# Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia

## Novel Variance-Component TWAS method for studying complex traits

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
E-mail: J.Y., jingjing.yang@emory.edu

**Keywords:** Transcriptome-wide association study; Variance component test; eQTL; Non-parametric Bayesian Dirichlet process regression; GWAS; Complex human diseases; 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 first impute gene expression by creating a weighted sum that aggregates SNPs with their corresponding cis-eQTL effects on reference transcriptome. Existing TWAS methods then employ 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. To increase TWAS robustness to this assumption, we propose a novel Variance-Component TWAS procedure (VC-TWAS) that assumes the effects of cis-eQTL SNPs on phenotype are random (with variance proportional to corresponding reference cis-eQTL effects) rather than fixed. VC-TWAS is applicable to both continuous and dichotomous phenotypes, as well as individual-level and summary-level GWAS data. Using simulated data, we show VC-TWAS is more powerful than traditional TWAS especially when eQTL genetic effects on test phenotype are no longer a linear function of their eQTL genetic effects on reference transcriptome. We further applied VC-TWAS to both individual-level (N=~3.4K) and summary-level (N=~54K) GWAS data to study Alzheimer's dementia (AD). With the individual-level data, we detected 13 significant risk genes including 6 known GWAS risk genes such as *TOMM40* that were missed by existing TWAS methods. With the summary-level data, we detected 57 significant risk genes considering only cis-SNPs and 71 significant genes considering both cis- and trans- SNPs; these findings also validated our findings with the individual-level GWAS data. Our VC-TWAS method is implemented in the TIGAR tool for public use.

## Author Summary

Existing Transcriptome-wide association studies (TWAS) tools make strong assumptions about the relationships among genetic variants, transcriptome, and phenotype that may be violated in practice, thereby substantially reducing the power. Here, we propose a novel variance-component TWAS method (VC-TWAS) that relaxes these assumptions and can be implemented with both individual-level and summary-level GWAS data. Our simulation studies showed that VC-TWAS achieved higher power compared to existing TWAS methods when the underlying assumptions required by existing TWAS tools were violated. We further applied VC-TWAS to both individual-level (N=~3.4K) and summary-level (N=~54K) GWAS data to study Alzheimer's dementia (AD). With individual-level data, we detected 13 significant risk genes including 6 known GWAS risk genes such as *TOMM40* that were missed by existing TWAS methods. Interestingly, 5 of these genes were shown to possess significant pleiotropic effects on AD pathology phenotypes, revealing possible biological mechanisms. With summary-level data of a larger sample size, we detected 57 significant risk genes considering only cis-SNPs and 71 significant genes considering both cis- and trans- SNPs, which also validated our findings with the individual-level GWAS data. In conclusion, VC-TWAS provides an important analytic tool for identifying risk genes whose effects on phenotypes might be mediated through transcriptomes.

## 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 SNP genotype data (generally cis-SNPs nearby the test gene) 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 on reference transcriptome. 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 reference transcriptome. Once created, existing TWAS tools test for association between the imputed GReX and traits of interest using GWAS data, where the test is based on a single variant 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].

Since existing TWAS tools derive GReX as the summation of eQTL SNP genotypes weighted by their corresponding effects on reference transcriptome, they essentially conduct weighted burden tests [8] to evaluate the association between GReX and outcome (referred to as Burden-TWAS in this paper). Burden-TWAS methods assume the effects of eQTL SNPs on

77 phenotype is a linear function of their corresponding estimated effects on reference transcriptome
78 (see Methods), which may not be true in practice. In particular, the effect of eQTL SNP on
79 phenotype may be a non-linear, rather than a linear, function of the eQTL effect on reference
80 transcriptome. Moreover, the assumed eQTL effects on transcriptome in the test dataset could
81 be mis-specified due to ancestral differences between the test and reference datasets. For
82 example, recent studies show that gene-expression prediction models trained in one population
83 perform poorly when applied to a different population [15]. The non-linearity assumption about
84 the effects of eQTL SNPs on the phenotype of interest does not affect the validity of TWAS but it
85 can reduce the power of Burden-TWAS methods.

86 Here, we derive a novel Variance-Component TWAS (VC-TWAS) method that relaxes the
87 linearity assumption of Burden-TWAS. This modification makes this method more robust to
88 misspecification of eQTL effect size estimates derived from the reference transcriptome, thus
89 improving its power compared to the Burden-TWAS methods. VC-TWAS is a TWAS analogue of
90 the Sequence Kernel Association Test (SKAT) [16-19] used for gene-based association studies.
91 Unlike Burden-TWAS methods, our VC-TWAS aggregates genetic information across test SNPs
92 using a kernel similarity function that allows upweighting or downweighing of specific variants in
93 the similarity score based on eQTL effect size magnitudes (which can be derived using DPR
94 methods in TIGAR or Elastic-Net penalized regression methods implemented in PrediXcan). The
95 test statistic can be thought of as a Variance-Component score statistic based on a mixed model
96 where each test variant has a random effect whose variance is a linear function of the squared
97 values of corresponding eQTL effect size.

98 By modeling variants with random effects, the technique relaxes the main assumption of
99 Burden-TWAS and is robust to misspecification of eQTL effect size estimates derived from the
100 reference transcriptome (in terms of the direction and magnitude). The variance component score
101 test employed by VC-TWAS also enjoys efficient computation and flexibility for studying both
102 quantitative and dichotomous phenotypes, which enables VC-TWAS to consider both cis- and

103 trans- eQTL SNPs. VC-TWAS is further applicable to both individual-level and summary-level
104 GWAS data.

105 We note that VC-TWAS is not the first TWAS method developed that relaxes the linearity
106 assumption of Burden-TWAS. A recently derived collaborative mixed model (CoMM) [20, 21]
107 likewise accounts for the uncertainty of eQTL effect size estimates from reference transcriptome
108 data by jointly modeling reference and test data within a linear mixed-model framework. However,
109 unlike VC-TWAS, CoMM is only applicable to quantitative phenotypes. Moreover, we show that
110 the likelihood ratio test approach used by CoMM is computationally far more expensive to run
111 than VC-TWAS in practice.

112 This manuscript is as organized as follows. First, we provide an overview of standard
113 Burden-TWAS as well as existing techniques for estimating cis-eQTL SNP effects from reference
114 transcriptome data. We then derive VC-TWAS approach for using both individual-level and
115 summary-level GWAS data and compare its performance with Burden-TWAS using simulated
116 data generated under different eQTL model assumptions. We then apply VC-TWAS to individual-
117 level GWAS data from Religious Order Study and Memory Aging Project (ROS/MAP) [22-25] and
118 Mayo Clinic Late-Onsite Alzheimer's disease (LOAD) [26, 27] cohorts, as well as the summary-
119 level GWAS data from the International Genomics of Alzheimer's Project (IGAP) [3], for studying
120 Alzheimer's dementia (AD). By considering only cis-eQTL effect sizes, VC-TWAS identified both
121 novel and known risk genes for AD within 2MB of the well-known major risk gene *APOE* of AD,
122 including the known risk gene *TOMM40* and *APOE*. Considering both cis- and trans- eQTL effect
123 sizes estimated by the Bayesian Genome-wide TWAS (BGW-TWAS) method [28], VC-TWAS
124 detected 71 risk genes for AD which complemented existing TWAS results using only cis-eQTL
125 SNP data. After describing the results, we provide a brief discussion summarizing our findings
126 and describing implementation of VC-TWAS into our previously developed software tool TIGAR
127 [8] for public use.
128

# 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,

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

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

$$\widehat{\boldsymbol{GReX}} = \boldsymbol{G_{new}}\widehat{\boldsymbol{w}}, \tag{2}$$

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:

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

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 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 this linear relationship between eQTL effect sizes on reference transcriptome and the SNP effect sizes on test phenotype may not be true when analyzing human data and limits the power of the existing Burden-TWAS methods.

## Estimation of eQTL effect sizes

Different methods can be used to estimate 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 $\boldsymbol{w}$ that only consider cis-eQTL. PrediXcan TWAS method [10] employs Elastic-Net penalized regression method [12] to estimate cis-eQTL effect sizes $\boldsymbol{w}$ from Equation (1) (Text S1). TIGAR[8] provides a more flexible approach to nonparametrically estimate cis-eQTL effect sizes by a Bayesian DPR method [13] (Text S2). Additionally, we also considered modeling gene expression using both cis- and trans- eQTL effect sizes estimated by the recently proposed Bayesian genome-wide TWAS (BGW-TWAS) method [28].

## VC-TWAS with individual-level GWAS data

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

$$Y_i = \boldsymbol{\beta}'\boldsymbol{G_i} + \boldsymbol{\alpha}'\boldsymbol{Z_i} + \varepsilon_i, \quad \beta_j' \sim N\left(0,\ w_j^2\tau\right), \quad \varepsilon_i \sim N\left(0,\ \sigma_\epsilon^2\right),$$

for continuous quantitative traits, and

$$logit\left(Prob(Y_i = 1)\right) = \boldsymbol{\beta}'\boldsymbol{G_i} + \boldsymbol{\alpha}'\boldsymbol{Z_i} + \varepsilon_i,\ \beta_j' \sim N\left(0, w_j^2\tau\right),\ \varepsilon_i \sim N\left(0,\ \sigma_\epsilon^2\right),$$

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 with respect to 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

$$Q = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}})'\boldsymbol{K}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}),\ \boldsymbol{K} = \boldsymbol{GWG}', \tag{4}$$

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 $j$th variant.

In contrast to SKAT methods, VC-TWAS takes eQTL effect size estimates from Equation (1) as variant weights ($w_j$). That is, the variances ($\tau w_j^2$) of SNP effect sizes on phenotype are assumed to be a linear function of eQTL effect size estimates, which is robust to both direction

179 and magnitude of eQTL effect size estimates. Since the Variance-Component score statistic $Q$
180 (Equation (4)) follows a mixture of chi-square distributions under the null hypothesis [29, 30], p-
181 value can be conveniently obtained from several approximation and exact methods like the
182 Davies exact method [31].

183 We note that our VC-TWAS method is computationally more complex than standard
184 Burden-TWAS given the need to perform eigen-decomposition of the kernel matrix $\boldsymbol{K}$ in Equation
185 4 to obtain an analytic p-value (Text S3). Such eigen-decomposition has computational
186 complexity $O(m^3)$ for considering $m$ SNPs with non-zero eQTL effect sizes. As DPR method
187 produces non-zero cis-eQTL effect size estimates for almost all test SNPs (with most cis-eQTL
188 effect size estimates being close to zero [8]), we explored an alternate VC-TWAS that considered
189 a reduced set of SNPs by filtering out those with cis-eQTL effect size estimates smaller than the
190 median cis-eQTL effect size estimate. As we show using simulated data, we can reduce up to 80%
191 computation time while having negligible impact on performance relative to using all SNPs with
192 non-zero cis-eQTL effect size estimates.

193 Compared to Burden-TWAS, VC-TWAS is based on a random effect model and test for if
194 the variation of random effects differs from 0. That is, the variance component score statistic
195 (Equation (4)) used by VC-TWAS does not directly model the directions of cis-eQTL effect sizes
196 ($w_j$) and is robust for mis-specifications of cis-eQTL effect sizes.

197

198 **VC-TWAS with summary-level GWAS data**

199 Since summary-level GWAS data generally provide SNP effect sizes on phenotype and
200 corresponding standard errors based on single variant tests (often meta-analysis), it is reasonable
201 to assume that the test phenotypes were adjusted for the confounding covariates. That is, in the
202 $Q$ statistic used by VC-TWAS as in Equation (4), we can assume the phenotype mean $\hat{\boldsymbol{\mu}}$ under
203 $\boldsymbol{H_0}$ is 0. This leads to a simplified formula [32, 33]

$$Q = \sum_{j=1}^{m} w_j^2 s_j^2, \quad (5)$$

where $s_j = \boldsymbol{G}_{\cdot j}'\boldsymbol{Y}/\widehat{\sigma_Y}^2$ is the single variant score statistic of the $j^{th}$ variant, $\widehat{\sigma_Y}^2$ is the estimated phenotype variance.

