A unified quantile framework reveals nonlinear heterogeneous transcriptome-wide associations

Tianying Wang,
Center for Statistical Science, Tsinghua University, Beijing, China, 100084

Iuliana Ionita-Laza*,
Department of Biostatistics, Columbia University, New York, NY 10032

Ying Wei*
Department of Biostatistics, Columbia University, New York, NY 10032

Abstract

Transcriptome-wide association studies (TWAS) are powerful tools for identifying putative causal genes by integrating genome-wide association studies and gene expression data. Most existing methods are based on linear models and therefore may miss or underestimate nonlinear associations. In this article, we propose a robust, quantile-based, unified framework to investigate nonlinear transcriptome-wide associations in a quantile process manner. Through extensive simulations and the analysis of multiple psychiatric and neurodegenerative disorders, we showed that the proposed framework gains substantial power over conventional approaches and leads to insightful discoveries on nonlinear associations between gene expression levels and traits, thereby providing a complementary approach to existing literature. In doing so, we applied the proposed method for 797 continuous traits from the UK Biobank, and the results are available in a public repository.

Some Key Words: Regression quantile process; Nonlinear association test; Transcriptome-Wide Association Studies.

Short title: Quantile TWAS

*Iuliana Ionita-Laza and Ying Wei have jointly supervised this work.

1 Introduction

Genome-wide association studies (GWAS) have been widely used to identify variants associated with complex diseases and traits. Identifying which genes are responsible for these association signals, however, remains challenging. Given that most of the associated variants are found in noncoding regions and may be involved in gene regulatory mechanisms, transcriptome-wide association studies (TWAS) (Gusev et al., 2016; Gamazon et al., 2015; Zhao et al., 2021; Xie, Shan, Zhao, and Hou, Xie et al.; Wainberg et al., 2019) have been proposed to leverage existing reference panels on genetic variation and gene expression data such as Genotype-Tissue Expression (GTEx) (Lonsdale et al., 2013; Consortium et al., 2020) to discover the potential target genes. Relative to GWAS, TWAS results identify significant genes, are less affected by linkage disequilibrium (Li et al., 2021), and provide more interpretable results, efficiently bringing us closer to a better understanding of the underlying causal mechanisms (Tang et al., 2021; Li et al., 2021).

TWAS is an integrative analysis that combines two distinct models – a gene expression model (Model A in Figure 1) that models gene expression as a function of eQTLs (i.e., SNPs that regulate gene expressions), and a GWAS model (Model B in Figure 1) that captures the associations between a trait and individual genetic variants. The two models are estimated separately and then combined to infer the association between genetically regulated gene expression and the phenotype (i.e. TWAS, represented by Model C in Figure 1). For example, the widely-used PrediXcan (Gamazon et al., 2015) first uses a sparse linear model such as elastic net or lasso to estimate the cis-eQTL effect sizes and impute gene expression levels. Then, in the second step, it formally tests the association between imputed gene expression and the trait of interest. S-PrediXcan (Barbeira et al., 2018) expands its

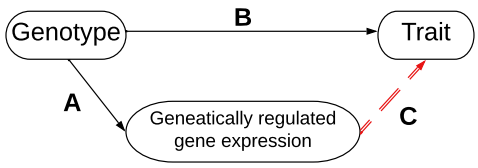


Figure 1: Quantile TWAS and classical linear TWAS models. Model A: model for SNP-gene expression association (eQTL model). Model B: model for SNP-Trait associations (GWAS model). Model C: model for expression-trait association (TWAS model).

application by only requiring GWAS summary statistics. Similarly, Gusev et al. (2016) and Nagpal et al. (2019) considered Bayesian gene expression models (Model A) instead of simple linear models in their TWAS approaches. To leverage multi-tissue data, UTMOST (Hu et al., 2019) proposed a multi-task learning method to jointly impute gene expression across tissues and then combine multiple tissue associations to obtain an overall gene-trait association.

Existing TWAS approaches are built on the fundamental assumption that both gene expression models (Model A, SNP-gene expression association) and genome-wide associations (Model B, SNP-trait association) are linear. However, there is ample evidence that suggests extensive heterogeneity of gene expression patterns, which is driven by genetic profiles, cellular and molecular heterogeneity, environmental, demographic, and technical factors (Leek and Storey, 2007; Somel et al., 2006; Budinska et al., 2013). Likewise, partly due to the existence of gene-gene and gene-environment interactions, several works including Umans et al. (2021) reported that many disease associated eQTLs are highly dynamic and context-specific in their gene regulations. Recognizing that existing linear models are insufficient to capture such heterogeneous associations, several nonlinear models have been proposed to identify heterogeneous eQTL-gene expression regulation mechanisms (Song et al., 2017; Wang et al., 2019; Geeleher et al., 2018). More recently, Lin et al. (2022) reported new TWAS discoveries by studying quadratic gene expression models.

We propose a quantile-based framework (Quantile TWAS or QTWAS) that allows us to identify nonlinear transcriptome-wide associations. It helps identify such dynamic eQTLs and uncover their links towards complex traits and disease. The key difference from the existing approach is to model the entire conditional quantile/distribution of gene expressions given their cis-variants. We then infer nonlinear TWAS association by combining such full-spectrum distributional SNP-gene expression association with GWAS summary statistics (Figure 1). We apply the proposed QTWAS framework to examine the transcriptome-wide association using the GTEx (v8) (Consortium et al., 2020) gene expression data for all protein-coding genes across 13 brain tissues, and the published GWAS summary statistics from 10 studies on psychiatric and neurodegenerative disorders. We also systematically evaluated the transcriptome-wide associations covering 797 continuous phenotypes in UK Biobank based on the gene expressions in 49 human tissues in GTEx (v8).

Our QTWAS framework has contributed to TWAS in several ways. First, it has allowed the inference of nonlinear transcriptome-wide associations. Second, the proposed framework generalizes and enhances the detection power of the existing linear-based TWAS approaches. Third, the proposed framework only requires GWAS summary statistics, and can therefore be widely applied to many studies. Fourth, the proposed framework is flexible and can assess both local and global associations. Finally, the proposed test statistics enjoy the advantages of quantile regression, such as invariance to trait transformation, leading to a more direct interpretation of the association patterns.

2 Model overview

The proposed QTWAS framework differs from existing linear model-based TWAS framework in that it tests based on imputing the distribution of gene-regulated expression, rather than

its conditional mean. Let $G \in \mathbb{R}^p$ be a set of genotypes of interest (i.e., p SNPs located within $\pm$1Mb from a gene's TSS), $E \in \mathbb{R}$ be the expression level of a target gene and E_G be its genetic component that can be explained by genetic variants, $T \in \mathbb{R}$ be the trait of interest, and $C \in \mathbb{R}^q$ be q demographic covariates such as age, gender, race, and ethnicity. Throughout the paper, we denote $Q_{E_G}(\tau)$ as the τth quantile of E_G. We propose to consider a more general model without assuming a linear association:

$$T = f(E_G) + \epsilon,$$

where $f(\cdot)$ is an unknown function and ϵ is the error term with mean zero. We are interested in testing whether E_G is significantly associated to T, that is, $H_0 : f(\cdot) = 0$ and $H_a : f(\cdot) \neq 0$. Though $f(\cdot)$ is unknown, it can be approximated by a piecewise linear function:

$$T = \sum_{k=1}^{K} (\gamma_{0,k} + \gamma_k E_G) 1\{E_G \in A_k\} + \epsilon, \tag{1}$$

where $A_k = \{Q_{E_G}(\tau_k), Q_{E_G}(\tau_{k+1})\}$ is the quantile region from the τ_kth quantile to the τ_{k+1}th quantile of E_G and $\cup_k A_k$ covers the range of E_G values. The slope coefficient γ_k assesses the localized gene-trait association within a quantile sub-region of E_G. Equivalently, we can test the null hypothesis above as follows:

$$H_0 : \gamma_k = 0 \text{ for } k = 1, ..., K; \quad H_a : \text{at least one } \gamma_k \neq 0.$$

We first train the quantile prediction model from GTEx to estimate the conditional quantile function of gene expression given a genotype profile, and then we construct test statistics that integrate GWAS summary statistics with the estimated quantiles of gene expressions (Figure 2).

Step 1: modeling the conditional quantile of $E_G \mid G$. Defining $Q_E(\tau \mid G, C)$ as the conditional quantile of gene expression E given genotypes G and covariates C, we assume

