Bayesian Genome-wide TWAS Method to Leverage both *cis*- and *trans*-eQTL Information through Summary Statistics

Justin M. Luningham,[1,5] Junyu Chen,[5] Shizhen Tang,[2,5] Philip L. De Jager,[3] David A. Bennett,[4] Aron S. Buchman,[4] and Jingjing Yang[5,*]

Summary

Transcriptome-wide association studies (TWASs) have been widely used to integrate gene expression and genetic data for studying complex traits. Due to the computational burden, existing TWAS methods do not assess distant *trans*-expression quantitative trait loci (eQTL) that are known to explain important expression variation for most genes. We propose a Bayesian genome-wide TWAS (BGW-TWAS) method that leverages both *cis*- and *trans*-eQTL information for a TWAS. Our BGW-TWAS method is based on Bayesian variable selection regression, which not only accounts for *cis*- and *trans*-eQTL of the target gene but also enables efficient computation by using summary statistics from standard eQTL analyses. Our simulation studies illustrated that BGW-TWASs achieved higher power compared to existing TWAS methods that do not assess *trans*-eQTL information. We further applied BWG-TWAS to individual-level GWAS data (N = ~3.3K), which identified significant associations between the genetically regulated gene expression (GReX) of *ZC3H12B* and Alzheimer dementia (AD) (p value = 5.42×10^{-13}), neurofibrillary tangle density (p value = 1.89×10^{-6}), and global measure of AD pathology (p value = 9.59×10^{-7}). These associations for *ZC3H12B* were completely driven by *trans*-eQTL. Additionally, the GReX of *KCTD12* was found to be significantly associated with β-amyloid (p value = 3.44×10^{-8}) which was driven by both *cis*- and *trans*-eQTL. Four of the top driven *trans*-eQTL of *ZC3H12B* are located within *APOC1*, a known major risk gene of AD and blood lipids. Additionally, by applying BGW-TWAS with summary-level GWAS data of AD (N = ~54K), we identified 13 significant genes including known GWAS risk genes *HLA-DRB1* and *APOC1*, as well as *ZC3H12B*.

Introduction

Although genome-wide association studies (GWASs) have identified thousands of variants associated with complex traits over the past decades,[1–5] most of these associations are located within noncoding regions and the underlying biological mechanisms by which these variants impact a phenotype are unknown.[6,7] Recent studies have shown that GWAS associations were enriched for regulatory elements such as expression quantitative trait loci (eQTL),[8–10] suggesting that integrating transcriptomic and genetic data could help identify key molecular mechanisms underlying complex traits.

One such integrative method is transcriptome-wide association study (TWAS),[11–13] which takes advantage of a reference panel with profiled transcriptomic and genetic data from the same individuals. A TWAS first utilizes such reference data to fit an imputation regression model for the expression quantitative trait of a target gene with nearby genotypes (e.g., *cis*-SNPs within 1 MB region of transcription starting site) as predictors, and then examines the gene-based association between the imputed genetically regulated gene expression (GReX) and the phenotype of interest. With fitted gene expression imputation models from reference data, TWASs can be conducted with test samples that have either individual-level or summary-level GWAS data.[12–14] The SNPs with non-zero effect sizes on reference transcriptome in the fitted imputation models are referred to as broad sense "eQTL" in TWASs. Examples of publicly available reference data include the Genotype-Tissue Expression (GTEx) project with transcriptomic data for 54 human tissues,[8] Genetic European Variation in Health and Disease (GEUVADIS) for lymphoblastoid cell lines,[15] and North American Brain Expression Consortium (NABEC) for cortex tissues.[16]

Essentially, a TWAS is equivalent to a burden type gene-based test taking "*cis*-eQTL effect sizes" that are non-zero coefficients of *cis*-SNPs from the fitted GReX imputation model as their corresponding burden weights.[11–13] By weighting genetic variants using *cis*-eQTL effect sizes, a TWAS assumes the effects of risk genes on the phenotype of interest are potentially mediated through their transcriptome variations. Recent studies of a wide range of complex traits such as schizophrenia, breast cancer, and Alzheimer dementia (AD)[17–21] using TWASs have identified additional risk genes besides known GWAS risk loci, demonstrating that additional significant associations can be identified by TWASs.

[1]Department of Population Health Sciences, Georgia State University School of Public Health, Atlanta, GA 30303, USA; [2]Department of Biostatistics and Bioinformatics, Emory University School of Public Health, Atlanta, GA 30322, USA; [3]Center for Translational and Computational Neuroimmunology, Department of Neurology and Taub Institute for Research on Alzheimer disease and the Aging Brain, Columbia University Irving Medical Center, New York, NY 10032, USA; [4]Rush Alzheimer disease Center, Rush University Medical Center, Chicago, IL 60612, USA; [5]Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA
*Correspondence: jingjing.yang@emory.edu

https://doi.org/10.1016/j.ajhg.2020.08.022.

However, existing TWAS methods only use genetic data of *cis*-SNPs of the target gene as predictors to fit the GReX imputation model.[11–13] As shown by recent studies, *trans*-SNPs (e.g., outside of the 1 MB region) of the target gene not only explain a significant amount of variation for most expression quantitative traits, but also often contain significant *trans*-eQTL that are likely to inform molecular mechanisms.[22,23] Thus, using both *cis*- and *trans*-SNPs is likely to increase the imputation accuracy of GReX and the power of TWASs. Nonetheless, the enormous computational cost required to fit ~20K GReX imputation models for genome-wide genes and ~10M genotypes per tissue type makes the routine use of existing TWAS methods impractical.

We propose a Bayesian genome-wide TWAS (BGW-TWAS) method that accounts for both *cis*- and *trans*-SNPs based on a Bayesian variable selection regression (BVSR) model[24] for imputing GReX. Our BGW-TWAS method circumvents the current computational burden impeding TWASs by enabling efficient computation via the scalable EM-MCMC algorithm[25] and the summary statistics of standard eQTL analyses based on single variant tests. First, we demonstrate the feasibility of this Bayesian approach by simulation studies with varying proportions of true causal *cis*- and *trans*-eQTL for expression quantitative traits. We compared BGW-TWAS with several existing TWAS methods including PrediXcan[11] and TIGAR[13] that assess only *cis*-SNPs. Then we applied BGW-TWAS to clinical and postmortem data from older adults with individual-level GWAS data (N = ~3.3K)[26] to study several clinical and pathological AD-related phenotypes including clinical diagnosis of AD, neurofibrillary tangle density, β-amyloid load, and a global summary measure of AD pathology. Further, we compared BGW-TWAS with alternative methods by using GWAS summary statistics for AD available from the International Genomics of Alzheimer's Project (IGAP)[27] (N = ~54K).

Our simulation studies revealed that BGW-TWAS achieved higher TWAS power by considering both *cis*- and *trans*-SNPs when *trans*-eQTL accounted for a non-negligible proportion of transcriptome variance. Our studies of human AD GWAS datasets identified several risk genes associated with AD phenotypes that were driven by *trans*-eQTL and thus not identified by alternative methods. The software for implementing BGW-TWAS is available freely on Github (see Web Resources).

Material and Methods

TWAS Procedure

The first step in a TWAS is to train an imputation model for profiled gene expression levels using genotype data as predictors on a per-gene basis.[11–13] The general imputation model based on linear regression is given by

$$\mathscr{E}_g = \boldsymbol{X}\boldsymbol{w} + \epsilon, \qquad \text{(Equation 1)}$$

where $\mathscr{E}_g$ denotes the expression quantitative trait of the target gene, centered and adjusted for non-genetic covariates; $\boldsymbol{X}$ denotes centered genotype data; $\boldsymbol{w}$ denotes the corresponding "eQTL" effect sizes for the target gene; and ϵ is an error term following a $N(0,\ \sigma_\epsilon^2 \boldsymbol{I})$ distribution. The intercept term is dropped for centering both response ($\mathscr{E}_g$) and explanatory ($\boldsymbol{X}$) variables. With $\widehat{\boldsymbol{w}}$ estimated from the training data (i.e., reference data) that have both transcriptomic and genetic data profiled for the same subjects, a TWAS will test the association between the phenotype of interest and the imputed GReX obtained from individual-level GWAS genotype data $\tilde{\boldsymbol{X}}$ of the test cohort as follows

$$\widehat{GReX} = \tilde{\boldsymbol{X}}\widehat{\boldsymbol{w}}.$$

Bayesian Variable Selection Regression

Existing TWAS methods only consider SNPs within 1 MB of the flanking 5′ and 3′ ends (*cis*-SNPs) in the gene expression imputation model (Equation 1).[11–13] In order to leverage additional information provided by *trans*-eQTL that are located outside the 1 MB flanking region of the target gene, we utilize the Bayesian Variable Selection Regression (BVSR)[24] model to account for both *cis*- and *trans*-SNPs as follows:

$$\mathscr{E}_g = \boldsymbol{X}_{cis}\boldsymbol{w}_{cis} + \boldsymbol{X}_{trans}\boldsymbol{w}_{trans} + \epsilon,\ \epsilon_i \sim N\left(0,\ \sigma_\epsilon^2\right). \qquad \text{(Equation 2)}$$

The BVSR model assumes a spike-and-slab prior distribution for w_i. That is, the prior on w_i is a mixture distribution of a normal distribution with zero mean and a point-mass density function at 0. In order to model potentially different distributions of the effect sizes for *cis*- and *trans*-SNPs, we assume the following respective priors,

$$w_{cis,i} \sim \pi_{cis} N\left(0,\ \sigma_{cis}^2\sigma_\epsilon^2\right) + (1-\pi_{cis})\delta_0(w_{cis,i});$$

$$w_{trans,i} \sim \pi_{trans} N\left(0,\ \sigma_{trans}^2\sigma_\epsilon^2\right) + (1-\pi_{trans})\delta_0(w_{cis,i}); \qquad \text{(Equation 3)}$$

where $(\pi_{cis},\ \pi_{trans})$ denote the respective probability that the coefficient is non-zero and normally distributed, and $\delta_0(w_i)$ is the point mass density function that takes value 0 when $w_i \neq 0$ and 1 when $w_i = 0$. Further, the following conjugate hyper prior distributions are respectively assumed for the *cis*- and *trans*-specific parameters,

$$\pi_{cis} \sim Beta(a_{cis}, b_{cis});\ \sigma_{cis}^2 \sim IG(k_1,\ k_2);$$

$$\pi_{trans} \sim Beta(a_{trans}, b_{trans});\ \sigma_{trans}^2 \sim IG(k_3,\ k_4); \qquad \text{(Equation 4)}$$

where *IG* indicates the Inverse Gamma distribution and hyper parameters $(a_{cis}, b_{cis}, a_{trans}, b_{trans},\ k_1,\ k_2, k_3,\ k_4)$ will be chosen to enable non-informative hyper prior distributions (see Supplemental Material and Methods for model details).

To facilitate computation, a latent indicator γ_i is assumed such that $w_i = 0$ if $\gamma_i = 0$, and w_i follows a normal distribution if $\gamma_i = 1$. Then the expected value of this indicator, $E[\gamma_i]$, represents the posterior probability (PP_i) for each individual SNP to have a non-zero effect size (i.e., to be an eQTL of the target gene).[24] Moreover, we propose a Bayesian approach to estimate GReX for test samples that can account for the uncertainty for each SNP to be an eQTL:

$$\widehat{GReX} = \sum_{i=1}^{p} X_i^{\sim}\left(\widehat{PP_i}\widehat{w}_i\right), \qquad \text{(Equation 5)}$$

where $\widetilde{X_i}$ represents the genotype data of variant *i* for test samples and $(\widehat{w}_i,\ \widehat{PP_i})$ denote the estimate of effect size and posterior probability (PP) of having a non-zero effect size from the BVSR model (Equations 2, 3, and 4) (see Supplemental Material and Methods for detailed Bayesian inference procedure). This Bayesian GReX estimate can then be used to conduct a TWAS with individual-level GWAS data by testing the association between the imputed GReX and the phenotype of interest.

BGW-TWAS with Summary-Level GWAS Data

With summary-level GWAS data that were generated by single variant tests, we employed the S-PrediXcan[14] approach to obtain a burden TWAS Z-score test statistic, including not only *cis*- but also *trans*-eQTL in the test. Let $\widehat{\beta_l}$ denote the SNP effect size of SNP l from GWAS, $SE(\widehat{\beta_l})$ denote the standard error of $\widehat{\beta_l}$, Z_l denote the Z-score statistic value by single variant test, $\widehat{\sigma_l}$ denote the estimated standard deviation of the genotype data of SNP l from reference panel, and $\widehat{\sigma_g}$ denote the estimated standard deviation of the imputed expression of gene g from reference panel. The burden TWAS Z-score test statistic for gene g is given by

$$Z_g = \sum_{l \in Model_g} \widehat{w_{lg}} \frac{\widehat{\sigma_l}}{\widehat{\sigma_g}} \frac{\widehat{\beta_l}}{SE\left(\widehat{\beta_l}\right)} = \sum_{l \in Model_g} \widehat{w_{lg}} \frac{\widehat{\sigma_l}}{\widehat{\sigma_g}} Z_l = \sum_{l \in Model_g} \frac{\left(\widehat{w_{lg}}\,\widehat{\sigma_l}\right) Z_l}{\sqrt{\widehat{w}' \boldsymbol{V} \widehat{w}}},$$

$$\widehat{\sigma_l^2} = Var(x_l),\ \widehat{\sigma_g^2} = \widehat{w}' \boldsymbol{V} \widehat{w},\ \boldsymbol{V} = Cov(\boldsymbol{X}),$$

where $\widehat{w_{lg}} = \widehat{PP_i}\widehat{w}_i$ is the product of posterior probability for SNP l to have non-zero eQTL effect size from the BVSR model (Equation 2). Here, $\boldsymbol{X}$ denotes the genotype matrix of analyzed SNPs from reference panels of the same ethnicity and $\boldsymbol{V}$ denotes the corresponding genotype covariance matrix.

Efficient Computation Techniques

In theory, the estimates of eQTL effect sizes and corresponding posterior probabilities $(\widehat{w}_i, \widehat{PP_i})$ can be obtained by using a standard Markov Chain Monte Carlo (MCMC)[28] algorithm. However, in practice, the computation burden for modeling genome-wide genotype data is nearly impossible because of enormous required memory capacity and slow convergence rate for MCMC. To circumvent these practical limitations, we employ several techniques to enable computational efficiency such that BGW-TWAS method can be deployed to leverage both *cis*- and *trans*-eQTL information in practice. In particular, we adapt a previously developed scalable expectation-maximization Markov Chain Monte Carlo (EM-MCMC) algorithm.[25] Unlike the original EM-MCMC algorithm requiring individual-level GWAS data, we can reduce up to 90% of the computation time by adapting the EM-MCMC algorithm to utilize only summary statistics, including the pre-calculated linkage disequilibrium (LD) coefficients and score statistics from standard eQTL analyses by single variant tests. Additionally, we prune genome-wide genotypes into a subset of genome regions that are approximately independent and contain either at least one *cis*-SNP or one *trans*-SNP with p value $< 1 \times 10^{-5}$ by standard eQTL analyses (technical details are provided in Supplemental Material and Methods).

Simulation Study Design

We conducted simulation studies to validate the performance of our proposed BGW-TWAS method through comparing with the alternative existing methods, e.g., PrediXcan, TIGAR, as well as BVSR using only *cis*-eQTL. To mimic real studies, we used real genotype data from the ROS/MAP study to simulate gene expression and phenotype data. We took 499 samples as our training data and 1,209 samples as our test data. GReX imputation models were fitted using the training data where "eQTL" effect sizes and the corresponding posterior probabilities were estimated. Given these fitted GReX imputation models, GReX data were imputed for a follow-up TWAS with the test data.

We arbitrarily selected five approximately independent genome blocks, including one "*cis*-" and four "*trans*-" genotype blocks (variants were filtered with minor allele frequency (MAF) > 5% and Hardy-Weinberg p value $> 10^{-5}$). With genotype matrix $\boldsymbol{X_g}$ of the randomly selected causal eQTL, we generated effect-sizes w_i to target a selected gene expression heritability h_e^2 and that all causal eQTL explain equal expression heritability. Gene expression levels were generated by $\mathscr{E}_{\boldsymbol{g}} = \boldsymbol{X_g w} + \epsilon_e$, with $\epsilon_e \sim N(0, (1 - h_e^2))$. Then we simulated phenotypes by $\boldsymbol{Y} = \beta \mathscr{E}_{\boldsymbol{g}} + \epsilon_p$, where β was selected with respect to a selected phenotype heritability h_p^2 and $\epsilon_p \sim N(0, (1 - h_p^2))$.

To mimic the complex genomic architecture of gene expression in practice, we considered two scenarios, one with 5 true causal eQTL representing the scenario with relatively large effect sizes and the other one with 22 true causal eQTL representing the scenario with relatively small effect sizes. For the scenario with 5 true causal eQTL, we considered three sub-scenarios with respect to how these true causal eQTL distributed over considered genome blocks: (1) all causal eQTL are from the *cis*-block; (2) two causal eQTL are from the *cis*-block explaining 70% of the specified h_e^2 while the other three causal eQTL are from the *trans*-blocks explaining the other 30% of h_e^2; (3) all causal QTL are from the *trans*-block. Similarly, for the scenario with 22 true causal eQTL, we considered three scenarios where 30%, 50%, and 70% of the causal eQTL were from *cis*-genome blocks. We also varied the total expression trait heritability and phenotype heritability in both scenarios, i.e., $(h_e^2, h_p^2) = ((5\%,\ 90\%),\ (10\%,\ 45\%),\ (20\%,\ 20\%),\ (50\%,\ 6\%))$ for the scenario with 5 true causal eQTL and $(h_e^2, h_p^2) = ((5\%,\ 99\%),\ (10\%,\ 80\%),\ (20\%,\ 35\%),\ (50\%,\ 8\%))$ for the scenario with 22 true causal eQTL. Here, different levels of phenotype heritability were arbitrarily selected to achieve similar levels of TWAS power across all scenarios.