As derived in the previous studies [34], we can estimate phenotype variance and then the score statistics by using only GWAS summary statistics including the single variant effect size estimate $\hat{\beta}_j$ and corresponding standard error $\widehat{\sigma_j}$ for $j^{th}$ SNP, sample size $n$, and a reference LD covariance matrix $\boldsymbol{\Sigma}$ of all test SNPs (Text S3). That is, the numerator and denominator of the score statistic can be estimated by,

$$\boldsymbol{G}_{\cdot j}'\boldsymbol{Y} = (n-1)\widehat{\beta_j}\Sigma_{j,j}\ ;\ \widehat{\sigma_Y}^2 = median\left(\Sigma_{j,j}\widehat{\sigma_j^2}(n-1) + \Sigma_{j,j}\hat{\beta}_j^{\ 2}; j = 1, \dots, m\right). \quad (6)$$

## ROS/MAP data

In this study we used clinical and postmortem data from older adults participating in two ongoing community-based cohorts studies the Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) [22-25] which document risk factors and chronic conditions of aging including dementia. 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.

Currently, we have microarray genotype data generated for 2,093 European-decent subjects from ROS/MAP [22-25], which are further imputed to the 1000 Genome Project Phase 3 [35] in our analysis [36]. 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 [37], which are used as reference data to train GReX prediction models in our application studies.

229 Using ROS/MAP data, we conducted TWAS for clinical diagnosis of late on-site
230 Alzheimer's dementia (LOAD) as well as three pathology indices of AD including quantified $\beta$-
231 amyloid load and PHFtau tangle density as well as a summary measure of AD pathology burden
232 (Text S4) [22, 23, 25].
233

234 **Mayo Clinic LOAD GWAS data**

235 Mayo Clinic LOAD GWAS data contain samples from two clinical AD Case-Control series:
236 Mayo Clinic Jacksonville (MCJ: 353 AD cases and 331 Controls), Mayo Clinic Rochester (MCR:
237 245 AD cases and 701 Controls) and a neuropathological series of autopsy-confirmed subjects
238 from the Mayo Clinic Brain Bank (MCBB: 246 AD cases and 223 non-AD Controls) [26, 27] . In
239 total, we have 844 cases with LOAD and 1,255 controls without a dementia diagnosis. Mayo Clinic
240 LOAD GWAS data have microarray genotype data profiled for 2,099 European-decent samples
241 that are further imputed to the 1000 Genome Project Phase 3 [35] in our analysis [36]. This cohort
242 only profiles the phenotype of clinical diagnosis of AD.
243

244 **IGAP GWAS summary statistics of AD**

245 We used the GWAS summary statistics of AD generated from the stage 1 of the
246 International Genomics of Alzheimer's Project (IGAP) with individuals of all European ancestry
247 [3]. This summary-level GWAS data was generated by meta-analysis (N=~54K) of four
248 previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and
249 37,154 controls (The European Alzheimer's disease Initiative – EADI the Alzheimer Disease
250 Genetics Consortium – ADGC The Cohorts for Heart and Aging Research in Genomic
251 Epidemiology consortium – CHARGE The Genetic and Environmental Risk in AD consortium –
252 GERAD).
253

## 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, as well as validate the VC-TWAS approach of using only summary-level GWAS data. We used the real genotype data from ROS/MAP [38] 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

$$\boldsymbol{E_g} = \boldsymbol{Gw} + \boldsymbol{\varepsilon_E}, \tag{7}$$

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 $\sigma_w^2$ 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

$$\boldsymbol{Y} = \phi \boldsymbol{E_g} + \boldsymbol{\varepsilon_Y} = r(\boldsymbol{Gw} + \boldsymbol{\varepsilon_E}) + \boldsymbol{\varepsilon_Y}, \quad \boldsymbol{\varepsilon_Y} \sim N(0, (1 - h_p^2)\boldsymbol{I}), \tag{8}$$

where $\boldsymbol{E_g}$ is the gene expression generated from Equation (7) and $r = \sqrt{h_p^2 / Var(\boldsymbol{E_g})}$ 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 the squares 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

$$\boldsymbol{Y} = \boldsymbol{G\beta} + \boldsymbol{\varepsilon_Y}\ , \boldsymbol{\varepsilon_Y} \sim N\ (0,\ (1 - h_p^2)\boldsymbol{I}\ ), \tag{9}$$

where $\boldsymbol{G}$ denotes the genotype matrix of randomly selected true causal SNPs that are also true causal cis-eQTL as in Equation (7), respective SNP effect sizes are generated from $\beta_i \sim N(0, rw_i{}^2)$ with corresponding cis-eQTL effect size $w_i$ as used in Equation (7) and *r* 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 $p_{causal}$ and $h_p^2$ were taken as $(p_{causal}, h_p^2) = ((0.001,0.2), (0.01,0.3), (0.1,0.4), (0.2,0.5))$ for simulating phenotypes from Model I, while taken as $(p_{causal}, h_p^2) = ((0.001,0.1), (0.01,0.1), (0.1,0.15), (0.2,0.15))$ for simulating phenotypes from Model II.

We used 499 ROS/MAP samples as training data that were also used as training data in our application studies, and randomly selected 1,232 ROS/MAP samples as test data. To show the power performance with respect to test sample size, we considered different test sample sizes (400, 800, 1232) under the scenario with $p_{causal} = 0.2$.

We estimated cis-eQTL effect sizes from training data by using PrediXcan and TIGAR DPR methods and then conducted Burden-TWAS and VC-TWAS with individual-level and summary-level test GWAS data. We note that Burden-TWAS using PrediXcan weights is equivalent to the PrediXcan method [10] while Burden-TWAS using DPR weights is equivalent to the TIGAR method [8]. We also compared Burden-TWAS and VC-TWAS with the CoMM method

303 under simulation settings with $p_{causal} = 0.2$ (we did not consider additional simulation settings
304 because of the computational demands of CoMM).

305 For each scenario, we repeated simulations for 1,000 times and obtained the power as
306 the proportion of simulations that had test p-value $< 2.5 \times 10^{-6}$ (genome-wide significance
307 threshold for gene-based test). Additionally, we simulated phenotype under the null hypothesis
308 $Y \sim N(0,1)$ for $10^6$ times and evaluated type I errors of Burden-TWAS and VC-TWAS, using
309 variant weights derived from PrediXcan and DPR methods.

310 For each scenario with VC-TWAS, we considered both the original form of the test as well
311 as the alternate computationally-efficient form that considered only the filtered set of variants with
312 cis-eQTL effect size estimates greater than the median effect size value in each simulation (about
313 50% SNPs).

314

315 # Results

316 ## Simulation results

317 We compared the performance of VC-TWAS and Burden-TWAS using PrediXcan weights
318 (cis-eQTL effect size estimates by Elastic-Net penalized regression) and DPR weights (cis-eQTL
319 effect size estimates by DPR) under various scenarios. We also evaluated the performance of
320 VC-TWAS and Burden-TWAS using filtered DPR weights as described in Methods. Under the
321 scenario with $p_{causal} = 0.2$, we considered different test sample sizes (400, 800, 1232), and
322 compared VC-TWAS with CoMM using both individual-level and summary level test data with
323 1232 test samples.

324 First, we compared TWAS power for studying phenotypes simulated from Model I that
325 assumed SNP effect sizes on phenotypes were of a linear function of their corresponding cis-
326 eQTL effect sizes. When $p_{causal} = (0.001, 0.01)$ with sparse true causal signals (Fig S1A), VC-
327 TWAS had comparable power with Burden-TWAS and both TWAS methods using PrediXcan

weights achieved higher power compared to using DPR weights. When $p_{causal} > 0.01$, we observed that both TWAS methods using DPR weights achieved higher power compared to using PrediXcan weights. These results are consistent with the previous TIGAR paper [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. As expected, with $\mathrm{p_{causal}} > 0.01$ (Fig S1A, Fig 1A), Burden-TWAS methods outperformed VC-TWAS under Model I which meets the assumptions by Burden-TWAS methods. Across all considered scenarios, all TWAS methods 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 Fig 1B and Fig S1B, we found that the VC-TWAS obtained higher power than Burden-TWAS across all scenarios. Especially, with $p_{causal} \geq 0.01$, the power by VC-TWAS using DPR weights was double of 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 $p_{causal} = 0.001$ and $h_e^2 \in (0.1,0.2)$ (Fig S1B), 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.

Third, we compared VC-TWAS and Burden-TWAS with different test sample sizes (400, 800, 1232) under the scenarios with ($p_{causal} = 0.2,\ h_e^2 = 0.2$). As shown in Fig 1(C,D), the power of VC-TWAS and Burden-TWAS increased with respect to the test sample size in both simulation models. Interestingly, for model II, the magnitude of the power difference between VC-TWAS and Burden-TWAS became more pronounced with increasing sample size.

352 Fourth, to validate the VC-TWAS approach for using only summary-level GWAS data, we
353 compared VC-TWAS with DPR weights by using individual-level and summary-level GWAS data,
354 under the scenarios with $p_{causal} = 0.2$. As shown in Fig 1(E, F), VC-TWAS using summary-level
355 GWAS data performed equivalent as using individual-level data under both simulation models.
356 Additionally, VC-TWAS using filtered DPR weights and summary-level GWAS data still had
357 similar performance to that using complete DPR weights.

358 Fifth, we also compared VC-TWAS with CoMM [20, 21] using individual-level and
359 summary-level GWAS data under the scenarios with $p_{causal} = 0.2$ (Fig S2). Although CoMM
360 outperformed VC-TWAS, CoMM required 600x more computation time on average than VC-
361 TWAS for testing a typical gene with ~5K test SNPs (Table S1; Fig S3). Take typical genes with
362 ~2K-5K test SNPs as examples, VC-TWAS costed ~20s versus ~38,037s by CoMM using
363 individual-level GWAS data, and VC-TWAS costed ~3s versus ~373s by CoMM using summary-
364 level GWAS data. As the number of the SNPs in a gene increases, the difference in computation
365 run times between VC-TWAS and CoMM become even more pronounced.

366 Last but not least, to evaluate type I error of Burden-TWAS and VC-TWAS, we conducted
367 $10^6$ times simulations under the null hypothesis where phenotypes were not associated with
368 genetic data of the test gene. Without loss of generality, we used gene expression data simulated
369 with $p_{causal} = 0.2,\ h_e^2 = 0.1$ and generated phenotypes randomly from a $N(0, 1)$ distribution. We
370 evaluated type I errors (Table 1) with multiple significant levels $(10^{-2}, 10^{-4}, 2.5 \times 10^{-6})$ ,
371 demonstrating that both TWAS approaches had type I errors well controlled across all considered
372 significance levels. We also presented the quantile-quantile (QQ) plots of p-values by all methods
373 in Fig S4.