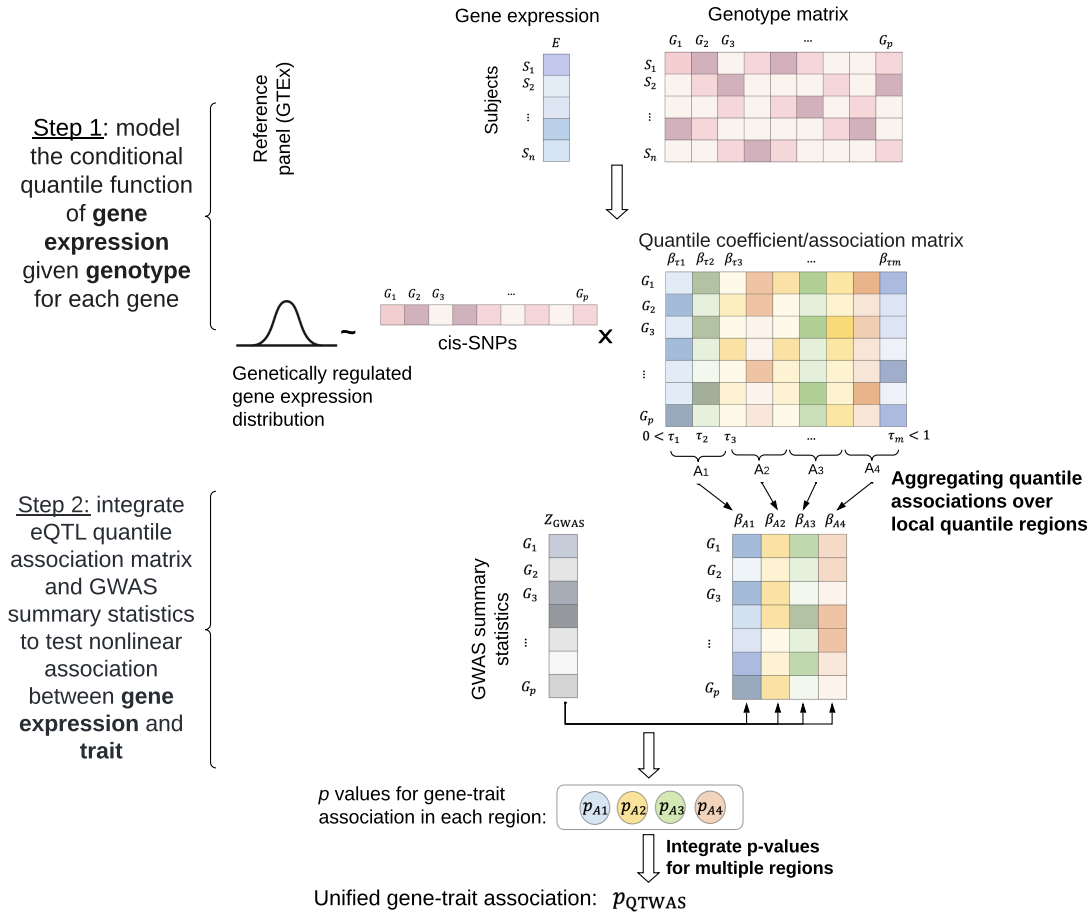


Figure 2: QTWAS flow chart.

the following linear quantile model for $Q_E(\tau \mid G, C)$ (Koenker and Bassett, 1978):

$$Q_E(\tau \mid G, C) = \alpha_0(\tau) + C^\top \alpha(\tau) + G^\top \beta(\tau) \quad \text{for all } \tau \in (0,1), \tag{2}$$

where $\alpha_0(\tau) \in \mathbb{R}$, $\alpha(\tau) \in \mathbb{R}^q$ and $\beta(\tau) \in \mathbb{R}^p$ are quantile-specific intercept and slopes for covariates and genotypes, respectively. With $U \sim \text{Unif}(0,1)$, it is important to observe from eq (2) that the gene-expression E can be viewed as a convolution of a genotype-related random variable $G^\top \beta(U)$ and an unrelated random variable $\alpha_0(U) + C^\top \alpha(U)$, and thus genotype-related gene expression E_G has the same distribution as $G^\top \beta(U)$. Denote the data in one tissue from GTEx as $\{E_i, G_i, C_i\}_{i=1}^n$ with sample size n. For a fine grid of $\tau \in (0,1)$,

we estimate $\beta(\tau)$ by solving

$$(\hat{\alpha}_{0,\tau}, \hat{\alpha}_\tau, \hat{\beta}_\tau) = \arg \min_{\alpha_{0,\tau}, \alpha_\tau, \beta_\tau} \sum_i^n \rho_\tau(E_i - \alpha_{0,\tau} - C_i\alpha_\tau - G_i\beta_\tau),$$

where $\rho_\tau(u) = |u|\{(1-\tau)I(u<0) + \tau I(u>0)\}$, $u \in \mathbb{R}$.

Step 2: testing quantile-stratified nonlinear gene-trait association. Denote $\hat{\gamma}_k$ as the estimate of γ_k for region A_k, $se(\hat{\gamma}_k)$ as the standard error of $\hat{\gamma}_k$, and N_{GWAS} as the sample size for GWAS data. Given the conditional independence assumption $T \perp\!\!\!\perp E_G \mid G$ (Barbeira et al., 2018; Hu et al., 2019), the z score for gene-trait association for region A_k is

$$Z_k = \frac{\hat{\gamma}_k}{se(\hat{\gamma}_k)} \approx \frac{\hat{\gamma}_k}{se(\hat{\gamma}_{A_k})} \approx \frac{\hat{\beta}_{A_k}^\top}{\sigma_{E_{A_k}}} \begin{pmatrix} \hat{\sigma}_1 & & \\ & \ddots & \\ & & \hat{\sigma}_p \end{pmatrix} \widetilde{Z},$$

where $\sigma_{E_{A_k}}$ is the standard deviation of imputed gene expression in region A_k, $\hat{\beta}_{\text{GWAS}}$ is the SNP-level estimated effect size available from GWAS summary statistics, $\hat{\sigma}_j$ is the estimated standard deviation of the j-th SNP, $\widetilde{Z}$ is the SNP-trait z scores, and $\hat{\beta}_{A_k} = \int_{\tau_k}^{\tau_{k+1}} \hat{\beta}(\tau) d\tau$, in which τ_k and τ_{k+1} define the range of A_k (Methods).

If there is no SNP-trait association, $\widetilde{Z} \approx N(0, D)$, where D is the linkage-disequilibrium matrix for SNPs, either estimated from the reference panel or external reference data. Hence, under $H_0 : \gamma_k = 0$ for $k = 1, ..., K$,

$$Z_k \approx N\left(0, \frac{1}{\sigma^2_{E_{A_k}}} \hat{\beta}_{A_k}^\top \text{diag}(\hat{\sigma}_1, ..., \hat{\sigma}_p) D \text{diag}(\hat{\sigma}_1, ..., \hat{\sigma}_p) \hat{\beta}_{A_k}\right).$$

After getting the p values for each region A_k, we combine them by the Cauchy combination method (Liu and Xie, 2020), which avoids time-consuming numerical permutations and provides an analytical form for combining correlated p-values.

Model implementation and evaluation

Variant screening procedure. Based on empirical evidence, E_G often depends on a sparse set of SNPs (Barbeira et al., 2018; Gamazon et al., 2015). Most existing TWAS approaches use penalized linear regression to select significant SNPs associated with the mean of E_G (Barbeira et al., 2018; Gamazon et al., 2015; Hu et al., 2019), which may not be optimal at identifying more local (quantile-stratified) associations. We introduce a new variant screening procedure to identify important SNPs separately for each region A_k. The new screening procedure is more effective at identifying heterogeneous distributional associations and non-normal outcomes. We outline the detailed algorithm below, and provide a flowchart of the screening procedure in the Appendix (Figure 1). Specifically, we aggregate multiple quantile rank score tests (Gutenbrunner et al., 1993) at selected quantile levels within A_k to select region-specific SNPs. For each variant located within ±1Mb of a gene's TSS, we first perform quantile rank score test (Gutenbrunner and Jurecková, 1992; Song et al., 2017) on several quantile levels selected from the target quantile intervals A_k's and then combine multiple p values using the Cauchy combination method (Liu and Xie, 2020). We select significant variants while controlling the false discovery rate at the 5% level using the method of Benjamini and Hochberg (1995). Among the selected SNPs, we further filtered out highly correlated SNPs via hierarchical clustering (more details are described in Appendix 1) and used the final set of SNPs as the G matrix used in model (2).

Evaluation of imputation accuracy. To evaluate the accuracy of the imputation model (2), we consider the goodness of fit criterion $R^Q(\tau) = 1 - \hat{V}(\tau)/\tilde{V}(\tau)$ (Koenker and Machado, 1999), a measurement of explained deviance by the quantile model associated to genetic effects at a fixed quantile level, where $\hat{V}(\tau) = \min \sum_i^n \rho_\tau(E_i - C_i^\top \alpha_\tau - G_i^\top \beta_\tau - \alpha_{0,\tau})$ and

$\tilde{V}(\tau) = \min \sum_i^n \rho_\tau(E_i - C_i^\top \alpha_\tau - \alpha_{0,\tau})$ are optimized quantile loss under the null and alternative models, respectively. It is a natural analog to R^2 in linear models. We integrate $R^Q(\tau)$ over the interval A_k, resulting in an average model adequacy in the quantile region A_k. The smaller the value, the less evidence for a meaningful SNP-expression association within the quantile region A_k. In practice, we screen out unreliable models (e.g., quantile models in a specific region A_k) with imputation accuracy $R^Q_{A_k} \leq 0.1$ and replace their p values by $u \sim \text{Unif}(0, 1)$. Then, the Cauchy combination method (Liu and Xie, 2020) is used to combine p values from multiple regions A_k's.

Implementation details in the GTEx data. We trained the gene expression prediction model for 49 tissues from the GTEx project (v8) as described below. Gene expression levels were normalized and adjusted for covariates, including sex, sequencing platform, and the top five principal components of genotype data, as well as the top 15 probabilistic estimation of expression residuals (PEER) factors (Hu et al., 2019; Stegle et al., 2010). We considered protein-coding genes, removed ambiguously stranded SNPs, and only considered ref/alt pairs A/T, C/G, T/A, G/C. SNPs with minor allele frequency less than 0.01 were excluded from the analyses. For each gene, we used SNPs between 1Mb upstream and downstream of the transcription start site. The LD matrix D is estimated from the GTEx data. The length of A_k and the number of regions (K) can be tuned depending on the application. For demonstration, we consider $K = 4$ regions: A_1 for $\tau \in (0.05, 0.3)$, A_2 for $\tau \in (0.3, 0.5)$, A_3 for $\tau \in (0.5, 0.7)$ and A_4 for $\tau \in (0.7, 0.95)$ such that $\cup_{k=1}^K A_k$ covers 5% percentile to 95% percentile of the value of E_G. We do not consider $\tau < 0.05$ and $\tau > 0.95$ as they are close to the tails, and the estimates tend to have larger variance owing to inflated zeros in G. In addition, we estimate β_{A_k} by a quantile region slightly larger than A_k for more numerically

stable results (e.g., using $\beta_\tau \mid \tau \in (0.05, 0.35)$ to estimate β_{A_1}). The length of the A_k and the number of regions K may affect the power of the QTWAS test statistic slightly in certain cases. But, given our experience, QTWAS is robust, and splitting the quantile range $(0, 1)$ into smaller regions does not help improve the power significantly.