In each simulation, with training data, we first fitted GReX imputation models by BVSR (BGW-TWAS) with both *cis*- and *trans*-genome blocks, as well as by Elastic-Net (PrediXcan) and nonparametric Bayesian Dirichlet process regression (TIGAR) with only *cis*-genome block. Then we conducted TWAS with imputed *GReX* by respective method. We also compared BGW-TWAS with using only *cis*-eQTL estimates from the same BVSR model. The performance was compared in terms of R^2 of the imputed GReX and TWAS power in test samples. Test R^2 was calculated as the squared correlation between imputed GReX and simulated gene expression values of the test samples. TWAS power was calculated as the proportion of 1,000 repeated simulations of each scenario with p value $< 2.5 \times 10^{-6}$ (genome-wide significance threshold for gene-based association studies).

ROS/MAP and Mayo Clinic GWAS Data of AD

Following simulation studies, we applied BGW-TWAS method to individual-level genomic and AD related phenotype data from older adults available from several studies. We used transcriptomic data, GWAS data, clinical diagnosis of AD and postmortem indices of AD pathology from the Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP)[29–31] and GWAS data from the Mayo Clinic Alzheimer's Disease Genetics Studies (MCADGS).[32–34] All participants from ROS/MAP sign an informed consent, an Anatomic Gift Act, and a consent for their data to be deposited in the Rush Alzheimer's Disease Center (RADC) repository. ROS/MAP studies were approved by the Institutional Review Board of Rush University Medical Center, Chicago, IL. MCADGS contains samples from two clinical AD case-control series (Mayo Clinic Jacksonville and Mayo Clinic Rochester) as well as a neuropathological series of autopsy-confirmed subjects from the Mayo Clinic Brain Bank.

Microarray genotype data generated for 2,093 European-decent subjects from ROS/MAP[35] and 2,099 European-decent subjects from MCADGS were further imputed to the 1000 Genomes Project Phase 3[36] in our analysis.[37]

Post-mortem brain samples from the dorsal lateral prefrontal cortex from ~30% of these ROS/MAP participants with assayed genotype data were profiled for transcriptomic data by next-generation RNA seqencing.[38] These data were used as reference data to train GReX prediction models in this study. We conducted TWASs for both clinical and pathological AD phenotypes. The clinical diagnosis of late-onset Alzheimer dementia was available for both ROS/MAP and MCADGS. Postmortem pathology indices of AD were only available for ROS/MAP and included PHFtau tangle density, β-amyloid load, and a global measure of AD pathology based on measures of neuritic and diffuse plaques and neurofibrillary tangles.[29–31] Additional details about the ongoing ROS/MAP cohort studies and how postmortem indices of tangles and β-amyloid load were quantified are included in prior publications[29–31] and summarized in the Supplemental Material and Methods.

Results

Simulation Results

For the scenario with five true causal eQTL and various expression heritability, our simulation studies showed that BGW-TWAS obtained the highest test R^2 for GReX and TWAS power than PrediXcan and TIGAR when any portion of the true causal eQTL are distributed over *trans*-genome blocks (Figures 1A and 1B). This is because BGW-TWAS leverages both *cis*- and *trans*-eQTL information while the alternative methods fail to account for *trans*-eQTL. Especially, when all true causal eQTL are from *trans*-genome regions, the alternative methods barely have any power to identify the TWAS association with nearly zero test R^2. As expected, BGW-TWAS and PrediXcan performed comparably when all causal eQTL were from the *cis*-genome block, while TIGAR performed slightly worse with sparse true causal eQTL (Figure 1A).

For the scenario with 22 mixed *cis*- and *trans*-eQTL, the performance comparison became more complicated with respect to various true expression heritability levels (Figures 1C and 1D). Particularly, when $h_e^2 = 0.05$, all methods had difficulties accurately estimating eQTL effect sizes and resulted in nearly zero test R^2. As expression heritability increased, the advantage of modeling both *cis*- and *trans*-genotype data by BGW-TWAS arisen and led to higher test R^2 and TWAS power. When $h_e^2 = (0.1, 0.2)$ and 70% of the true causal eQTL were *cis*-, BGW-TWAS was less effective than PrediXcan and TIGAR while TIGAR achieved the best performance. This is likely due to the fact that the nonparametric Bayesian Dirichlet process regression model used by TIGAR is preferred when true causal eQTL manifest relatively small effect sizes, which is consistent with previous findings.[13]

In contrast, when true causal eQTL signals have relatively large effect sizes and are distributed outside the *cis*-region of the target gene, BGW-TWAS method is preferred due to the improved accuracy for GReX prediction by leveraging *trans*-SNP data. By comparing with using only BVSR estimates of *cis*-eQTL, we showed that a significant proportion of transcriptome variation due to *trans*-eQTL was missed and the follow-up TWAS was underpowered.

TWAS of AD-Related Phenotypes with Individual-Level GWAS Data

Next, we applied BGW-TWAS to the individual-level GWAS data from ROS/MAP[26,31] and MCADGS.[32] First, we trained the BVSR GReX imputation models using samples ($n = 499$) from the ROS/MAP cohort that contained both profiled genotype data and transcriptomic data obtained from the dorsal lateral prefrontal cortex. All expression quantitative traits were normalized and corrected for age at death, sex, postmortem interval (PMI), study (ROS or MAP), batch effects, RNA integrity number (RIN), top three principal components derived from genome-wide genotype data, and cell type proportions (oligodendrocytes, astrocytes, microglia, neurons). The cell type proportions were derived by using CIBERSORT pipeline[39] with single-cell RNA-seq transcriptome profiles from human brain tissues as in Darmanis et al. [40] to de-convolute bulk RNA-seq data.[41]

When we applied BGW-TWAS, we obtained GReX imputation models for 14,156 genes, compared to respective 6,011 genes and 14,214 genes by PrediXcan and TIGAR that have at least one *cis*-eQTL with nonzero effect size on expression quantitative trait. Across the 6,011 genes with GReX imputation models by PrediXcan, our BGW-TWAS approach had a smaller train R^2 (squared correlation between fitted and profiled gene expression values) value for expression quantitative traits for only 855 genes (Figure S1A). While TIGAR and BGW-TWAS yielded a similar number of GReX imputation models, BGW method is expected to result in higher imputation accuracy when *trans*-eQTL play an important role in affecting gene expression levels as shown by our simulation results. Of 13,142 genes that had imputation models fitted by both TIGAR and BGW-TWAS, BGW-TWAS had smaller train R^2 for only 3,304 genes. That is, BGW-TWAS would be preferred for genes that have sparse eQTL, especially *trans*-eQTL, while TIGAR would be preferred for genes that have less sparse eQTL that are mostly *cis*-eQTL (Figure S1B).

We imputed Bayesian GReX values for all remaining individuals with genotype data in ROS/MAP and MCADGS by using Equation 4. We then conducted TWASs by testing the association between the standardized GReX values (with unit variance) and both clinical and pathological AD phenotypes. The TWASs for these phenotypes controlled for age at death, sex, smoking, ROS or MAP study, education level, and top three principal components derived from genome-wide genotype data.

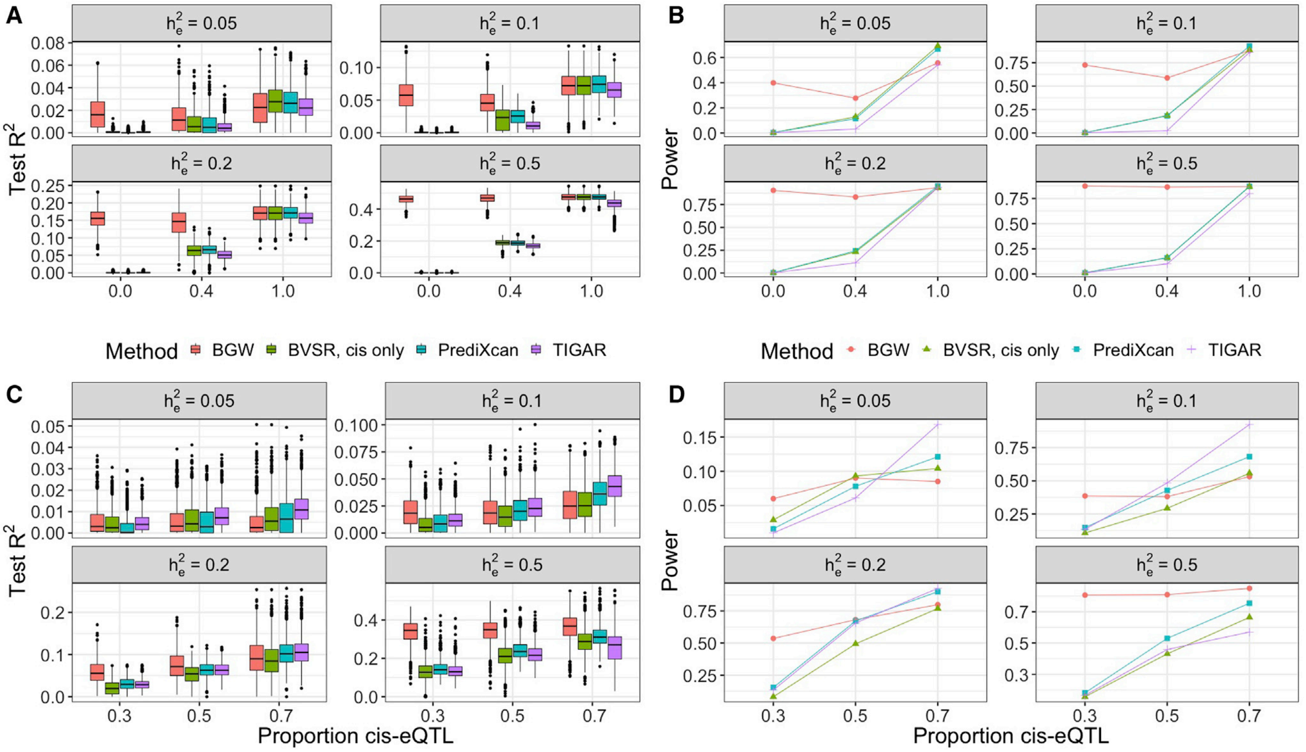


Figure 1. Simulation TWASs Comparing BGW-TWAS, BVSR with *cis*-eQTL only, PrediXcan, and TIGAR Methods
Simulation studies used various gene expression heritability $h_e^2 = (0.05,\ 0.1,\ 0.2,\ 0.5)$ and various true causal *cis*-eQTL proportions. Test R^2 was calculated as the squared correlation between imputed GReX and simulated gene expression values of the test samples.
(A and B) Test R^2 and TWAS power comparison with 5 true causal eQTL. BGW-TWAS was found to out-perform the alternative methods when a non-negligible proportion of true causal eQTL were from *trans*-genome regions.
(C and D) Test R^2 and TWAS power comparison with 22 true causal eQTL. BGW-TWAS was found to out-perform alternative method when >50% of true causal eQTL were from *trans*-genome regions and $h_e^2 > 0.1$.

For the dichotomous phenotype of clinical diagnosis of AD, the case/control status was determined by different rules and the available confounding variables were different for ROS/MAP and MCADGS. Cognitive status at death for individuals from the ROS/MAP cohort is based on the review of all longitudinal clinical data available at the time of death blinded to all pathologic data. Individuals were classified as having no cognitive impairment, mild cognitive impairment, or AD. In this study, samples from individuals with AD were taken as case subjects and samples from individuals without dementia (i.e., either with no cognitive impairment or mild cognitive impairment) were taken as control subjects. For the MCADGS samples, case subjects were determined for samples with a medical history of late-onset AD diagnosis, and available confounding variables included only age, sex, and top three principal components derived from GWAS data. Therefore, we meta-analyzed these two cohorts for AD clinical diagnosis by applying the inverse-variance weighting method[42] to summary statistics obtained by TWAS per cohort. We compared the meta-TWAS results obtained with BGW-TWAS to alternative TWAS methods.

BGW-TWAS identified *ZC3H12B* (located on chromosome X) whose GReX values were associated with AD with effect size $\beta = 0.265$, p value 5.42×10^{-13}, and FDR $= 3.07 \times 10^{-8}$ (Figure 2A; Table 1). Both within-cohort TWASs obtained positive effect sizes ($\beta = 0.22,\ 0.29$) for ROS/MAP and MCADGS, with respective p value $= 2 \times 10^{-4},\ 4.12 \times 10^{-10}$. On the other hand, this gene was not identified by either PrediXcan or TIGAR because the association of this gene is completely driven by *trans*-eQTL (Figures S2 and S4).

TWASs of pathological AD phenotypes were restricted to ROS/MAP from whom postmortem AD indices were collected. TWASs were conducted for individuals with GWAS genotype data and AD pathology indices: tangles (n = 1,121), β-amyloid (n = 1,114), and global AD pathology (n = 1,139). These results are shown in the Manhattan plots in Figures 2B, 3A, 3B, S2, S3, and S5– S7.

Using BGW-TWAS, *ZC3H12B* was identified to be associated with global AD pathology with p value $= 9.59 \times 10^{-7}$ (Figure 2B; Table 1) as well as neurofibrillary tangle density with p value 1.89×10^{-6} (Figure 3A; Table 1). *KCTD12* located on chromosome 12 was identified to be significantly associated with β-amyloid load with p value $= 3.44 \times 10^{-8}$(Figure 3B; Table 1).

We show the BVSR posterior probabilities for considered SNPs to be eQTL for *ZC3H12B* and *KCTD12* in Figure 4 and the standard eQTL analyses results for these two genes in Figure S4. These data suggest that the association between

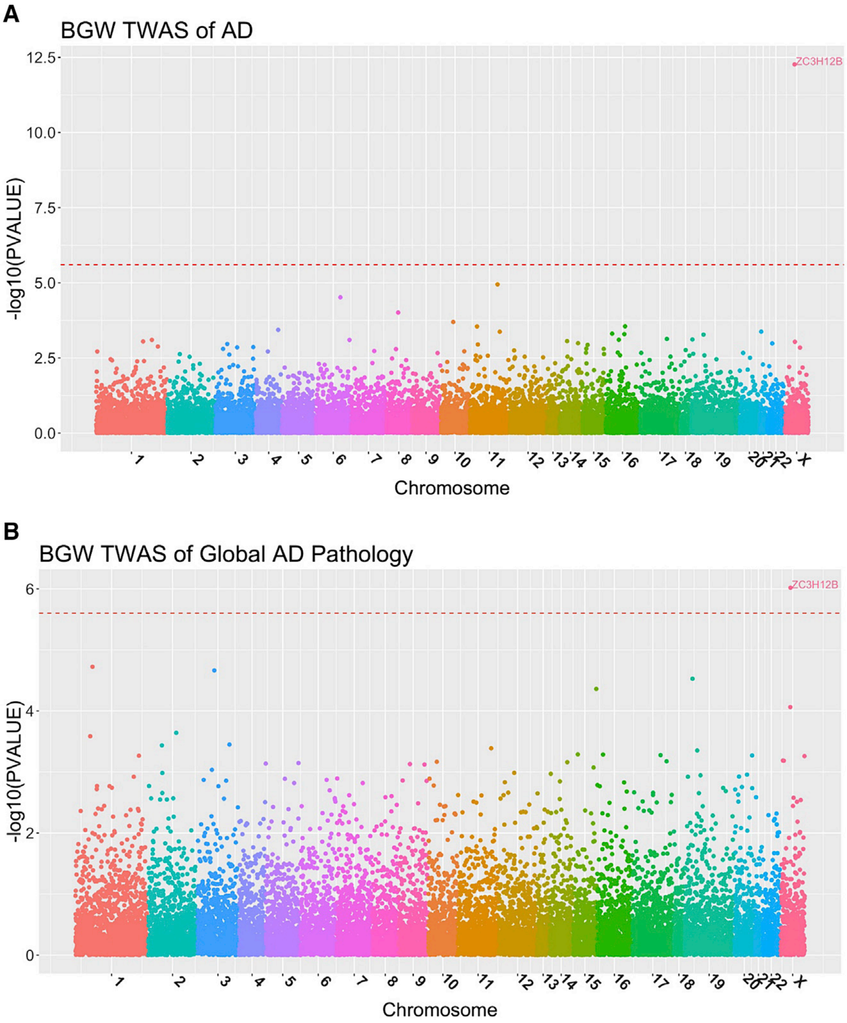


Figure 2. Manhattan Plots of BGW-TWAS Results of AD Clinical Diagnosis and Global AD Pathology

Here, -log10(p values) by BGW-TWAS were plotted and red lines denote genome-wide significant threshold (2.5×10^{-6}) for gene-based association studies. ZC3H12B was found to be significantly associated with both AD clinical diagnosis (A) and global AD pathology (B).

the imputed GReX values of *ZC3H12B* and AD phenotypes is completely driven by *trans*-eQTL, while the association between the GReX values of *KCTD12* and β-amyloid load is driven by both *cis*- and *trans*-eQTL. Four of the top driven *trans*-eQTL (*rs12721051, rs4420638, rs56131196, rs157592;* Table 2) for *ZC3H12B* are located in *APOC1*, a known risk gene for AD[43] and blood lipids,[44–46] which is <12 KB away from the well-known AD risk gene *APOE*.[47] Specifically, *rs12721051* located in the 3′ UTR region of *APOC1* was identified as a GWAS signal of total cholesterol levels;[46] *rs4420638* located in the downstream of *APOC1* is in linkage disequilibrium (LD) with the *APOE-E4* allele (*rs429358*) and was identified to be a GWAS signal of various blood lipids measurements (i.e., low-density lipoprotein cholesterol measurement, C-reactive protein measurement, triglyceride measurement, and total cholesterol measurement)[44] and AD;[48] *rs56131196* located in the downstream and *rs157592* located in the regulatory region of *APOC1* were identified as GWAS signals of AD and independent of *APOE-E4*.[49] Additionally, *ZC3H12B* was found to regulate pro-inflammatory activation of macrophages[50] and has higher expression in brain, spinal cord, and thymus tissue types compared to other tissues.[51] These results showed that the effects of these known GWAS signals (*rs4420638, rs56131196, rs157592*) of AD could be mediated through the expression levels of *ZC3H12B*.

TWAS of AD with Summary-Level GWAS Data

To validate our findings using individual-level GWAS data from ROS/MAP and MCADGS, we conducted a TWAS of AD using the publicly available IGAP GWAS summary statistics.[27] Specifically, we used the GWAS summary statistics that were generated by meta-analysis of four consortia (~17K case subjects and ~37K control subjects, Europeans): the Alzheimer's Disease Genetic Consortium (ADGC), the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, the European Alzheimer's Disease Initiative (EADI), and the Genetic and Environmental Risk in Alzheimer's Disease (GERAD) Consortium.

