_{causal}} = (0.1, 0.2)$. 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 were plotted in this figure. SNPs with GWAS p-value $< 5 \times 10^{-8}$ were colored in red in the plots of cis-eQTL effect sizes (B,D), top significant SNPs by GWAS in (A,C) were colored in blue in the plots of cis-eQTL effect sizes (B,D).

Supplemental Data

Supplementary Figures

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
MayoClinicLOAD data, https://www.synapse.org/#!Synapse:syn2910256

Declaration of Interests

The authors declare no competing interests.

Acknowledgement

J.Y. was supported by the startup funding from Department of Human Genetics at Emory University School of Medicine. M.P.E. was supported by National Institutes of Health grant award R01GM117946. ROS/MAP study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, R01AG56352, U01AG32984, U01AG46152, U01AG61356, the Illinois Department of Public Health, and the Translational Genomics Research Institute. The MCADGC led by Dr. Nilüfer Ertekin-Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL uses samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. MCADGC data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation.

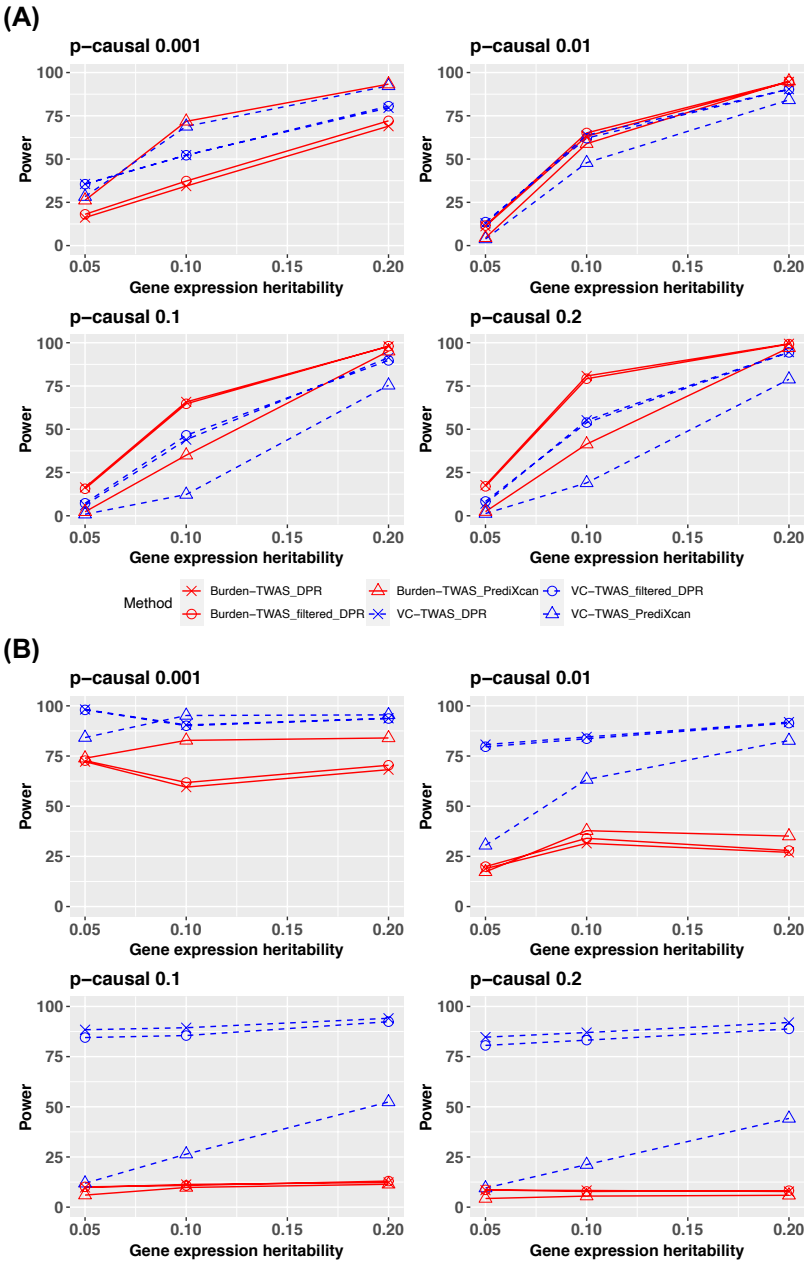
References

1. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9, 356-369.
2. Kettunen, J., Tukiainen, T., Sarin, A.P., Ortega-Alonso, A., Tikkanen, E., Lyytikainen, L.P., Kangas, A.J., Soininen, P., Wurtz, P., Silander, K., et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat Genet 44, 269-276.
3. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45, 1452-1458.
4. 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.
5. 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. Am J Hum Genet 92, 667-680.
6. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6, e1000888.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat Commun 8, 456.
14. Li, Y.I., Wong, G., Humphrey, J., and Raj, T. (2019). Prioritizing Parkinson's disease genes using population-scale transcriptomic data. Nat Commun 10, 994.