3 Simulation studies

3.1 Simulation settings

For simulations, we use genotype data on $n = 670$ individuals (from whole blood tissue in GTEx v8) in order to generate gene expression levels, and then we resample (from whole blood tissue in GTEx v8) $n = 1000$ subjects to generate the trait data from their genotypes.

Null model. We generate the gene expression E from the model: $E = G^\top \beta + C^\top \alpha + e$, in which β is estimated based on true GTEx data via elastic net with the tuning parameter set as 0.5. For each gene, G includes all SNPs within $\pm 1Mb$ from its TSS. The trait T is generated by $T = C^\top \eta + \epsilon$. Both error terms e and ϵ follow standard normal distribution. Gene expression E is normalized before analysis. The set of covariates C is provided in GTEx data, including the top five principal components, top 15 PEER factors, platform, and sex. α and η are vectors with each element randomly drawn from Unif$(0, 1)$. This null model preserves the gene regulation from GTEx data with no gene-trait association. A similar setting has been applied in Hu et al. (2019).

Alternative models. For power analyses, we consider three different SNP-gene expression models in the reference panel, and we assume a linear SNP-Trait model in GWAS data (since we use available GWAS summary statistics).

SNP-Expression models. We consider the following three models:

(a) Location shift: $E = G^\top \beta + C^\top \alpha + e$;

(b) Location-scale: $E = G^\top \beta + C^\top \alpha + (1 + 0.5 G^\top \beta) e$;

(c) Local signal: $Q_E(\tau > 0.7) = 5\frac{\tau - 0.7}{1 - 0.7} G^\top \beta + C^\top \alpha + F_e^{-1}(\tau)$.

In the location shift model (a), the genetic variants G only affect the mean of E, while in the location-scale model (b), the genetic variants G affect both the mean and variance of E. In the local signal model (c), the variants G only affect part of the distribution of E (i.e. G only affect the upper quantile (> 0.7th quantile) of E). In each of the three scenarios, we consider two error distributions for e: standard normal and Cauchy distributions. Under these alternative models, we show that TWAS is nonlinear when the SNP-expression association is heterogeneous (Appendix Figure 2).

SNP-Trait model. We consider a simple linear model $T = G^\top \zeta + C^\top \eta + \epsilon$, where ϵ follows a standard normal distribution, and genetic variants G are generated by resampling from GTEx v8 genotype data.

To illustrate the performance in different scenarios, we randomly select 1% of SNPs from the 2Mb region around TSS to be causal (i.e., with non-zero effect sizes β and ζ). We set $\zeta = \mathbf{1}_p$ and $\beta = 2 \cdot \mathbf{1}_p$ for local signal model, $\zeta = 0.2 \cdot \mathbf{1}_p$ and $\beta = 0.4 \cdot \mathbf{1}_p$ for location-scale model, and $\zeta = 0.1 \cdot \mathbf{1}_p$ and $\beta = 0.2 \cdot \mathbf{1}_p$ for location shift model, where $\mathbf{1}_p$ represents a column vector with all elements being 1. Gene expression E is normalized before analysis. α and η are vectors with each element randomly drawn from $\text{Unif}(0, 1)$. C is the same set of covariates as in the null model.

Similar to Hu et al. (2019), we randomly select 500 genes, and generate expression data and traits independently for each gene, as described above. For power analyses, we repeat the data generation procedure two times per gene, and report the statistical power based on

1,000 replicates at the significance threshold $\alpha = 2.5e\text{-}6$. For type I error analysis, we repeat the procedure for each gene 20,000 times, and report the results based on 10^7 replicates at different significance thresholds ranging from 0.05 to $2.5e\text{-}6$. In addition, we compare the proposed framework with our own implementation of S-PrediXcan (note that we have re-implemented S-PrediXcan as it needs to be trained based on simulated data, and we denote it as "S-PrediXcan*"). For our method, we report results integrated across quantile levels (denoted as "QTWAS") as well as quantile region stratified results, such as "QTWAS$_{A_k}$" for $k = 1, ..., 4$. The choices of $A_1, ..., A_4$ are described in the implementation details in Section 6. For both integrated and region-specific QTWAS, we only report results for genes with $R^Q_{A_k} > 0.1$, as previously mentioned. Similarly, for S-PrediXcan*, we only report results for genes with imputation $R^2 > 0.1$. We randomly generate $p \sim \text{Unif}(0, 1)$ if the elastic net model in S-PrediXcan* does not select any variables, or if none of the four regions A_1-A_4 in QTWAS has valid p value (e.g., no variant is selected or the quantile model does not pass the imputation accuracy check).

3.2 Simulation results

Type I error analysis. The type I error for both QTWAS and S-PrediXcan* are controlled at all significance levels (Table 1).

α	S-PrediXcan*	QTWAS	QTWAS$_{A1}$	QTWAS$_{A2}$	QTWAS$_{A3}$	QTWAS$_{A4}$
0.05	5.031E-02	5.001E-02	5.005E-02	5.000E-02	4.990E-02	4.996E-02
1e-2	1.008E-02	1.002E-02	1.003E-02	1.002E-02	9.969E-03	1.000E-02
1e-3	1.004E-03	1.005E-03	1.022E-03	1.029E-03	1.012E-03	1.016E-03
1e-4	1.009E-04	1.022E-04	1.044E-04	1.054E-04	1.034E-04	1.066E-04
1e-5	8.100E-06	1.060E-05	1.000E-05	9.800E-06	1.060E-05	1.170E-05
2.5e-6	2.100E-06	2.400E-06	2.200E-06	3.100E-06	2.800E-06	2.400E-06

Table 1: Type I error results for S-PrediXcan* and QTWAS as well as for quantile region stratified QTWAS with A_1-A_4 based on 10^7 replicates.

Power analyses. QTWAS, combining different quantile intervals, has improved power in most scenarios compared to S-PrediXcan* (Figure 3). For the location shift models, S-PrediXcan* and QTWAS have comparable power with normal errors, while QTWAS has slightly higher power than S-PrediXcan* with Cauchy errors. For the location-scale models, QTWAS showed substantially improved power over S-PrediXcan*, with its power gain stemming from the stratified association QTWAS_{A_4} in the setting of normal errors and from QTWAS_{A_3} and QTWAS_{A_4} in the setting of Cauchy errors. For local signal models, we observed dominant power boost for QTWAS, owing to the power of QTWAS_{A_4}, corresponding to the true signals simulated at the top quantiles (Appendix Figure 2 (c)). Therefore, the region-specific quantile test statistic can reveal more complex and detailed association patterns.

3.3 Additional comparisons with S-PrediXcan

Imputation accuracy. To compare the imputation accuracy for QTWAS and S-PrediXcan*, we plotted the average of $R^Q_{A_k}$ over four regions against R^2 (Figure 4). Except for the location shift model with normal error, QTWAS generally explained more deviance than S-PrediXcan*. Specifically, in the location-scale and local signal models, S-PrediXcan* explains low deviance, indicating poor goodness of fit using linear models.

Evaluation of variant screening procedure. To evaluate the quantile variant screening, we measure the canonical correlation between selected sets and the causal set in the three alternative models with normal errors. The proportion of replicates with correlation greater than 0.95 is reported in Table 2 based on 1,000 replicates. For the location shift model, both QTWAS and S-PrediXcan* select SNP sets highly correlated with the true causal set. In the location-scale model, the proportion of replicates selecting highly correlated SNPs

