Novel Bayesian transcriptome-wide association study method leveraging 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], Jingjing Yang[5*]

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's Disease and the Aging Brain, Columbia University Irving Medical Center, New York, NY10032, USA
4. Rush Alzheimer's 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 author: jingjing.yang@emory.edu

Keywords: Transcriptome-wide association study; Bayesian variable selection model; cis-eQTL; trans- eQTL; Summary statistics; Alzheimer's dementia

Abstract

Transcriptome-wide association studies (TWAS) have been widely used to integrate gene expression and genetic data for studying complex traits. Due to the computation burden, existing TWAS methods neglect distant trans- expression quantitative trait loci (eQTL) that are known to explain a significant proportion of the variation for most expression quantitative traits. To leverage both cis- and trans- eQTL information for TWAS, we propose a novel TWAS approach based on Bayesian variable selection regression model, which not only accounts for both cis- and trans- SNPs of the target gene but also enables efficient computation by using summary statistics of standard eQTL analyses and a scalable EM-MCMC algorithm. Simulation studies illustrate that our Bayesian approach achieved higher TWAS power compared to existing methods. By application studies, we identified gene *ZC3H12B* whose GReX is associated with both Alzheimer's dementia (AD) (p-value=2.15×10^{-11}) and a global measure of AD pathology (p-value=2.438×10^{-6}), which is completely driven by trans-eQTL. We also identified gene *KCTD12* whose GReX is associated with β-amyloid load (p-value=7.63×10^{-8}), which is driven by both cis- and trans- eQTL. Particularly, four of the top driven trans-eQTL of *ZC3H12B* are located in gene *APOC1* (<12KB away from the well-known risk gene *APOE* of AD) and are also known GWAS signals of AD and blood lipids. Free software for implementing our proposed Bayesian TWAS approach is available on Github.

Introduction

Genome-wide association studies (GWAS) have identified thousands of variants associated with complex diseases and traits over the past decades[1-5]. However, most of these GWAS hits are located within noncoding regions, and the underlying biological mechanisms by which these variants impact a phenotype remain largely 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 can help identify key biological mechanisms of complex traits.

One such integrative method is transcriptome-wide association study (TWAS)[11-13] that takes advantage of reference panel with profiled transcriptomic and genetic data of the same subjects. Specifically, TWAS first utilizes such reference data to fit a prediction regression model for expression quantitative trait of the target gene with nearby genotypes (e.g., cis-SNPs within 1MB region) as predictors, and then tests the gene-based genetic association between the predicted genetically regulated gene expression (GReX) and the phenotype of interest with additional test cohort. The SNPs with non-zero effect sizes on gene expression in the prediction model are referred to as broad sense of "eQTL" in TWAS. Example 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[14], and the North American Brain Expression Consortium (NABEC) for cortex tissues[15]. With fitted prediction models from reference data, TWAS can be conducted with test samples that have either individual-level or summary-level GWAS data[12; 13; 16].

Essentially, 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 prediction model as their corresponding burden weights[11-13]. By weighting genetic variants by cis-eQTL effect sizes, TWAS assumes the effects of risk genes on the phenotype of interest are potentially mediated through their transcriptome variations. Recent TWAS studies have identified both known GWAS risk genes and novel risk genes for a wide range of complex phenotypes, including schizophrenia, breast cancer, and Alzheimer's dementia[17-21].

However, existing TWAS methods only use genetic data of cis-SNPs of the target gene as predictors to fit the GReX prediction model[11-13]. As shown by recent studies, trans-SNPs (e.g., outside of the 1MB 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 with

important biological interpretations[22; 23]. Thus, using both cis- and trans- SNPs as predictors is expected to increase the prediction accuracy of GReX and increase the power of TWAS, where the biggest barrier will be the enormous computation power required to fit ~20K GReX prediction models for genome-wide genes.

To circumvent these computational limitations, we propose a novel Bayesian TWAS approach that accounts for both cis- and trans- SNPs in a Bayesian variable selection regression (BVSR) model[24] for predicting GReX, and enables efficient computation by using summary statistics of standard eQTL analyses based on single variant tests and a scalable EM-MCMC algorithm[25]. We demonstrate the feasibility of this Bayesian approach in simulation studies with varying proportions of true causal cis- and trans- eQTL for expression quantitative traits. Specifically, we compare this Bayesian approach to the previously proposed PrediXcan method that uses only cis-SNPs by Elastic-Net penalized regression method[11]. We apply this Bayesian approach to clinical and postmortem data from older adults, to conduct TWAS for Alzheimer's dementia (AD) as well as postmortem indices of AD pathology including neurofibrillary tangle density (tangles), β-amyloid load (β-amyloid), and a global measure of AD pathology (gpath).

In particular, our simulation studies revealed that this Bayesian approach achieved higher TWAS power by considering both cis- and trans- SNPs for predicting GReX. Our application studies identified two novel risk genes associated with AD phenotypes —— *ZC3H12B* associated with both clinical diagnosis of AD (p-value=2.15×10^{-11}) and global measure of AD pathology (p-value=2.438×10^{-6}) which is completely driven by trans-eQTL, and *KCTD12* associated with β-amyloid load (p-value=7.63×10^{-8}) which is driven by both cis- and trans- eQTL. Especially, four of the top driven trans-eQTL of genes *ZC3H12B* (*rs12721051, rs4420638, rs56131196, rs157592* located in gene *APOC1*) are located <12KB away from the well-known risk gene *APOE* of AD[30], and are also known GWAS signals of AD[26] and blood

lipids[27-29]. Free software for implementing our proposed Bayesian TWAS approach is available on Github.