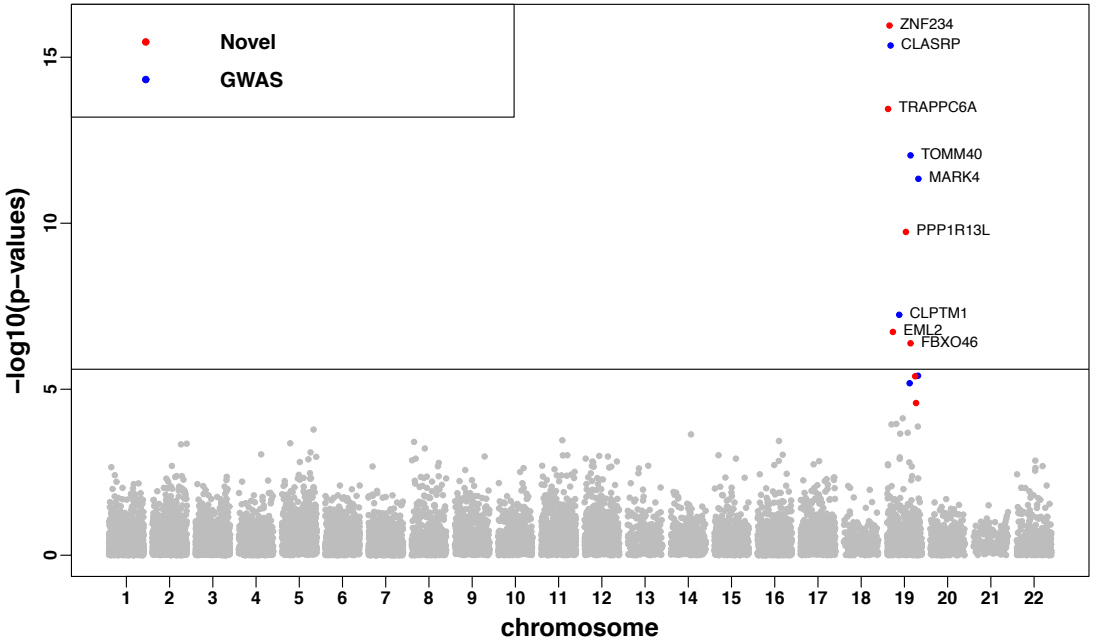
15. Keys, K.L., Mak, A.C.Y., White, M.J., Eckalbar, W.L., Dahl, A.W., Mefford, J., Mikhaylova, A.V., Contreras, M.G., Elhawary, J.R., Eng, C., et al. (2019). On the cross-population generalizability of gene expression prediction models. bioRxiv, 552042.
16. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89, 82-93.
17. 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.
18. 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.
19. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet 92, 841-853.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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. Nat Genet 41, 192-198.
25. 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 Genet 8, e1002707.
26. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological) 58, 267-288.
27. Hoerl, A.E., and Kennard, R.W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 42, 80-86.
28. Muller, P., and Mitra, R. (2013). Bayesian Nonparametric Inference - Why and How. Bayesian Anal 8.
29. 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.
30. 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.

31. 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.
32. 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.
33. 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.
34. 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.
35. 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. Nat Genet 48, 1284-1287.
36. 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.
37. 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.
38. 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. Nat Genet 50, 968-978.
39. 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.
40. 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. Nat Genet 51, 404-413.
41. Marioni, R.E., Harris, S.E., Zhang, Q., McRae, A.F., Hagenaars, S.P., Hill, W.D., Davies, G., Ritchie, C.W., Gale, C.R., Starr, J.M., et al. (2018). GWAS on family history of Alzheimer's disease. Transl Psychiatry 8, 99.
42. Andaleon, A., Mogil, L.S., and Wheeler, H.E. (2019). Genetically regulated gene expression underlies lipid traits in Hispanic cohorts. PLoS One 14, e0220827.
43. Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., Robinson, M.R., McGrath, J.J., Visscher, P.M., Wray, N.R., et al. (2018). Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun 9, 224.
44. 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.
45. 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.