BGW-TWAS identified 13 significant genes located in chromosome 3, 6, 7, 10, 11, 19, and X, including known GWAS risk genes *HLA-DRB1*[52] and *APOC1*, and *ZC3H12B* that was identified using individual-level GWAS data from

Table 1. Significant Risk Genes Identified by BGW-TWAS using Individual-Level GWAS Data from ROS/MAP and MCADGS Cohorts

Gene	CHR	Position	Train R^2	p Value	Effect Size (SD)	Phenotype
ZC3H12B	X	64,708,614	0.24	5.42×10^{-13}	0.265 (0.037)	AD
ZC3H12B	X	64,708,614	0.24	9.59×10^{-7}	0.142 (0.029)	global AD pathology
ZC3H12B	X	64,708,614	0.24	1.89×10^{-6}	0.138 (0.029)	tangles
KCTD12	13	77,454,311	0.09	3.44×10^{-8}	0.143 (0.026)	β-amyloid

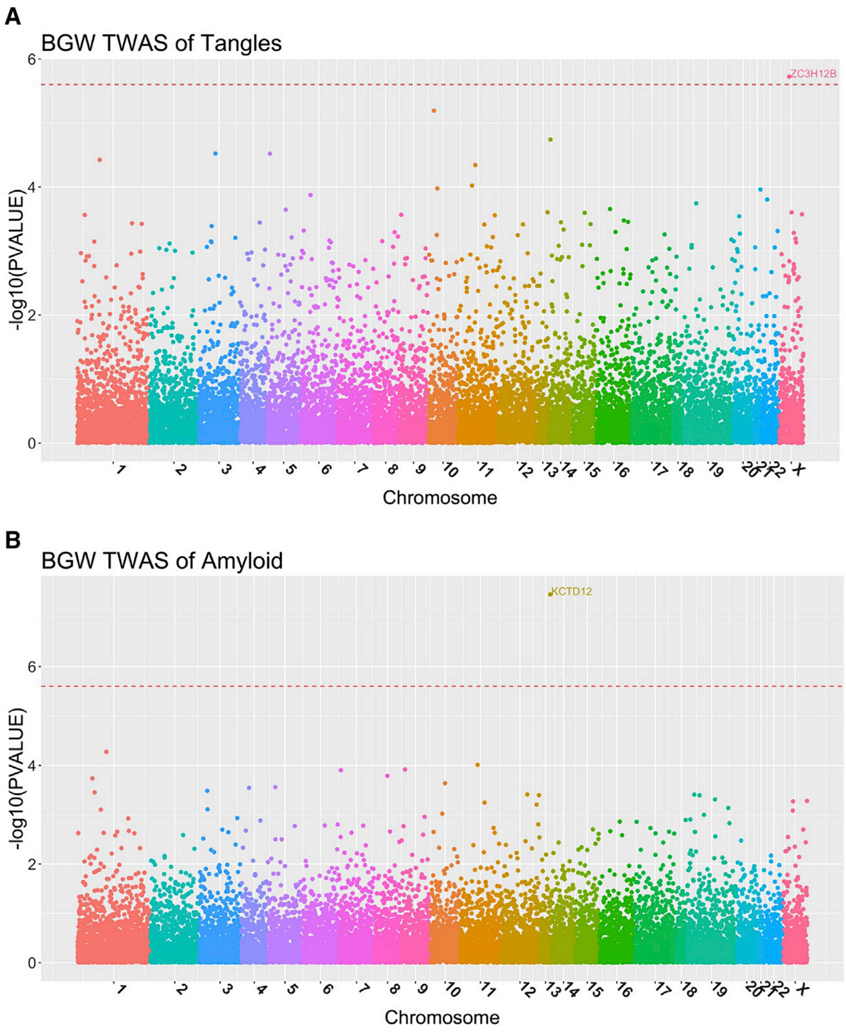


Figure 3. Manhattan Plots of BGW-TWAS Results of Neurofibrillary Tangle Density and β-Amyloid Load

Here, -log10(p values) by BGW-TWAS were plotted and Red lines denote genome-wide significant threshold (2.5×10^{-6}) for gene-based association studies. ZC3H12B was found to be significantly associated with neurofibrillary tangle density (A). KCTD12 was found to be significantly associated with and **β**-amyloid load (B).

ROS/MAP and MCADGS (Table 3). Moreover, seven of these genes (including *HLA-DRB1* and *APOC1*) were also identified when we only considered *cis*-eQTL estimates by BVSR in the TWAS. *CEACAM19* near the well-known GWAS risk gene *APOE* was also identified by S-PrediXcan and TIGAR. Known GWAS risk gene *HLA-DRB1*[52] was also identified by S-PrediXcan (Table 3; Tables S1–S3).

Our results showed that by using BVSR estimates of *cis*- and *trans*-eQTL (BGW-TWAS), most independent risk loci were identified including loci driven by *trans*-eQTL. For those significant genes driven mainly by *cis*-eQTL, a TWAS using BVSR estimates of *cis*-eQTL still identified more independent significant risk loci (distributed over chromosomes 2, 3, 6, 11, and 19) than S-PrediXcan and TIGAR, including all 4 significant genes (*HLA-DRB1, SLC39A13, PVR, CEACAM19*) identified by S-PrediXcan and 4 out of 21 significant genes (*ZNF227, ZFP112, PVR, CEACAM19*) identified by TIGAR (Tables S1–S3). Although TIGAR identified the most significant TWAS genes (21), these genes are from chromosomes 11 and 19, which are likely to be driven by the same *cis*-eQTL from two independent loci.

These TWAS results using summary-level GWAS data with a much larger sample size validated our findings obtained with BGW-TWAS using individual-level GWAS data from ROS/MAP and MCADGS.

Insights about eQTL Genetic Architecture

In addition to imputing Bayesian GReX values (Equation 5), the posterior probabilities of having non-zero eQTL effect sizes estimated by BVSR also provide insights into the genetic architecture of eQTL, especially about how potential eQTL are distributed across the genome. Note that the posterior probability obtained from the BVSR model (Equations 2, 3, and 4) is essentially the expected probability for a SNP to be an eQTL. Therefore, the sum of posterior probabilities of having non-zero eQTL effect sizes represents the expected number of eQTL.

From our simulation studies, we observed that the expected proportions of *cis*-eQTL were consistent with the true proportions of causal *cis*-eQTL. The expected number of eQTL obtained across simulation scenarios is presented in Table 4, where two out of five (40%) causal eQTL and 11 out of 22 (50%) causal eQTL are from *cis*-genome regions. We can see that, with higher true expression heritability, the expected number of eQTL is closer to the true number of causal eQTL. We can also see that the expected number of eQTL is more accurate for the scenario with 5 true causal eQTL than with 22 true causal eQTL, which is due to the fact that the BVSR model prefers relatively larger eQTL effect sizes. These simulation results demonstrated the validity of our BGW-TWAS method based on the BVSR model as well as the usefulness of the sum of posterior probabilities of having non-zero eQTL effect sizes.

For 14,156 genes with fitted GReX prediction models by BVSR using the ROS/MAP data, after excluding 19 outlier genes with >100 expected eQTL, we obtained the average number of expected eQTL as 2.44 (SD = 5.70) across genome-wide regions, 0.25 (SD = 1.24) for *cis*-eQTL, and 2.48 (SD = 5.49) for *trans*-eQTL. That is, on overage, 88% of eQTL were from *trans*-genome regions with respect to

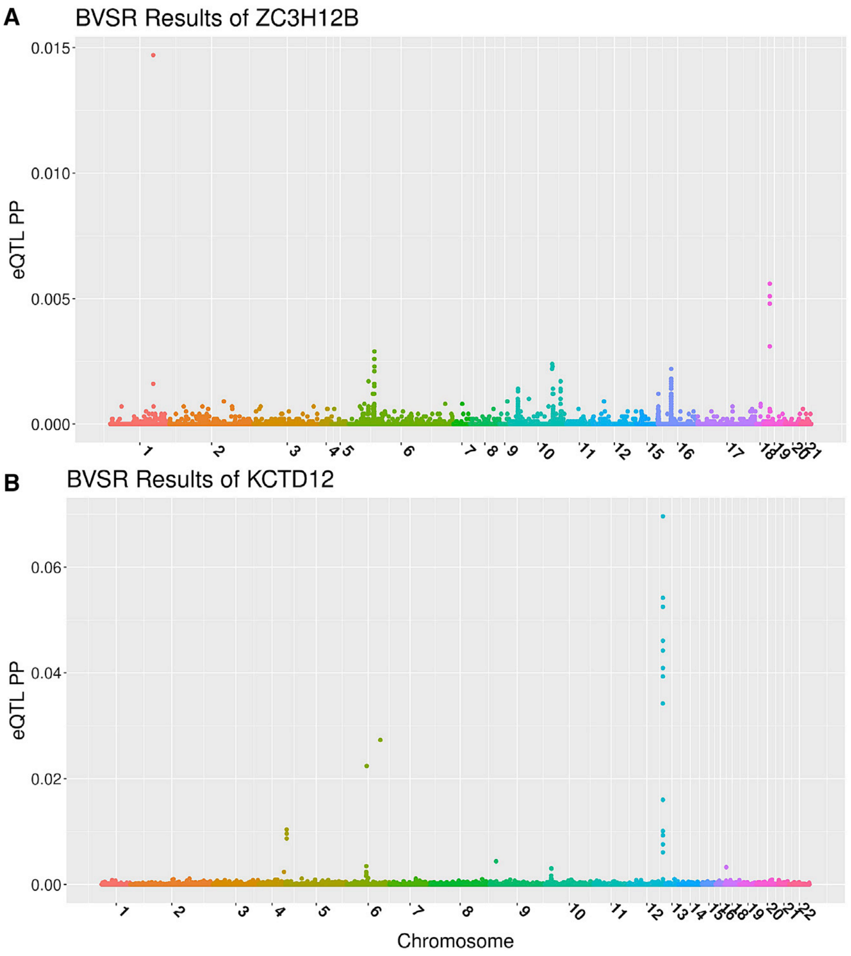


Figure 4. BVSR Posterior Probabilities (PP) of Having Non-zero eQTL Effect Sizes for Analyzed *cis*- and *trans*-SNPs, with Target Genes ZC3H12B and KCTD12

(A) ZC3H12B located on chromosome X has top *trans*-eQTL from chromosomes 1, 6, and 19, where all eQTL are of *trans*-eQTL.

(B) KCTD12 located on chromosome 12 has top *cis*-eQTL from chromosome 12 and *trans*-eQTL from chromosomes 4 and 6.

the target gene. We can see that ~90% genes with train R^2 > 0.05 have ~2–3 average expected eQTL, and ~10% genes with train $R^2 < 0.05$ have >5 average expected eQTL (Table 4). By linking these findings with our simulation studies where train R^2 is likely to be >0.05 when true expression heritability is >0.1, we can conclude that ~90% genes are likely to have true expression heritability >0.1.

Additionally, from our Bayesian estimates of the *cis*- and *trans*-specific posterior probabilities of having non-zero eQTL effect sizes (i.e., π_{cis}, π_{trans} in Equation 3) for genome-wide genes using ROS/MAP data (Figure S11), we can see that π_{cis} and π_{trans} clearly follow different distributions. This also validates our assumptions of respective prior distribution for *cis*- and *trans*-hyper parameters.

Discussion

In this paper, we proposed and validated a Bayesian genome-wide TWAS (BGW-TWAS) method based on the BVSR[24] model to leverage the information of both *cis*- and *trans*-eQTL. We derived an efficient computational approach to fit the BVSR model with large-scale genomic data, by pruning genome regions that contain either at least one *cis*-SNP or one potential *trans*-eQTL and adapting the previously developed scalable EM-MCMC algorithm[25] with pre-calculated LD coefficients and summary statistics from standard eQTL analyses. BGW-TWAS extends previous TWAS methods[11–13] that utilize only partial genotype information from a small window of *cis*- SNPs to train the GReX imputation model.

Genotype data of *trans*-eQTL have been shown to explain a significant amount of variation of expression quantitative traits and provide important molecular mechanisms underlying known GWAS loci of complex diseases.[22,23] The results from our simulation and application studies demonstrated that BGW-TWAS improves the yield of a TWAS by levering both *cis*- and *trans*-eQTL information. For example, higher precision of GReX prediction and power of TWASs were obtained in our simulation studies when true causal *trans*-eQTL existed. These results showed that BGW-TWAS has a greater advantage for scenarios where eQTL have relatively large effect sizes for the expression quantitative traits (e.g., 5 versus 22 true

Table 2. *trans*-SNPs with Top Five Posterior Probability (PP) > 0.003 of Having Non-zero eQTL Effect Sizes for *ZC3H12B*

CHR	POS	rsID	Function	MAF	PP	w	p Value
1	159,135,282	*rs3026946*	intergenic	0.213	0.0147	−0.071	6.25×10^{-7}
19	45,422,160	*rs12721051*	3′ UTR (*APOC1*)	0.161	0.0031	0.071	3.94×10^{-6}
19	45,422,846	*rs56131196*	downstream (*APOC1*)	0.173	0.0048	0.069	1.75×10^{-6}
19	45,422,946	*rs4420638*	downstream (*APOC1*)	0.173	0.0051	0.068	1.77×10^{-6}
19	45,424,514	*rs157592*	regulatory region (*APOC1*)	0.181	0.0056	0.075	1.43×10^{-6}

Table 3. Significant Genes Identified by BGW-TWAS using IGAP GWAS Summary Statistics of AD

			TWAS p Value			
Gene	**CHR**	**Position**	**BGW-TWAS**	**BVSR *cis*-eQTL**	**PrediXcan**	**TIGAR**
GPX1[a]	3	49,394,608	2.45×10^{-98}	2.45×10^{-98}	–	3.15×10^{-1}
FAM86DP	3	75,484,261	1.55×10^{-13}	4.81×10^{-1}	5.38×10^{-1}	9.63×10^{-1}
BTN3A2[a]	6	26,378,546	1.59×10^{-26}	1.56×10^{-26}	3.17×10^{-1}	5.04×10^{-1}
ZNF192[a]	6	28,124,089	1.26×10^{-32}	1.25×10^{-32}	8.56×10^{-2}	2.07×10^{-1}
AL022393.7[a]	6	28,144,452	3.25×10^{-178}	2.24×10^{-178}	1.50×10^{-1}	8.36×10^{-2}
HLA-DRB1[a,b]	6	32,557,625	1.02×10^{-12}	8.99×10^{-13}	2.06×10^{-6}	–
AEBP1	7	44,154,161	5.55×10^{-220}	8.62×10^{-1}	6.69×10^{-1}	4.19×10^{-1}
BUB3	10	124,924,886	6.64×10^{-18}	1.05×10^{-2}	–	4.76×10^{-1}
FBXO3	11	33,796,089	1.48×10^{-9}	6.88×10^{-1}	–	1.13×10^{-1}
CEACAM19[a,b,c]	19	45,187,631	4.7×10^{-13}	2.54×10^{-13}	3.60×10^{-12}	2.83×10^{-16}
APOC1[a]	19	45,422,606	8.9×10^{-11}	1.11×10^{-10}	3.18×10^{-6}	7.2×10^{-3}
ZC3H12B	X	64,727,767	2.08×10^{-37}	–	–	–
CXorf56	X	118,699,397	6.02×10^{-07}	–	–	–

TWAS p values by alternative methods, i.e., using BVSR *cis*-eQTL estimates only, PrediXcan, and TIGAR are also provided. p values for genes that were missed by TWAS were indicated as "–." *ZC3H12B* that was identified by BGW-TWAS using individual-level GWAS data from ROS/MAP and MCADGS was also identified by BGW-TWAS using IGAP summary-level GWAS statistics. *CEACAM19* from chromosome 19 was identified by all TWAS methods, and *HLA-DRB1* from chromosome 6 is a known GWAS risk locus.

[a] Genes that were also identified as significant by using BVSR *cis*-eQTL estimates.

[b] Genes that were also identified by PrediXcan.

[c] Genes that were also identified by TIGAR.

causal eQTL with the same expression heritability). This is because variable selection by the BVSR model is designed to select sparse signals with relatively large effect sizes as shown in previous GWASs.[24,25]

By applying our BGW-TWAS method to several human AD datasets, we identified a risk gene (*ZC3H12B*) with GReX values that were significantly associated with both clinical diagnosis of AD and postmortem AD pathology indices (neurofibrillary tangle density and global measure of AD pathology). This association was not identified by existing TWAS methods because this gene is shown to be completely driven by *trans*-eQTL. Importantly, a potential biological mechanism was revealed by showing that the top driven *trans*-eQTL of *ZC3H12B* are known GWAS signals of AD[43] and blood lipids[44–46] and <12 KB away from the well-known AD risk gene (*APOE*).[47] Thus, we expect BGW-TWAS leveraging both *cis*- and *trans*-eQTL has potential for making a large impact on advancing our understanding of complex human diseases and traits.

By fitting BVSR models using both *cis*- and *trans*-eQTL, we not only can account for the uncertainty for a SNP to be an eQTL to predict GReX (Equation 5), but also can use the sum of posterior probabilities of having non-zero eQTL effect sizes to estimate the expected number of eQTL.[24,25] The distribution of expected eQTL can also help characterize the underlying genetic architecture of expression quantitative traits.

The current study has several limitations. First, while BGW-TWAS reduces the computational burden for modeling both *cis*- and *trans*-eQTL, its computing costs are still substantial to train GReX prediction models for genome-wide genes (~20K) per tissue type. It requires approximately 30 min of computation time and 3 GB memory per gene (with parallel computation implemented in 4 CPU cores). Parallel computation can be employed to make use of high-performance computation clusters with multiple cores to reduce computation time. Second, our current method is designed to use pre-calculated in-sample LD coefficients and summary statistics from single variant eQTL analyses; further work is required to expand this approach to use approximate LD coefficients generated from reference samples of the same ethnicity. Third, our simulation studies showed that the non-parametric Bayesian method TIGAR performed best when all causal eQTL are *cis*- with relatively small effect sizes (e.g., 22 true causal *cis*-eQTL). Our TWAS results of AD using the IGAP summary statistics demonstrated that TIGAR and BGW-TWAS yield complementary findings. These results highlight the potential utility of leveraging both methods especially for studies in which the true distributions of *cis*- and *trans*-eQTL of the test genes are generally unknown.