Method

TWAS Procedure

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

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

where $\boldsymbol{\mathcal{E}}_g$ denotes gene expression levels (i.e., expression quantitative trait) for a 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 $\boldsymbol{\epsilon}$ is an error term following a $N(0, \sigma_\epsilon^2 \boldsymbol{I})$ distribution. The intercept term is dropped due to centering both response ($\boldsymbol{\mathcal{E}}_g$) and explanatory variables ($\boldsymbol{X}$). With $\widehat{\boldsymbol{w}}$ estimated from the training data (e.g., reference data) that have both transcriptomic and genetic data profiled for the same subjects, TWAS will test the association between the phenotype of interest and the predicted GReX that can be estimated from genotype data $\widetilde{\boldsymbol{X}}$ of the test cohort as follows

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

Bayesian Variable Selection Regression Model

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

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

Specifically, 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,

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

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

where $(\pi_{cis},\ \pi_{trans})$ denote the respective probability that the coefficient is 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,

$$\pi_{cis} \sim Beta(a_{cis}, b_{cis});\ \sigma_{cis}^2 \sim IG(k_1, k_2); \qquad \text{(Equation 3)}$$

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

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 Supplementary Text 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, $\mathrm{E}[\gamma_i]$, represents the posterior probability (PP_i) for each individual SNP to have non-zero effect sizes (i.e., to be an eQTL of the target gene)[24]. Moreover, we propose a novel approach to estimate GReX for test samples that can account for the uncertainty for each SNP to be a broad sense eQTL with a non-zero effect size:

$$\widehat{GReX} = \sum_{i=1}^{p} \widetilde{X}_i (\widehat{PP}_i \, \widehat{w}_i) \qquad \text{(Equation 4)}$$

where $\widetilde{X}_i$ represents the genotype data of variant i for test samples and ($\widehat{w}_i$, $\widehat{PP}_i$) denote estimates of effect size and posterior causal probability from the BVSR model (Equations 2-4) (see Supplementary Text for detailed Bayesian inference procedure). This Bayesian GReX estimate can then be used to conduct TWAS by testing the association between the predicted GReX and the phenotype of interest.

Efficient Computation Techniques

In theory the estimates of eQTL effect sizes and corresponding posterior inclusion probabilities ($\widehat{w}_i$, $\widehat{PP}_i$) can be obtained by using a standard Markov Chain Monte Carlo (MCMC)[33] 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. To circumvent these practical limitations, we employ several computational techniques to enable computation efficiency such that this Bayesian TWAS method can be deployed to leverage both cis- and trans- eQTL information in practice. First, we employ a previously developed scalable expectation-maximization Markov Chain Monte Carlo (EM-MCMC) algorithm[25] to estimate (w_i, PP_i). Unlike the original EM-MCMC algorithm requiring individual-level GWAS data, we can reduce up to 90% of the computation time by enabling 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. Second, to further reduce the computation burden, we apply the BVSR model to a subset of pre-selected 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.

Simulation Study Design

We conducted simulation studies to validate the performance of our proposed Bayesian TWAS method leveraging both cis- and trans- eQTL information and compared with the existing PrediXcan method considering 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 1,200 samples as our test data. GReX prediction models were fitted using the training data where "eQTL" effect sizes and the corresponding posterior probabilities were estimated. Given these fitted GReX prediction models, GReX data were imputed for 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 ensure targeted gene expression heritability h_e^2 and that all causal eQTL explain equal expression heritability. Gene expression levels were generated by $\boldsymbol{\mathcal{E}_g} = \boldsymbol{X_g w} + \epsilon_e$, with $\epsilon_e \sim N(0, (1 - h_e^2))$. Then we simulated phenotypes by $\boldsymbol{Y} = \beta \boldsymbol{\mathcal{E}_g} + \epsilon_p$, where β was selected with respect to 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: i) all causal eQTL are from the cis-block; ii) 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; iii) 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., $\left(h_e^2, h_p^2\right) = \left((5\%, 80\%),\ (10\%, 50\%), (20\%, 25\%), (50\%, 10\%)\right)$ for the scenario with 5 true causal eQTL and $\left(h_e^2, h_p^2\right) = ((5\%, 98\%), (10\%, 85\%), (20\%, 30\%), (50\%, 5\%)$ for the scenario with 22 true causal eQTL. Here, different levels of phenotype heritability were arbitrarily chosen to achieve similar levels of TWAS power across all scenarios.

In each simulation, we first fitted GReX prediction models by BVSR with both *cis*- and *trans*- genome blocks and by Elastic-Net (as used by PrediXcan) with only *cis*- genome block using training data. Then we conducted TWAS with predicted *GReX* by both methods. The performance of our Bayesian approach was compared to PrediXcan in terms of R^2 of the predicted GReX and TWAS power in test samples. Test R^2 was calculated as the squared correlation between predicted 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).

<u>ROS/MAP and Mayo Clinic GWAS data of AD</u>

Following simulation studies, we applied our proposed Bayesian TWAS method to genomic and AD phenotype data from older adults 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) [34-36] and GWAS data from the Mayo Clinic Alzheimer's Disease Genetics Studies (MCADGS)[37-39].

ROS and MAP are prospective cohort studies of aging and dementia, which recruit older adults without known dementia at enrollment who agree to annual evaluation 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 RADC

repository to be repurposed by other investigators. Both studies were approved by an Institutional Review Board of Rush University Medical Center, Chicago, IL. The Mayo Clinic late-onsite Alzheimer's Dementia GWAS study 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[40] and 2,099 European-decent subjects from MCADGS, were further imputed to the 1000 Genome Project Phase 3[41] in our analysis[42].

Post-mortem brain samples (gray matter of the dorsal lateral prefrontal cortex) from ~30% these ROS/MAP participants with assayed genotype data were profiled for transcriptomic data by next-generation RNA-seqencing[43], which were used as reference data to train GReX prediction models. In this paper, we conducted TWAS for clinical diagnosis of late-onsite Alzheimer's dementia (available for both ROS/MAP and MCADGS) as well as postmortem pathology indices of AD (only available for ROS/MAP) quantified with β-antibody specific immunostains, including PHFtau tangle density, β-amyloid load, and a global measure of AD pathology (a combination of neuritic and diffuse plaques and neurofibrillary tangles)[34-36]. The tangle density quantifies the average PHFtau tangle density within two or more 20µm sections from eight brain regions — hippocampus, entorhinal cortex, midfrontal cortex, inferior temporal, angular gyrus, calcarine cortex, anterior cingulate cortex, and superior frontal cortex. The β-amyloid load quantifies the average percent area of cortex occupied by β-amyloid protein in adjacent sections from the same eight brain regions. These two are based on immunohistochemistry. The global measure of AD pathology is based on counts of neuritic and diffuse plaques and neurofibrillary tangles (15 counts) on 6µm sections stained with modified Bielschowsky[34-36].