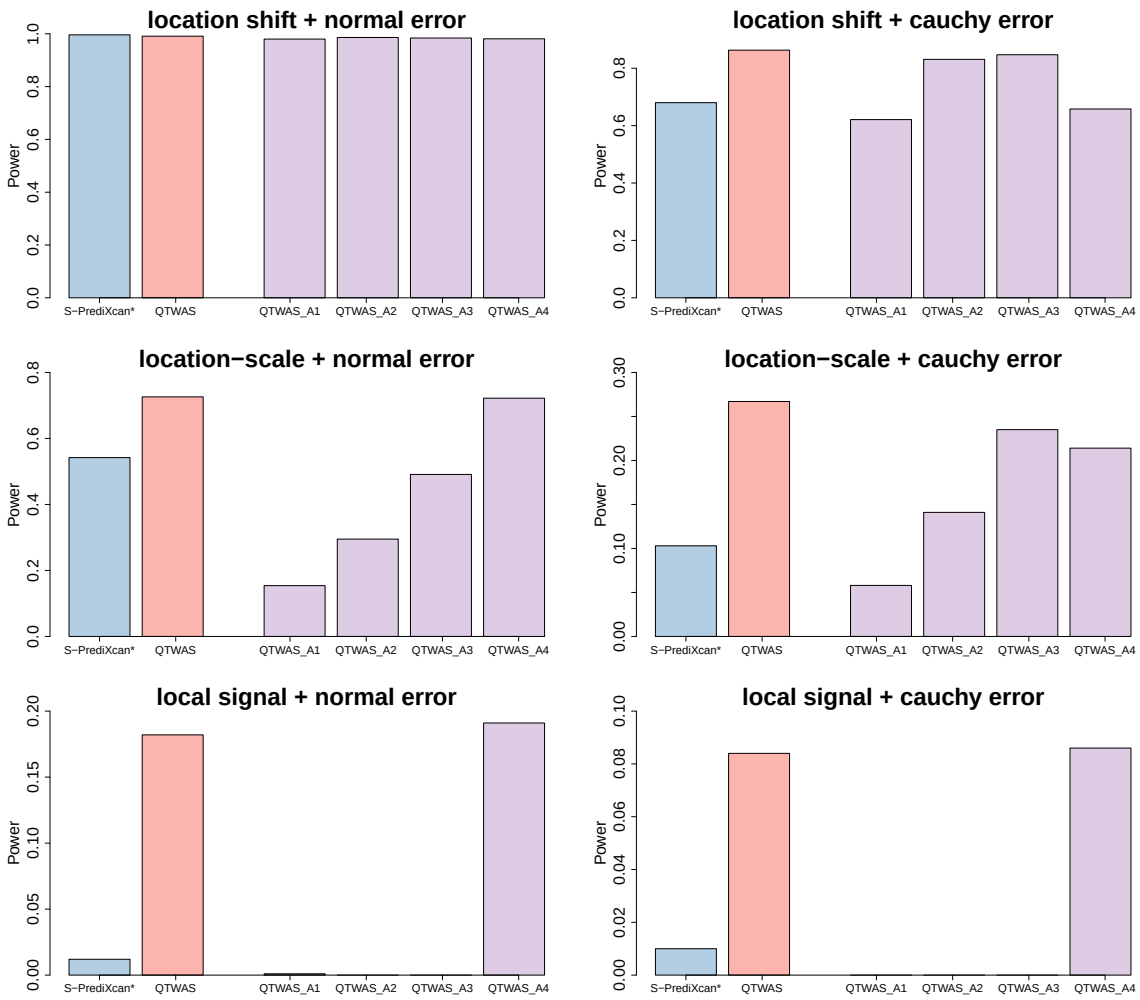


Figure 3: Power of S-PrediXcan* and QTWAS in location shift (top panel), location-scale (middle panel), and local signal (bottom panel) models with normal error (left column) and Cauchy error (right column), respectively. In each panel power is shown for S-PrediXcan*, QTWAS, and quantile region stratified QTWAS with A_1-A_4. The significance threshold is $\alpha = 2.5e$-6.

is increasing with the quantile levels and is comparable to S-PrediXcan* for interval A_4, consistent with the power results (Figure 3). In the local shift model, QTWAS selected a set of SNPs with high correlation with the true causal set in the upper quantile interval A_4 more often than S-PrediXcan*, as expected. Thus, we have demonstrated that QTWAS can select variants that are more correlated to the underlying causal variants.

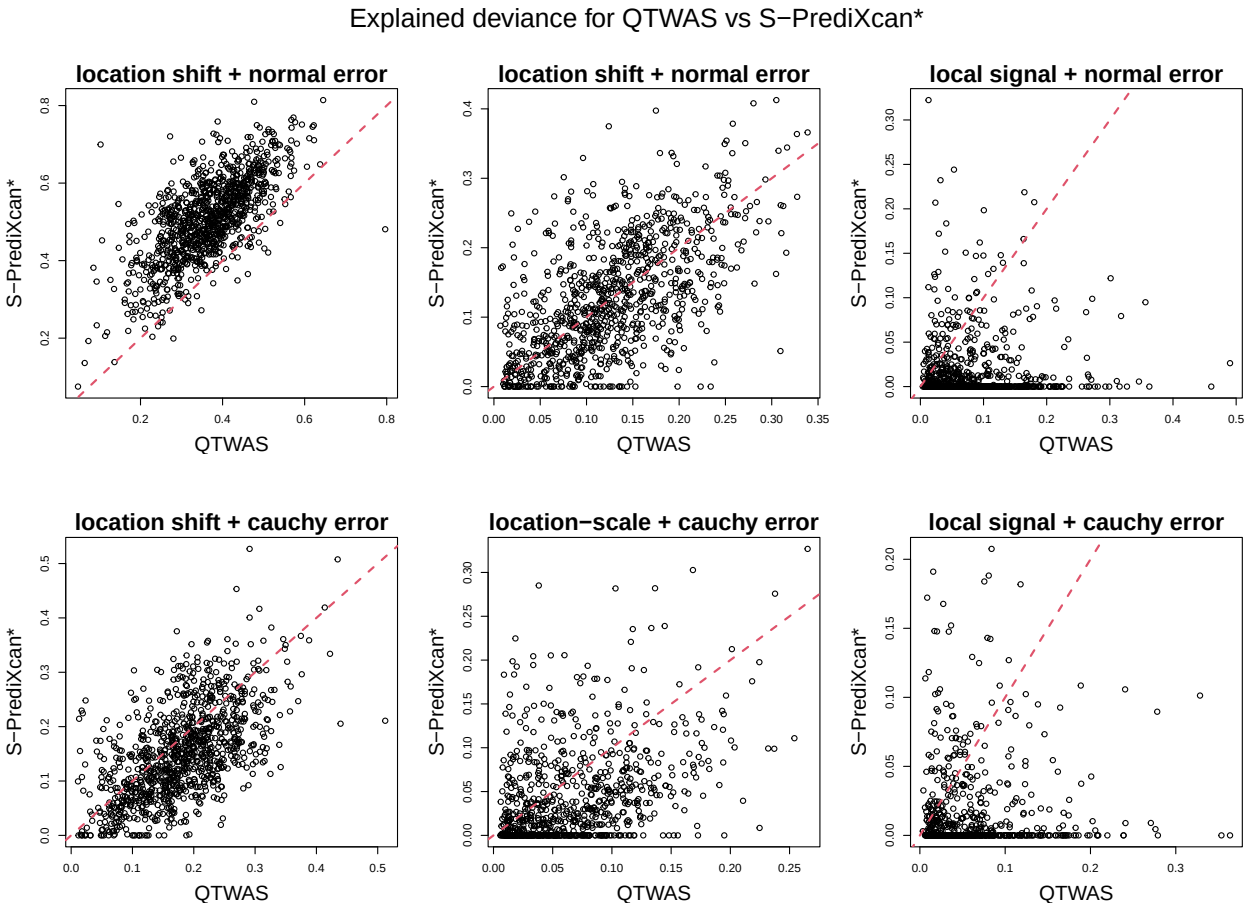


Figure 4: Explained deviance for QTWAS and S-PrediXcan*. Location shift, location-scale and local signal models are presented, with normal errors and Cauchy errors.

Model	QTWAS$_{A1}$	QTWAS$_{A2}$	QTWAS$_{A3}$	QTWAS$_{A4}$	S-PrediXcan*
location shift	97.9%	99.1%	99.1%	99.0%	99.8%
location-scale	31.3%	54.2%	70.4%	83.7%	85.4%
local shift	4.5%	0.5%	0.5%	39.1%	15.4%

Table 2: The proportion of replicates (out of 1,000) with canonical correlation values between the selected variable set and the true causal set greater than 0.95.

4 Applications to psychiatric/neurodegenerative disorders

We apply QTWAS and S-PrediXcan to summary statistics from ten GWAS studies on brain disorders, including five neuropsychiatric traits: schizophrenia (SCZ, (Pardiñas et al., 2018)),

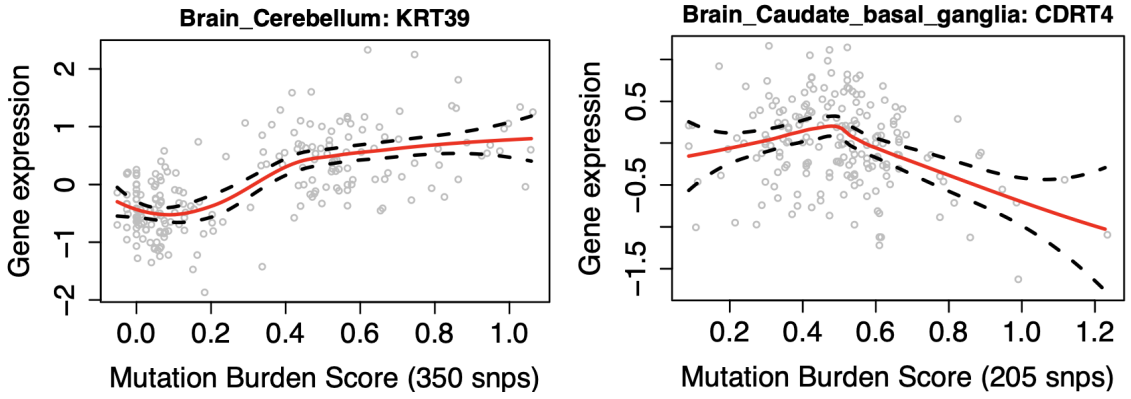
attention-deficit/hyperactivity disorder (ADHD, (Demontis et al., 2019)), bipolar disorder (BD, (Stahl et al., 2019)), autism spectrum disorder (ASD, (Grove et al., 2019)) and major depressive disorder (MDD, (Howard et al., 2019)); and four neurodegenerative traits: Alzheimer's disease (AD_Kunkle, (Kunkle et al., 2019); AD_Jansen, (Jansen et al., 2019)), Parkinson's disease (PD, (Nalls et al., 2019)), multiple sclerosis (MS, (Andlauer et al., 2016)) and amyotrophic lateral sclerosis (ALS, (Van Rheenen et al., 2016)). Sample information for those studies is summarized in the Appendix (Table 1).

For each trait, we applied S-PrediXcan and QTWAS to 13 brain tissues from GTEx v8 (namely Brain Cortex, Hippocampus, Cerebellum, Frontal Cortex (BA9), Hypothalamus, Nucleus accumbens (basal ganglia), Spinal cord (cervical c-1), Substantia nigra, Amygdala, Anterior cingulate cortex (BA24), Caudate (basal ganglia), Cerebellum Hemisphere and Putamen (basal ganglia)). We focus on genes with $R^Q_{A_k} > 0.1$ for the imputation accuracy. For S-PrediXcan, we use the pre-trained models directly downloaded from the PredictDB website. Then, we use the Cauchy combination method (Liu and Xie, 2020) to combine the 13 individual tissue p values. We validate our discoveries using a spatiotemporal transcriptomic dataset (Kang et al., 2011), and a single-cell transcriptomic dataset (Wang et al., 2018).