In conclusion, the BGW-TWAS method presented herein provides a framework for leveraging information from both *cis*- and *trans*-eQTL to conduct gene-based association studies. Because *trans*-QTL are common for other quantitative omics traits, e.g., epigenetic, proteomic, and metabonomic, our proposed computational procedure would be to investigate other quantitative omics traits in gene-based association studies. Integrative method developments will

Table 4. Average Sums of Posterior Probabilities of Having Non-zero eQTL Effect Sizes that Are Stratified Based on Gene Expression Heritability (Either True Simulated Heritability in Simulation Studies or the Range of Train R^2 of the Fitted BVSR Models with ROS/MAP Data

	Sum of Posterior Probabilities		
Gene Expression Heritability	**Whole Genome**	***cis*-Region**	***trans*-Region**
5 True Causal eQTL			
0.05	0.79	0.46	0.33
0.1	2.28	1.13	1.15
0.2	3.72	1.44	2.28
0.5	4.91	1.56	3.35
22 True Causal eQTL			
0.05	0.05	0.02	0.03
0.1	0.21	0.11	0.10
0.2	1.43	0.87	0.56
0.5	6.46	3.89	2.57
ROS/MAP			
(0, 0.05) 1,504 genes	6.63	0.60	6.23
(0.05, 0.1) 1,964 genes	1.45	0.13	1.32
(0.1, 0.25) 6,617 genes	2.00	0.17	1.83
(0.25, 0.5) 3,224 genes	2.66	0.22	2.44
(0.5, 1) 474 genes	3.04	0.31	2.73

The simulation scenarios presented here are those with 2 of 5 and 11 of 22 true causal eQTL from the *cis*-regions.

stand to benefit from our BGW-TWAS method, especially the perspectives of leveraging information from *trans*-QTL and efficient computation techniques derived from this paper. In addition, BGW-TWAS can be applied to study other complex human phenotypes to identify potential risk genes that could be targeted in further drug discovery.

Data and Code Availability

ROS/MAP data can be requested through Rush Alzheimer's Disease Center and Synapse. MCADGS data can be requested through Synapse. IGAP summary statistics are available online. Summary statistics generated from our BGW-TWAS methods for studying AD are publicly available through Synapse. Source code of BGW-TWAS is available through Github. See Web Resources for URLs.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.08.022.

Acknowledgment

J.Y. was supported by the startup funding from Department of Human Genetics at Emory University School of Medicine. 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, U01AG32984, U01AG46152, and 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 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, and R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation.

Declaration of Interests

The authors declare no competing interests.

Received: March 5, 2020
Accepted: August 24, 2020
Published: September 21, 2020

Web Resources

BGW-TWAS, https://github.com/yanglab-emory/BGW-TWAS

BGW-TWAS summary statistics on Synapse.org, https://www.synapse.org/#!Synapse:syn22316791

IGAP, http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php

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

PrediXcan, https://github.com/hakyim/PrediXcan

ROS/MAP data on Synapse.org, https://www.synapse.org/#!Synapse:syn3219045

Rush Alzheimer's Disease Center, https://www.radc.rush.edu/

TIGAR, https://github.com/yanglab-emory/TIGAR

References

1. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.
2. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
3. 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.
4. Nikpay, M., Goel, A., Won, H.H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P., Hopewell, J.C., et al. (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat. Genet. *47*, 1121–1130.
5. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.
6. Huang, Q. (2015). Genetic study of complex diseases in the post-GWAS era. J. Genet. Genomics *42*, 87–98.
7. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The Post-GWAS Era: From Association to Function. Am. J. Hum. Genet. *102*, 717–730.
8. Battle, A., Brown, C.D., Engelhardt, B.E., Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; and eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.
9. 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.
10. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.
11. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
12. 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. Nat. Genet. *48*, 245–252.
13. 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., and Yang, J. (2019). TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. Am. J. Hum. Genet. *105*, 258–266.
14. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., et al.; GTEx Consortium (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. *9*, 1825.
15. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
16. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. *6*, e1000952.
17. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. Am. J. Hum. Genet. *100*, 473–487.
18. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B.M., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet. *50*, 538–548.
19. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.O., Lu, Y., Cai, Q., et al.; NBCS Collaborators; and kConFab/AOCS Investigators (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat. Genet. *50*, 968–978.
20. Raj, T., Li, Y.I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.C., Ng, B., Gupta, I., Haroutunian, V., et al. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. Nat. Genet. *50*, 1584–1592.
21. 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.
22. Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., et al. (2017). The Genetic Architecture of Gene Expression in Peripheral Blood. Am. J. Hum. Genet. *100*, 371.
23. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. biorXiv. https://doi.org/10.1101/447367.

24. Guan, Y.T., and Stephens, M. (2011). Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems. Ann. Appl. Stat. *5*, 1780–1815.
25. Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G.; and International Age-Related Macular Degeneration Genomics Consortium (2017). A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. Am. J. Hum. Genet. *101*, 404–416.
26. De Jager, P.L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B.N., Felsky, D., Klein, H.U., White, C.C., Peters, M.A., Lodgson, B., et al. (2018). A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci. Data *5*, 180142.
27. 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.; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; and Cohorts for Heart and Aging Research in Genomic Epidemiology (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. *45*, 1452–1458.
28. Casella, G. (2001). Empirical Bayes Gibbs sampling. Biostatistics *2*, 485–500.
29. 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.
30. 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.
31. 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* (s1), S161–S189.
32. 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.
33. 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.; Alzheimer's Disease Genetics Consortium (2012). Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. PLoS Genet. *8*, e1002707.
34. Allen, M., Carrasquillo, M.M., Funk, C., Heavner, B.D., Zou, F., Younkin, C.S., Burgess, J.D., Chai, H.S., Crook, J., Eddy, J.A., et al. (2016). Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci. Data *3*, 160089.
35. 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.; Alzheimer's Disease Neuroimaging Initiative (2012). A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol. Aging *33*, 1017.e1–1017.e15.
36. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
37. Das, S., Forer, L., Schönherr, 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.
38. 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.
39. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.
40. Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Hayden Gephart, M.G., Barres, B.A., and Quake, S.R. (2015). A survey of human brain transcriptome diversity at the single cell level. Proc. Natl. Acad. Sci. USA *112*, 7285–7290.
41. Wingo, T.S., Yang, J., Fan, W., Min Canon, S., Gerasimov, E.S., Lori, A., Logsdon, B., Yao, B., Seyfried, N.T., Lah, J.J., et al. (2020). Brain microRNAs associated with late-life depressive symptoms are also associated with cognitive trajectory and dementia. NPJ Genom. Med. *5*, 6.
42. Hartung, J., Knapp, G., and Sinha, B.K. (2011). Statistical meta-analysis with applications (John Wiley & Sons).
43. 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.
44. Ligthart, S., Vaez, A., Hsu, Y.H., Stolk, R., Uitterlinden, A.G., Hofman, A., Alizadeh, B.Z., Franco, O.H., Dehghan, A.; Inflammation Working Group of the CHARGE Consortium; PMI-WG-XCP; and LifeLines Cohort Study (2016). Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation. BMC Genomics *17*, 443.
45. Ligthart, S., Vaez, A., Võsa, U., Stathopoulou, M.G., de Vries, P.S., Prins, B.P., Van der Most, P.J., Tanaka, T., Naderi, E., Rose, L.M., et al.; LifeLines Cohort Study; and CHARGE Inflammation Working Group (2018). Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. Am. J. Hum. Genet. *103*, 691–706.
46. Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., et al.; Global Lipids Genetics Consortium; Myocardial Infarction Genetics (MIGen) Consortium; Geisinger-Regeneron DiscovEHR Collaboration; and VA Million Veteran Program (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet. *50*, 1514–1523.
47. Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., Bullido, M.J., Engelborghs, S., De Deyn, P., Berr, C., et al. (2011). APOE and Alzheimer disease: a major gene with semi-dominant inheritance. Mol. Psychiatry *16*, 903–907.
48. Kamboh, M.I., Demirci, F.Y., Wang, X., Minster, R.L., Carrasquillo, M.M., Pankratz, V.S., Younkin, S.G., Saykin, A.J., Jun, G., Baldwin, C., et al.; Alzheimer's Disease Neuroimaging Initiative (2012). Genome-wide association study of Alzheimer's disease. Transl. Psychiatry *2*, e117.
49. Zhou, X., Chen, Y., Mok, K.Y., Kwok, T.C.Y., Mok, V.C.T., Guo, Q., Ip, F.C., Chen, Y., Mullapudi, N., Giusti-Rodríguez, P., et al.;

Alzheimer's Disease Neuroimaging Initiative (2019). Non-coding variability at the APOE locus contributes to the Alzheimer's risk. Nat. Commun. *10*, 3310.

50. Liang, J., Wang, J., Azfer, A., Song, W., Tromp, G., Kolattukudy, P.E., and Fu, M. (2008). A novel CCCH-zinc finger protein family regulates proinflammatory activation of macrophages. J. Biol. Chem. *283*, 6337–6346.

51. Fagerberg, L., Hallström, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol. Cell. Proteomics *13*, 397–406.

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

The American Journal of Human Genetics, Volume *107*

Supplemental Data

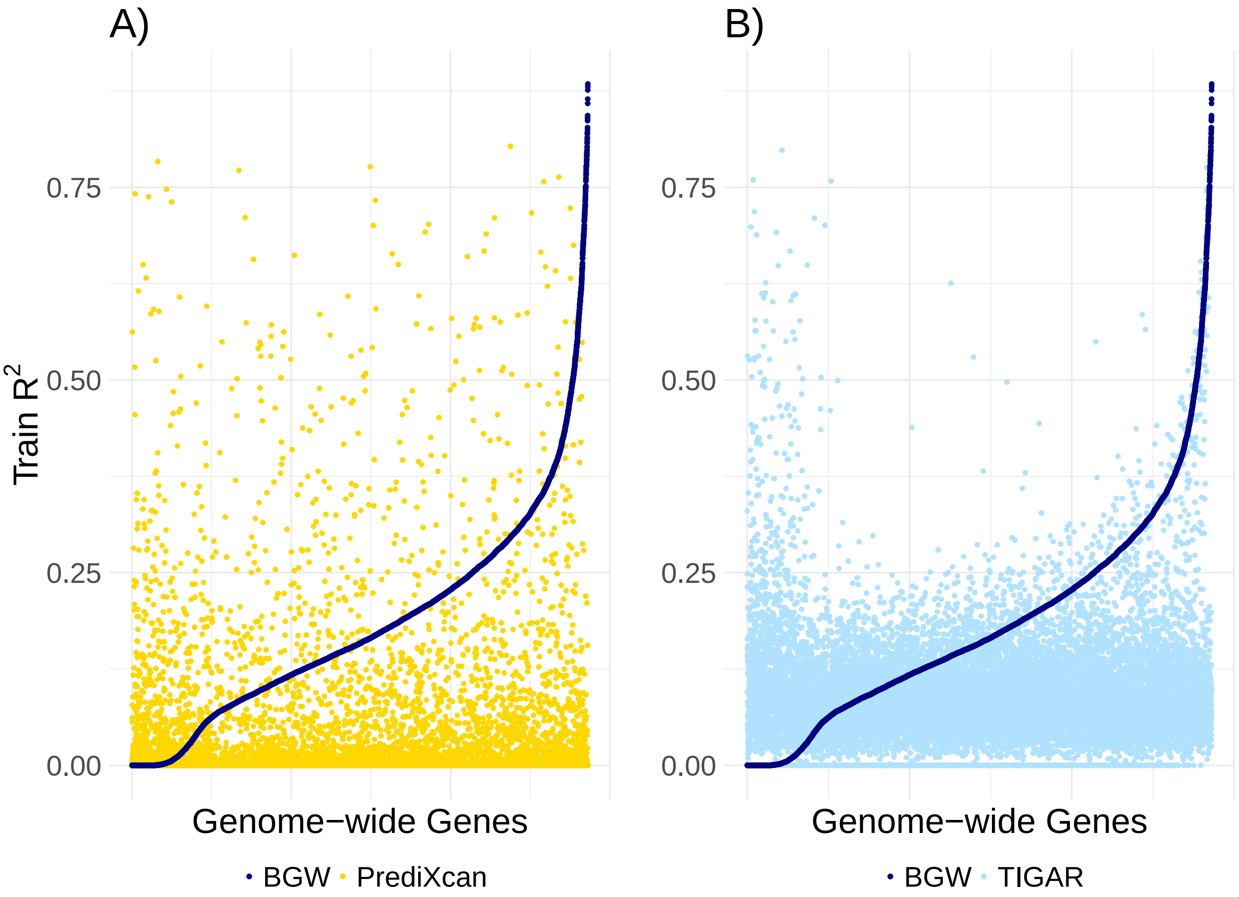
Bayesian Genome-wide TWAS Method to Leverage both *cis*- and *trans*-eQTL Information through Summary Statistics

Justin M. Luningham, Junyu Chen, Shizhen Tang, Philip L. De Jager, David A. Bennett, Aron S. Buchman, and Jingjing Yang

1 Supplemental Figures

1.1 Training R^2 comparison

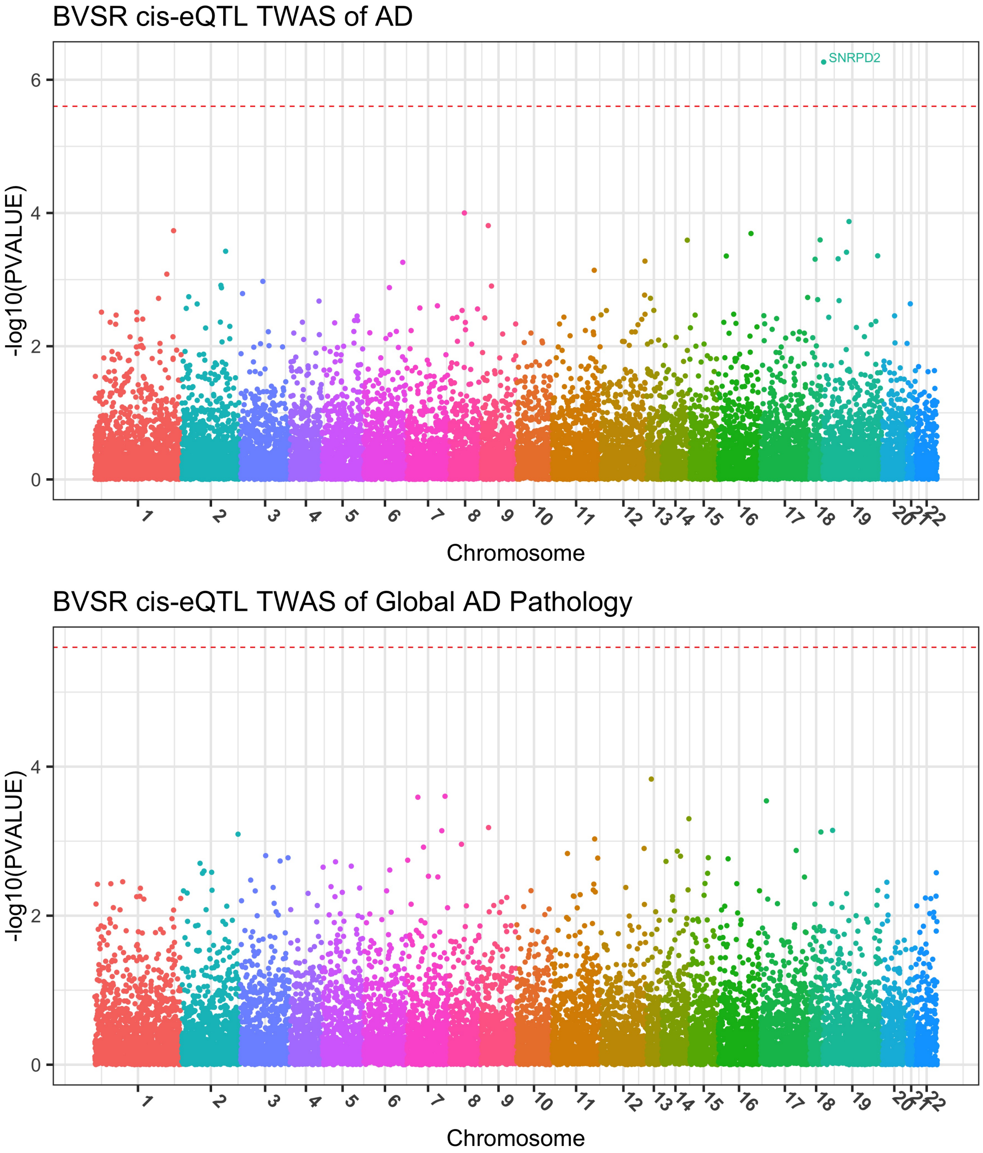
Figure S 1: Compare train R^2 obtained by our BGW method, PrediXcan, and TIGAR.



Train R^2 value per gene obtained by our BGW method (dark blue), PrediXcan (yellow), and TIGAR (baby blue) for genome-wide genes are plotted, where genes were ranked in the increasing order of R^2 by BGW.

1.2 TWAS results using individual-level GWAS data of ROS/MAP and MCADGS

Figure S 2: Manhattan plots of TWAS results of AD clinical diagnosis and global AD pathology by using BVSR cis-eQTL estimates only.



BVSR cis-eQTL TWAS of AD

BVSR cis-eQTL TWAS of Global AD Pathology

Figure S 3: Manhattan plots of TWAS results of neurofibrillary tangle density (tangles) and β-amyloid load (amyloid) by using BVSR cis-eQTL estimates only.