Results

Simulation Results

For the scenario with 5 true causal eQTL and various expression heritability, our simulation studies showed that our Bayesian approach based on the BVSR model lead to higher test R^2 for GReX and TWAS power than the PrediXcan method when a portion of the true causal eQTL are distributed over trans- genome blocks (Figure 1 (A,B)). This is because our Bayesian approach leverages both cis- and trans- eQTL information while the PrediXcan method does not account for trans- eQTL. As expected, the performance of our Bayesian approach was similar to PrediXcan when all causal eQTL were from the cis- genome block, and the PrediXcan method failed with nearly zero test R^2 when all causal eQTL were from trans- genome blocks (Figure 1A).

For the scenario with 22 mixed cis- and trans- eQTL, the performance comparison between our Bayesian approach and PrediXcan method became more complicated with respect to various true expression heritability levels (Figure 1 (C, D)). Particularly, when $h_e^2 = 0.05$, both 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 BVSR model was observed and led to higher test R^2 and TWAS power. When $h_e^2 = 0.1$ and 70% of the true causal eQTL were cis-, our Bayesian approach was less effective than the PrediXcan method. This is likely due to the fact that the Elastic-Net regression model used by PrediXcan is preferred than the BVSR model when true causal eQTL manifest relatively small effect sizes.

In contrast, when true causal eQTL signals have relatively large effect sizes and are distributed outside the cis-region of the target gene, our Bayesian TWAS method based on the BVSR model is preferred due to the improved accuracy for GReX prediction by leveraging trans- SNP data.

TWAS of Clinical and Postmortem Alzheimer's Dementia Related Phenotypes

Next, we applied our Bayesian approach to conduct TWAS using the GWAS data from ROS/MAP[36; 44] and MCADGS[37]. First, we trained the GReX prediction model by applying the BVSR model to samples (N=499) from the ROS/MAP cohort that possess both profiled genotype data and transcriptomic data of dorsal lateral prefrontal cortex tissues. All expression quantitative traits were normalized and corrected for age at death, sex, PMI, study (ROS or MAP), Batch Effects, RIN Scores, and Cell type proportions (oligodendrocytes, astrocytes, microglia, neurons).

Using our Bayesian TWAS approach, we obtained GReX prediction models for 14,156 genes as compared to 6,011 valid prediction models using PrediXcan (with at least one cis-SNP with nonzero effect size on expression quantitative trait). Across the 6,011 genes with valid prediction models by PrediXcan, our Bayesian 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 (Supplementary Figure 1).

We imputed GReX values for all remaining individuals who only possessed genotype data in ROS/MAP and MCADGS. We then conducted TWAS by testing the association between the GReX values (fitted values of training samples and imputed values for the remaining samples) and phenotypes of AD clinical diagnosis as well as quantitative pathology indices of AD including neurofibrillary tangle density, β-amyloid load, and a global measure of AD pathology.

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. Specifically, samples from the ROS/MAP cohort have their AD status determined by review of all clinical data at the time of death blinded to all pathologic data. Individuals were determined to have dementia or mild cognitive impairment, with the remaining persons classified as no cognitive impairment. We assigned samples with AD as cases and samples with mild cognitive impairment or no cognitive impairment as controls. Confounding

variables that were adjusted for during TWAS with the ROS/MAP cohort included age at death, sex, smoking, ROS or MAP study, education level, and top three principal components derived from GWAS data. Whereas, for the Mayo Clinic GWAS samples, cases were determined for samples with a medical history of late onsite AD diagnosis, and available confounding variables only included 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[45] to summary statistics generated by TWAS within each cohort. We also compared the meta-TWAS results by our Bayesian approach to results generated using PrediXcan (see Manhattan plots in Figures 2A and Supplementary Figure 2A).

Our Bayesian approach identified gene *ZC3H12B* (located on chromosome X) whose GReX values are associated with AD clinical diagnosis with effect size $\beta = 133.26$ and p-value 2.15×10^{-11} (Figure 2A; Table 1). Both within-cohort TWAS obtained positive effect sizes ($\beta = 86.5, 159.5$) for ROS/MAP and Mayo Clinic cohorts, with respective p-value $= 0.04$, 4.12×10^{-10}. On the other hand, this gene was missed by the PrediXcan method, because the association of this gene is completely drive by trans-eQTL that were neglected by PrediXcan (Supplementary Figure 2A).

TWAS of postmortem pathology phenotypes of AD was performed only in decedents from ROS/MAP from whom postmortem indices were collected. Specifically, TWAS were conducted for 1,121, 1,114, and 1,139 ROS/MAP samples with GWAS genotype data and respective phenotype data for tangles, β-amyloid, and global measure of AD pathology (see Manhattan plots in Figures 2B, 3(A, B) and Supplementary Figures 2B, 3(A, B)). Confounding variables including age at death, sex, smoking, ROS or MAP study, education level, and top three principal components derived from genome-wide genotype data were adjusted for in TWAS of these phenotypes.

Interestingly, by our Bayesian approach, *ZC3H12B* was also identified to be associated with the quantitative global pathology index with genome-wide significant p-value 2.438×10^{-6} (Figure 2B; Table 1), and marginally significantly associated with neurofibrillary tangle density with p-value 2.61×10^{-6} (Figure 3A; Table 1). Gene *KCTD12* located on chromosome 12 was identified to be significantly associated with β-amyloid load with p-value 7.63×10^{-8}(Figure 3B; Table 1).