Single-tissue results for AD. We use the trait AD_Jansen (Jansen et al., 2019) to illustrate the performance based on single-tissue data. Through the Quantile-Quantile (QQ) plots per tissue (Appendix Figure 3), we show that QTWAS is more powerful than S-PrediXcan while controlling the type I error. We further summarize the number of genes identified by QTWAS and S-PrediXcan in each tissue with the imputation accuracy thresholds $R^Q_{A_k} > 0.1$ (default), 0.2, and 0.3 (Table 2). For all tissues, QTWAS with $R^Q_{A_k} > 0.1$ identified more significant gene-trait associations compared to S-PrediXcan. QQ plots for other traits are

shown in the Appendix (Figure 4-12). Full results on significant genes for all ten traits per tissue can be found at `https://tianyingw.github.io/QTWAS/`.

Visualizing the nonlinear association in GTEx data. Owing to the lack of individual genotype in GWAS data, we propose to investigate the genotype-expression model (model B in Figure 1) to gain more insights on the nonlinear gene-trait associations. Specifically, we compute a mutation burden score per gene for each individual by adding up the minor alleles of reported cis-QTL from the GTEx portal website (denoted as X) and the residuals from the linear model $E \sim C$ (denoted as Y), and then we fit a lowess smoothing model on Y and X. For illustration purposes, we present two genes only identified by QTWAS for the trait AD_Jansen, namely *KRT39* and *CDRT4*. The association between Alzheimer's Disease and *CDRT4* has been previously reported in literature (Liu et al., 2021). *KRT39* has also been found as an aging-related protein (Ma et al., 2020). After accounting for covariate effects, lowess estimated curves clearly show nonlinear associations between the gene expression and the mutation burden score (Figure 5).



(a) (b)

Figure 5: Examples of nonlinear SNP-expression associations: (a) gene *KRT39* and brain cerebellum tissue, and (b) gene *CDRT4* and brain caudate basal ganglia tissue. The red line is the lowess fitting curve; black dotted lines are 95% confidence band; grey points are observed data.

Combining multiple-tissue signals. As the sample sizes in GTEx data vary across different tissues, the tissues providing the smallest p values are not necessarily the most relevant for a given trait. Thus, we combined results from 13 brain tissues using the Cauchy combination approach for a more comprehensive analysis. QQ plots for all ten traits based on 13 brain tissues results are presented in the Appendix (Figure 13). For all traits, 95% of the distributions of $-\log_{10} p_{\text{QTWAS}}$ and of $-\log_{10} p_{\text{S-PrediXcan}}$ are consistent with each other, whereas QTWAS is more powerful than S-PrediXcan on the top 5% genes. We found that multi-tissue results are more robust, with a substantially larger number of significant genes identified by both QTWAS and S-PrediXcan (Figure 6). Again, we use AD_Jansen for illustration. Both QTWAS and S-PrediXcan identified strong associations on chromosome 19, while QTWAS also reported relatively strong signals on chromosome 6 and new significant genes at several other locations (Figure 7). Similarly, QTWAS shows improved power in detecting gene-trait associations for other traits (Appendix Figure 14-22).

QTWAS identified more loci than S-PrediXcan. Based on the GWAS summary statistics from the ten traits we investigated, we select genome-wide significant SNPs (p values $< 2.5e$-08) and define loci as $\pm$ 500kb centered at significant SNPs (to limit overlap, only the locus corresponding to the most significant SNP is retained). Next, we report the number of loci identified by QTWAS and S-PrediXcan, i.e. loci with a significant gene (TSS) within the locus, based on multi-tissue results from 13 brain tissues. We observe that QTWAS identified more new loci (Figure 8).

Validating the new gene discoveries by QTWAS

Next we show validation results for the set of genes identified by QTWAS only.

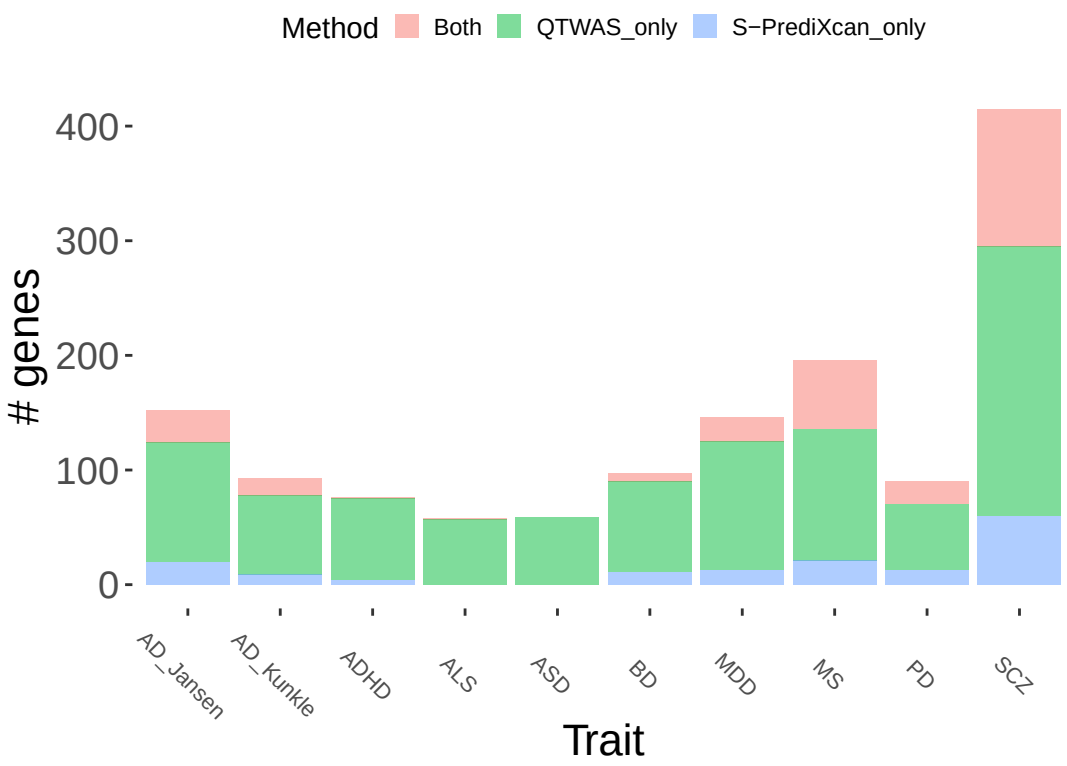


Figure 6: Number of genes identified by QTWAS and S-PrediXcan for ten traits based on 13 brain tissues. "QTWAS_only" represents the number of genes only identified by QTWAS; "S-PrediXcan_only" represents the number of genes only identified by S-PrediXcan; "Both" represents the number of genes identified by both methods. The significance threshold is 2.5e-6.

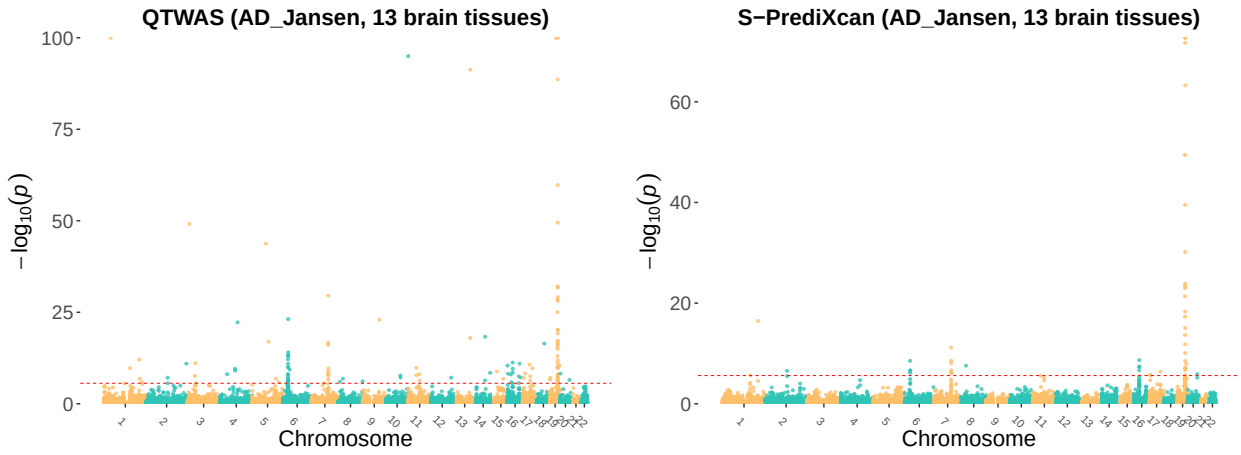


Figure 7: Manhattan plots for QTWAS and S-PrediXcan based on 13 brain tissues for study AD_Jansen. The significance threshold is 2.5e-6.