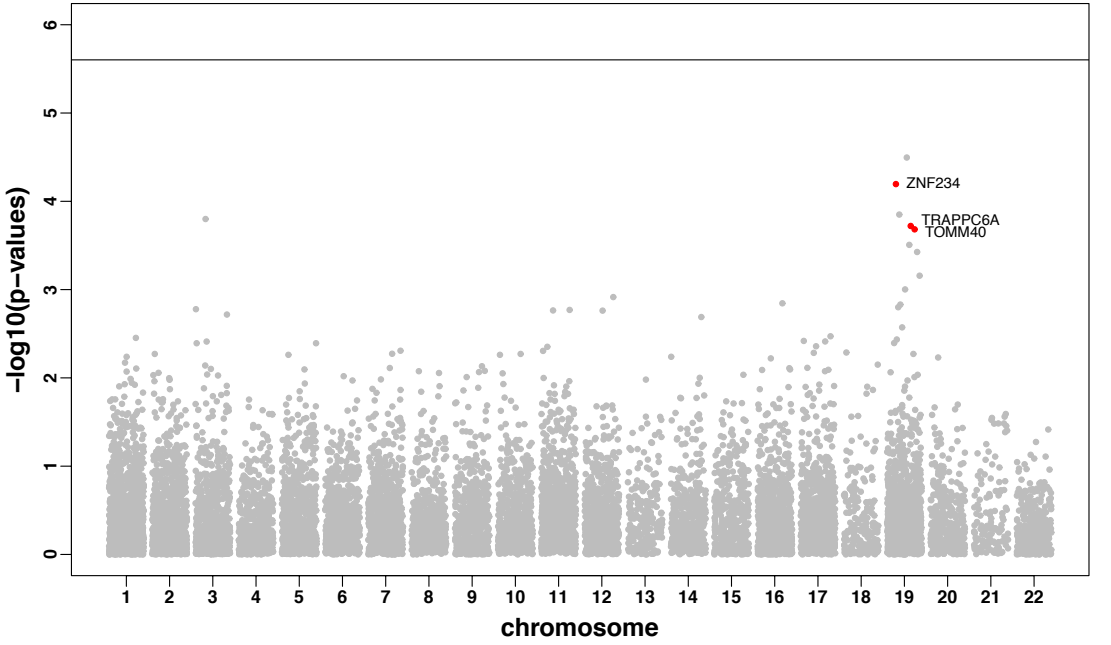
46. Xi, Z.Q., Sun, J.J., Wang, X.F., Li, M.W., Liu, X.Z., Wang, L.Y., Zhu, X., Xiao, F., Li, J.M., Gong, Y., et al. (2007). HSPBAP1 is found extensively in the anterior temporal neocortex of patients with intractable epilepsy. Synapse 61, 741-747.
47. Albers, M.W., Gilmore, G.C., Kaye, J., Murphy, C., Wingfield, A., Bennett, D.A., Boxer, A.L., Buchman, A.S., Cruickshanks, K.J., Devanand, D.P., et al. (2015). At the interface of sensory and motor dysfunctions and Alzheimer's disease. Alzheimers Dement 11, 70-98.
48. Saha, P., and Sen, N. (2019). Tauopathy: A common mechanism for neurodegeneration and brain aging. Mech Ageing Dev 178, 72-79.
49. Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson, W.C., Im, H.K., Liu, Y., and Wheeler, H.E. (2018). Genetic architecture of gene expression traits across diverse populations. PLOS Genetics 14, e1007586.
50. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nat Genet 51, 592-599.
51. Lee, S., Teslovich, T.M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet 93, 42-53.