We presented 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 Supplementary Figure 4. We noted that the association between the GReX values of *ZC3H12B* and AD phenotypes was completely driven by trans-eQTL, while the association between the GReX values of *KCTD12* and β-amyloid load was 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 gene *APOC1* that is a known risk gene for AD[26] and blood lipids[27-29], which is <12KB away from the well-known risk gene *APOE* of AD[30]. Particularly, *rs12721051* located in the 3' UTR region of *APOC1* was identified as a GWAS signal of total cholesterol levels[29]; *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)[27] and AD[31]; *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*[32]. Additionally, gene *ZC3H12B* was found to regulate proinflammatory activation of macrophages[46] and express higher in brain, spinal cord and thymus tissue types comparing to others[47]. These results showed that the effects of these known GWAS signals (*rs4420638, rs56131196, rs157592*) of AD could be mediated through the gene expression levels of *ZC3H12B*.

<u>Insights about eQTL Genetic Architecture</u>

In addition to imputing GReX values (Equation 5), the posterior probabilities estimated by BVSR also provide insights into the genetic architecture of eQTL, especially about how eQTL are distributed across the genome. Note that the posterior probability obtained from the BVSR model (Equations 2-4) is essentially the expected probability for a SNP to be an eQTL. Therefore, the sum of posterior inclusion probabilities 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 3. 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 BVSR model prefers relatively larger eQTL effect sizes. These simulation results demonstrated the validity of our Bayesian approach based on the BVSR model as well as the usefulness of the sum of posterior inclusion probabilities.

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 with non-zero estimated effect sizes by the BVSR model were from *trans*- genome regions with respect to the target gene. We can see that ~90% genes with train $R^2 > 0.05$ have ~2-3 average expected eQTL, and ~10% genes have train $R^2 < 0.05$ have >5 average expected eQTL (Table 3). 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.

Discussion

In this paper, we proposed a novel Bayesian TWAS approach by using 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 pre-selecting genome regions that contain either at least one cis-SNPs or one potential trans-eQTL and implementing the previously developed scalable EM-MCMC algorithm[25] with pre-calculated LD coefficients and summary statistics from standard eQTL analyses. This novel approach extends previous TWAS methods[11-13] that only utilize partial genotype information from a small window of *cis*- SNPs to train the *GReX* prediction model.

Genotype data of trans-eQTL have been shown to explain a significant amount of variation of expression quantitative traits and possess important biological functions underlying known GWAS loci of complex diseases[22; 23]. The results from our simulation and application studies demonstrated that our Bayesian approach improves the yield of TWAS by levering both cis- and trans- eQTL information. For example, higher precision of GReX prediction and power of TWAS were obtained in our simulation studies when true causal trans-eQTL existed. These results showed that our Bayesian approach has a greater advantage for scenarios where eQTL have relatively large effect sizes for the expression quantitative traits (e.g., 5 vs. 22 true 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 GWAS[24; 25].

By applying our Bayesian approach to the studies of AD using ROS/MAP and MCADGS cohorts, we identified a novel risk gene (*ZC3H12B*) whose GReX values are associated with both clinical diagnosis of AD and postmortem AD pathology indices (neurofibrillary tangle density and global measure of AD pathology). This association was missed by the PrediXcan method[11] because it was 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[26] and blood lipids[27-29] and <12KB away from the famous well-known risk gene of AD (*APOE*)[30]. Since recent studies generally identified a handful of eQTL with relatively large effect sizes for majority genes and trans-eQTL are often presented[8; 48], we expect our Bayesian approach that selects sparse signals and leverages trans-eQTL information will have potential for making great impact in practice.

In addition, by fitting BVSR models using both cis- and trans- eQTL, we can not only account for the uncertainty for a SNP to be an eQTL (posterior inclusion probability) to predict GReX (Equation 5), but also use the sum of posterior inclusion probabilities of all eQTL to estimate the expected number of eQTL[24; 25]. The distribution of expected eQTL could provide useful insight for the underlying genetic architecture of expression quantitative traits.

The current study is not without limitations. First, a large amount of computing resource is required to use our Bayesian TWAS method to train GReX prediction models for genome-wide genes (~20K) per tissue type. Even with the efficient computational approach we developed, approximately 30 minutes of computation time (with 4 CPU cores) and 3GB memory were needed per gene. Parallel computation can be employed to make use of 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 investigation 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 Elastic-Net penalized regression model[49] as used by the PrediXcan[11] method has advantages when all causal eQTL are cis- with relatively small effect sizes (e.g., 22 true causal cis-eQTL). Further work is needed to develop a similar Bayesian TWAS procedure based on Bayesian Elastic-Net regression model (with a mixture of Gaussian and Laplacian prior for the effect sizes)[50], which can include PrediXcan as a special case when only considering cis-eQTL. However, how to efficiently fit such Bayesian prediction models for

GReX with large-scale genomic data is still an open question. These open questions will be part of our continuing research.

In conclusion, our Bayesian TWAS approach 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 Bayesian approach would be important for developing similar frameworks for integrating other quantitative omics traits in gene-based association studies. Such integrative method developments stand to benefit from our Bayesian TWAS approach, especially the approach of leveraging information from trans- QTL and efficient computation techniques derived from this paper. In addition, our Bayesian TWAS approach could be extended to studies of other complex human phenotypes to identify novel risk genes that could be targeted in further drug discovery.

Tables

Table 1. Significantly associated genes identified by our Bayesian TWAS approach.

Gene	Chr	Position	Train R^2	p-value	Effect size (SD)	Phenotype
ZC3H12B	X	64,708,614	0.24	7.31×10^{-13}	139.84 (19.49)	AD
ZC3H12B	X	64,708,614	0.24	9.59×10^{-7}	72.28 (14.67)	Global measure of AD pathology
KCTD12	13	77,454,311	0.09	3.44×10^{-8}	5.43 (0.98)	β-Amyloid

Table 2. Trans-eQTL with top five PP > 0.003 for gene 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. Average sums of posterior inclusion probabilities that are stratified based on train R^2 of the fitted BVSR models (simulation studies) for GReX prediction for conditions with 2 of 5 eQTL being cis-SNPs and 11 of 22 eQTL being cis-SNPs for the different causal eQTL conditions, respectively.

h_e^2	**Sum of Posterior Inclusion Probabilities**			**Number of Causal eQTL**
	Whole Genome	**Cis- Region**	**Trans-Region**	
0.05	0.79	0.46	0.33	5
	0.05	0.02	0.03	22
0.1	2.28	1.13	1.15	5
	0.21	0.11	0.10	22
0.2	3.72	1.44	2.28	5
	1.43	0.87	0.56	22
0.5	4.91	1.56	3.35	5
	6.46	3.89	2.57	22

Table 4. Average sums of posterior inclusion probabilities that are stratified based on train R^2 of the fitted BVSR models (using ROS/MAP data) for GReX prediction.