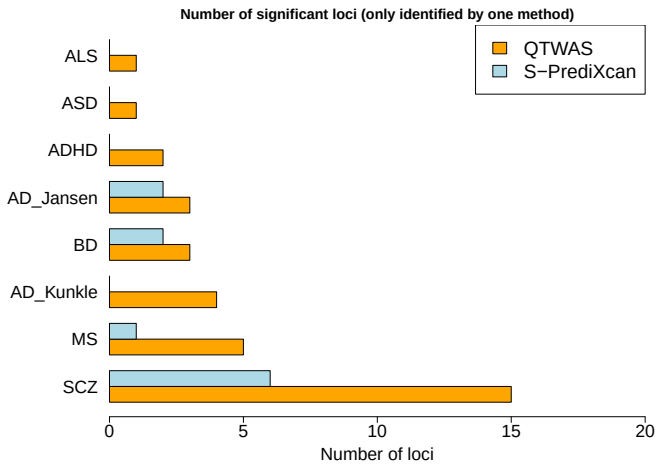


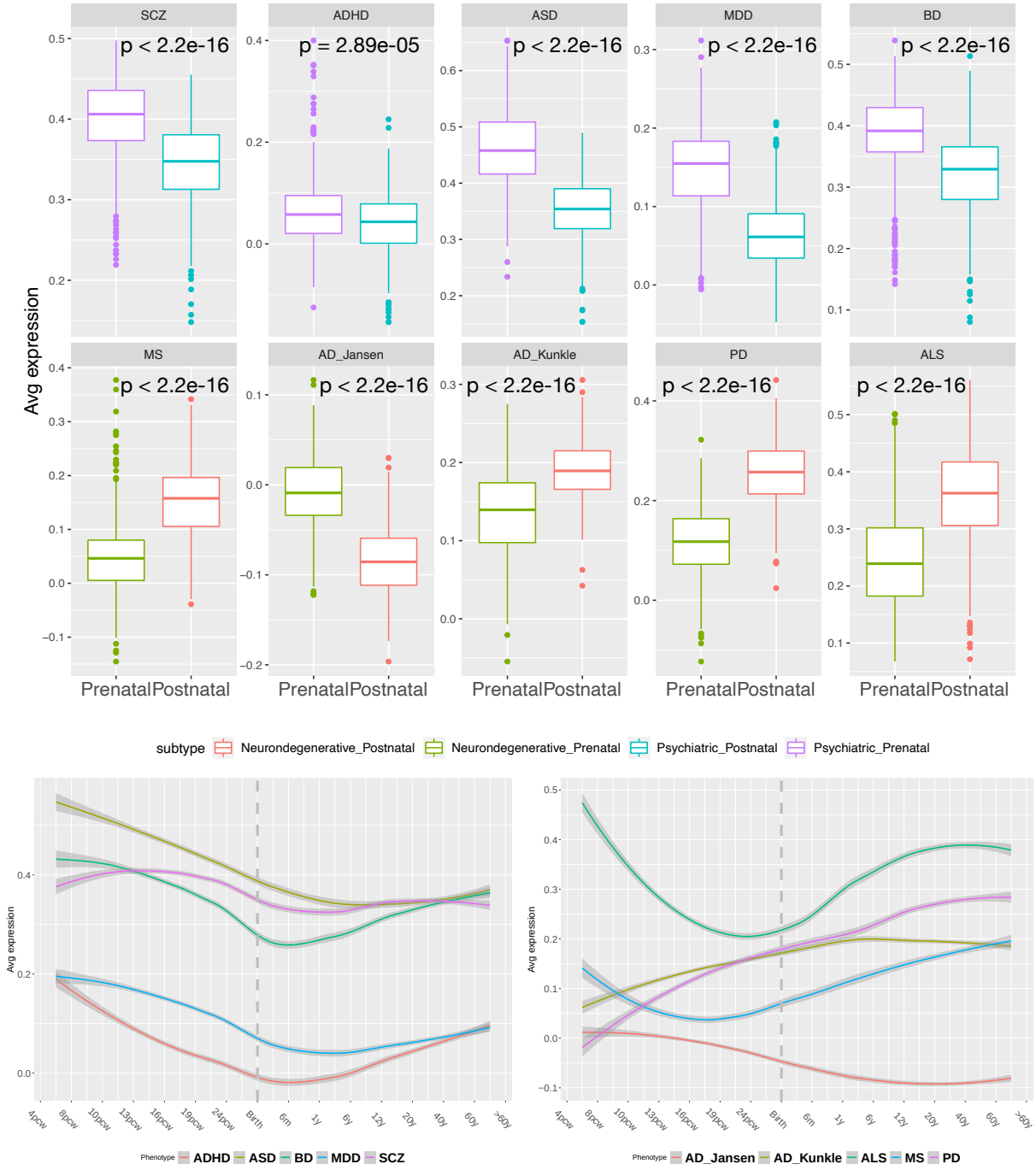
Figure 8: The number of loci only identified by either QTWAS or by S-PrediXcan.

Developmental transcriptomic data. We use spatiotemporal brain transcriptomic data collected from clinically unremarkable donors without signs of large-scale genomic abnormalities (Kang et al., 2011). The data are available for six brain regions: neocortex (NCX), mediodorsal nucleus of the thalamus (MD), cerebellar cortex (CBC), hippocampus (HIP), amygdala (AMY), and striatum (STR). Here we focus on the cortical expression profiles (NCX area), as in Ma et al. (2021), because of its relatively large sample size with 410 prenatal samples and 526 postnatal samples. The prenatal samples correspond to measurements collected from 4 postconceptional weeks (PCW) till birth, and the postnatal samples correspond to those collected from birth to age ≥ 60 years (Kang et al., 2011). The data have been quantile normalized, and we further center the gene expression within each sample. As some GWAS studies are underpowered, a liberal significance threshold of 10^{-4} is used. We then compute the average gene expression levels of significant genes only identified by QTWAS based on multi-tissue results (13 brain tissues) and compare the expression patterns between prenatal and postnatal stages for each trait. We observe significantly higher prenatal expressions among all five psychiatric disorders, and higher postnatal expressions among all five neurodegenerative disorders except for AD_Jansen (Figure 9), consistent with results

reported in Sey et al. (2020) and Ma et al. (2021). When performing the same analyses but with all genes identified by QTWAS, we find similar patterns (Appendix Figure 24). These results support the validity of the new discoveries from QTWAS.

Single cell expression profiles. We further explore the cell-type specific expression profile for 285 single cells from 6 adult brain cell-types, including neurons (131 cells), astrocytes (62 cells), microglia (16 cells), endothelial (20 cells), oligodendrocytes (38 cells) and oligodendrocyte progenitor cells (OPC, 18 cells) from Wang et al. (2018). We standardized the expressions for the six adult cell-types, then calculated the average expression level based on significant genes across multiple cells within each cell type. We focused on those genes only identified by QTWAS (not identified by S-PrediXcan). We observed different cellular signatures and cell-type specificity, consistent with existing studies (Figure 9). Specifically, neurons are important for all psychiatric traits. Similarly, microglia are important for AD (Sarlus et al., 2017; Ma et al., 2021; Sey et al., 2020; Doens and Fernández, 2014); neurons and oligodendrocytes have been recognized for PD (Azevedo et al., 2022; Sey et al., 2020); and neurons are important for ALS (Sey et al., 2020). We also perform the same analyses with all genes identified by QTWAS and obtain similar results (Appendix Figure 23).

Browser for results on UKBiobank traits In addition, we have applied QTWAS to 797 UK Biobank continuous phenotypes with their summary statistics for 28 million imputed variants (available at https://pan.ukbb.broadinstitute.org). We provide phenome-wide results for genes, genome-wide gene-based results for each trait with respect to all 49 tissues in GTEx. Results can be obtained from https://tianyingw.github.io/QTWAS/.



(a) Psychiatric disorders (b) Neurodegenerative disorders

Figure 9: Human brain developmental expression of significant genes identified by **QTWAS only**. P-values of Wilcoxon rank sum tests comparing prenatal and postnatal expression levels are also shown.

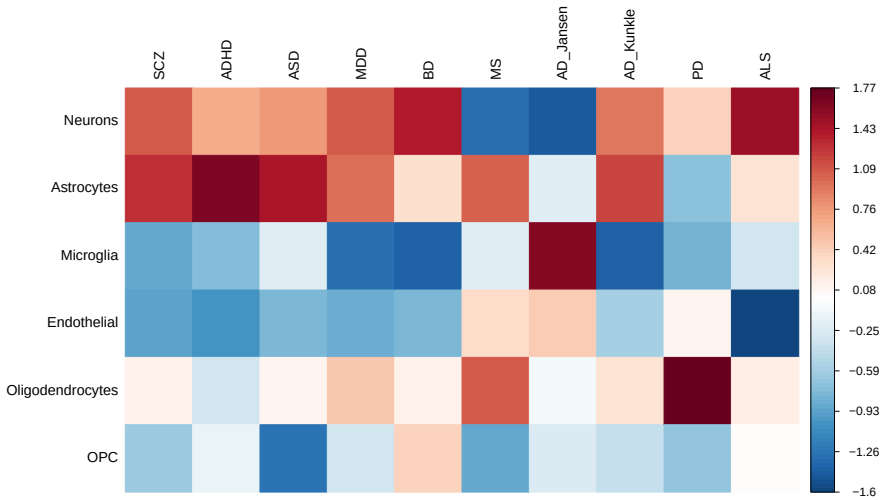


Figure 10: Cell-type expression profiles of significant genes identified by **QTWAS only** based on 13 brain tissues. Each row represents a cell type, and each column represents a trait.

5 Discussion

We have proposed a novel quantile-based TWAS approach to effectively leverage expression reference panels and discover nonlinear gene-trait associations. Quantile regression tools have been previously proposed in the context of genetic association studies (Song et al., 2017; Wang et al., 2019). However, this is the first quantile-process-based method in TWAS. A key reason for its appeal is the promise of prioritizing candidate causal genes whose expression levels mediate the phenotypic effects in a dynamic and non-linear manner. As shown in both simulations and applications, such quantile models are able to identify more associations and provide more insights on how gene expression levels regulate phenotypes. In addition, the interval-based variant screening through quantile rank test leads to a more accurate and detailed imputation of gene expression in specific quantile intervals.