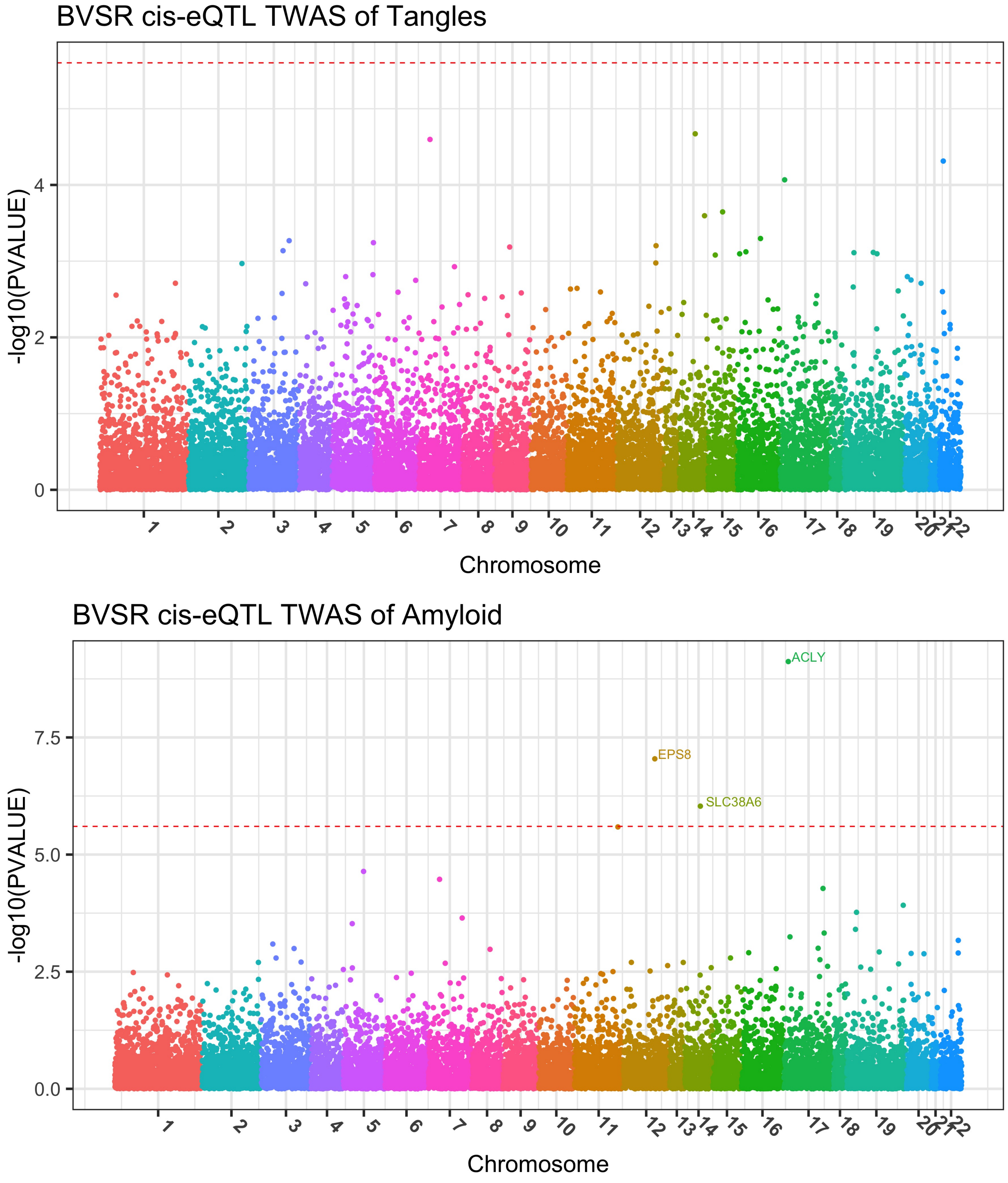
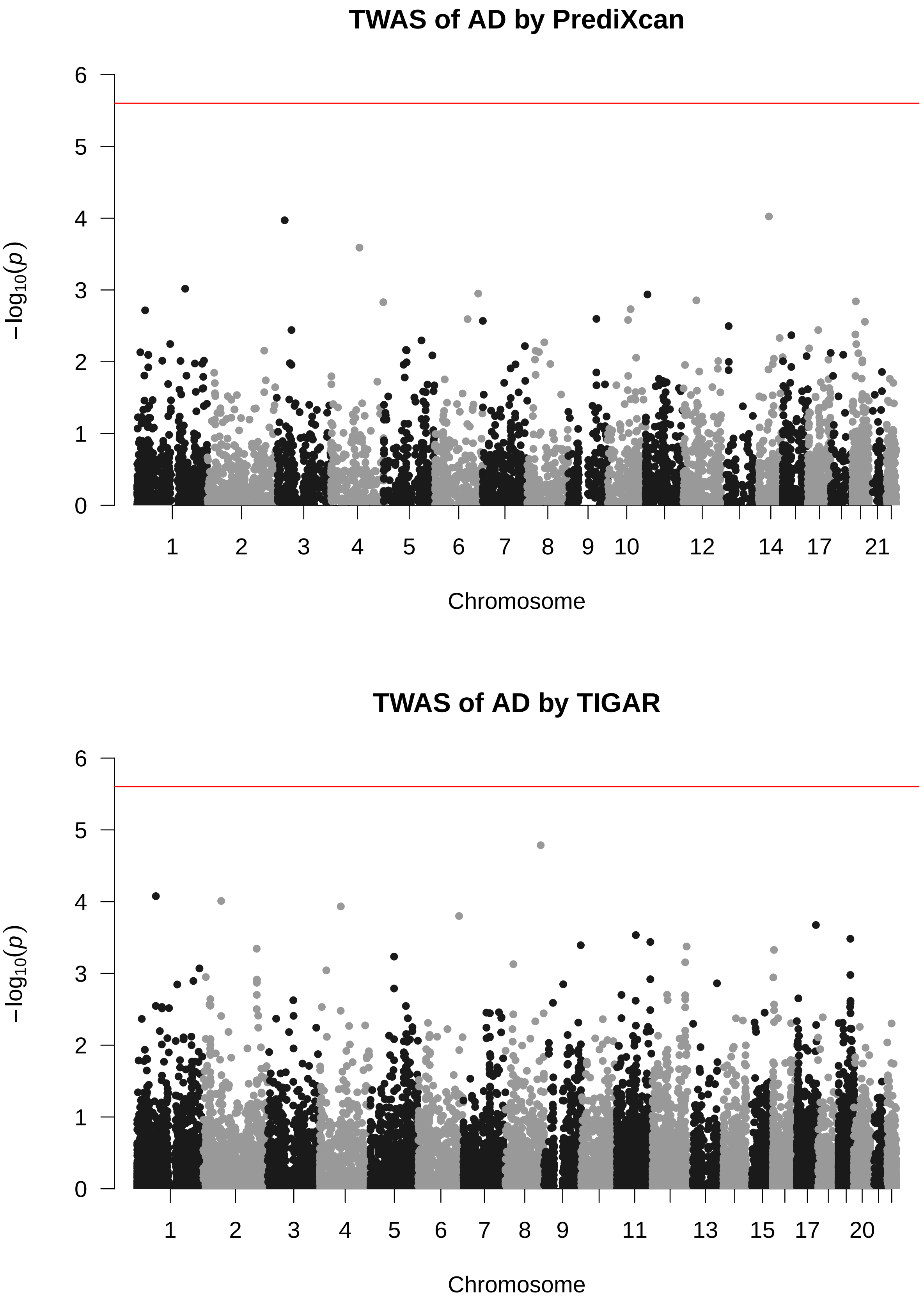


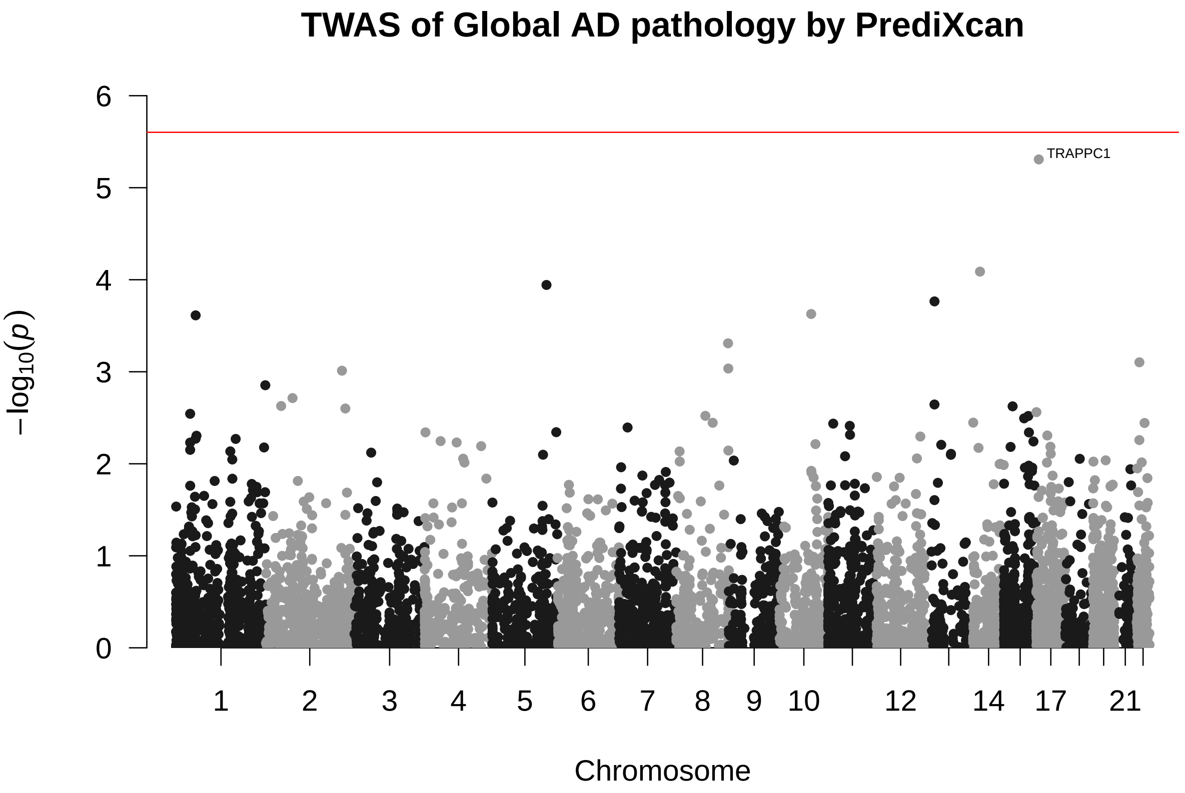
Figure S 4: Manhattan plots of TWAS results of AD clinical diagnosis by PrediXcan and TIGAR.



TWAS of AD by PrediXcan

TWAS of AD by TIGAR

Figure S 5: Manhattan plots of TWAS results of global AD pathology by PrediXcan and TIGAR.



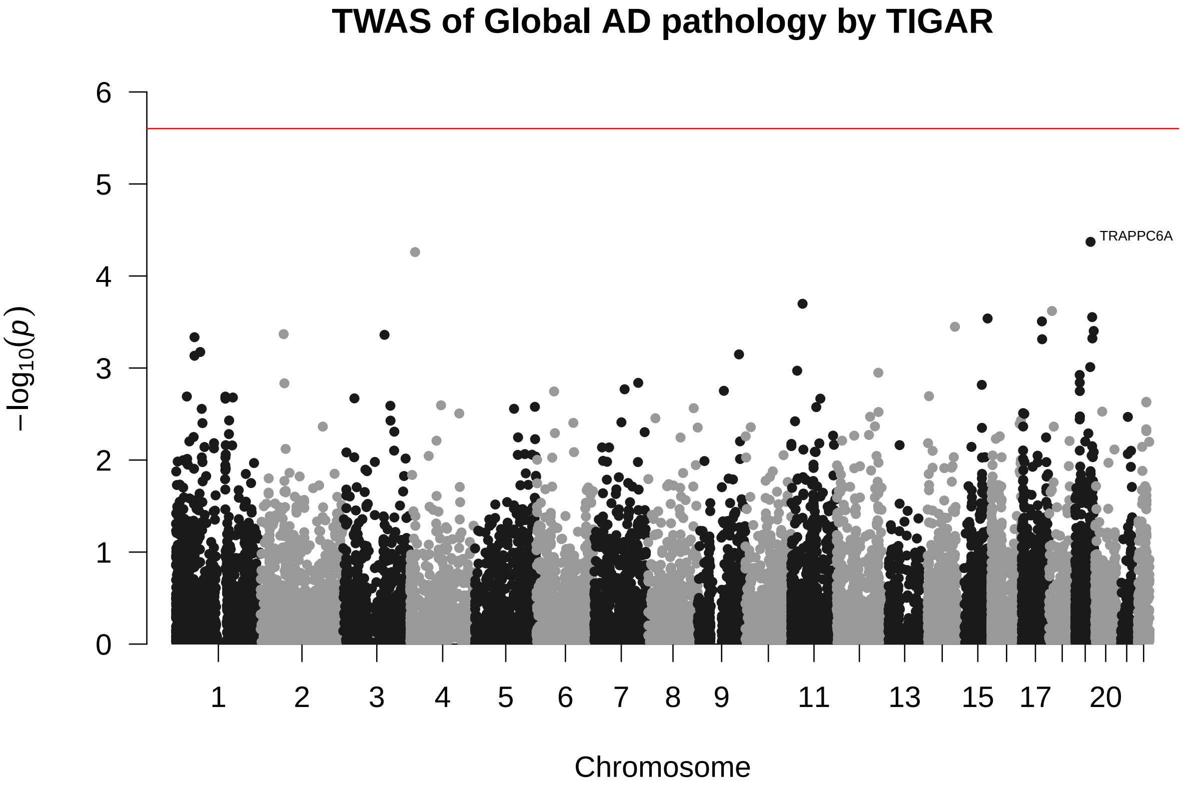
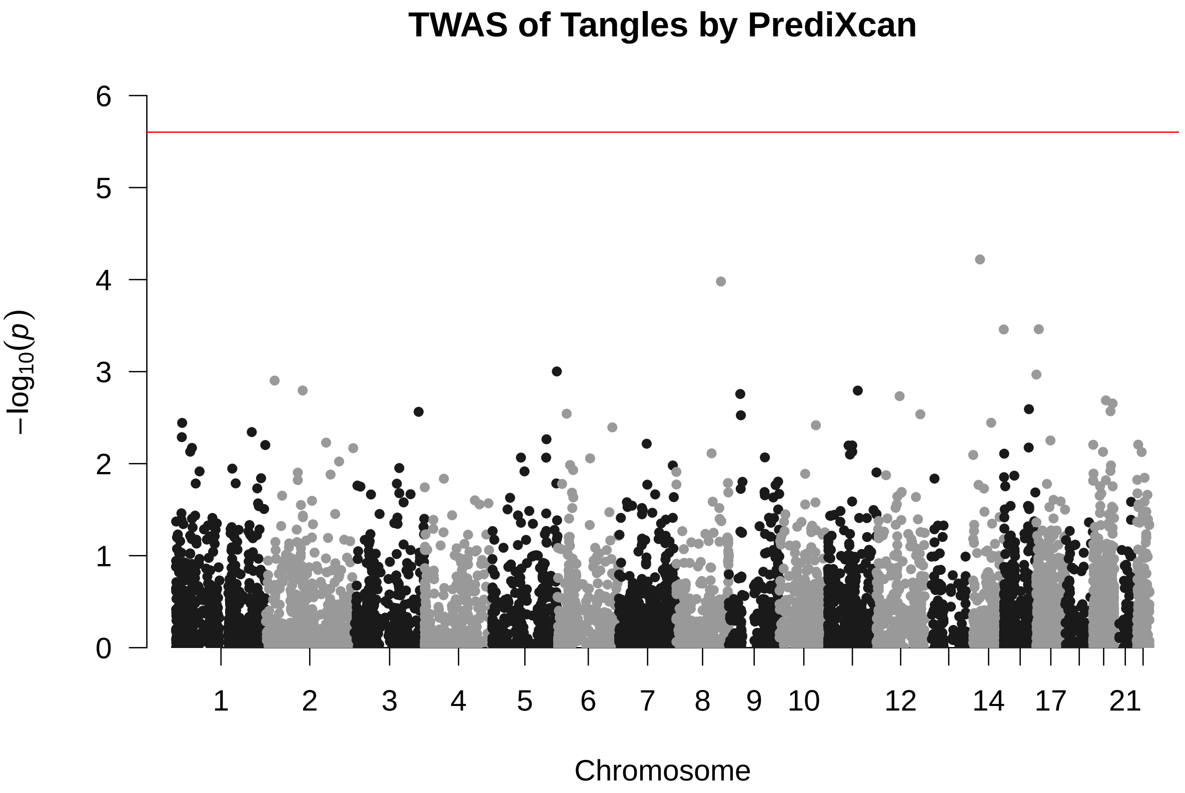


Figure S 6: Manhattan plots of TWAS results of neurofibrillary tangle density (tangles) by PrediXcan and TIGAR.