Train R^2	**Sum of Posterior Inclusion Probabilities**			**Number of Genes**
	Whole Genome	**Cis- Region**	**Trans-Region**	
(0, 0.05)	6.63	0.60	6.23	1,504
(0.05, 0.1)	1.45	0.13	1.32	1,964
(0.1, 0.25)	2.00	0.17	1.83	6,617
(0.25, 0.5)	2.66	0.22	2.44	3,224
(0.5, 1)	3.04	0.31	2.73	474

Figure Legends

Figure 1. Simulation TWAS Studies Comparing Bayesian and PrediXcan 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. (A, B) Test R^2 and TWAS power comparison when there are 5 true causal eQTL. Our Bayesian method was found to out-perform PrediXcan method when a certain proportion of true causal eQTL are from trans- genome regions. (C, D) Test R^2 and TWAS power comparison when there are 22 true causal eQTL. Our Bayesian method was found to out-perform PrediXcan method when >50% of true causal eQTL are from trans-genome regions.

Figure 2. Manhattan plots of our Bayesian TWAS results of AD clinical diagnosis (A) and a global measure of AD pathology (B). *The red lines denote genome-wide significant threshold (2.5×10^{-6}) for gene-based association studies. Gene ZC3H12B was found to be significantly associated with both AD clinical diagnosis and a global measurement of AD pathology.*

Figure 3. ***Manhattan plots of our Bayesian TWAS results of neurofibrillary tangle density (A) and β-amyloid load (B).*** *The red lines denote genome-wide significant threshold (2.5×10^{-6}) for gene-based association studies. Gene ZC3H12B was found to be significantly associated with neurofibrillary tangle density. Gene KCTD12 was found to be significantly associated with and β-amyloid load.*

Figure 4. ***Plots of BVSR posterior probabilities (PP) for analyzed cis- and trans- SNPs to be an eQTL for gene ZC3H12B (A) and KCTD12 (B).*** *Gene ZC3H12B located on chromosome X has top trans-eQTL from chromosomes 1, 6, and 19. Gene KCTD12 located on chromosome 12 has top cis-eQTL and trans-eQTL from chromosomes 4 and 6.*

Supplemental Data

Supplementary Text and Figures

Declaration of Interests

The authors declare no competing interests.

Acknowledgement

J.Y. was supported by the startup funding from Department of Human Genetics at Emory University School of Medicine. 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, U01AG61356, the Illinois Department of Public Health, and the Translational Genomics Research Institute. The MCADGC led by Dr. Nilüfer Ertekin-Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL uses samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. MCADGC data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation.