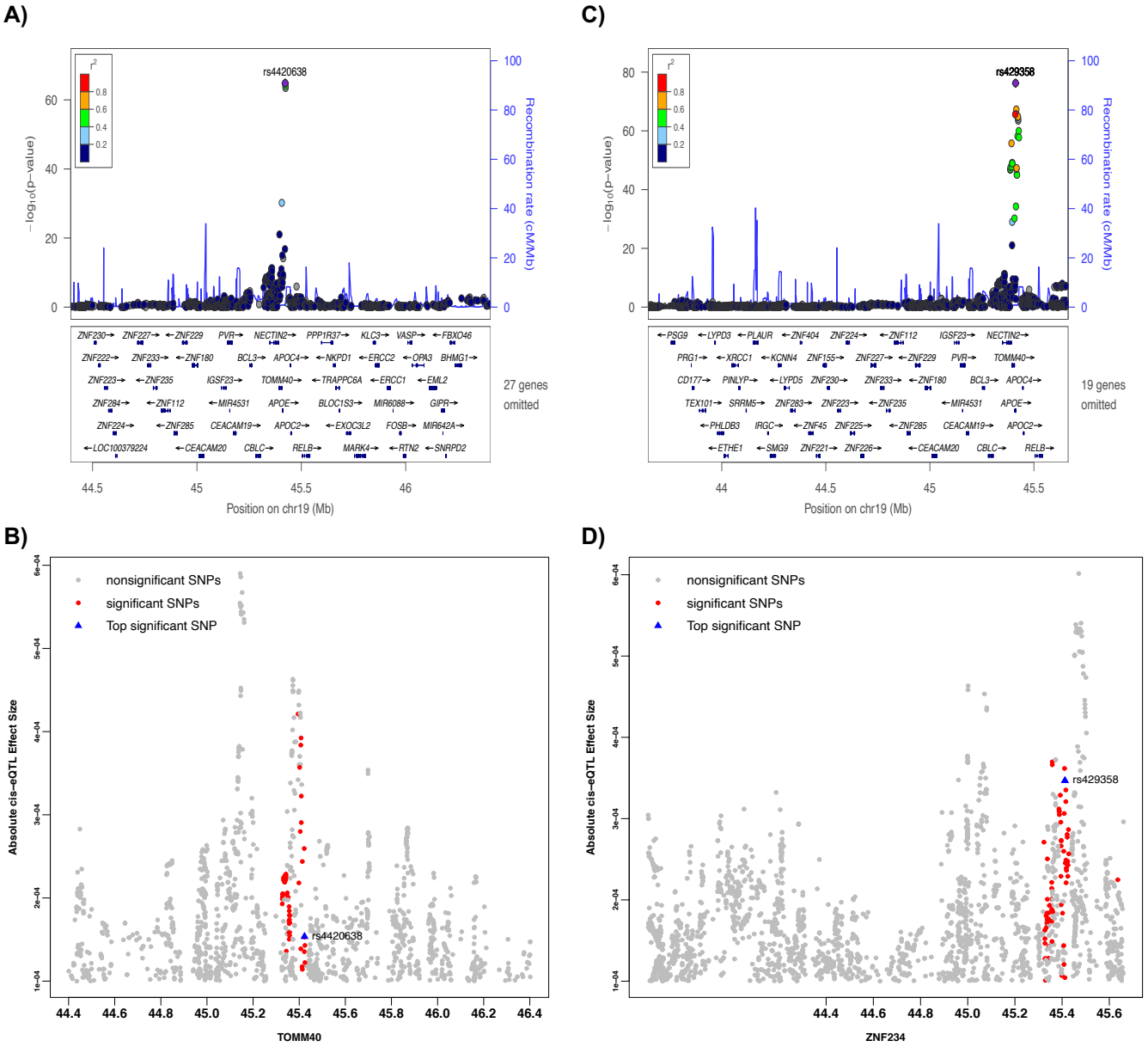

(A)
p−causal 0.001
p−causal 0.01
p−causal 0.1
p−causal 0.2
Power
Gene expression heritability
0
25
50
75
100
0.05
0.10
0.15
0.20
Method
Burden−TWAS_DPR
Burden−TWAS_filtered_DPR
Burden−TWAS_PrediXcan
VC−TWAS_DPR
VC−TWAS_filtered_DPR
VC−TWAS_PrediXcan
(B)
p−causal 0.001
p−causal 0.01
p−causal 0.1
p−causal 0.2
Power
Gene expression heritability

A)


Novel
GWAS
ZNF234
CLASRP
TRAPPC6A
TOMM40
MARK4
PPP1R13L
CLPTM1
EML2
FBXO46
−log10(p−values)
0
5
10
15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
chromosome

B)


ZNF234
TRAPPC6A
TOMM40
−log10(p−values)
0
1
2
3
4
5
6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
chromosome


A)
rs4420638
r^2
0.8
0.6
0.4
0.2
$-\log_{10}$(p−value)
Recombination rate (cM/Mb)
27 genes omitted
Position on chr19 (Mb)
C)
rs429358
19 genes omitted
Position on chr19 (Mb)
B)
nonsignificant SNPs
significant SNPs
Top significant SNP
Absolute cis−eQTL Effect Size
rs4420638
TOMM40
D)
nonsignificant SNPs
significant SNPs
Top significant SNP
Absolute cis−eQTL Effect Size
rs429358
ZNF234