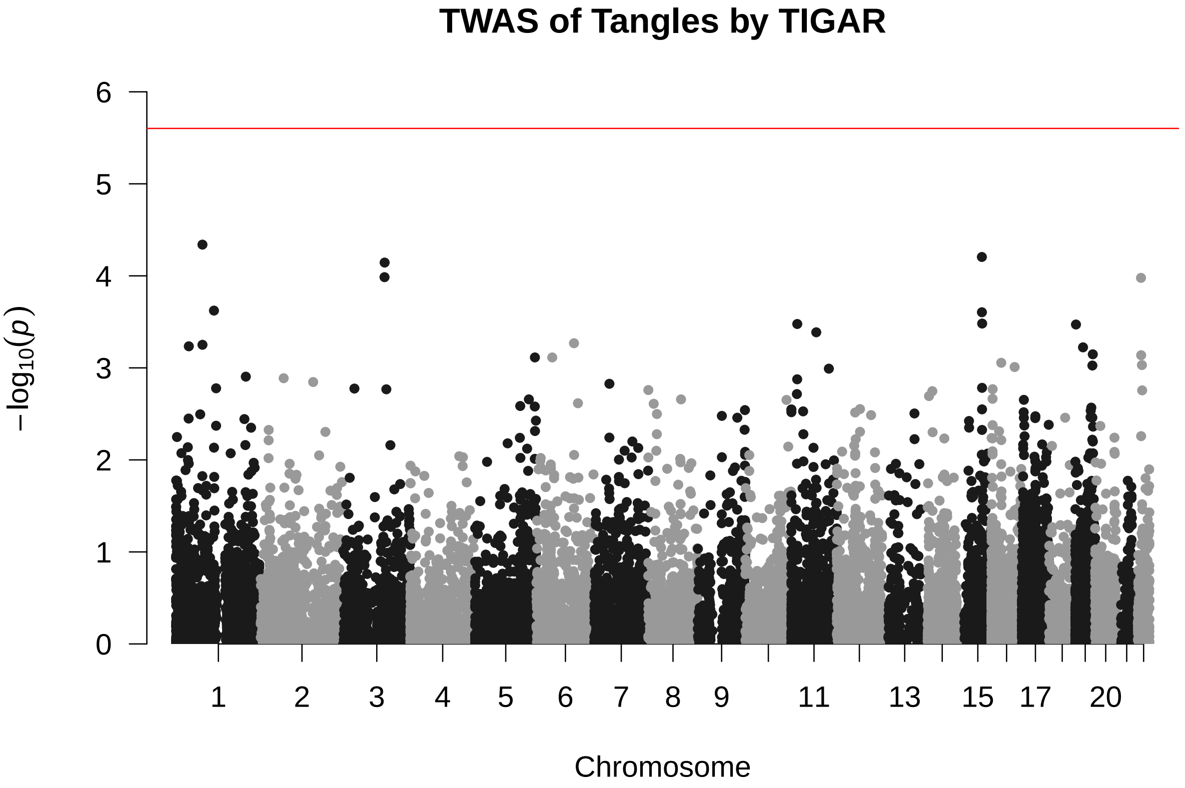
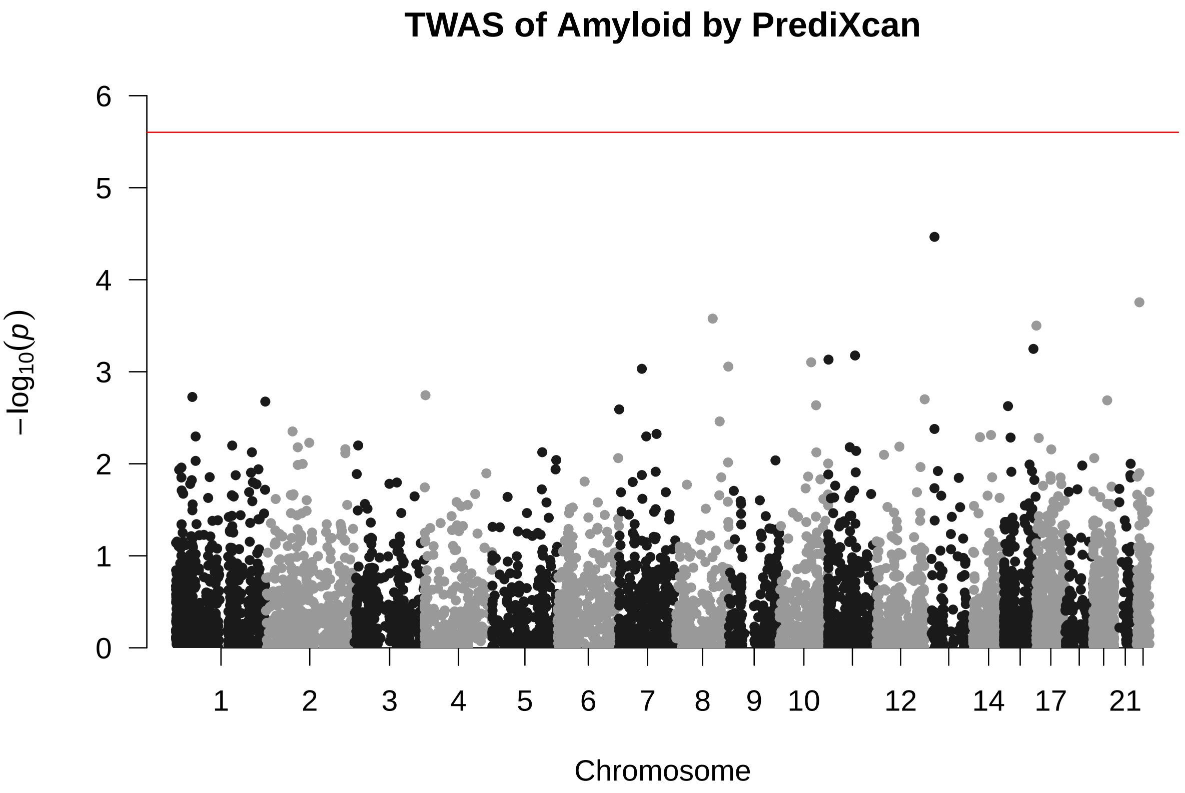


Figure S 7: Manhattan plots of TWAS results of β-amyloid load (amyloid) by PrediXcan and TIGAR.



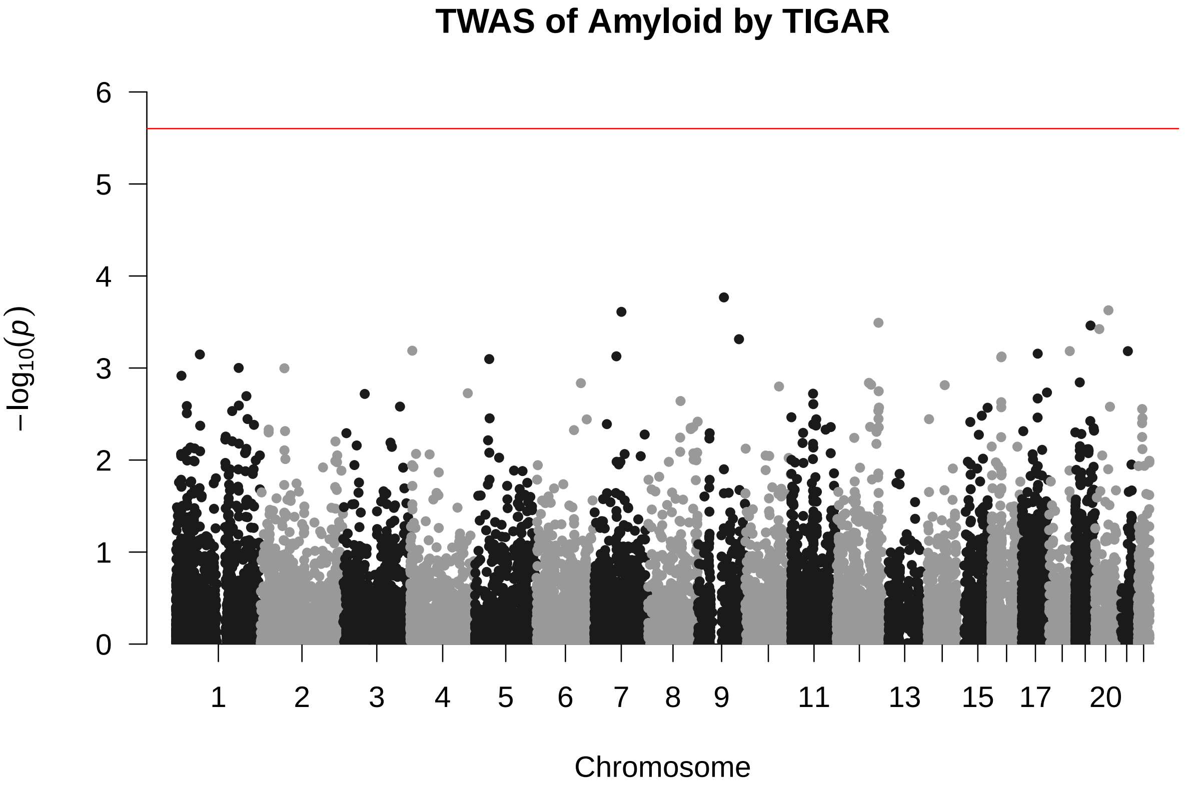
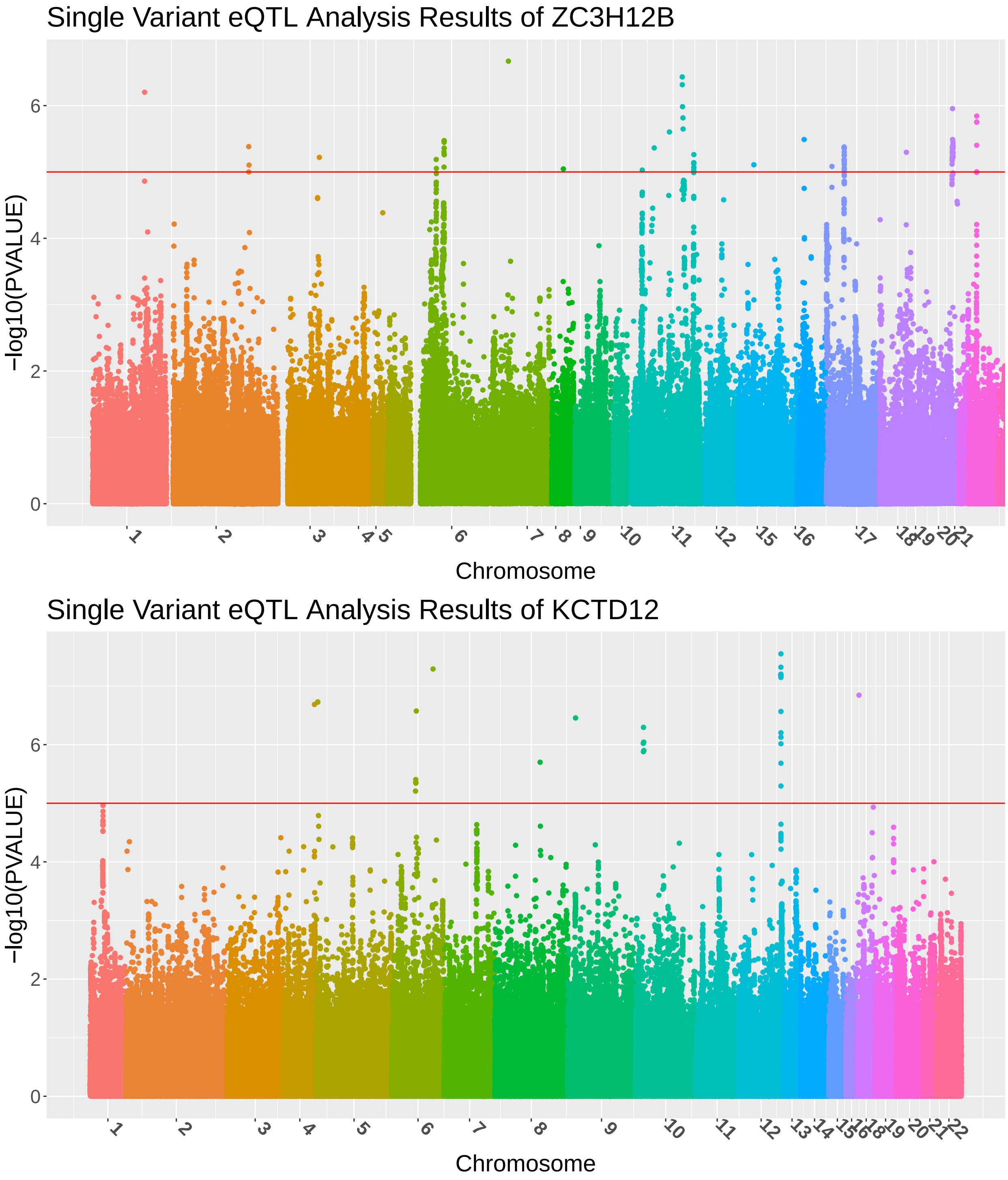


Figure S 8: Manhattan plots of standard eQTL analysis results of genes *ZC3H12B* and *KCTD12* by single variant tests.



Single Variant eQTL Analysis Results of ZC3H12B

Single Variant eQTL Analysis Results of KCTD12

1.3 TWAS results using IGAP summary-level GWAS data of AD

Figure S 9: Manhattan plots of TWAS results using IGAP summary statistics of AD by using both cis-eQTL and trans-eQTL (BGW-TWAS) and only cis-eQTL estimates obtained by BVSR.

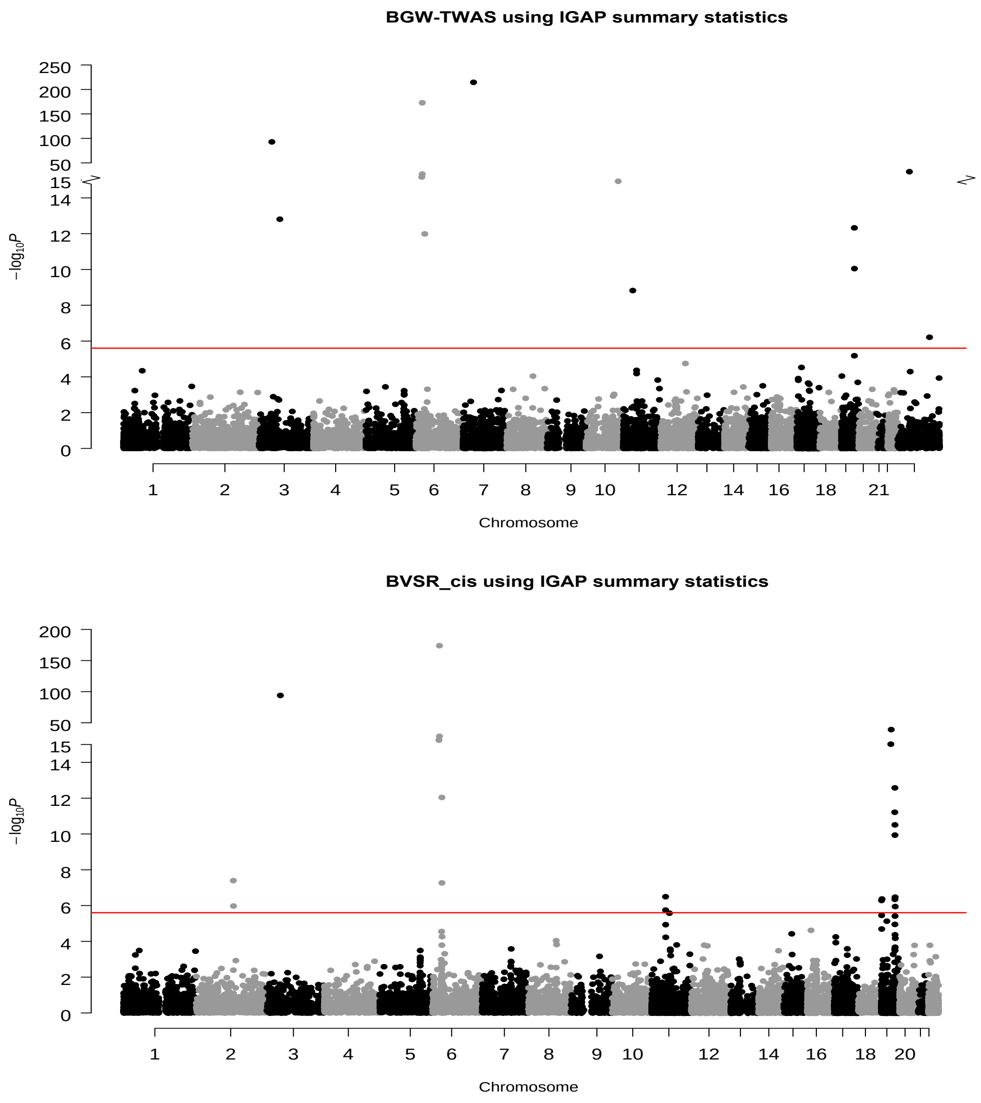
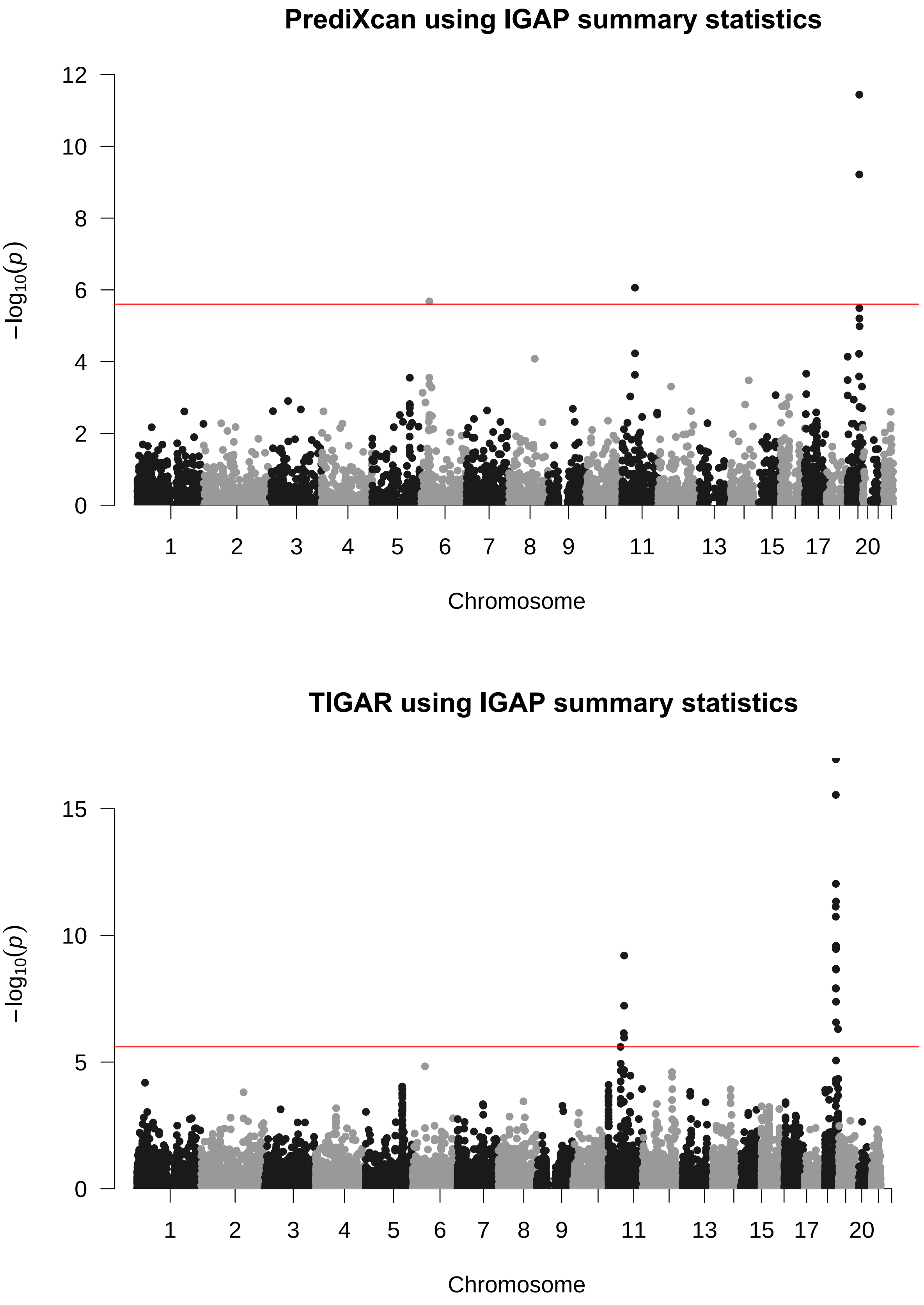