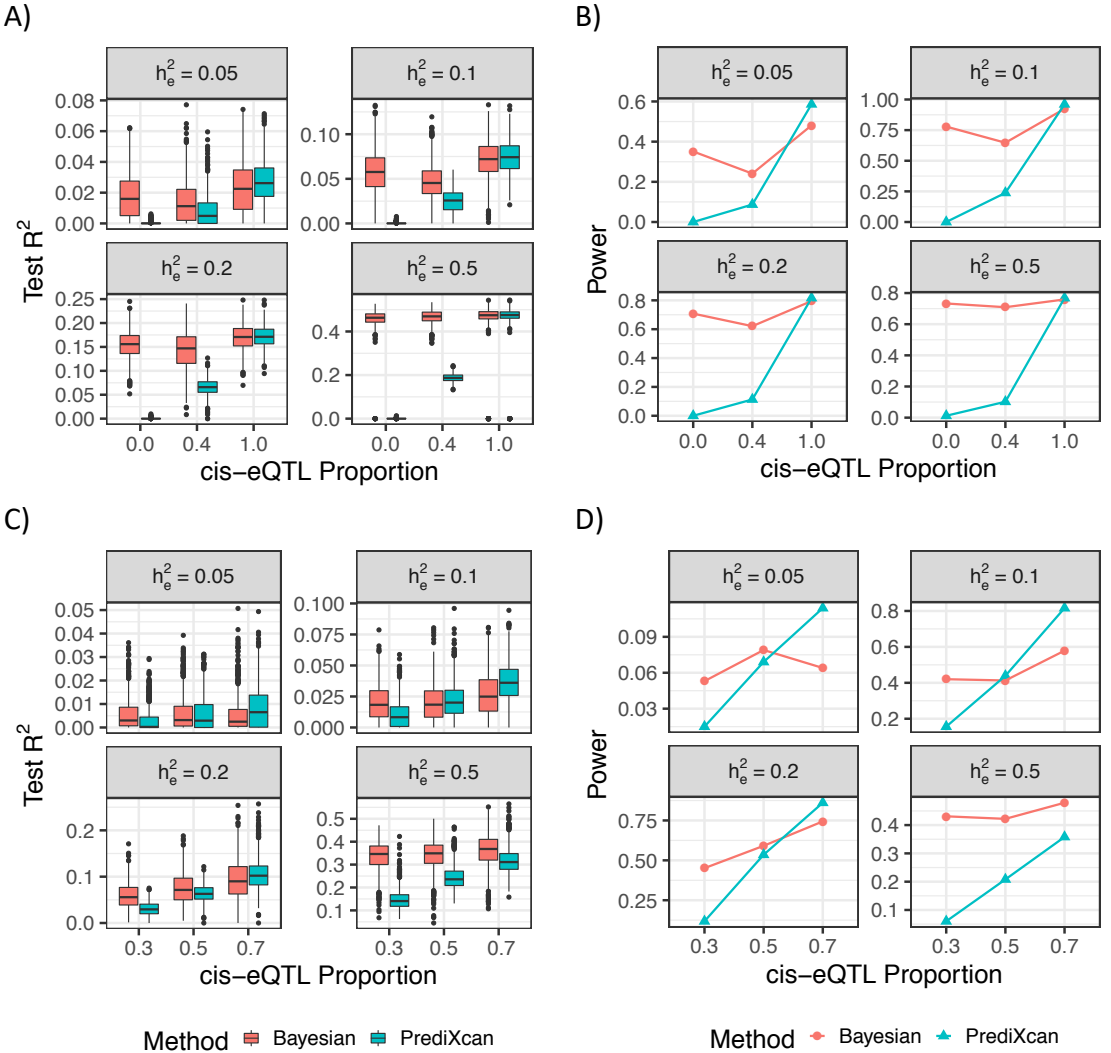
References

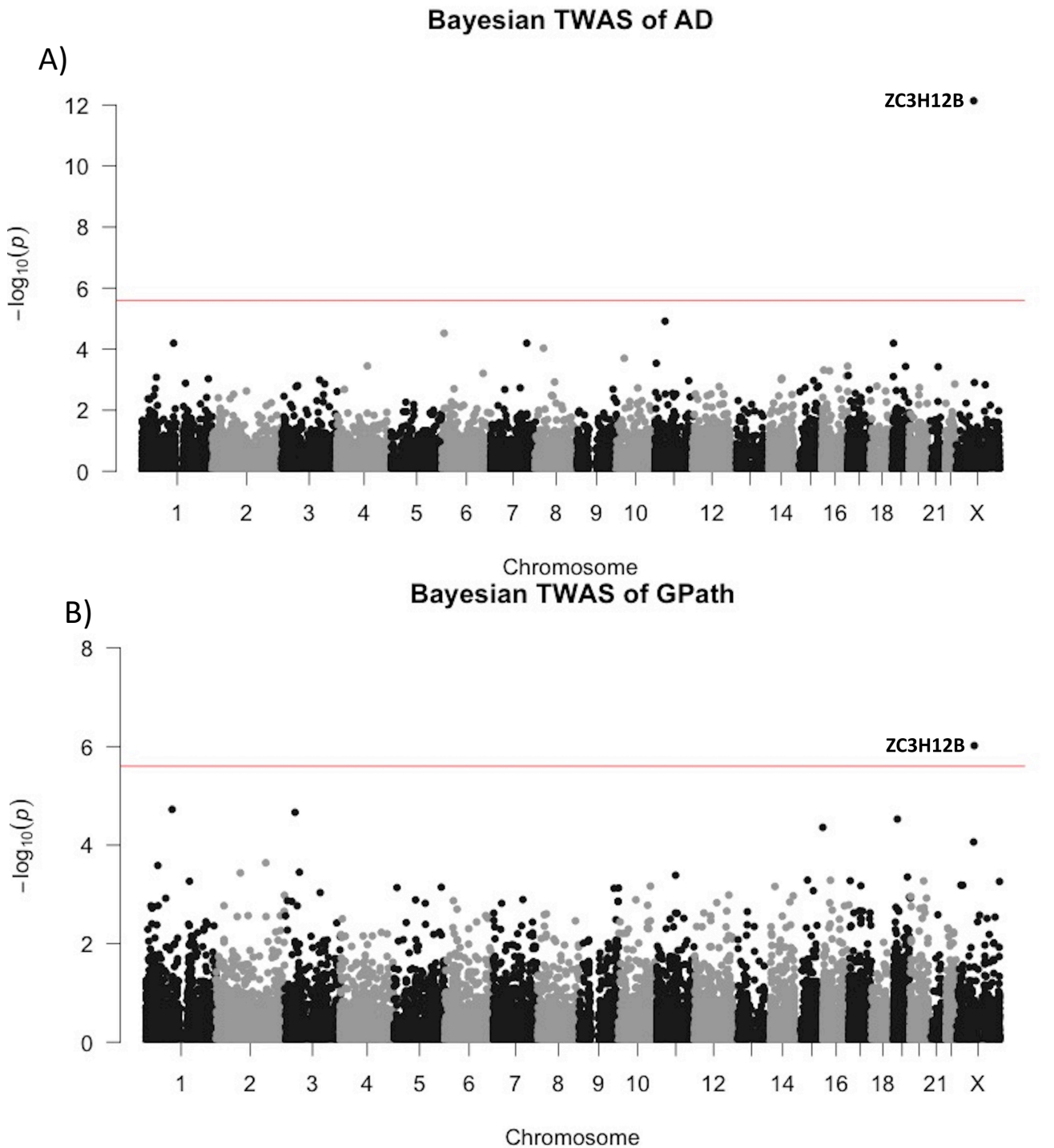
1. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nature reviews Genetics 6, 95-108.
2. Wellcome Trust Case Control, C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 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. Nature reviews Genetics 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. Nature genetics 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. American journal of human genetics 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. American journal of human genetics 102, 717-730.
8. GTEx Consortium. (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 genetics 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., GTEx Consortium, Nicolae, D.L., et al. (2015). A gene-