Compared to the mean-based approach S-PrediXcan, QTWAS showed more robust and promising discoveries across different traits. In general, a large number of genes discovered by

S-PrediXcan are also identified by QTWAS. Through validation analyses, we show that the novel genes identified by QTWAS are likely to be functional and relevant to the trait under study. In addition, QTWAS is more powerful, discovering more new loci and more genes per shared locus. Note that, like other TWAS approaches, QTWAS inevitably identifies multiple hit genes per locus due to LD confounding and co-regulation (Wainberg et al., 2019), and hence fine-mapping such loci can further prioritize relevant genes at each locus.

Several emerging topics in TWAS are left for future work. Constructing imputation models based on multi-tissues, such as UTMOST (Hu et al., 2019), can effectively increase imputation accuracy and power. Multi-tissue quantile modeling may allow investigations of more comprehensive nonlinear graphical associations across tissues. Furthermore, the current QTWAS framework can be better developed when individual GWAS data are available, which would allow nonparametric approaches to estimate higher resolution nonlinear gene-trait associations.

6 Methods

Problem statement

Denote $G \in \mathbb{R}^p$ as a set of genotypes of interest (i.e., p SNPs located within $\pm$1Mb from a gene's TSS), $E \in \mathbb{R}$ as the expression level of a target gene, $T \in \mathbb{R}$ as the trait of interest, $C \in \mathbb{R}^q$ as q demographic covariates, and $Q_{E_G}(\tau)$ as the τth quantile of E_G. Assuming that a gene expression level E is determined by both the genetic profile and other demographic and environmental confounders, we decompose it by $E = E_G + E_C$, where E_G is the genetic component of gene-expression that can be explained by genetic variants, while E_C is the non-genetic component. The general TWAS goal is to establish an association between the phenotype T and E_G.

Typical TWAS approaches start by building a linear prediction model of gene expression given G and C and then by testing the association between a trait and the predicted E_G in GWAS data. Such approaches implicitly assume linear associations in TWAS. Here, we consider the following model:

$$T = \sum_{k=1}^{K} (\gamma_{0,k} + \gamma_k E_G) 1\{E_G \in A_k\} + \epsilon, \tag{3}$$

where $A_k = \{Q_{E_G}(\tau_k), Q_{E_G}(\tau_{k+1})\}$ is the quantile region from the τ_kth quantile to the τ_{k+1}th quantile of E_G. Eq (3) is a piecewise linear function approximating $T = f(E_G) + \epsilon$ with an unknown $f(\cdot)$.

By design, we approximate $f(x)$ in the interval A_k by a linear function $\gamma_{0,k} + \gamma_k x$, such that $(\gamma_{0,k}, \gamma_k) = \arg\min_{\boldsymbol{\gamma}} \mathbb{E}_T (T - \gamma_0 - \gamma_1 E_G)^2 1\{E_G \in A_k\}$. The slope coefficient γ_k assesses the localized gene-trait association within a quantile sub-region of E_G, which can be written as

$$\gamma_k = \frac{cov(E_G, T \mid E_G \in A_k)}{var(E_G \mid E_G \in A_k)}, \text{ for } k = 1, ..., K, \tag{4}$$

and the null hypothesis is $H_0 : \gamma_k = 0$ for $k = 1, ..., K$; H_a : at least one $\gamma_k \neq 0$. The coefficient γ_k can be estimated by leveraging a conditional quantile process model of the gene expression and GWAS summary statistics. The key idea is to decompose the covariance $cov(E_G, T \mid E_G \in A_k)$ in (4) by the law of total variance:

$$\begin{aligned} cov(E_G, T \mid E_G \in A_k) &= \mathbb{E}\{cov(E_G, T \mid G, E_G \in A_k)\} \\ &\quad + cov\{\mathbb{E}(E_G \mid G, E_G \in A_k), \mathbb{E}(T \mid G, E_G \in A_k)\}. \end{aligned} \tag{5}$$

Assuming the conditional independence $T \perp\!\!\!\perp E_G \mid G$ as in Barbeira et al. (2018) and Hu et al. (2019), the covariance is determined by the correlations between $\mathbb{E}(E_G \mid G, E_G \in A_k)$ and $\mathbb{E}(T \mid G, E_G \in A_k)$.

Step 1: modeling the conditional quantile of $E_G \mid G$

Defining $Q_E(\tau \mid G, C)$ as the conditional quantile of gene expression E given genotypes G and covariates C, we assume the following linear quantile model for $Q_E(\tau \mid G, C)$ (Koenker and Bassett, 1978):

$$Q_E(\tau \mid G, C) = \alpha_0(\tau) + C^\top \alpha(\tau) + G^\top \beta(\tau) \quad \text{for all } \tau \in (0, 1), \tag{6}$$

where $\alpha_0(\tau) \in \mathbb{R}$, $\alpha(\tau) \in \mathbb{R}^q$ and $\beta(\tau) \in \mathbb{R}^p$ are quantile specific intercept and slopes for covariates and genotypes, respectively. Denote the data in one tissue from GTEx as $\{E_i, G_i, C_i\}_{i=1}^n$ with sample size n, where $E_i \in \mathbb{R}$ is the gene expression, $G_i \in \mathbb{R}^p$ is the $p-$dimensional genotype data, and $C_i \in \mathbb{R}^q$ is the $q-$dimensional covariates, including sex, sequencing platform, top three principal components of genotype data, and the top probabilistic estimation of expression residuals (PEER) factors. Thus, for a fine grid of $\tau \in (0, 1)$, we estimate $\beta(\tau)$ by solving

$$(\hat{\alpha}_{0,\tau}, \hat{\alpha}_\tau, \hat{\beta}_\tau) = \arg \min_{\alpha_{0,\tau}, \alpha_\tau, \beta_\tau} \sum_i^n \rho_\tau(E_i - \alpha_{0,\tau} - C_i \alpha_\tau - G_i \beta_\tau),$$

where $\rho_\tau(u) = |u|\{(1-\tau)I(u < 0) + \tau I(u > 0)\}$, $u \in \mathbb{R}$.

In eq (6), the gene expression E has the same distribution as $\alpha_0(U) + C^\top \alpha(U) + G^\top \beta(U)$, where $U \sim \text{Unif}(0, 1)$. One can view the gene-expression E as a convolution of a genotype-related random variable $G^\top \beta(U)$ and an unrelated random variable $\alpha_0(U) + C^\top \alpha(U)$. It follows that genotype-related gene expression E_G, as defined in the decomposition $E = E_G + E_C$, has the same distribution as $G^\top \beta(U)$. Thus,

$$\mathbb{E}(E_G \mid G, E_G \in A_k) = \int_{\tau_k}^{\tau_{k+1}} G^\top \beta(u) du = G^\top \beta_{A_k},$$

where $\beta_{A_k} = \int_{\tau_k}^{\tau_{k+1}} \beta(u) du$. Accordingly, we have $\hat{\beta}_{A_k} = \int_{\tau_k}^{\tau_{k+1}} \hat{\beta}(\tau) d\tau$, in which τ_k and τ_{k+1} define the range of A_k.

Step 2: testing quantile-stratified nonlinear gene-trait association

Denote $\hat{\gamma}_k$ as the estimate of γ_k for region A_k, $se(\hat{\gamma}_k)$ as the standard error of $\hat{\gamma}_k$, and N_{GWAS} as the sample size for GWAS data. The z score for gene-trait association for region A_k is $Z_k = \hat{\gamma}_k / se(\hat{\gamma}_k)$. Given eq (4)-(5) and the conditional independence assumption $T \perp\!\!\!\perp E_G \mid G$, we have that $\mathbb{E}\{cov(E_G, T \mid G, E_G \in A_k)\} = 0$, and $\mathbb{E}(T \mid G, E_G \in A_k) = G^\top \beta_{\text{GWAS}}$, where β_{GWAS} is the SNP-level effect size. With σ_j representing the standard deviation of the j-th SNP,

$$cov\{\mathbb{E}(E_G \mid G, E_G \in A_k), \mathbb{E}(T \mid G, E_G \in A_k)\} = cov(G^\top \beta_{A_k}, G^\top \beta_{\text{GWAS}}) = \beta_{A_k}^\top \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_p^2 \end{pmatrix} \beta_{\text{GWAS}},$$

such that

$$\hat{\gamma}_k = \frac{\hat{\beta}_{A_k}^\top}{\sigma_{E_{A_k}}^2} \begin{pmatrix} \hat{\sigma}_1^2 & & \\ & \ddots & \\ & & \hat{\sigma}_p^2 \end{pmatrix} \hat{\beta}_{\text{GWAS}},$$

where $\sigma_{E_{A_k}}$ is the standard deviation of imputed gene expression in region A_k, $\hat{\beta}_{\text{GWAS}}$ is the SNP-level estimated effect size available from GWAS summary statistics, and $\hat{\sigma}_j$ is the estimated standard deviation of the j-th SNP. Regarding $se(\hat{\gamma}_k)$, we have

$$se(\hat{\gamma}_k) = \sqrt{\frac{var(\epsilon)}{N_{\text{GWAS}} var(E_G \mid E_G \in A_k)}} \approx \frac{\hat{\sigma}_T}{\sqrt{N_{\text{GWAS}}}\sigma_{E_{A_k}}},$$

where $\hat{\sigma}_T$ is the estimated standard deviation of trait T. We have $var(\epsilon) \approx \sigma_T^2$ based on the empirical observation that only a very small proportion of variability of the trait can be explained by one gene (Hu et al., 2019; O'Connor et al., 2017). We then use the same argument for the SNP level association statistics in GWAS data. For the jth SNP in the model, its z score can be denoted as

$$\widetilde{Z}_j = \frac{\hat{\beta}_{\text{GWAS,j}}}{se(\hat{\beta}_{\text{GWAS,j}})} \approx \frac{\sqrt{N_{\text{GWAS}}}\hat{\sigma}_j \hat{\beta}_{\text{GWAS,j}}}{\hat{\sigma}_T},$$

and the matrix form GWAS z score is

$$\widetilde{Z} \approx \frac{\sqrt{N_{\text{GWAS}}}}{\hat{\sigma}_T} \begin{pmatrix} \hat{\sigma}_1 & & \\ & \ddots & \\ & & \hat{\sigma}_p \end{pmatrix} \hat{\beta}_{\text{GWAS}}.$$