374 **Table 1. Type I errors under null simulation studies for Burden-TWAS and VC-TWAS with**
375 **$p_{causal}$ = 0.2, $h_e^2$ = 0.1.**
376

| Significance Level | Burden-TWAS | | | VC-TWAS | | |
|---|---|---|---|---|---|---|
| | DPR | Filtered DPR | PrediXcan | DPR | Filtered DPR | PrediXcan |

| | | | | | | |
|---|---|---|---|---|---|---|
| $1.00 \times 10^{-2}$ | $9.82 \times 10^{-3}$ | $9.86 \times 10^{-3}$ | $9.27 \times 10^{-3}$ | $9.43 \times 10^{-3}$ | $9.46 \times 10^{-3}$ | $9.23 \times 10^{-3}$ |
| $1.00 \times 10^{-4}$ | $8.64 \times 10^{-5}$ | $8.44 \times 10^{-5}$ | $9.95 \times 10^{-5}$ | $8.64 \times 10^{-5}$ | $9.05 \times 10^{-5}$ | $8.24 \times 10^{-5}$ |
| $2.50 \times 10^{-6}$ | $2.00 \times 10^{-6}$ | $2.00 \times 10^{-6}$ | $2.00 \times 10^{-6}$ | $2.00 \times 10^{-6}$ | $1.00 \times 10^{-6}$ | $6.00 \times 10^{-6}$ |

To summarize, VC-TWAS performed similarly to Burden-TWAS for studying phenotypes simulated from Model I with sparse true causal eQTL, but 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 a function of cis-eQTL effect sizes but not explicitly a linear relationship (leading to more general models of effect). Generally, 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, VC-TWAS using filtered DPR weights achieved similar power as using complete DPR weights, while saving up to 80% of computation time.

## Application studies of AD with individual-level GWAS data

We applied VC-TWAS to the individual-level 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, 39], 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. Here, we roughly chose the threshold $10^{-4}$ to filter DPR weights in our VC-TWAS such that on average the variance component test considered ~50% SNPs from the test gene region. 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 (Fig S3).

401 Leveraging the clinical and postmortem AD data available in ROS/MAP, we were able to
402 apply VC-TWAS to four clinical and pathologic AD phenotypes (Text S4). We examined final
403 cognitive status diagnosis of 1,436 decedents (AD Dementia (N=609) versus No Dementia
404 (N=827)), as well as three postmortem AD phenotypes including: continuous outcomes of $\beta$-
405 amyloid load (N=1,294), PHFtau tangle density (tangles, N=1,303), and a global AD pathology
406 (N=1,329). In the VC-TWAS of all four AD phenotypes, we adjusted for covariates of age, smoking
407 status, sex, study group (ROS or MAP), education, and the top three principal components of
408 ancestry.

409 With Mayo Clinic cohort, we conducted VC-TWAS for AD clinical diagnosis with 844 cases
410 diagnosed with LOAD and 1,255 controls showed no signal of dementia, while adjusting for
411 covariates age, sex, and top three principal components of ancestry. Since only the phenotype of
412 AD clinical diagnosis was profiled by both ROS/MAP and Mayo Clinic cohorts (under different
413 diagnosis criteria) and different sets of covariates were adjusted in VC-TWAS, we conducted
414 meta-analysis with VC-TWAS summary statistics for each study by using Fisher's method (meta
415 VC-TWAS) [40] to increase power with a larger sample size.

416 By meta VC-TWAS, we detected 13 significant risk genes with FDR < 0.05 that were
417 located within ~2MB region around the well-known AD risk gene *APOE* on chromosome 19 (Fig
418 2A; Table 2). Seven of those significant genes were known risk genes of AD by previous GWAS
419 (*CLASRP, TOMM40, MARK4, CLPTM1, CEACAM19, RELB*) [41, 42] and Burden-TWAS using
420 DPR weights (*TRAPPC6A*) [8].

421 **Table 2. Significant genes for phenotype AD clinical diagnosis by meta VC-TWAS with**
422 **filtered DPR weights.** Significant genes were identified with FDR < 0.05 by meta-TWAS with
423 ROS/MAP and Mayo Clinic cohorts. AD risk genes identified by previous GWAS are shaded in
424 grey.
425 a: Genes with significant p-values <0.0013 (Bonferroni correction with respect to 13 genes and 3
426 phenotypes) for at least one AD pathology phenotype
427

| Gene name | CHR | Start | End | P-value | FDR |
|---|---|---|---|---|---|
| *ZNF234*[a] | 19 | 44,645,710 | 44,664,462 | $1.11 \times 10^{-16}$ | $1.40 \times 10^{-12}$ |
| *CLASRP*[a] | 19 | 45,542,298 | 45,574,214 | $4.44 \times 10^{-16}$ | $2.81 \times 10^{-12}$ |

| | | | | | |
|---|---|---|---|---|---|
| *TRAPPC6A* [a] | 19 | 45,666,187 | 45,681,485 | $3.60 \times 10^{-14}$ | $1.52 \times 10^{-10}$ |
| *TOMM40* [a] | 19 | 45,394,477 | 45,406,935 | $9.05 \times 10^{-13}$ | $2.86 \times 10^{-9}$ |
| *MARK4* | 19 | 45,754,550 | 45,808,541 | $4.62 \times 10^{-16}$ | $1.17 \times 10^{-8}$ |
| *PPP1R13L* | 19 | 45,882,892 | 45,909,607 | $1.82 \times 10^{-10}$ | $3.84 \times 10^{-7}$ |
| *CLPTM1* | 19 | 45,457,848 | 45,496,598 | $5.71 \times 10^{-8}$ | $1.03 \times 10^{-4}$ |
| *EML2* | 19 | 46,112,660 | 46,148,726 | $1.88 \times 10^{-7}$ | $2.97 \times 10^{-4}$ |
| *FBXO46* | 19 | 46,213,887 | 46,234,151 | $4.13 \times 10^{-7}$ | $5.80 \times 10^{-4}$ |
| *CEACAM19* [a] | 19 | 45,174,724 | 45,187,631 | $3.93 \times 10^{-6}$ | $4.68 \times 10^{-3}$ |
| *GIPR* | 19 | 46,171,502 | 46,185,704 | $4.07 \times 10^{-6}$ | $4.68 \times 10^{-3}$ |
| *RELB* | 19 | 45,504,695 | 45,541,452 | $6.63 \times 10^{-6}$ | $6.99 \times 10^{-3}$ |
| *ZNF225* | 19 | 44,617,548 | 44,637,255 | $2.59 \times 10^{-5}$ | $2.51 \times 10^{-2}$ |

As clinical AD diagnosis can be associated with AD related pathologies, in further analyses we investigated whether the genes found to be associated with clinical AD diagnosis (Table 2) were also associated with pathologic AD phenotypes. We examined the VC-TWAS p-values of these significant genes with respect to AD pathology phenotypes ($\beta$-amyloid, tangles and global AD pathology) (Table S2; Fig 2B; Figs S5-S6). Interestingly, 5 out of these 13 genes had at least one VC-TWAS p-value <0.0013 (Bonferroni correction with respect to 13 genes and 3 AD pathology phenotypes) for one of the three AD pathology phenotypes. Particularly, genes *ZNF234* and *CLASRP* have VC-TWAS p-value <0.0013 for all three AD pathology phenotypes, which are likely to be involved in the biological mechanisms of both $\beta$-amyloid and tangles. The other three genes (TRAPPC6A, *TOMM40, CEACAM19*) have VC-TWAS p-value <0.0013 for $\beta$-amyloid and global AD pathology, which are likely to be involved only in the biological mechanism of $\beta$-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 $\beta$-amyloid, the second most significant gene (p-value = $6.39 \times 10^{-5}$) by VC-TWAS of global AD pathology, 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 could be potentially mediated through its gene expression, and the expression of this gene could be also potentially involved in the mechanisms of both AD pathology indices of $\beta$-amyloid and tangles. Besides AD, gene *ZNF234* is also a known risk gene for lipid traits [43]. The genetically regulated gene expression of this

448 gene might also affect lipid traits, thus leading to a pleiotropy phenomenon of AD and lipid traits
449 [44]. Additionally, *ZNF234* is known to be involved in the super pathway of gene expression and
450 is annotated with the Gene Ontology term of nucleic acid binding and DNA-binding transcription
451 factor activity [45].

452 Another significant gene of interest is *TOMM40*, which has FDR $= 2.86 \times 10^{-9}$ by meta
453 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
454 $\beta$-amyloid, tangles, and global AD pathology, respectively. These findings suggest that the
455 association of this well-known AD risk gene *TOMM40* [46] could be mediated through its gene
456 expression via $\beta$-amyloid but not tangles.

457 For all SNPs considered by meta VC-TWAS of genes *ZNF234* and *TOMM40* for studying
458 AD clinical diagnosis, we colocalized the meta GWAS results for AD clinical diagnosis with
459 ROS/MAP and Mayo Clinic cohorts and the corresponding DPR weight (i.e., cis-eQTL effect size)
460 magnitude. Interestingly, we found that the VC-TWAS association of these two genes were likely
461 to be driven by SNPs around *APOE/TOMM40* loci that also possessed major cis-eQTL effect size
462 magnitudes (Fig S7).

463 In addition, our VC-TWAS identified a significant gene *HSPBAP1* (FDR = 0.058) for
464 tangles (Fig S5B). As shown by previous studies, mRNA of gene *HSPBAP1* was abnormally
465 expressed in the anterior temporal neocortex of patients with intractable epilepsy [47]. Based on
466 our VC-TWAS results, gene *HSPBAP1* might not have a significant genetic effect on AD dementia,
467 but might have a significant effect on tangle pathology (p-value = $4.57 \times 10^{-6}$) and may account
468 for the increasing recognition of non-cognitive AD phenotypes [48]. This provides support about
469 that gene *HSPBAP1* could be involved in the mechanism of brain pathology tangles and other
470 neurological diseases such as intractable epilepsy [47, 49].

471 We also applied VC-TWAS with PrediXcan weights, but no significant genes were
472 identified with FDR <0.05 (Figs S8-10), neither the genes with smallest p-values were proximal

473 to *APOE*. In contrast, our results by VC-TWAS with filtered DPR weights provided potential
474 biological interpretations for several known AD risk genes via gene expression and for their
475 associations with both clinical and pathologic AD phenotype.
476
477 **Application studies of AD with summary-level GWAS data**
478 Next, we applied VC-TWAS to the stage1 summary-level GWAS data of AD from IGAP
479 [3], which has a much larger sample size (~54K with 17,008 AD cases and 37,154 controls). We
480 considered filtered cis-eQTL DPR weights as used in the above applications with individual-level
481 GWAS data, as well as the SNP weights of both cis- and trans- eQTL generated by the BGW-
482 TWAS method [28]. LD covariance matrices from ROS/MAP individual-level GWAS data were
483 used for implementing VC-TWAS with summary-level GWAS data.
484
485 Using filtered cis-eQTL DPR weights
486 By using the filtered cis-eQTL DPR weights, we identified 57 significant risk genes with
487 FDR <0.05 by VC-TWAS (Fig 3A; Table S3, Fig S11A), including the most well-known AD risk
488 genes *TOMM40* and *APOE,* along with 45 genes that were located within ~2MB region around
489 those two risk genes. Other significant genes are distributed on chromosome 6, 8, 11 and 20 (Fig
490 3A). Among these 57 significant genes, 20 genes were known risk genes by previous GWAS [41,
491 42] (Table S3). Genes detected by VC-TWAS using the individual-level GWAS data of ROS/MAP
492 and Mayo Clinic cohorts were all replicated (Table S3). Burden-TWAS using IGAP summary data
493 and filtered cis-eQTL DPR weights (Table S3) also identified 22 out of 57 of these significant
494 genes.
495 Since most significant genes are nearby genes on chromosome 11 and 19, we made
496 analogous TWAS locus zoom plots for these two major loci in Fig 3(B,C), where the between
497 gene $R^2$ were calculated with respect to the predicted GReX values using the individual-level

498 GWAS data of ROS/MAP and Mayo Clinic cohorts. We observed that most significant genes do
499 have highly correlated GReX values, showing that nearby significant TWAS associations are likely
500 to be not independent. For the locus on chromosome 19 (Fig 3C), we can see that the $R^2$ between
501 gene *MARK4* (and a few other genes shown in orange color) and the top significant gene *CLASRP*
502 is greater than 0.4. The significant genes in blue colors are likely to be independent associations
503 from the top significant gene *CLASRP*. For the locus on chromosome 11 (Fig 3B), significant
504 genes *TMEM132A, OSBP, STX3, and TMEM109* are highly correlated with the top significant
505 gene *PRPF19,* while gene *ME3* tend to be another independent association.
506
507 Using cis- and trans- eQTL weights generated by BGW-TWAS
508 To provide a complementary list of significant genes by considering both cis- and trans-
509 eQTL, we conducted VC-TWAS with the IGAP summary data using the cis- and trans- eQTL
510 weights generated by the BGW-TWAS method [28]. We detected total 71 significant genes with
511 FDR < 0.05 (Fig 4; Table S4; Fig S11B), among which 6 genes were identified by GWAS [42, 50-
512 53] and 22 genes were shown to be related with AD or other neurological diseases by previous
513 studies and (Table 3).