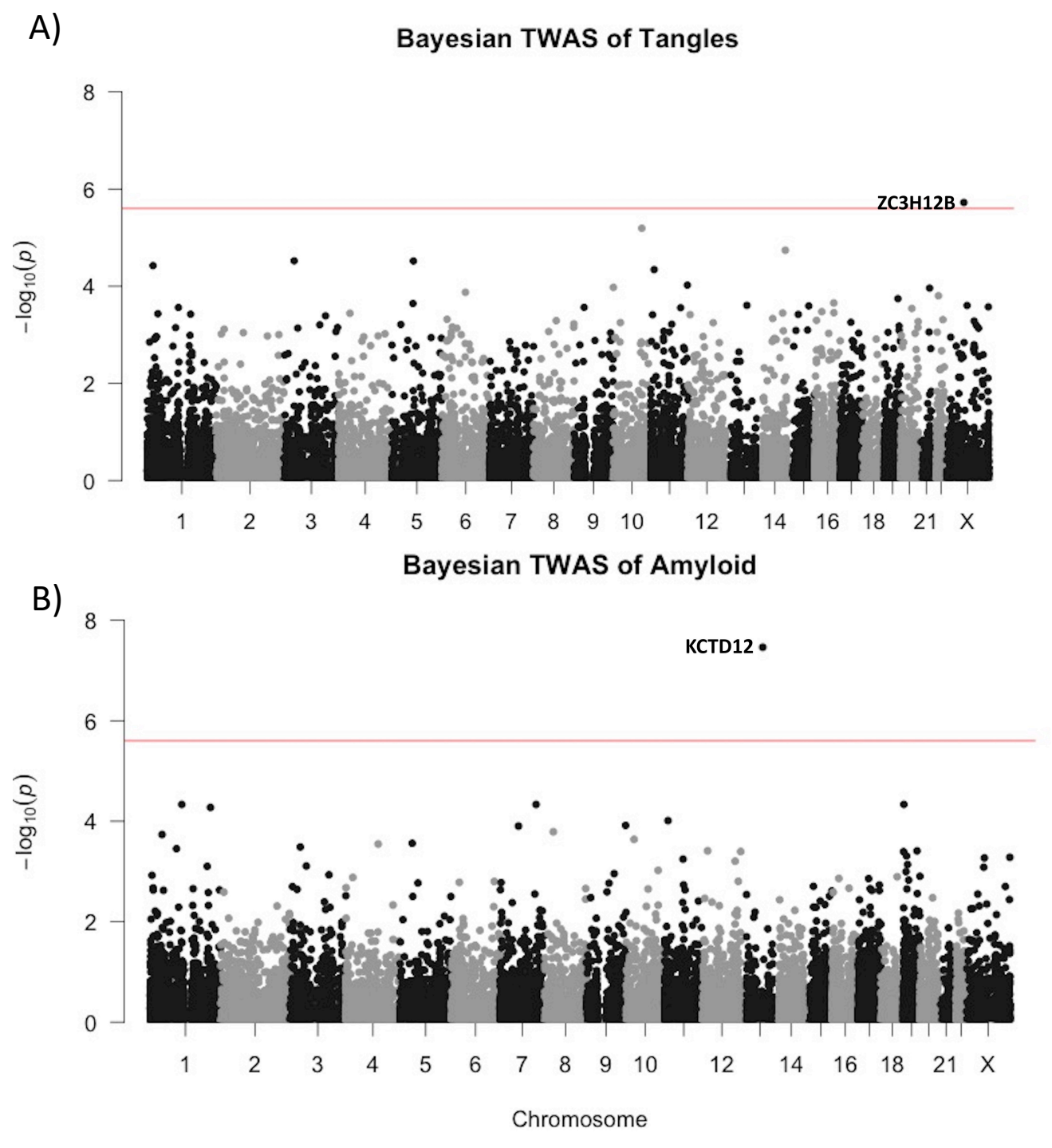
based association method for mapping traits using reference transcriptome data. Nature genetics 47, 1091-1098.
12. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics 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., et al. (2019). TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. American journal of human genetics 105, 258-266.
14. Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-511.
15. 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 genetics 6, e1000952.
16. 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. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 9, 1825.
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. American journal of human genetics 100, 473-487.
18. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Schizophrenia Working Group of the Psychiatric Genomics, C., McCarroll, S., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nature genetics 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. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nature genetics 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. Nature genetics 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. Nature genetics 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. American journal of human genetics 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. 447367.
24. Guan, Y.T., and Stephens, M. (2011). Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems. Annals of Applied Statistics 5, 1780-1815.
25. Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G., and International Age-Related Macular Degeneration Genomics, C. (2017). A Scalable Bayesian Method for Integrating