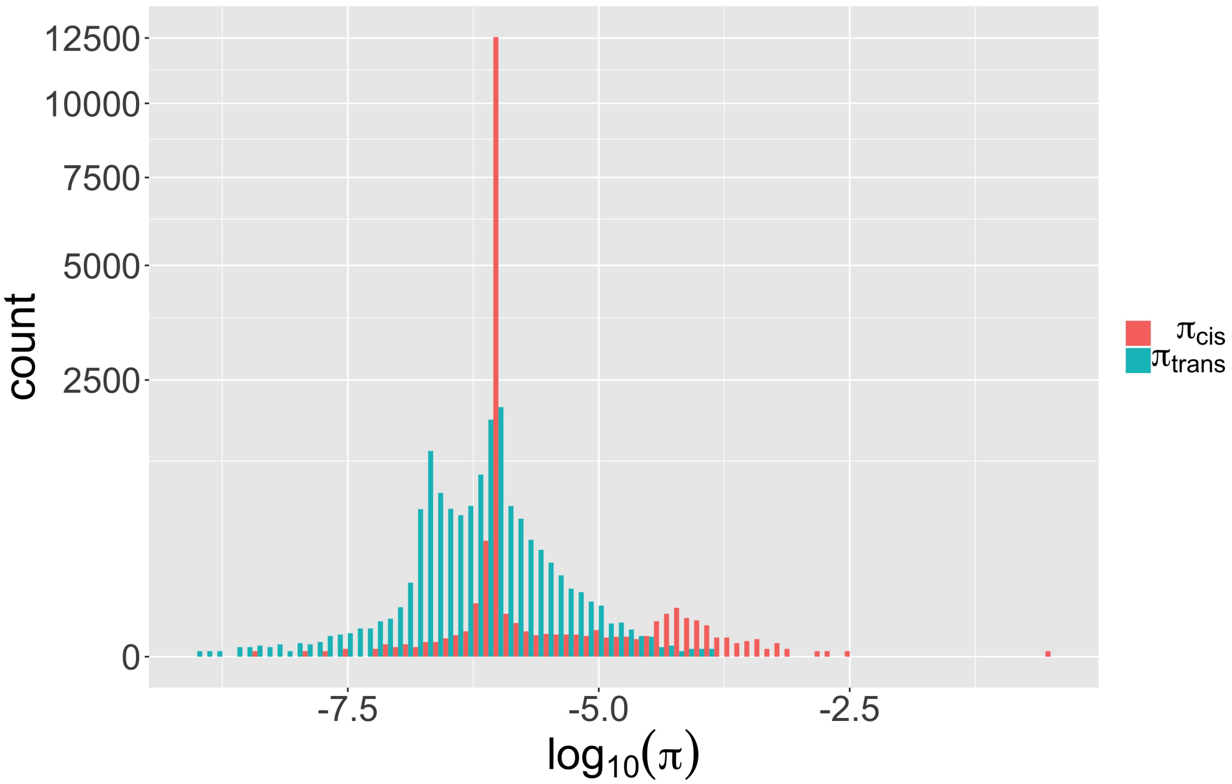
Figure S 10: Manhattan plots of TWAS results using IGAP summary statistics of AD by S-PrediXcan and TIGAR .



PrediXcan using IGAP summary statistics

TIGAR using IGAP summary statistics

Figure S 11: Histogram of log_{10} Bayesian estimates for cis- and trans- specific probabilities of being an eQTL (π_{cis}, π_{trans}), over genome-wide genes.



2 Supplemental Tables

Gene	CHR	Position	P-VALUE	Z-score
WDR33	2	128,458,595	4.01e-08	-5.49
SAP130	2	128,698,790	1.05e-06	-4.88
GPX1[a]	3	493,96,033	2.45e-98	21.04
BTN3A2[a]	6	26,365,386	1.56e-26	10.66
ZNF192[a]	6	28,109,715	1.25e-32	-11.89
AL022393.7[a]	6	28,143,965	2.24e-178	-28.47
HLA-DRB1[a,b]	6	32,546,545	8.99e-13	7.14
HLA-DQB1	6	32,627,243	5.33e-08	-5.43
RP11-750H9.5	11	4,740,4698	3.08e-07	-5.11
SLC39A13[b]	11	47,428,682	1.73e-06	-4.78
FSTL3	19	676,388	4.98e-07	-5.02
MKNK2	19	2,037,480	4.12e-07	-5.06
ZNF227[c]	19	44,716,690	5.87e-12	6.88
ZFP112[c]	19	44,830,705	5.79e-20	9.14
PVR[a,b,c]	19	45,147,097	4.3e-07	-5.05
CEACAM19[a,b,c]	19	45,174,723	2.54e-13	7.31
TOMM40	19	45,394,476	2.97e-11	-6.64
APOC1[a]	19	45,417,920	1.11e-10	6.45
FOSB	19	45,971,252	3.26e-07	-5.10
SNRPD2	19	46,190,712	1.97e-43	-13.81
FBXO46	19	46,213,886	1.09e-06	-4.87

Table S 1: Significant genes identified by TWAS using BVSR cis-eQTL estimates only.

a. Genes that were also identified by BGW-TWAS.

b. Genes that were also identified by PrediXcan.

c. Genes that were also identified by TIGAR.

Gene	CHR	Position	P-VALUE	Z-score
HLA-DRB1[a]	6	32546545	2.06e-06	4.75
SLC39A13[a]	11	47428682	8.57e-07	-4.92
PVR[a,b]	19	45147097	6.02e-10	-6.19
CEACAM19[a,b]	19	45174723	3.6e-12	6.95

Table S 2: Significant genes identified by S-PrediXcan.

a. Genes that were also identified by BVSR estimates of cis-eQTL and trans-eQTL (BGW-TWAS) or cis-eQTL only.

b. Genes that were also identified by TIGAR.

Gene	**CHR**	**Position**	**P-VALUE**	**Z-score**
STX3	11	59,480,928	7.28e-07	-4.95
PRPF19	11	60,658,201	5.98e-08	5.42
TMEM109	11	60,681,345	1.09e-06	-4.88
TMEM132A	11	60,691,934	6.26e-10	-6.18
ZNF227[a]	19	44,716,690	1.21e-08	5.7
ZFP112[a]	19	44,830,705	7.32e-12	6.85
PVR[a,b]	19	45,147,097	1.82e-11	-6.72
CEACAM19[a,b]	19	45,174,723	2.83e-16	8.18
CLPTM1	19	45,457,847	9.17e-13	-7.14
CLASRP	19	45,542,297	2.08e-09	5.99
ZNF296	19	45,574,758	3.45e-10	-6.28
TRAPPC6A	19	45,666,186	1.11e-17	-8.56
MARK4	19	45,754,549	3.41e-10	6.28
RTN2	19	45,988,549	2.7e-07	5.14
PPM1N	19	45,992,034	1.24e-08	5.69
OPA3	19	46,030,684	2.55e-10	-6.32
EML2	19	46,112,659	2.22e-09	5.98
GIPR	19	46,171,501	4.58e-12	6.92
SNRPD2	19	46,190,712	2.69e-10	6.32
DMWD	19	46,286,204	4.15e-08	-5.48
TTYH1	19	54,926,372	4.98e-07	-5.03

Table S 3: Significant genes identified by TIGAR.

a. Genes that were also identified by BVSR estimates of cis-eQTL and trans-eQTL (BGW-TWAS) or cis-eQTL only.

b. Genes that were also identified by S-PrediXcan.

3 Supplemental Methods

3.1 Bayesian Variable Selection Regression Model

Consider the following Bayesian variable selection regression (BVSR) model [1] for quantitative gene expression traits:

$$\mathcal{E}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{w}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1},\ w_i \sim \pi N(0, \sigma_w^2\sigma_\epsilon^2) + (1-\pi)\delta_0(\cdot),\ \epsilon_i \sim N(0, \sigma_\epsilon^2), \tag{1}$$

where $\mathcal{E}_{n\times 1}$ denotes the vector of centered quantitative expression levels for n samples; $\boldsymbol{X}_{n\times p}$ denotes the centered genotype matrix of p genetic variants; ϵ_i denotes the residual error independently and identically distributed (i.i.d.) with normal distribution $N(0, \sigma_\epsilon^2)$; and the broad sense "eQTL" effect size w_i follows a spike-and-slab prior distribution [1, 2, 3] — that is, w_i follows the normal distribution $N(0, \sigma_w^2\sigma_\epsilon^2)$ with probability π and the point-mass density function $\delta_0(\cdot)$ at 0 with probability $(1-\pi)$. Basically, $\delta_0(w_i) = 1$ if $w_i = 0$, otherwise $\delta_0(w_i) = 0$. The effect size variance $(\sigma_w^2\sigma_\epsilon^2)$ is assumed to be scaled by the error term variance (σ_ϵ^2) for the purpose of computational convenience.

As is typical of SNP-based genome-wide analyses, the genotype matrix $\boldsymbol{X}_{n\times p}$ contains either dosage data within range $[0, 2]$ or genotype data with values $\{0, 1, 2\}$ denoting the expected or genotyped number of minor alleles. The assumption of the spike-and-slab prior for w_i enforces variable selection in the regression model (1). Assuming both $\mathcal{E}_{n\times 1}$ and columns of $\boldsymbol{X}_{n\times p}$ are centered, the intercept term is omitted from the regression model.

3.1.1 Model cis- and trans- eQTL

In this paper, we employ this BVSR model (1) to account for both cis- and trans- eQTL genotype data (cis- and trans- are defined based on SNP proximity to the target gene) for modeling quantitative gene expression traits. Particularly, we extend the BVSR model to allow for respective prior distributions for the effect sizes of cis- and trans- SNPs (i.e., eQTL) as follows:

$$\begin{aligned}
\mathcal{E}_g &= \boldsymbol{X}_{cis}\boldsymbol{w}_{cis} + \boldsymbol{X}_{trans}\boldsymbol{w}_{trans} + \boldsymbol{\epsilon} \\
w_{cis,i} &\sim \pi_{cis}N(0, \sigma_{cis}^2\sigma_\epsilon^2) + (1-\pi_{cis})\delta_0(w_{cis,i}) \\
w_{trans,i} &\sim \pi_{trans}N(0, \sigma_{trans}^2\sigma_\epsilon^2) + (1-\pi_{trans})\delta_0(w_{trans,i}), \\
\epsilon_i &\sim N(0, \sigma_\epsilon^2).
\end{aligned} \tag{2}$$

Generally, SNPs within ±1MB of the flanking 5' and 3' ends of the target gene are considered as cis-SNPs, and other SNPs on the genome are considered as trans-SNPs.

Note that cis- and trans- can be viewed as two non-overlapped annotations for SNPs in the BVSR model (1), which makes model (2) a special case of the previously developed Bayesian Functional GWAS (BFGWAS) method [4]. Similarly, the following independent and conjugate hyper priors are assumed for hyper parameters in model (2):

$$\begin{aligned} \pi_{cis} \sim Beta(a_{cis}, b_{cis}),\ \sigma^2_{cis} \sim IG(k_1, k_2), \\ \pi_{trans} \sim Beta(a_{trans}, b_{trans}),\ \sigma^2_{trans} \sim IG(k_3, k_4), \\ \sigma^2_{\epsilon} \sim IG(k_5, k_6) \end{aligned} \tag{3}$$

where $Beta(a_q, b_q)$ denotes a Beta distribution with positive shape parameters a_q and b_q for $q = \{cis, trans\}$, and an Inverse-Gamma distribution $IG(k_s, k_l)$ with shape parameters k_s and scale parameters k_l is assumed for $(\sigma^2_{cis}, \sigma^2_{trans}, \sigma^2_{\epsilon})$ with respective shape and scale parameters.

Hyper prior values are chosen for a_q and b_q to enforce a sparse model, such that the mean of the Beta distribution $\frac{a_q}{a_q+b_q} = 10^{-6}$ with $(a_q + b_q)$ equal to the total number of variants of respective annotation $q = \{cis, trans\}$. The hyper priors for Inverse Gamma are taken as $k_1 = k_2 = k_3 = k_4 = k_5 = k_6 = 0.1$ to induce non-informative priors for $(\sigma^2_{cis}, \sigma^2_{trans}, \sigma^2_{\epsilon})$. Thus, the posterior estimates of $(\pi_{cis}, \sigma^2_{cis}, \pi_{trans}, \sigma^2_{trans})$ will mainly be driven by data likelihood.

3.1.2 Latent Indicator Variable

To facilitate computation, a latent indicator vector $\boldsymbol{\gamma}_{p\times 1}$ is introduced [2] into the model (2), where each element $\gamma_i \in \{0, 1\}$ indicates whether the corresponding ith effect $w_{q,i}$ equals to 0 with $\gamma_i = 0$ or follows the $N(0, \sigma^2_q \sigma^2_{\epsilon})$ distribution with $\gamma_i = 1$. Equivalently,

$$\gamma_i \sim Bernoulli(\pi_i),\ \boldsymbol{w}_{-\boldsymbol{\gamma}} \sim \delta_0(\cdot),\ \boldsymbol{w}_{\boldsymbol{\gamma}} \sim MVN_{|\boldsymbol{\gamma}|}(0, \sigma^2_{\epsilon} \boldsymbol{V}_{\boldsymbol{\gamma}}), \tag{4}$$

where $|\boldsymbol{\gamma}|$ denotes the number of non-zero entries in $\boldsymbol{\gamma}$; $\boldsymbol{w}_{-\boldsymbol{\gamma}}$ denotes the sub-vector of $\boldsymbol{w}_{p\times 1}$ corresponding to variants with $\gamma_i = 0$; $\boldsymbol{w}_{\boldsymbol{\gamma}}$ denotes the sub-vector of $\boldsymbol{w}_{p\times 1}$ corresponding to the variants with $\{\gamma_j = 1; j = 1, \cdots, |\boldsymbol{\gamma}|\}$ that follows a multivariate normal distribution (MVN) with mean 0 and covariance $\sigma^2_{\epsilon} \boldsymbol{V}_{\boldsymbol{\gamma}}$; and $\boldsymbol{V}_{\gamma}$ is the corresponding sub-covariance-matrix of SNPs with $\gamma_i = 1$, $\boldsymbol{V}_{p\times p} = diag(\sigma^2_{q,1}, \cdots, \sigma^2_{q,p})$, where $\sigma^2_{q,i} = \sigma^2_{cis}$ if the ith SNP is of cis- annotation and $\sigma^2_{q,i} = \sigma^2_{trans}$ if the ith SNP is of trans- annotation. The expectation of the latent indicator variable $(E[\gamma_i])$ is the posterior probability (PP_i) for the ith SNP to be an eQTL with effect size w_i.

3.1.3 Bayesian Inference

From the above assumed BVSR model (2, 3, 4), the posterior joint distribution of $(\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\pi}, \tau)$ is proportional to the product of likelihood and prior density functions,

$$\begin{aligned} P(\boldsymbol{w}, \gamma, \boldsymbol{\sigma^2}, \boldsymbol{\pi}, \sigma_\epsilon^2 | \mathcal{E}_g, \boldsymbol{X}, \boldsymbol{A}) \propto & P(\mathcal{E}_g | \boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\gamma}, \sigma_\epsilon^2) P(\boldsymbol{w} | \boldsymbol{A}, \boldsymbol{\pi}, \boldsymbol{\sigma^2}, \sigma_\epsilon^2) P(\gamma | \boldsymbol{\pi}) \\ & P(\boldsymbol{\pi}) P(\boldsymbol{\sigma^2}) P(\sigma_\epsilon^2), \end{aligned} \tag{5}$$

where $\boldsymbol{\pi} = (\pi_{cis}, \pi_{trans})$, $\boldsymbol{\sigma^2} = (\sigma_{cis}^2, \sigma_{trans}^2)$, and $\boldsymbol{A}$ is a $p \times 2$ matrix with binary values denoting if the analyzed SNPs are categorized as cis- or trans-.

The main challenge for modeling trans-SNPs in addition of cis-SNPs in the regression model of expression quantitative traits (2, 3, 4) is the computation burden of making Bayesian inference for $(\boldsymbol{w}, E[\gamma] = \boldsymbol{PP})$ with respect to genome-wide genes ($\sim$20K). In order to make the Bayesian inference feasible in practice, we utilize the scalable Expectation-Maximization Markov chain Monte Carlo (EM-MCMC) algorithm proposed for BFGWAS [4]. Specifically, we first segment considered genotype data into approximately independent blocks based on the block-wise linkage disequilibrium (LD) structure of the human genome, i.e., $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_K\}$. Then we can write the likelihood function (5) as a product of likelihood functions for $\boldsymbol{X}_k$,

$$P(\mathcal{E}_g | \boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\gamma}, \sigma_\epsilon^2) = \prod_{k=1}^{K} P_k(\mathcal{E}_g | \boldsymbol{X}_k, \boldsymbol{w}_k, \boldsymbol{\gamma}_k, \sigma_\epsilon^2), \tag{6}$$

where $(\mathcal{E}_g | \boldsymbol{X}_k, \boldsymbol{w}_k, \boldsymbol{\gamma}_k, \tau) \sim MVN_{|\boldsymbol{\gamma_k}|}(\boldsymbol{X}_k \boldsymbol{w}_k, \sigma_\epsilon^2 \boldsymbol{I}_{|\boldsymbol{\gamma_k}|})$.