514 **Table 3. Significant genes identified by VC-TWAS using IGAP summary statistics data with**
515 **both BGW weights, which are either known GWAS loci or shown to be related with AD or**
516 **other neurological diseases by previous studies.** AD risk genes identified by previous GWAS
517 are shaded in grey.
518 a: Genes shown to be related with AD or other neurological diseases by previous studies

| Gene Name | CHROM | Start | End | P-value | FDR |
|---|---|---|---|---|---|
| *ARHGEF2*[a] | 1 | 155,916,644 | 155,966,129 | $1.75 \times 10^{-16}$ | $2.46 \times 10^{-13}$ |
| *GAS5*[a] | 1 | 173,833,037 | 173,838,020 | $1.33 \times 10^{-22}$ | $3.11 \times 10^{-19}$ |
| *NCOA1*[a] | 2 | 24,714,782 | 24,993,571 | $2.42 \times 10^{-25}$ | $8.51 \times 10^{-22}$ |
| *CENPO*[a] | 2 | 25,016,004 | 25,045,245 | $4.07 \times 10^{-7}$ | $1.85 \times 10^{-4}$ |
| *SLC4A10*[a] | 2 | 162,280,842 | 162,841,792 | $8.80 \times 10^{-6}$ | $2.58 \times 10^{-3}$ |
| *MAN2A1* | 5 | 109,025,066 | 109,205,326 | $6.30 \times 10^{-8}$ | $3.70 \times 10^{-5}$ |
| *ZMAT2*[a] | 5 | 140,079,746 | 140,086,266 | $1.83 \times 10^{-6}$ | $6.43 \times 10^{-4}$ |
| *HLA-DRB5*[a] | 6 | 32,485,119 | 32,498,064 | $7.81 \times 10^{-7}$ | $3.23 \times 10^{-4}$ |
| *AHR*[a] | 7 | 17,338,245 | 17,385,776 | $2.11 \times 10^{-5}$ | $5.62 \times 10^{-3}$ |

| | | | | | |
|---|---|---|---|---|---|
| *JAZF1* | 7 | 27,870,191 | 28,220,362 | $7.49 \times 10^{-8}$ | $4.22 \times 10^{-5}$ |
| *ACHE* [a] | 7 | 100,487,614 | 100,494,594 | $5.93 \times 10^{-5}$ | $1.39 \times 10^{-2}$ |
| *DLGAP2* [a] | 8 | 1,449,530 | 1,656,642 | $1.18 \times 10^{-5}$ | $3.38 \times 10^{-3}$ |
| *RFX3* [a] | 9 | 3,218,296 | 3,526,004 | $4.69 \times 10^{-5}$ | $1.16 \times 10^{-2}$ |
| *SLC25A28* [a] | 10 | 101,370,281 | 101,380,535 | $1.55 \times 10^{-7}$ | $7.79 \times 10^{-5}$ |
| *PDCD11* [a] | 10 | 105,156,404 | 105,206,049 | $2.52 \times 10^{-8}$ | $1.69 \times 10^{-5}$ |
| *ROBO4* [a] | 11 | 124,753,586 | 124,768,396 | $1.13 \times 10^{-6}$ | $4.30 \times 10^{-4}$ |
| *USP30* [a] | 12 | 109,460,893 | 109,525,831 | $1.50 \times 10^{-10}$ | $1.41 \times 10^{-7}$ |
| *PPP1R3E* [a] | 14 | 23,765,111 | 23,772,057 | $2.11 \times 10^{-5}$ | $5.62 \times 10^{-3}$ |
| *MTFMT* [a] | 15 | 65,294,844 | 65,321,977 | $3.10 \times 10^{-5}$ | $8.09 \times 10^{-3}$ |
| *PARP6* | 15 | 72,533,521 | 72,565,340 | $6.77 \times 10^{-6}$ | $2.07 \times 10^{-3}$ |
| *GRIN2A* [a] | 16 | 9,852,375 | 10,276,611 | $2.22 \times 10^{-8}$ | $1.56 \times 10^{-5}$ |
| *SETD6* [a] | 16 | 58,549,382 | 58,554,431 | $3.44 \times 10^{-5}$ | $8.81 \times 10^{-3}$ |
| *NUP88* [a] | 17 | 5,264,257 | 5,323,480 | $2.65 \times 10^{-7}$ | $1.24 \times 10^{-4}$ |
| *NOS2* [a] | 17 | 26,083,791 | 26,127,555 | $9.55 \times 10^{-5}$ | $2.07 \times 10^{-2}$ |
| *CEACAM19* | 19 | 45,174,723 | 45,187,631 | $1.46 \times 10^{-5}$ | $4.03 \times 10^{-3}$ |
| *APOC1* [a] | 19 | 45,417,920 | 45,422,606 | $2.80 \times 10^{-6}$ | $9.63 \times 10^{-4}$ |

For example, gene *ARHGEF2* on chromosome 1(with FDR $= 2.46 \times 10^{-13}$) is shown to be interact with all four members of the *MARK* family including *MARK4* on chromosome 19 which has an emerging tole in the phosphorylation of MAPT/TAU in Alzheimer's disease and identified by VC-TWAS using cis-eQTL DPR weights [54]. Gene *GAS5* (with FDR $= 3.11 \times 10^{-19}$) was shown to have a novel role in microglial polarization and the pathogenesis of demyelinating diseases which suggested the potential therapeutic benefit of targeting *GAS5* for the treatment of neurological disorders including AD [55]. Gene *NCOA1* (with FDR $8.51 \times 10^{-22}$) was shown to have pivotal roles in memory and learning [56]. SNPs in Gene *PDCD11* (with FDR $= 1.69 \times 10^{-5}$) have been showed to be significantly associated with AD [57]. Gene *APOC1* was detected by several GWASs [41, 42], even when conditioning on *APOE* gene [58]. The presence of gene *ACHE* (with FDR $= 1.39 \times 10^{-2}$) is a common feature in AD brain [59]. We also detected genes *NUP88* [60], *ROBO4* [61], *DLGAP2* [62], *AHR* [63], *PPP1R3E* [64], *RFX3* [65] and *NOS2* [66] that were shown to possess biological link with AD. Additional

533 identified genes *USP30* [67], *GRIN2A* [68], *SLC25A28* [69], *HLA-DRB5* [70], *ZMAT2* [71],
534 *SLC4A10* [72], *MTFMT* [73] and *SETD6* [74] were shown to be related with neurological diseases
535 by previous studies.

536 Overall, we show that VC-TWAS using BGW weights provide a complementary list of risk
537 genes to those previously identified by considering additional trans- eQTL information.

538

539 **Discussion**

540 In this paper, we propose a novel variance-component TWAS (VC-TWAS) method that
541 leverages eQTL effect sizes from reference transcriptome but does not make the assumption of
542 a linear relationship between SNP effect sizes on phenotypes and cis-eQTL effect sizes. VC-
543 TWAS is applicable to both quantitative and dichotomous outcomes and can further handle both
544 individual-level and summary-level GWAS data. By implementing this VC-TWAS with cis-eQTL
545 effect sizes estimated by DPR method [8, 13], we created a powerful test statistic that had good
546 performance in simulation studies and obtained biologically meaningful TWAS results for both
547 clinical and pathologic AD phenotypes. In particular, with the individual-level GWAS data of ~3.4K
548 samples, we detected 13 TWAS genes for AD dementia, including the well-known GWAS risk
549 gene *TOMM40* and previously identified TWAS gene *TRAPPC6A* [8]. Moreover, 6 out of these
550 13 genes were identified by previous GWAS [41]. The pleiotropy effects of 5 of these genes with
551 respect to AD dementia and indices of AD pathology demonstrated the possible biological
552 mechanisms linking AD risk genes via $\beta$-amyloid and tangles with AD dementia.

553 By applying VC-TWAS with summary-level GWAS data of AD with a much larger sample
554 size, we not only validated our findings with the individual-level GWAS data but also detected
555 additional novel risk genes. In particular, by applying VC-TWAS with both cis- and trans- eQTL
556 effect sizes estimated by the BGW-TWAS method, we identified a list of significant AD risk genes
557 that complement risk genes identified by considering only cis-eQTL information.

558 To help users conduct our VC-TWAS method conveniently and efficiently, we added this
559 method into our previously developed tool — Transcriptome Integrated Genetic Association
560 Resource (TIGAR) [8]. The user has the option of estimating cis-eQTL effect sizes by either the
561 PrediXcan method (i.e., Elastic-Net) [12] or nonparametric Bayesian DPR method [13], or using
562 eQTL weights generated by previous studies. The VC-TWAS function works for using individual-
563 level GWAS data to study continuous or dichotomous phenotypes, as well as using summary-
564 level GWAS data.

565 Since the variance component test statistic used by VC-TWAS involves calculating and
566 performing an eigen-decomposition of a genotypic kernel matrix, efficient computation is required
567 (even when filtering variants to include only those variants with relatively larger magnitude of
568 eQTL estimates) for obtaining the corresponding p-values in practice. Our TIGAR tool implements
569 multi-threaded computation to take advantage of high-performance cloud computing clusters and
570 enable practical computation for testing genome-wide genes with respect to reference
571 transcriptome data of multiple tissue types. A typical VC-TWAS with ~20K genes can be
572 accomplished within ~80 hours by using 1 CPU core with 32 GB memory.

573 Of course, current TWAS methods including our VC-TWAS still have their limitations.
574 Because of genetic and transcriptomic heterogeneities across different ethnicities, one may have
575 difficulty in translating eQTL effect size estimates across cohorts with different ethnicities [75].
576 That being said, VC-TWAS is likely more robust to this phenomenon than Burden-TWAS given
577 the former relaxes the assumption of the latter of a linear relationship between the SNP effect
578 size and eQTL effect in the transcriptome. Nevertheless, reference panels with diverse ethnicities
579 and multiple tissue types are certainly needed to enhance TWAS to study complex diseases
580 across different ancestral groups. We should note though that both ROS/MAP and Mayo Clinic
581 cohorts that we considered in this work have similar (European) origins.

582 Nevertheless, our novel technique allows a more flexible framework to account for the
583 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.

587 **Acknowledgements**

588 ROS/MAP study data were provided by the Rush Alzheimer's Disease Center, Rush University
589 Medical Center, Chicago, IL. Data collection was supported through funding by NIH/NIA grants,
590 the Illinois Department of Public Health, and the Translational Genomics Research Institute. The
591 MCADGC led by Dr. Nilüfer Ertekin-Taner and Dr. Steven G. Younkin, Mayo Clinic,
592 Jacksonville, FL uses samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's
593 Disease Research Center, and the Mayo Clinic Brain Bank.