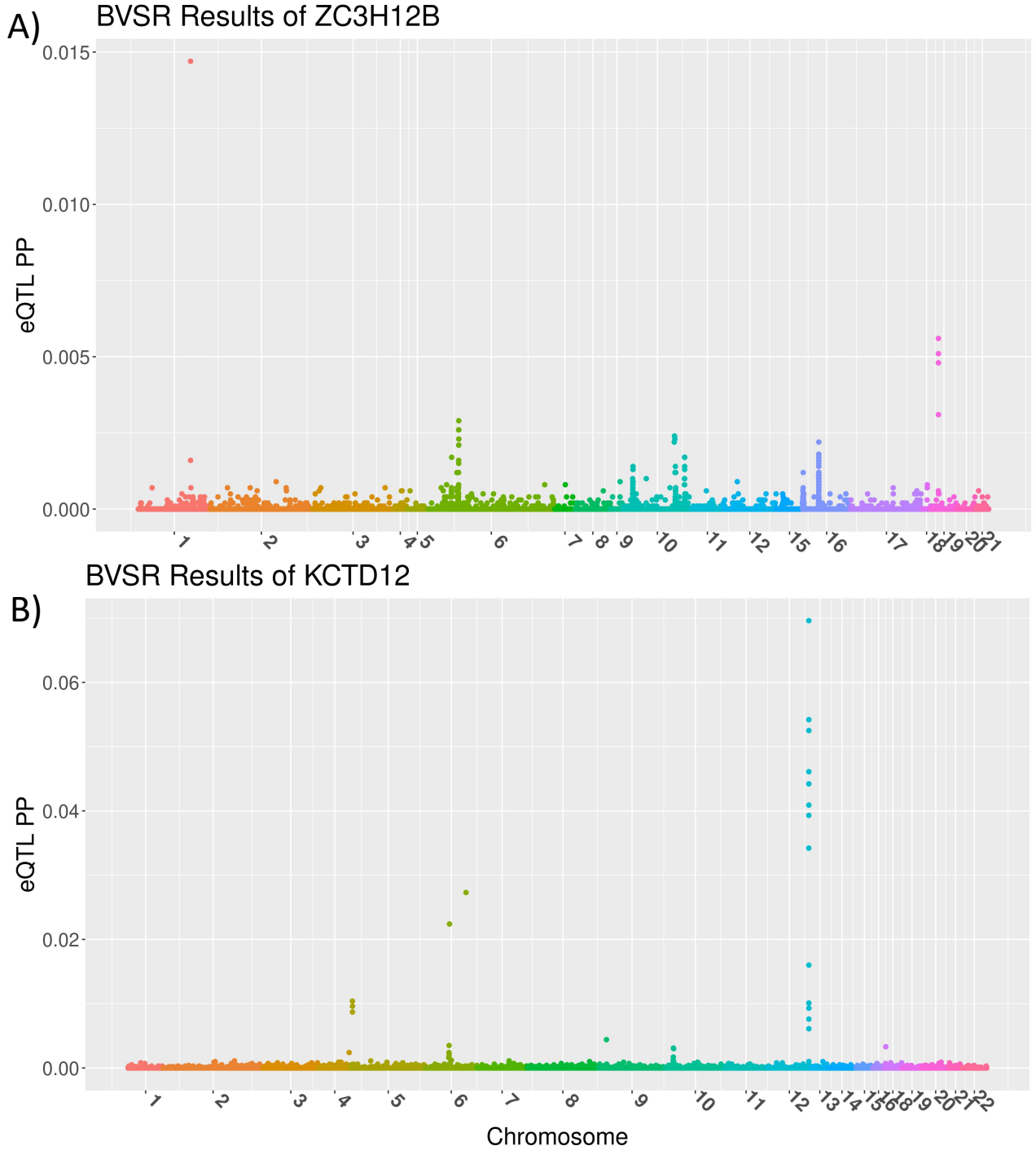
Functional Information in Genome-wide Association Studies. American journal of human genetics 101, 404-416.
26. Marioni, R.E., Harris, S.E., Zhang, Q., McRae, A.F., Hagenaars, S.P., Hill, W.D., Davies, G., Ritchie, C.W., Gale, C.R., Starr, J.M., et al. (2018). GWAS on family history of Alzheimer's disease. Transl Psychiatry 8, 99.
27. Ligthart, S., Vaez, A., Hsu, Y.H., Inflammation Working Group of the, C.C., Pmi Wg, X.C.P., LifeLines Cohort, S., Stolk, R., Uitterlinden, A.G., Hofman, A., Alizadeh, B.Z., et al. (2016). Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation. BMC genomics 17, 443.
28. Ligthart, S., Vaez, A., Vosa, 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. (2018). Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. American journal of human genetics 103, 691-706.
29. 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. (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nature genetics 50, 1514-1523.
30. 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.
31. Kamboh, M.I., Demirci, F.Y., Wang, X., Minster, R.L., Carrasquillo, M.M., Pankratz, V.S., Younkin, S.G., Saykin, A.J., Alzheimer's Disease Neuroimaging, I., Jun, G., et al. (2012). Genome-wide association study of Alzheimer's disease. Transl Psychiatry 2, e117.
32. Zhou, X., Chen, Y., Mok, K.Y., Kwok, T.C.Y., Mok, V.C.T., Guo, Q., Ip, F.C., Chen, Y., Mullapudi, N., Alzheimer's Disease Neuroimaging, I., et al. (2019). Non-coding variability at the APOE locus contributes to the Alzheimer's risk. Nat Commun 10, 3310.
33. Casella, G. (2001). Empirical Bayes Gibbs sampling. Biostatistics 2, 485-500.
34. 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.
35. 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.
36. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., and Schneider, J.A. (2018). Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis 64, S161-S189.
37. Carrasquillo, M.M., Zou, F., Pankratz, V.S., Wilcox, S.L., Ma, L., Walker, L.P., Younkin, S.G., Younkin, C.S., Younkin, L.H., Bisceglio, G.D., et al. (2009). Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nature genetics 41, 192-198.
38. Zou, F., Chai, H.S., Younkin, C.S., Allen, M., Crook, J., Pankratz, V.S., Carrasquillo, M.M., Rowley, C.N., Nair, A.A., Middha, S., et al. (2012). Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. PLoS genetics 8, e1002707.
39. 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.
40. De Jager, P.L., Shulman, J.M., Chibnik, L.B., Keenan, B.T., Raj, T., Wilson, R.S., Yu, L., Leurgans, S.E., Tran, D., Aubin, C., et al. (2012). A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol Aging 33, 1017 e1011-1015.

41. 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
42. Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nature genetics 48, 1284-1287.
43. 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.
44. 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.
45. Hartung, J., Knapp, G., and Sinha, B.K. (2011). Statistical meta-analysis with applications.(John Wiley & Sons).
46. 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.
47. Fagerberg, L., Hallstrom, 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.
48. Aguet, F., Barbeira, A.N., Bonazzola, R., Brown, A., Castel, S.E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., et al. (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv, 787903.
49. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301-320.
50. Li, Q., and Lin, N. (2010). The Bayesian elastic net. Bayesian Anal 5, 151-170.


A)
B)
C)
D)
$h_e^2 = 0.05$
$h_e^2 = 0.1$
$h_e^2 = 0.2$
$h_e^2 = 0.5$
Test R^2
Power
cis−eQTL Proportion
Method
Bayesian
PrediXcan


Bayesian TWAS of AD
A)
ZC3H12B
12
10
8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
12
14
16
18
21
X
Chromosome
Bayesian TWAS of GPath
B)
ZC3H12B
8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
12
14
16
18
21
X
Chromosome


A)
Bayesian TWAS of Tangles
ZC3H12B
$-\log_{10}(p)$
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
12
14
16
18
21
X
B)
Bayesian TWAS of Amyloid
KCTD12
$-\log_{10}(p)$
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
12
14
16
18
21
X
Chromosome


A)
BVSR Results of ZC3H12B
eQTL PP
0.015
0.010
0.005
0.000
1
2
3
4
5
6
7
8
9
10
11
12
15
16
17
18
19
20
21
B)
BVSR Results of KCTD12
eQTL PP
0.06
0.04
0.02
0.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
19
20
21
22
Chromosome