Combining the derivations for Z_k and $\widetilde{Z}$, we have the gene-level quantile-stratified z score

$$Z_k = \frac{\hat{\gamma}_k}{se(\hat{\gamma}_{A_k})} \approx \frac{\hat{\beta}_{A_k}^\top}{\sigma_{E_{A_k}}} \begin{pmatrix} \hat{\sigma}_1 & & \\ & \ddots & \\ & & \hat{\sigma}_p \end{pmatrix} \widetilde{Z}.$$

Under the null hypothesis, $\widetilde{Z} \approx N(0, D)$, where D is the linkage-disequilibrium matrix for SNPs, either estimated from the reference panel or external reference data. Hence, under $H_0 : \gamma_k = 0$ for $k = 1, ..., K$,

$$Z_k \approx N\left(0, \frac{1}{\sigma^2_{E_{A_k}}} \hat{\beta}_{A_k}^\top \text{diag}(\hat{\sigma}_1, ..., \hat{\sigma}_p) D \text{diag}(\hat{\sigma}_1, ..., \hat{\sigma}_p) \hat{\beta}_{A_k}\right).$$

After acquiring the p value for each region A_k, we combine them by the Cauchy combination method (Liu and Xie, 2020). For those genes measured in multiple tissues, one can further combine multiple p-values across tissues by the same method.

Data availability and web resources: We use data from existing studies from GTEx (v8, https://gtexportal.org/home/) and summary level GWAS results on neuropsychiatric and neurodegenerative traits (Pardiñas et al., 2018; Demontis et al., 2019; Stahl et al., 2019; Grove et al., 2019; Howard et al., 2019; Kunkle et al., 2019; Jansen et al., 2019; Nalls et al., 2019; Andlauer et al., 2016; Van Rheenen et al., 2016). In data analyses, we use pre-trained models for S-PrediXcan. The models are available at the PredictDB website (https://predictdb.org/post/2021/07/21/gtex-v8-models-on-eqtl-and-sqtl/).

Code availability: We have implemented the proposed QTWAS framework in an R package `QTWAS` (https://github.com/tianyingw/QTWAS).

References

Andlauer, T. F., Buck, D., Antony, G., Bayas, A., Bechmann, L., Berthele, A., Chan, A., Gasperi, C., Gold, R., Graetz, C., et al. (2016). Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Science advances*, 2, e1501678.

Azevedo, C., Teku, G., Pomeshchik, Y., Reyes, J. F., Chumarina, M., Russ, K., Savchenko, E., Hammarberg, A., Lamas, N. J., Collin, A., et al. (2022). Parkinson's disease and multiple system atrophy patient ipsc-derived oligodendrocytes exhibit alpha-synuclein–induced changes in maturation and immune reactive properties. *Proceedings of the National Academy of Sciences*, 119, e2111405119.

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. *Nature communications*, 9, 1825.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K. O., Di Narzo, A. F., Yan, P., Hodgson, J. G., Weinrich, S., et al. (2013). Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal of pathology*, 231, 63–76.

Consortium, G. et al. (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369, 1318–1330.

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Bækvad-Hansen, M., et al. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics*, 51, 63–75.

Doens, D. and Fernández, P. L. (2014). Microglia receptors and their implications in the response to amyloid β for alzheimer's disease pathogenesis. *Journal of neuroinflammation*, 11, 1–14.

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., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47, 1091.

Geeleher, P., Nath, A., Wang, F., Zhang, Z., Barbeira, A. N., Fessler, J., Grossman, R. L., Seoighe, C., and Stephanie Huang, R. (2018). Cancer expression quantitative trait loci (eqtls) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. *Genome biology*, 19, 1–14.

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O. A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, 51, 431–444.

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.

Gutenbrunner, C. and Jurecková, J. (1992). Regression rank scores and regression quantiles. *The Annals of Statistics*, pages 305–330.

Gutenbrunner, C., Jurečková, J., Koenker, R., and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *Journaltitle of Nonparametric Statistics*, 2, 307–331.

Howard, D. M., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., Shirali, M., Coleman, J. R., Hagenaars, S. P., Ward, J., Wigmore, E. M., et al. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience*, 22, 343–352.

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., Yu, Z., Li, B., Gu, J., Muchnik, S., et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *bioRxiv*, page 286013.

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. *Nature genetics*, 51, 404–413.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, 478, 483–489.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94, 1296–1310.

Koenker, R. W. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., Boland, A., Vronskaya, M., Van Der Lee, S. J., Amlie-Wolf, A., et al. (2019). Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates aβ, tau, immunity and lipid processing. *Nature genetics*, 51, 414–430.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3, e161.

Li, B., Veturi, Y., Verma, A., Bradford, Y., Daar, E. S., Gulick, R. M., Riddler, S. A., Robbins, G. K., Lennox, J. L., Haas, D. W., et al. (2021). Tissue specificity-aware twas (tsa-twas) framework identifies novel associations with metabolic, immunologic, and virologic traits in hiv-positive adults. *PLoS genetics*, 17, e1009464.

Li, D., Liu, Q., and Schnable, P. S. (2021). Twas results are complementary to and less affected by linkage disequilibrium than gwas. *Plant Physiology*, .

Lin, Z., Xue, H., Malakhov, M. M., Knutson, K. A., and Pan, W. (2022). Accounting for nonlinear effects of gene expression identifies additional associated genes in transcriptomewide association studies. *Human molecular genetics*, .

Liu, D., Wang, Y., Jing, H., Meng, Q., and Yang, J. (2021). Mendelian randomization integrating gwas and mqtl data identified novel pleiotropic dna methylation loci for neuropathology of alzheimer's disease. *Neurobiology of Aging*, 97, 18–27.

Liu, Y. and Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115, 393–402.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, 45, 580–585. PMID: 23715323.

Ma, J., Liu, M., Wang, Y., Xin, C., Zhang, H., Chen, S., Zheng, X., Zhang, X., Xiao, F., and Yang, S. (2020). Quantitative proteomics analysis of young and elderly skin with dia mass spectrometry reveals new skin aging-related proteins. *Aging (Albany NY)*, 12, 13529.

Ma, S., Dalgleish, J. L., Lee, J., Wang, C., Liu, L., Gill, R., Buxbaum, J. D., Chung, W., Aschard, H., Silverman, E. K., et al. (2021). Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. *medRxiv*, .

Nagpal, S., Meng, X., Epstein, M. P., Tsoi, L. C., Patrick, M., Gibson, G., De Jager, P. L., Bennett, D. A., Wingo, A. P., Wingo, T. S., et al. (2019). Tigar: An improved bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *The American Journal of Human Genetics*, 105, 258–266.

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D. A., Noyce, A. J., Xue, A., et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology*, 18, 1091–1102.

O'Connor, L. J., Gusev, A., Liu, X., Loh, P.-R., Finucane, H. K., and Price, A. L. (2017). Estimating the proportion of disease heritability mediated by gene expression levels. *BioRxiv*, page 118018.

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., et al. (2018). Common

schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics*, 50, 381–389.

Sarlus, H., Heneka, M. T., et al. (2017). Microglia in alzheimer's disease. *The Journal of clinical investigation*, 127, 3240–3249.

Sey, N. Y., Hu, B., Mah, W., Fauni, H., McAfee, J. C., Rajarajan, P., Brennand, K. J., Akbarian, S., and Won, H. (2020). A computational tool (h-magma) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nature Neuroscience*, 23, 583–593.

Somel, M., Khaitovich, P., Bahn, S., Pääbo, S., and Lachmann, M. (2006). Gene expression becomes heterogeneous with age. *Current Biology*, 16, R359–R360.

Song, X., Li, G., Zhou, Z., Wang, X., Ionita-Laza, I., and Wei, Y. (2017). Qrank: a novel quantile regression tool for eqtl discovery. *Bioinformatics*, 33, 2123–2130.

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., Mattheisen, M., Wang, Y., Coleman, J. R., Gaspar, H. A., et al. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics*, 51, 793–803.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS computational biology*, 6, e1000770.

Tang, S., Buchman, A. S., De Jager, P. L., Bennett, D. A., Epstein, M. P., and Yang, J. (2021). Novel variance-component twas method for studying complex human diseases with applications to alzheimer's dementia. *PLoS genetics*, 17, e1009482.

Umans, B. D., Battle, A., and Gilad, Y. (2021). Where are the disease-associated eqtls? *Trends in Genetics*, 37, 109–124.

Van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., Van Der Spek, R. A., Võsa, U., De Jong, S., Robinson, M. R., et al. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature genetics*, 48, 1043–1048.

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, J. L. M., Im, H. K., Pasaniuc, B., Rivas, M. A., and Kundaje, A. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, 51, 592–599.

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.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C., Clarke, D., Gu, M., Emani, P., Yang, Y. T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362.

Wang, T., Ionita-Laza, I., and Wei, Y. (2019). Integrated quantile rank test (iqrat) for gene-level associations in sequencing studies. *arXiv preprint arXiv:1910.10102*, .

Xie, Y., Shan, N., Zhao, H., and Hou, L. Transcriptome wide association studies: general framework and methods. *Quantitative Biology*, page 0.

Zhao, B., Shan, Y., Yang, Y., Yu, Z., Li, T., Wang, X., Luo, T., Zhu, Z., Sullivan, P., Zhao, H., et al. (2021). Transcriptome-wide association analysis of brain structures yields

insights into pleiotropy with complex neuropsychiatric traits. *Nature communications*, 12, 1–11.