## Figure Captions

**Figure 1. Power comparison of Burden-TWAS and VC-TWAS methods under simulation scenarios with 20% true causal eQTL for gene expression (i.e., $p_{causal}$ = 0.2) in the test gene.** Phenotypes simulated from Model I (A, C, E) and Model II (B, D, F) were considered. Various types of SNP weights were considered, including those derived from PrediXcan method, DPR method, and filtered DPR weights (A, B). Various test sample sizes (400, 800, 1232) were considered (B, D). The VC-TWAS approach of using only summary-level GWAS data were validated (E, F).

**Figure 2. Manhattan plots of meta-analysis VC-TWAS for AD clinical diagnosis (A) and VC-TWAS of global AD pathology (B) with filtered DPR weights.** Genes with FDR < 0.05 are colored in (A), with red for novel risk genes and blue for known AD risk genes. Genes with FDR < 0.05 in meta VC-TWAS of AD clinical diagnosis and p-value < 0.0013 in VC-TWAS of global AD pathology are colored in red in (B).

**Figure 3. Manhattan plots of VC-TWAS using IGAP summary data with filtered cis-eQTL DPR weights (A) and TWAS locus zoom plots for the loci on chromosome 11 (B) and chromosome 19 (C).** Significant genes with FDR < 0.05 are colored, with red for significant ones only identified by using IGAP summary data and blue for the ones replicating our VC-TWAS findings using the individual-level GWAS data of ROS/MAP and Mayo Clinic cohort. Between gene $R^2$ in (B, C) were calculated with respect to GReX values. The $R^2$ in locus zoom plot is the squared correlation of GReX between the most significant gene and other neighborhood genes. The heatmap is based on the squared correlation matrix of GReX.

**Figure 4. Manhattan plots of VC-TWAS using IGAP summary data with BGW cis- and trans-eQTL weights.** Significant genes with FDR < 0.05 are colored.

619 ## References


620 1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al.
621 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat
622 Rev Genet. 2008;9(5):356-69. Epub 2008/04/10. doi: 10.1038/nrg2344. PubMed PMID:
623 18398418.
624 2. Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikainen LP, et al.
625 Genome-wide association study identifies multiple loci influencing human serum metabolite
626 levels. Nat Genet. 2012;44(3):269-76. Epub 2012/01/31. doi: 10.1038/ng.1073. PubMed PMID:
627 22286219; PubMed Central PMCID: PMCPMC3605033.
628 3. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-
629 analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.
630 Nature genetics. 2013;45(12):1452-8. Epub 2013/10/29. doi: 10.1038/ng.2802. PubMed PMID:
631 24162737; PubMed Central PMCID: PMCPMC3896259.
632 4. Cannon ME, Mohlke KL. Deciphering the Emerging Complexities of Molecular
633 Mechanisms at GWAS Loci. Am J Hum Genet. 2018;103(5):637-53. Epub 2018/11/06. doi:
634 10.1016/j.ajhg.2018.10.001. PubMed PMID: 30388398; PubMed Central PMCID:
635 PMCPMC6218604.
636 5. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: detecting gene-
637 disease associations by matching patterns of expression QTL and GWAS. Am J Hum Genet.
638 2013;92(5):667-80. Epub 2013/05/07. doi: 10.1016/j.ajhg.2013.03.022. PubMed PMID:
639 23643380; PubMed Central PMCID: PMCPMC3644637.
640 6. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs
641 are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet.
642 2010;6(4):e1000888. Epub 2010/04/07. doi: 10.1371/journal.pgen.1000888. PubMed PMID:
643 20369019; PubMed Central PMCID: PMCPMC2848547.
644 7. Gamazon ER, Huang RS, Cox NJ, Dolan ME. Chemotherapeutic drug susceptibility
645 associated SNPs are enriched in expression quantitative trait loci. Proc Natl Acad Sci U S A.
646 2010;107(20):9287-92. Epub 2010/05/06. doi: 10.1073/pnas.1001827107. PubMed PMID:
647 20442332; PubMed Central PMCID: PMCPMC2889115.
648 8. Nagpal S, Meng X, Epstein MP, Tsoi LC, Patrick M, Gibson G, et al. TIGAR: An
649 Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of
650 Complex Traits. The American Journal of Human Genetics. 2019;105(2):258-66. doi:
651 https://doi.org/10.1016/j.ajhg.2019.05.018.
652 9. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical
653 Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, et al. Genetic effects on gene
654 expression across human tissues. Nature. 2017;550(7675):204-13. Epub 2017/10/13. doi:
655 10.1038/nature24277. PubMed PMID: 29022597; PubMed Central PMCID: PMCPMC5776756.
656 10. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et
657 al. A gene-based association method for mapping traits using reference transcriptome data. Nat
658 Genet. 2015;47(9):1091-8. Epub 2015/08/11. doi: 10.1038/ng.3367. PubMed PMID: 26258848;
659 PubMed Central PMCID: PMCPMC4552594.
660 11. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, et al. Integrative approaches
661 for large-scale transcriptome-wide association studies. Nature genetics. 2016;48(3):245-52. Epub
662 2016/02/09. doi: 10.1038/ng.3506. PubMed PMID: 26854917; PubMed Central PMCID:
663 PMCPMC4767558.

12. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2005;67(2):301-20.
13. Zeng P, Zhou X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat Commun. 2017;8(1):456. Epub 2017/09/08. doi: 10.1038/s41467-017-00470-2. PubMed PMID: 28878256; PubMed Central PMCID: PMCPMC5587666.
14. Li YI, Wong G, Humphrey J, Raj T. Prioritizing Parkinson's disease genes using population-scale transcriptomic data. Nat Commun. 2019;10(1):994. Epub 2019/03/03. doi: 10.1038/s41467-019-08912-9. PubMed PMID: 30824768; PubMed Central PMCID: PMCPMC6397174.
15. Keys KL, Mak ACY, White MJ, Eckalbar WL, Dahl AW, Mefford J, et al. On the cross-population generalizability of gene expression prediction models. bioRxiv. 2019:552042. doi: 10.1101/552042.
16. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82-93. Epub 2011/07/09. doi: 10.1016/j.ajhg.2011.05.029. PubMed PMID: 21737059; PubMed Central PMCID: PMCPMC3135811.
17. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet. 2008;82(2):386-97. Epub 2008/02/07. doi: 10.1016/j.ajhg.2007.10.010. PubMed PMID: 18252219; PubMed Central PMCID: PMCPMC2664991.
18. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010;86(6):929-42. Epub 2010/06/22. doi: 10.1016/j.ajhg.2010.05.002. PubMed PMID: 20560208; PubMed Central PMCID: PMCPMC3032061.
19. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92(6):841-53. Epub 2013/05/21. doi: 10.1016/j.ajhg.2013.04.015. PubMed PMID: 23684009; PubMed Central PMCID: PMCPMC3675243.
20. Yeung K-F, Yang Y, Yang C, Liu J. CoMM: A Collaborative Mixed Model That Integrates GWAS and eQTL Data Sets to Investigate the Genetic Architecture of Complex Traits. Bioinformatics and Biology Insights. 2019;13:117793221988143. doi: 10.1177/1177932219881435.
21. Yang Y, Shi X, Jiao Y, Huang J, Chen M, Zhou X, et al. CoMM-S2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. Bioinformatics. 2019;36(7):2009-16. doi: 10.1093/bioinformatics/btz880.
22. Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. Curr Alzheimer Res. 2012;9(6):628-45. Epub 2012/04/05. doi: 10.2174/156720512801322573. PubMed PMID: 22471860; PubMed Central PMCID: PMCPMC3409291.
23. Bennett DA, Schneider JA, Buchman AS, Barnes LL, Boyle PA, Wilson RS. Overview and findings from the rush Memory and Aging Project. Curr Alzheimer Res. 2012;9(6):646-63. Epub 2012/04/05. doi: 10.2174/156720512801322663. PubMed PMID: 22471867; PubMed Central PMCID: PMCPMC3439198.
24. Ng B, White CC, Klein HU, Sieberts SK, McCabe C, Patrick E, et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat

710 Neurosci. 2017;20(10):1418-26. Epub 2017/09/05. doi: 10.1038/nn.4632. PubMed PMID:
711 28869584; PubMed Central PMCID: PMCPMC5785926.
712 25. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious
713 Orders Study and Rush Memory and Aging Project. J Alzheimers Dis. 2018;64(s1):S161-S89.
714 Epub 2018/06/06. doi: 10.3233/JAD-179939. PubMed PMID: 29865057; PubMed Central
715 PMCID: PMCPMC6380522.
716 26. Carrasquillo MM, Zou F, Pankratz VS, Wilcox SL, Ma L, Walker LP, et al. Genetic
717 variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nat
718 Genet. 2009;41(2):192-8. Epub 2009/01/13. doi: 10.1038/ng.305. PubMed PMID: 19136949;
719 PubMed Central PMCID: PMCPMC2873177.
720 27. Zou F, Chai HS, Younkin CS, Allen M, Crook J, Pankratz VS, et al. Brain Expression
721 Genome-Wide Association Study (eGWAS) Identifies Human Disease-Associated Variants.
722 PLOS Genetics. 2012;8(6):e1002707. doi: 10.1371/journal.pgen.1002707.
723 28. Luningham JM, Chen J, Tang S, De Jager PL, Bennett DA, Buchman AS, et al. Bayesian
724 Genome-wide TWAS method to leverage both cis- and trans- eQTL information through
725 summary statistics. bioRxiv. 2020:2020.03.05.979187. doi: 10.1101/2020.03.05.979187.
726 29. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway
727 data: least-squares kernel machines and linear mixed models. Biometrics. 2007;63(4):1079-88.
728 Epub 2007/12/15. doi: 10.1111/j.1541-0420.2007.00799.x. PubMed PMID: 18078480; PubMed
729 Central PMCID: PMCPMC2665800.
730 30. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a
731 disease outcome using logistic kernel machine regression via logistic mixed models. BMC
732 Bioinformatics. 2008;9:292. Epub 2008/06/26. doi: 10.1186/1471-2105-9-292. PubMed PMID:
733 18577223; PubMed Central PMCID: PMCPMC2483287.
734 31. Moschopoulos PG, Canada WB. The distribution function of a linear combination of chi-
735 squares. Computers & Mathematics with Applications. 1984;10(4):383-6. doi:
736 https://doi.org/10.1016/0898-1221(84)90066-X.
737 32. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare
738 variants in sequencing association studies. Am J Hum Genet. 2013;93(1):42-53. Epub
739 2013/06/19. doi: 10.1016/j.ajhg.2013.05.010. PubMed PMID: 23768515; PubMed Central
740 PMCID: PMCPMC3710762.
741 33. Feng S, Liu D, Zhan X, Wing MK, Abecasis GR. RAREMETAL: fast and powerful
742 meta-analysis for rare variants. Bioinformatics. 2014;30(19):2828-9. Epub 2014/06/05. doi:
743 10.1093/bioinformatics/btu367. PubMed PMID: 24894501; PubMed Central PMCID:
744 PMCPMC4173011.
745 34. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication
746 DIG, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies
747 additional variants influencing complex traits. Nature genetics. 2012;44(4):369-75, S1-3. Epub
748 2012/03/20. doi: 10.1038/ng.2213. PubMed PMID: 22426310; PubMed Central PMCID:
749 PMCPMC3593158.
750 35. Buchanan CC, Torstenson ES, Bush WS, Ritchie MD. A comparison of cataloged
751 variation between International HapMap Consortium and 1000 Genomes Project data. J Am Med
752 Inform Assoc. 2012;19(2):289-94. Epub 2012/02/10. doi: 10.1136/amiajnl-2011-000652.
753 PubMed PMID: 22319179; PubMed Central PMCID: PMCPMC3277631.