For the convenience of implementing MCMC algorithm per genome block in parallel with given shared hyper parameters $\boldsymbol{\pi} = (\pi_{cis}, \pi_{trans})$, $\boldsymbol{\sigma^2} = (\sigma_{cis}^2, \sigma_{trans}^2)$, σ_ϵ^2 is fixed across all genome blocks as the variance of $\mathcal{E}_g$. As shown by previous studies [4], this assumption only results in slightly conserved estimates for $(\boldsymbol{w}, \boldsymbol{PP})$ but saves the hassle of estimating a block-specific error variance. By EM-MCMC algorithm, we estimate $(\boldsymbol{w}_k, \boldsymbol{PP}_k)$ by implementing MCMC algorithm [5, 6] per block (i.e., Expectation step (E-step)) with given values of $(\boldsymbol{\pi}, \boldsymbol{\sigma^2})$; and then update the estimates of $(\boldsymbol{\pi}, \boldsymbol{\sigma^2})$ by maximizing the corresponding expected posterior likelihood function [7] (Maximization step (M-step)) given the estimates of $(\boldsymbol{w}, \boldsymbol{PP})$ from the previous E-step. A few such EM iterations will be run until the estimates of $(\boldsymbol{\pi}, \boldsymbol{\sigma^2})$ converge, which generally requires $\sim$5 EM iterations [4]. The estimates $(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{PP}})$ from the last E-step will be used to calculate Bayesian genetically regulated gene expression (GReX) levels for additional samples with GWAS genotype data $\widetilde{\boldsymbol{X}}$,

$$\widehat{GReX} = \sum_{i=1}^{p} \widetilde{X}_i (\widehat{PP}_i \widehat{w}_i). \tag{7}$$

Detailed derivations about the conditional posterior distributions of $\boldsymbol{w}_k$ and $\boldsymbol{\gamma_k}$ as well as the EM-MCMC algorithm are referred to the supplementary note of the BFGWAS paper [4]. To further reduce the computation time for fitting such tissue-gene-specific Bayesian model for genome-wide genes, we propose a novel computational technique to implement the MCMC algorithm using only the pre-calculated LD coefficients and summary statistics from standard eQTL analyses based on single variant tests. Below, we briefly outline steps of the EM-MCMC algorithm using only summary statistics.

3.1.4 Fast MCMC using Only Summary Statistics

As shown in the BFGWAS supplementary note [4], the posterior distribution for $(\boldsymbol{w}_k, \boldsymbol{\gamma}_k)$ of block k is

$$P(\boldsymbol{w}_k, \boldsymbol{\gamma}_k | \boldsymbol{X}_k, \boldsymbol{\mathcal{E}}_g, \boldsymbol{\pi}, \boldsymbol{\sigma^2}, \epsilon_\epsilon^2) \propto P(\boldsymbol{\mathcal{E}}_g | \boldsymbol{X}_k, \boldsymbol{w}_k, \boldsymbol{\gamma}_k, \epsilon_\epsilon^2) P(\boldsymbol{w}_k | \boldsymbol{\gamma}_k, \boldsymbol{\sigma^2}, \epsilon_\epsilon^2) P(\boldsymbol{\gamma}_k | \boldsymbol{\pi}). \tag{8}$$

The conditional posterior distribution of $\boldsymbol{w}_{|\boldsymbol{\gamma_k}|}$ is given by

$$\begin{aligned} &P(\boldsymbol{w}_{|\boldsymbol{\gamma_k}|} | \boldsymbol{X}_{|\boldsymbol{\gamma_k}|}, \mathcal{E}_g, \boldsymbol{\gamma}_k, \boldsymbol{\sigma^2}, \sigma_\epsilon^2) \sim \\ &\qquad MVN_{|\gamma_k|} \left((\boldsymbol{X}'_{|\boldsymbol{\gamma_k}|} \boldsymbol{X}_{|\boldsymbol{\gamma_k}|} + \boldsymbol{V}^{-1}_{|\boldsymbol{\gamma_k}|})^{-1} (\boldsymbol{X}'_{|\boldsymbol{\gamma_k}|} \mathcal{E}_g),\ \sigma_\epsilon^2 (\boldsymbol{X}'_{|\boldsymbol{\gamma_k}|} \boldsymbol{X}_{|\boldsymbol{\gamma_k}|} + \boldsymbol{V}^{-1}_{|\boldsymbol{\gamma_k}|})^{-1} \right). \end{aligned} \tag{9}$$

After integrating $\boldsymbol{w}_k$ out from (8), the marginal conditional posterior distribution for $\boldsymbol{\gamma}_k$ is given by

$$\begin{aligned} P(\boldsymbol{\gamma}_k | \boldsymbol{X}_k, \boldsymbol{\mathcal{E}}_g, \boldsymbol{\pi}, \boldsymbol{\sigma^2}, \sigma_\epsilon^2) &\propto \int_{\boldsymbol{w}_k} P_k(\boldsymbol{\mathcal{E}}_g | \boldsymbol{X}_k, \boldsymbol{w}_k, \boldsymbol{\gamma}_k, \sigma_\epsilon^2) P(\boldsymbol{w}_k | \boldsymbol{\gamma}_k, \boldsymbol{\sigma^2}, \sigma_\epsilon^2) P(\boldsymbol{\gamma}_k | \boldsymbol{\pi}) d\boldsymbol{w}_k \\ &\propto |\boldsymbol{\Omega}_{|\boldsymbol{\gamma_k}|}|^{-1/2} \exp\left\{ \frac{1}{2\sigma_\epsilon^2} (\boldsymbol{\mathcal{E}}'_g \boldsymbol{X}_{|\boldsymbol{\gamma_k}|}) \boldsymbol{V}_{|\boldsymbol{\gamma_k}|} \boldsymbol{\Omega}^{-1}_{|\boldsymbol{\gamma_k}|} (\boldsymbol{X}'_{|\boldsymbol{\gamma_k}|} \boldsymbol{\mathcal{E}}_g) \right\} P(\boldsymbol{\gamma}_k | \boldsymbol{\pi}), \end{aligned} \tag{10}$$

where $\boldsymbol{\Omega}_{|\boldsymbol{\gamma_k}|} = \boldsymbol{V}_{|\boldsymbol{\gamma_k}|} (\boldsymbol{X}'_{|\boldsymbol{\gamma_k}|} \boldsymbol{X}_{|\boldsymbol{\gamma_k}|}) + \boldsymbol{I}_{|\boldsymbol{\gamma_k}|}$, σ_ϵ^2 will be taken as the variance of $\boldsymbol{\mathcal{E}}_g$, and the subscript $|\gamma_k|$ indicates sub-matrices or sub-vectors corresponding to variants with nonzero indicator variables. That is, $\boldsymbol{V}_{|\boldsymbol{\gamma_k}|}$ is a diagonal matrix with $(\boldsymbol{V}_{|\boldsymbol{\gamma_k}|})_{jj} = \sigma^2_{cis}$ or σ^2_{trans} given the jth SNP is cis- or trans-.

As discussed in the BFGWAS paper [4], majority computation time is spent on implementing the MCMC algorithm per genome block (E-step), because $> 10{,}000$ MCMC iterations are required for obtaining converged estimates for $(\boldsymbol{w_k}, \boldsymbol{\gamma}_k)$. Particularly, most computation resource is spent on evaluating the posterior likelihood (9) and (10) per MCMC iteration, where calculating $\boldsymbol{X}'_{|\boldsymbol{\gamma_k}|} \boldsymbol{X}_{|\boldsymbol{\gamma_k}|}$ and $\boldsymbol{X}'_{|\boldsymbol{\gamma_k}|} \mathcal{E}_g$ costs most computation time.

Therefore, by deriving values of $\boldsymbol{X}'_{|\gamma_k|}\boldsymbol{X}_{|\gamma_k|}$ and $\boldsymbol{X}'_{|\gamma_k|}\mathcal{E}_g$ from pre-calculated LD coefficients and summary statistics [8], up to 90% computation time can be saved.

Consider the single variant regression model that is generally used for standard eQTL analyses,

$$\mathcal{E}_g = X_i w_i + \boldsymbol{\epsilon}, \tag{11}$$

where X_i denotes the genotype vector of the ith SNP. The least square estimator $\widetilde{w}_i$ is given by $\widetilde{w}_i = (X_i'X_i)^{-1}X_i'\mathcal{E}_g$. Then the elements of vector $\boldsymbol{X}'_{|\gamma_k|}\mathcal{E}_g$ are given by

$$(\boldsymbol{X}'_{|\gamma_k|}\mathcal{E}_g)_i = \widetilde{w}_i(X_i'X_i), \tag{12}$$

where $X_i'X_i = (n-1)Var(X_i)$ can be either pre-calculated using individual-level genotype data or approximated by $2nf_i(1-f_i)$ with sample size n and minor allele frequency f_i.

Note that $\{X_i'X_i;\ i = 1, ..., p\}$ are also the diagonal values of $\boldsymbol{X}'_{|\gamma_k|}\boldsymbol{X}_{|\gamma_k|}$. For the off-diagonal values, $[\boldsymbol{X}'_{|\gamma_k|}\boldsymbol{X}_{|\gamma_k|}]_{(i,j)}$ can be derived from the LD coefficient between the ith and jth SNPs, $r_{ij} = \frac{X_i'X_j}{\sqrt{(X_i'X_i)(X_j'X_j)}}$. That is,

$$[\boldsymbol{X}'_{|\gamma_k|}\boldsymbol{X}_{|\gamma_k|}]_{ij} = r_{ij}\left(\sqrt{(X_i'X_i)(X_j'X_j)}\right). \tag{13}$$

Thus, given the summary statistics of the variance of quantitative gene expression trait $\mathcal{E}_g$, sample size n, either SNP genotype variances $\{Var(X_i)\}$ or minor allele frequencies $\{f_i\}$, pre-calculated LD coefficients $\{r_{i,j};\, i, j = 1, \cdots, p\}$, and least square estimates of effect sizes $\{\widetilde{w}_i\}$ from single variant regression models (12), we can obtain values of $\boldsymbol{X}'_{|\gamma_k|}\boldsymbol{X}_{|\gamma_k|}$ and $\boldsymbol{X}'_{|\gamma_k|}\mathcal{E}_g$ with manageable computation cost to evaluate the posterior mean of $\boldsymbol{w}_k$ (9) and the posterior likelihood for $\boldsymbol{\gamma}_k$ (10). Moreover, if individual level genotype and expression quantitative trait data are not available, the MCMC can still be implemented by using summary statistics where minor allele frequencies and LD coefficients could be approximated by corresponding values generated from reference panels of the same ethnicity.

3.1.5 Adapted EM-MCMC Algorithm

To further reduce computation burden, instead of considering all segmented genome blocks for genome-wide SNPs, we only fit model (2) with pruned genome blocks that either contain at least one cis-SNP or at least one potential trans-eQTL with p-value $< 1 \times 10^{-5}$ by single variant test (i.e., testing $H_0 : w_i = 0$ with model (12)).

The steps of our adapted EM-MCMC algorithm per gene per tissue type are as follows:

(i) Generate summary statistics: either obtain from individual level genotype data and single variant analyses for genome-wide SNPs, or obtain SNP genotype variances and LD coefficients from reference panel and other summary statistics from previous standard eQTL analyses based on single variant tests;

(ii) Prune genome blocks for applying model (2):

   (a) Consider blocks that contain either at least one cis-SNP or at least one potential trans-eQTL with single variant test p-value $< 1 \times 10^{-5}$;

   (b) Select up to B blocks with minimal p-values within block from smallest to largest. For example, $B = 100$ was used in our application studies. This number can be tuned by users to reduce total computation time accordingly;

   (c) Select any remaining blocks containing cis-SNPs that were not selected in (b);

(iii) Apply EM-MCMC algorithm to the pruned blocks:

   (a) Fix σ_ϵ^2 at the variance of $\mathcal{E}_g$;

   (b) Set initial values for $(\boldsymbol{\pi}, \boldsymbol{\sigma^2})$, e.g., $\pi_{cis} = \pi_{trans} = 1 \times 10^{-6}$ and $\sigma_{cis}^2 = \sigma_{trans}^2 = 0.1$;

   (c) E-step: Conditioning on the most recent estimates of $(\boldsymbol{\pi}, \boldsymbol{\sigma^2})$, estimate $(\boldsymbol{w}, \boldsymbol{PP})$ by implementing MCMC algorithm per block using only summary statistics;

   (d) M-step: Conditioning on the estimates of $(\boldsymbol{w}, \boldsymbol{PP})$ from the previous E-step, update $(\boldsymbol{\pi}, \boldsymbol{\sigma^2})$ by their maximum a posteriori estimates (MAPs), maximizing the expected log-posterior-likelihood functions [7];

   (e) Repeat the EM-steps (c) and (d) for a few times until the MAPs of $(\boldsymbol{\pi}, \boldsymbol{\sigma^2})$ converge, e.g., 5 EM steps.

(iv) Estimates of $(\boldsymbol{w}, \boldsymbol{PP})$ from the last E-step will be used to impute Bayesian GReX from GWAS genotype data of new samples by (7).

See the supplementary note of the BFGWAS paper [4] for details of the EM-MCMC algorithm (iii).

3.2 Software

Software implementing this BGW-TWAS is now available at GitHub (`https://github.com/yanglab-emory/BGW-TWAS`). All steps are wrapped together (enabling parallel computation) through submitting jobs by a Makefile that is generated by a Perl script, which wraps together generating summary statistics by single variant tests with individual-level genotype and gene expression data, pruning genome blocks, implementing adapted EM-MCMC algorithm, and calculating Bayesian $\widehat{GReX}$ for TWAS.

3.3 ROS/MAP Data Description

3.3.1 Study Design

ROS and MAP are prospective cohort studies of aging and dementia, which recruit older adults without known dementia at enrollment who agree to annual clinical testing and brain donation with structured collection of postmortem brain indices at the time of death. All participants sign an informed consent, an Anatomic Gift Act and a consent for their data deposited in the Rush Alzheimer's Disease Center (RADC) repository to be re-purposed by other investigators. Both studies were approved by an Institutional Review Board of Rush University Medical Center, Chicago, IL. Both studies employ harmonized clinical and postmortem data collection facilitating joint analyses.

3.3.2 Cognitive testing and Cognitive Status Diagnosis

Trained technicians administered 17 cognitive tests annually as described previously from which a composite measure of global cognition was constructed[9]. Cognitive status was determined in a three-step process. Annual cognitive testing was scored by a computer and reviewed by a neuro-psychologist to diagnose cognitive impairment and reviewed by a physician. At the time of death, a physician used all cognitive and clinical data collected prior to death, blinded to all postmortem data, to classify persons with respect to dementia, mild cognitive impairment and no cognitive impairment as previously described[10].

3.4 Postmortem Assessment

Brain removal, tissue sectioning and preservation, and a uniform gross and microscopic examination with quantification of post-mortem indices followed a standard protocol [11, 12, 13]. AD pathology (i.e., neuritic plaques, diffuse plaques and neurofibrillary tangles) was visualized using a modified Bielschowsky silver stain on sections from five brain regions and a global summary measure was constructed as described in prior publications[14]. β-amyloid load and paired helical filament tau immunoreactive neuronal neurofibrillary tangles (tangles) were quantified in 8 brain regions (anterior cingulate cortex, superior frontal cortex, mid frontal cortex, inferior temporal cortex, hippocampus, entorhinal cortex, angular gyrus/supramarginal gyrus, and calcarine cortex). Overall β-amyloid load was calculated through averaging mean percent area of β-amyloid deposition per region, across multiple brain regions. Likewise, tangles densities were derived by averaging tangles densities across corresponding brain regions. The global measure of AD pathology is based on counts of neuritic and diffuse plaques and neurofibrillary tangles (15 counts) on 6m sections stained with modified Bielschowsky.

Supplemental References

[1] Yongtao Guan and Matthew Stephens. “Bayesian variable selection regression for genome-wide association studies and other large-scale problems”. In: *Ann. Appl. Stat.* 5.3 (Sept. 2011), pp. 1780–1815. DOI: 10.1214/11-AOAS455.

[2] Edward I. George and Robert E. McCulloch. “Variable Selection via Gibbs Sampling”. In: *Journal of the American Statistical Association* 88.423 (1993), pp. 881–889. DOI: 10.1080/01621459.1993.10476353.

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

[4] Jingjing Yang et al. “A scalable Bayesian method for integrating functional information in genome-wide association studies”. In: *The American Journal of Human Genetics* 101.3 (2017), pp. 404–416.

[5] George Casella. “Empirical Bayes Gibbs sampling”. In: *Biostatistics* 2.4 (2001), pp. 485–500. ISSN: 1465-4644 (Print) 1465-4644 (Linking). DOI: 10.1093/biostatistics/2.4.485.

[6] Sameer Singh, Michael Wick, and Andrew McCallum. *Monte Carlo MCMC: efficient inference by approximate sampling*. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012, pp. 1104–1113.

[7] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. English. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38. ISSN: 00359246.

[8] Jian Yang et al. “Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits”. In: *Nat Genet* 44.4 (2012), 369–75, S1–3. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking). DOI: 10.1038/ng.2213.

[9] Robert S Wilson et al. “Early and late life cognitive activity and cognitive systems in old age”. In: *Journal of the International Neuropsychological Society: JINS* 11.4 (2005), p. 400.

[10] PA Boyle et al. “Mild cognitive impairment: risk of Alzheimer disease and rate of cognitive decline”. In: *Neurology* 67.3 (2006), pp. 441–445.

[11] David A Bennett et al. “Overview and findings from the religious orders study”. In: *Current Alzheimer Research* 9.6 (2012), pp. 628–645.

[12] David A Bennett et al. “Overview and findings from the rush Memory and Aging Project”. In: *Current Alzheimer Research* 9.6 (2012), pp. 646–663.

[13] David A Bennett et al. “Religious orders study and rush memory and aging project”. In: *Journal of Alzheimer’s Disease* 64.s1 (2018), S161–S189.

[14] David A Bennett et al. “Relation of neuropathology to cognition in persons without cognitive impairment”. In: *Annals of neurology* 72.4 (2012), pp. 599–609.