754 36. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation
755 genotype imputation service and methods. Nat Genet. 2016;48(10):1284-7. Epub 2016/08/30.
756 doi: 10.1038/ng.3656. PubMed PMID: 27571263; PubMed Central PMCID: PMCPMC5157836.
757 37. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, et al.
758 Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and
759 other loci. Nat Neurosci. 2014;17(9):1156-63. Epub 2014/08/19. doi: 10.1038/nn.3786. PubMed
760 PMID: 25129075; PubMed Central PMCID: PMCPMC4292795.
761 38. De Jager PL, Shulman JM, Chibnik LB, Keenan BT, Raj T, Wilson RS, et al. A genome-
762 wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol
763 Aging. 2012;33(5):1017 e1-15. Epub 2011/11/08. doi: 10.1016/j.neurobiolaging.2011.09.033.
764 PubMed PMID: 22054870; PubMed Central PMCID: PMCPMC3307898.
765 39. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide
766 association study of 229,000 women identifies new candidate susceptibility genes for breast
767 cancer. Nat Genet. 2018;50(7):968-78. Epub 2018/06/20. doi: 10.1038/s41588-018-0132-x.
768 PubMed PMID: 29915430; PubMed Central PMCID: PMCPMC6314198.
769 40. Rödel E. Fisher, R. A.: Statistical Methods for Research Workers, 14. Aufl., Oliver &
770 Boyd, Edinburgh, London 1970. XIII, 362 S., 12 Abb., 74 Tab., 40 s. Biometrische Zeitschrift.
771 1971;13(6):429-30. doi: 10.1002/bimj.19710130623.
772 41. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-
773 wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease
774 risk. Nat Genet. 2019;51(3):404-13. Epub 2019/01/09. doi: 10.1038/s41588-018-0311-9.
775 PubMed PMID: 30617256; PubMed Central PMCID: PMCPMC6836675.
776 42. Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on
777 family history of Alzheimer's disease. Transl Psychiatry. 2018;8(1):99. Epub 2018/05/20. doi:
778 10.1038/s41398-018-0150-6. PubMed PMID: 29777097; PubMed Central PMCID:
779 PMCPMC5959890.
780 43. Andaleon A, Mogil LS, Wheeler HE. Genetically regulated gene expression underlies
781 lipid traits in Hispanic cohorts. PLoS One. 2019;14(8):e0220827. Epub 2019/08/09. doi:
782 10.1371/journal.pone.0220827. PubMed PMID: 31393916; PubMed Central PMCID:
783 PMCPMC6687110.
784 44. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. Causal associations
785 between risk factors and common diseases inferred from GWAS summary data. Nat Commun.
786 2018;9(1):224. Epub 2018/01/18. doi: 10.1038/s41467-017-02317-2. PubMed PMID: 29335400;
787 PubMed Central PMCID: PMCPMC5768719.
788 45. Tripathi S, Christie KR, Balakrishnan R, Huntley R, Hill DP, Thommesen L, et al. Gene
789 Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage
790 for a large-scale curation effort. Database (Oxford). 2013;2013:bat062. Epub 2013/08/29. doi:
791 10.1093/database/bat062. PubMed PMID: 23981286; PubMed Central PMCID:
792 PMCPMC3753819.
793 46. Chiba-Falek O, Gottschalk WK, Lutz MW. The effects of the TOMM40 poly-T alleles on
794 Alzheimer's disease phenotypes. Alzheimers Dement. 2018;14(5):692-8. Epub 2018/03/11. doi:
795 10.1016/j.jalz.2018.01.015. PubMed PMID: 29524426; PubMed Central PMCID:
796 PMCPMC5938113.
797 47. Xi ZQ, Sun JJ, Wang XF, Li MW, Liu XZ, Wang LY, et al. HSPBAP1 is found
798 extensively in the anterior temporal neocortex of patients with intractable epilepsy. Synapse.
799 2007;61(9):741-7. Epub 2007/06/15. doi: 10.1002/syn.20417. PubMed PMID: 17568411.

48. Albers MW, Gilmore GC, Kaye J, Murphy C, Wingfield A, Bennett DA, et al. At the interface of sensory and motor dysfunctions and Alzheimer's disease. Alzheimers Dement. 2015;11(1):70-98. Epub 2014/07/16. doi: 10.1016/j.jalz.2014.04.514. PubMed PMID: 25022540; PubMed Central PMCID: PMCPMC4287457.
49. Saha P, Sen N. Tauopathy: A common mechanism for neurodegeneration and brain aging. Mech Ageing Dev. 2019;178:72-9. Epub 2019/01/23. doi: 10.1016/j.mad.2019.01.007. PubMed PMID: 30668956; PubMed Central PMCID: PMCPMC6377302.
50. Sherva R, Tripodis Y, Bennett DA, Chibnik LB, Crane PK, de Jager PL, et al. Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. Alzheimer's & dementia : the journal of the Alzheimer's Association. 2014;10(1):45-52. Epub 2013/03/25. doi: 10.1016/j.jalz.2013.01.008. PubMed PMID: 23535033.
51. Kamboh MI, Fan KH, Yan Q, Beer JC, Snitz BE, Wang X, et al. Population-based genome-wide association study of cognitive decline in older adults free of dementia: identification of a novel locus for the attention domain. Neurobiol Aging. 2019;84:239.e15-.e24. Epub 2019/04/08. doi: 10.1016/j.neurobiolaging.2019.02.024. PubMed PMID: 30954325; PubMed Central PMCID: PMCPMC6739197.
52. Beecham GW, Hamilton K, Naj AC, Martin ER, Huentelman M, Myers AJ, et al. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. PLoS Genet. 2014;10(9):e1004606. Epub 2014/09/05. doi: 10.1371/journal.pgen.1004606. PubMed PMID: 25188341; PubMed Central PMCID: PMCPMC4154667.
53. Hu X, Pickering EH, Hall SK, Naik S, Liu YC, Soares H, et al. Genome-wide association study identifies multiple novel loci associated with disease progression in subjects with mild cognitive impairment. Transl Psychiatry. 2011;1(11):e54. Epub 2011/01/01. doi: 10.1038/tp.2011.50. PubMed PMID: 22833209; PubMed Central PMCID: PMCPMC3309471.
54. Sandí M-J, Marshall CB, Balan M, Coyaud É, Zhou M, Monson DM, et al. MARK3-mediated phosphorylation of ARHGEF2 couples microtubules to the actin cytoskeleton to establish cell polarity. Science Signaling. 2017;10(503):eaan3286. doi: 10.1126/scisignal.aan3286.
55. Sun D, Yu Z, Fang X, Liu M, Pu Y, Shao Q, et al. LncRNA GAS5 inhibits microglial M2 polarization and exacerbates demyelination. EMBO Rep. 2017;18(10):1801-16. Epub 2017/08/14. doi: 10.15252/embr.201643668. PubMed PMID: 28808113.
56. Sun Z, Xu Y. Nuclear Receptor Coactivators (NCOAs) and Corepressors (NCORs) in the Brain. Endocrinology. 2020;161(8). doi: 10.1210/endocr/bqaa083.
57. Yang R, Zhan M, Guo M, Yuan H, Wang Y, Zhang Y, et al. Yolk sac-derived Pdcd11-positive cells modulate zebrafish microglia differentiation through the NF-κB-Tgfβ1 pathway. Cell Death & Differentiation. 2020. doi: 10.1038/s41418-020-0591-3.
58. Lucatelli JF, Barros AC, Silva VK, Machado Fda S, Constantin PC, Dias AA, et al. Genetic influences on Alzheimer's disease: evidence of interactions between the genes APOE, APOC1 and ACE in a sample population from the South of Brazil. Neurochem Res. 2011;36(8):1533-9. Epub 2011/05/03. doi: 10.1007/s11064-011-0481-7. PubMed PMID: 21533863.
59. García-Ayllón M-S, Small DH, Avila J, Sáez-Valero J. Revisiting the Role of Acetylcholinesterase in Alzheimer's Disease: Cross-Talk with P-tau and β-Amyloid. Front Mol Neurosci. 2011;4:22-. doi: 10.3389/fnmol.2011.00022. PubMed PMID: 21949503.

845 60. Sheffield LG, Miskiewicz HB, Tannenbaum LB, Mirra SS. Nuclear pore complex
846 proteins in Alzheimer disease. J Neuropathol Exp Neurol. 2006;65(1):45-54. Epub 2006/01/18.
847 doi: 10.1097/01.jnen.0000195939.40410.08. PubMed PMID: 16410748.
848 61. Reinikainen KJ, Riekkinen PJ, Halonen T, Laakso M. Decreased muscarinic receptor
849 binding in cerebral cortex and hippocampus in alzheimer's disease. Life Sciences.
850 1987;41(4):453-61. doi: https://doi.org/10.1016/0024-3205(87)90221-9.
851 62. Ouellette AR, Neuner SM, Dumitrescu L, Anderson LC, Gatti DM, Mahoney ER, et al.
852 Cross-Species Analyses Identify Dlgap2 as a Regulator of Age-Related Cognitive Decline and
853 Alzheimer's Dementia. Cell Reports. 2020;32(9):108091. doi:
854 https://doi.org/10.1016/j.celrep.2020.108091.
855 63. Ramos-García NA, Orozco-Ibarra M, Estudillo E, Elizondo G, Gómez Apo E, Chávez
856 Macías LG, et al. Aryl Hydrocarbon Receptor in Post-Mortem Hippocampus and in Serum from
857 Young, Elder, and Alzheimer's Patients. Int J Mol Sci. 2020;21(6):1983. doi:
858 10.3390/ijms21061983. PubMed PMID: 32183254.
859 64. Gong CX, Singh TJ, Grundke-Iqbal I, Iqbal K. Phosphoprotein phosphatase activities in
860 Alzheimer disease brain. J Neurochem. 1993;61(3):921-7. Epub 1993/09/01. doi:
861 10.1111/j.1471-4159.1993.tb03603.x. PubMed PMID: 8395566.
862 65. Remnestål J. Expression and distribution of transcription factors NPAS3 och RFX3 in
863 Alzheimer's disease [Student thesis]2015.
864 66. Colton CA, Vitek MP, Wink DA, Xu Q, Cantillana V, Previti ML, et al. NO synthase 2
865 (NOS2) deletion promotes multiple pathologies in a mouse model of Alzheimer's disease.
866 Proceedings of the National Academy of Sciences. 2006;103(34):12867-72. doi:
867 10.1073/pnas.0601075103.
868 67. Rusilowicz-Jones E, Jardine J, Kallinos A, Pinto-Fernandez A, Guenther F, Giurrandino
869 M, et al. A novel USP30 inhibitor recapitulates genetic loss of USP30 and sets the trigger for
870 PINK1-PARKIN amplification of mitochondrial ubiquitylation. bioRxiv.
871 2020:2020.04.16.044206. doi: 10.1101/2020.04.16.044206.
872 68. Myers S, Yuan H, Kang J, Tan F, Traynelis S, Low C. Distinct roles of GRIN2A and
873 GRIN2B variants in neurological conditions [version 1; peer review: 2 approved].
874 F1000Research. 2019;8(1940). doi: 10.12688/f1000research.18949.1.
875 69. Babenko VN, Smagin DA, Galyamina AG, Kovalenko IL, Kudryavtseva NN. Altered
876 Slc25 family gene expression as markers of mitochondrial dysfunction in brain regions under
877 experimental mixed anxiety/depression-like disorder. BMC Neurosci. 2018;19(1):79. Epub
878 2018/12/13. doi: 10.1186/s12868-018-0480-6. PubMed PMID: 30537945; PubMed Central
879 PMCID: PMCPMC6288882.
880 70. Yokoyama JS, Wang Y, Schork AJ, Thompson WK, Karch CM, Cruchaga C, et al.
881 Association Between Genetic Traits for Immune-Mediated Diseases and Alzheimer Disease.
882 JAMA Neurol. 2016;73(6):691-7. doi: 10.1001/jamaneurol.2016.0150. PubMed PMID:
883 27088644.
884 71. Russ J. Systematic interaction mapping reveals novel modifiers of neurodegenerative
885 disease processes 2012.
886 72. Wostyn P, Audenaert K, De Deyn PP. Choroidal Proteins Involved in Cerebrospinal
887 Fluid Production may be Potential Drug Targets for Alzheimer's Disease Therapy. Perspect
888 Medicin Chem. 2011;5:11-7. doi: 10.4137/PMC.S6509. PubMed PMID: 21487536.
889 73. La Piana R, Weraarpachai W, Ospina LH, Tetreault M, Majewski J, Bruce Pike G, et al.
890 Identification and functional characterization of a novel MTFMT mutation associated with

891 selective vulnerability of the visual pathway and a mild neurological phenotype. neurogenetics.
892 2017;18(2):97-103. doi: 10.1007/s10048-016-0506-0.
893 74. Jeromin A, Lasseter HC, Provost AC, Daskalakis NP, Etkin A, Gehrman P, et al. Driving
894 Progress in Posttraumatic Stress Disorder Biomarkers. Biol Psychiatry. 2020;87(6):e13-e4. Epub
895 2019/10/19. doi: 10.1016/j.biopsych.2019.07.036. PubMed PMID: 31623824.
896 75. Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JI, et al. Genetic
897 architecture of gene expression traits across diverse populations. PLOS Genetics.
898 2018;14(8):e1007586. doi: 10.1371/journal.pgen.1007586.
899

# Supporting information captions

## Supplementary Texts

**Text S1. Details of PrediXcan's approach of estimating cis-eQTL effect sizes**

**Text S2. Details of TIGAR's approach of estimating cis-eQTL effect sizes**

**Text S3. Details of VC-TWAS approach with summary-level GWAS data and P-value calculation**

**Text S4. Details of ROS/MAP data**

## Supplementary Figures

**Fig S1.** 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. In Model I, the combinations of causal probability and phenotype heritability are $\left(p_{causal}, h_p^2\right) = ((0.001,0.2), (0.01,0.3), (0.1,0.4), (0.2,0.5))$. In Model II, the combinations of causal probability and phenotype heritability are $\left(p_{causal}, h_p^2\right) = ((0.001,0.1), (0.01,0.1), (0.1,0.15), (0.2,0.15))$.

**Fig S2.** TWAS power comparison for VC-TWAS and CoMM with phenotypes simulated from Model I (A) and Model II (B) using individual-level and summary-level data under the scenarios with $p_{causal} = 0.2$.

**Fig S3**. Box plot of the number of test SNPs considered by VC-TWAS of all genome-wide genes in the application studies of AD, with complete DPR weights and filtered DPR weights derived from the ROS/MAP training data.

**Fig S4.** 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.

**Fig S5.** Manhattan plots of VC-TWAS results with filtered DPR weights for studying quantitative AD pathology of $\beta$-Amyloid (A) and tangles (B). Genes with FDA < 0.05 by meta VC-TWAS for studying AD clinical diagnosis are colored in red in (A) and top significant gene for studying tangles phenotype with FDR = 0.058 is colored in red in (B).

**Fig S6.** Q-Q plots of VC-TWAS results with filtered DPR weights for studying $\beta$-amyloid, tangles, and global AD pathology with ROS/MAP cohort, as well as meta VC-TWAS results with filtered DPR weights for studying AD clinical diagnosis with ROS/MAP and Mayo Clinic cohorts.

**Fig S7.** Locus zoom plots of GWAS results and the magnitude (i.e., absolute value) of cis-eQTL effect size estimates by DPR method for SNPs that were considered by VC-TWAS of genes *TOMM40* (A, B) and *ZNF2334* (C, D). Filtered test SNPs with the cis-eQTL effect size magnitude > $10^{-4}$ were plotted here. SNPs with GWAS p-value $< 5 \times 10^{-8}$ were colored in red in (B,D), top significant SNPs by GWAS in (A,C) were shown as the blue triangle in (B,D).

**Fig S8.** Manhattan plots of VC-TWAS results with PrediXcan weights for studying AD clinical diagnosis (A) and global AD pathology (B).

**Fig S9.** Manhattan plots of VC-TWAS results with PrediXcan weights for studying quantitative AD pathology of β-Amyloid (A) and tangles (B).

**Fig S10.** Q-Q plots of VC-TWAS results with PrediXcan weights for studying $\beta$-amyloid, tangles, and global AD pathology 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.

**Fig S11.** Q-Q plots of VC-TWAS results with cis-eQTL DPR filtered weights and BGW weights on IGAP GWAS summary statistics.

## Supplementary Tables

**Table S1.** Computation time (in the unit of second) by Burden-TWAS, VC-TWAS and CoMM with individual and summary-level GWAS data by using with 1 CPU core with 32GB memory, for example genes that have different test SNP numbers.

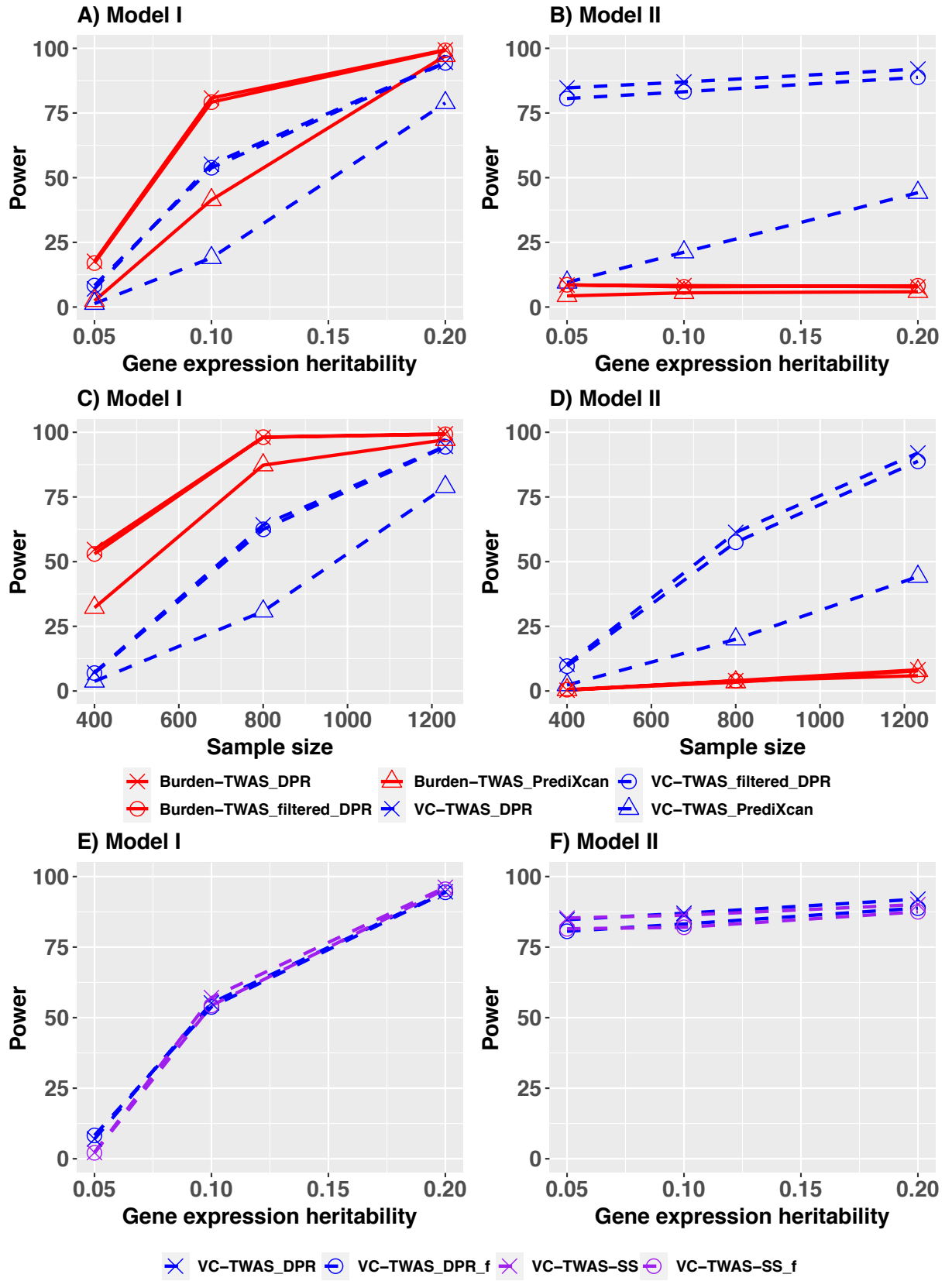
**Table S2.** Genes with VC-TWAS p-value <0.0013 with respect to at least one AD pathology phenotype and FDR <0.05 by meta VC-TWAS of AD clinical diagnosis. AD risk genes identified by previous GWAS are shaded in grey.

**Table S3.** Significant genes identified by VC-TWAS using IGAP summary statistics with filtered cis-eQTL DPR weights. Significant genes were identified with FDR < 0.05. AD risk genes identified by previous GWAS are shaded in grey.

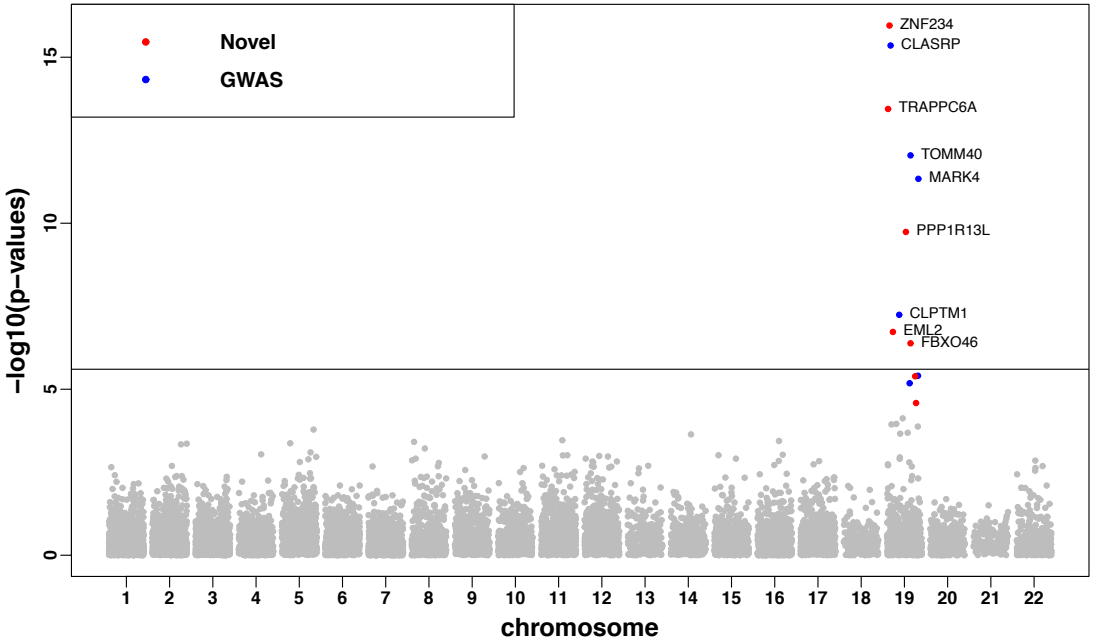
**Table S4:** Novel significant genes identified by VC-TWAS using summary statistics with BGW weights on IGAP summary statistics.

## Web Resources

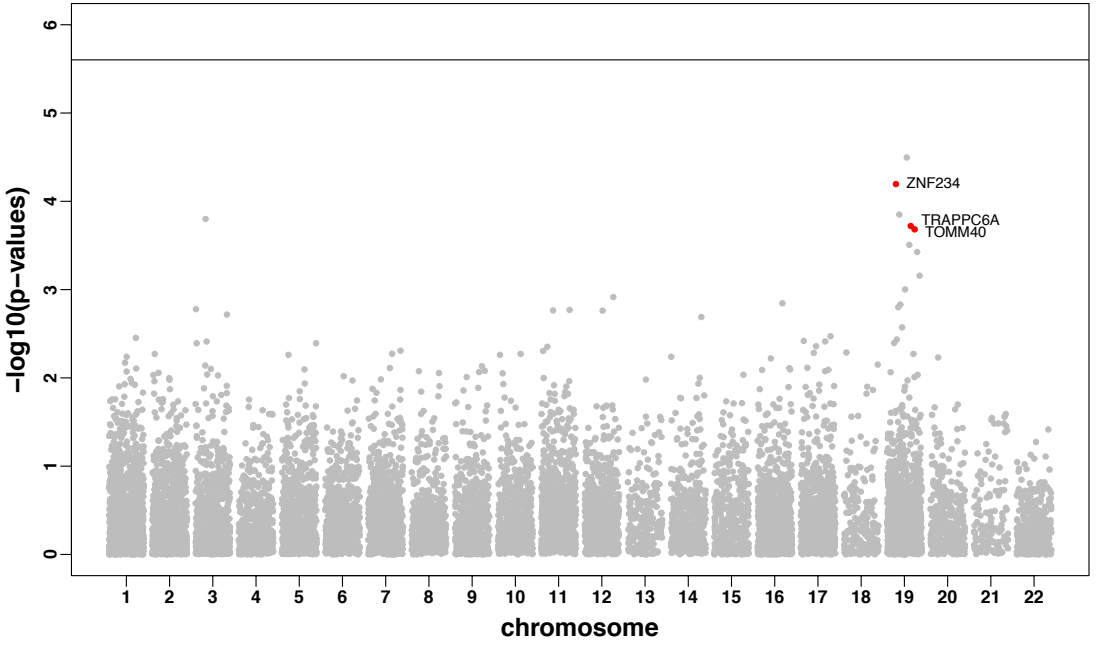
TIGAR, https://github.com/yanglab-emory/TIGAR
PrediXcan, https://github.com/hakyim/PrediXcan
RADC Research Resource Sharing Hub, http://www.radc.rush.edu/
ROS/MAP data, https://www.synapse.org/#!Synapse:syn3219045
MayoClinicLOAD data, https://www.synapse.org/#!Synapse:syn2910256
IGAP summary data, https://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php
BGW weights of brain frontal cortex tissue, https://www.synapse.org/#!Synapse:syn22316792

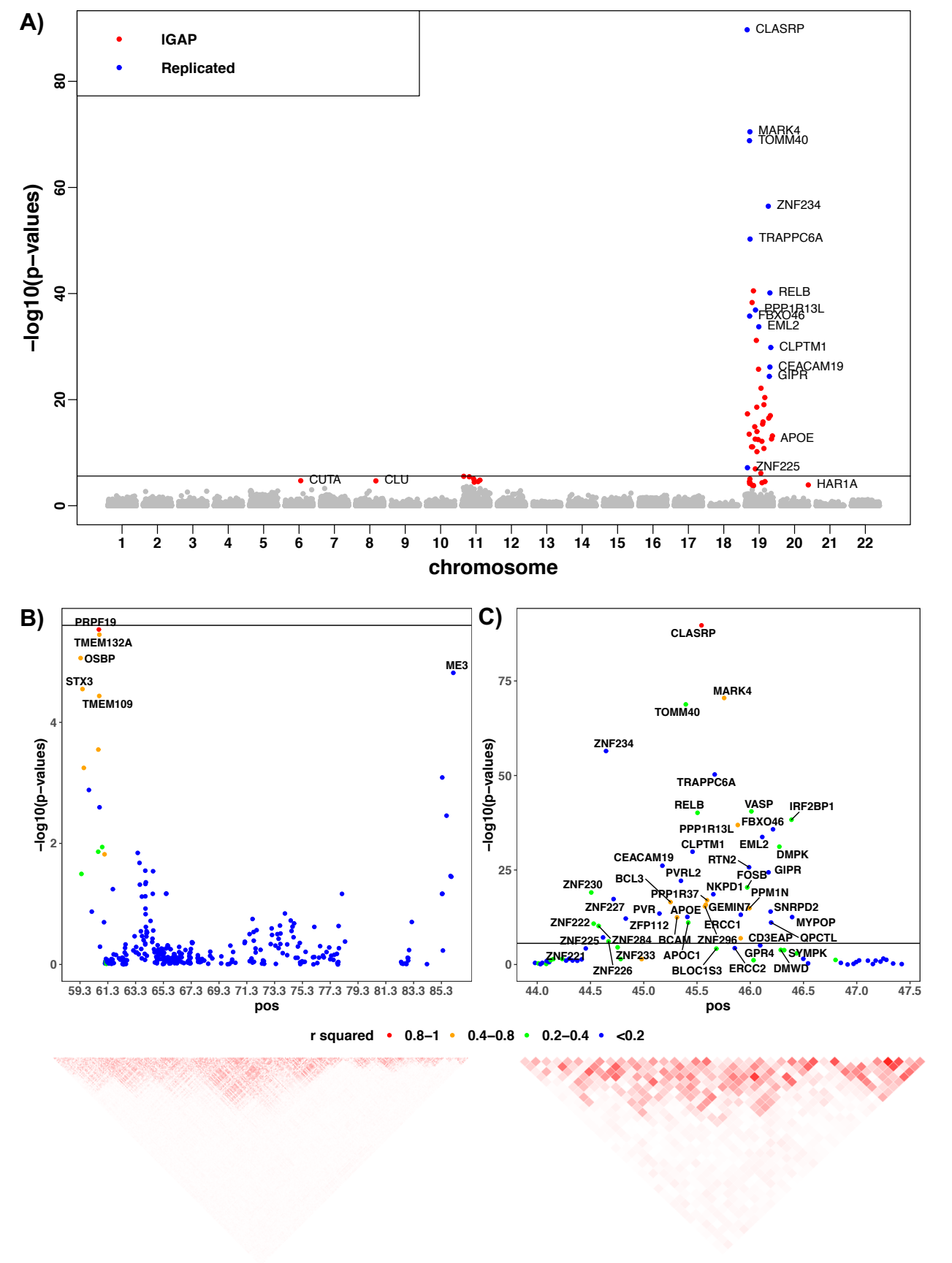

A) Model I
B) Model II
C) Model I
D) Model II
E) Model I
F) Model II
Power
Gene expression heritability
Sample size
0
25
50
75
100
0.05
0.10
0.15
0.20
400
600
800
1000
1200
Burden−TWAS_DPR
Burden−TWAS_filtered_DPR
Burden−TWAS_PrediXcan
VC−TWAS_DPR
VC−TWAS_filtered_DPR
VC−TWAS_PrediXcan
VC−TWAS_DPR
VC−TWAS_DPR_f
VC−TWAS−SS
VC−TWAS−SS_f

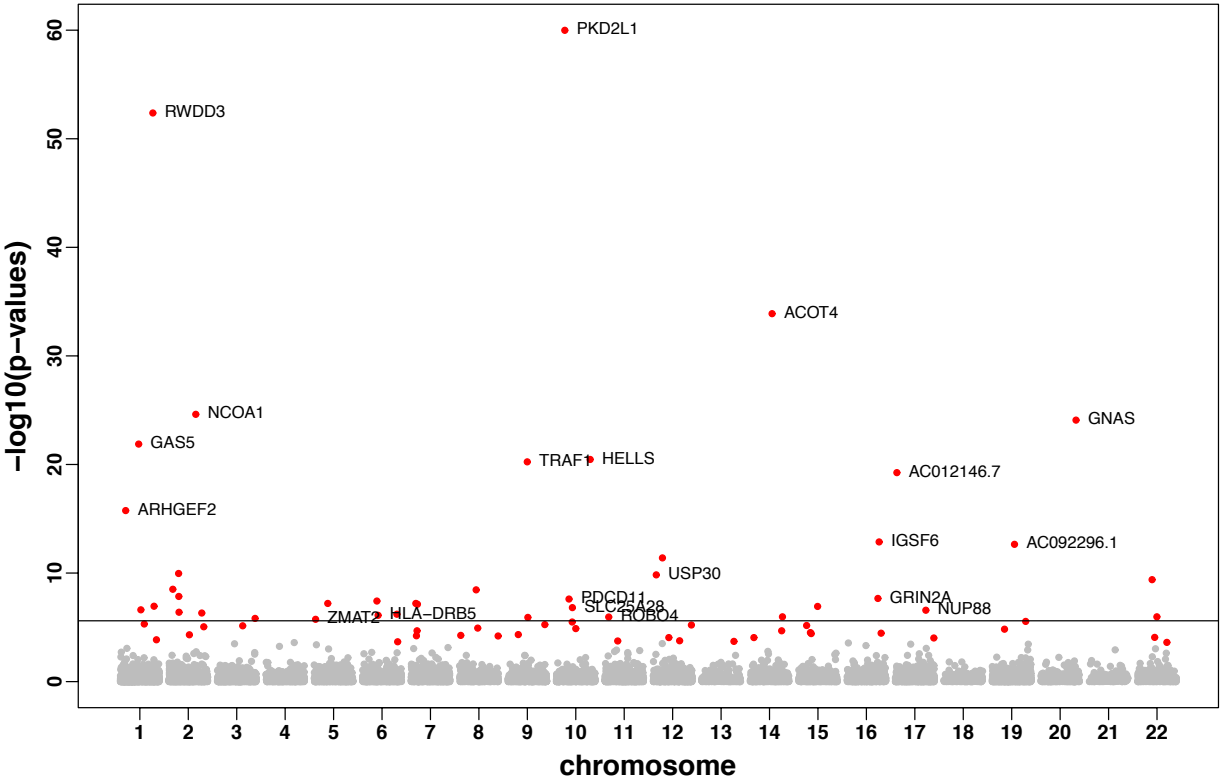
A)


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)
IGAP
Replicated
−log10(p−values)
chromosome
CLASRP
MARK4
TOMM40
ZNF234
TRAPPC6A
RELB
PPP1R13L
FBXO46
EML2
CLPTM1
CEACAM19
GIPR
APOE
ZNF225
HAR1A
CUTA
CLU
B)
PRPF19
TMEM132A
OSBP
STX3
TMEM109
ME3
−log10(p−values)
pos
C)
CLASRP
MARK4
TOMM40
ZNF234
TRAPPC6A
RELB
VASP
IRF2BP1
FBXO46
PPP1R13L
CLPTM1
EML2
DMPK
CEACAM19
RTN2
GIPR
PVRL2
FOSB
BCL3
ZNF230
NKPD1
PPM1N
PRP1R37
ZNF227
PVR
APOE
GEMIN7
SNRPD2
ZNF222
ZFP112
ERCC1
MYPOP
ZNF225
ZNF284
BCAM
ZNF296
CD3EAP
QPCTL
ZNF221
ZNF233
APOC1
GPR4
SYMPK
ZNF226
BLOC1S3
ERCC2
DMWD
−log10(p−values)
pos
r squared
0.8–1
0.4–0.8
0.2–0.4
<0.2


PKD2L1
RWDD3
ACOT4
NCOA1
GAS5
GNAS
TRAF1
HELLS
AC012146.7
ARHGEF2
IGSF6
AC092296.1
USP30
PDCD11
SLC25A28
GRIN2A
NUP88
ZMAT2
HLA−DRB5
ROBO4
−log10(p−values)
0
10
20
30
40
50
60
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
chromosome