

Accounting for nonlinear effects of gene expression identifies additional associated genes in transcriptome-wide association studies

Zhaotong Lin , Haoran Xue, Mykhaylo M. Malakhov , Katherine A. Knutson and Wei Pan *

Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA

*To whom correspondence should be addressed at: A460 Mayo Building, 420 Delaware St SE, Minneapolis, MN 55455, USA. Tel: (612)626-2705; Fax: (612)626-0660; Email: panxx014@umn.edu

Abstract

Transcriptome-wide association studies (TWAS) integrate genome-wide association study (GWAS) data with gene expression (GE) data to identify (putative) causal genes for complex traits. There are two stages in TWAS: in Stage 1, a model is built to impute gene expression from genotypes, and in Stage 2, gene–trait association is tested using imputed gene expression. Despite many successes with TWAS, in the current practice, one only assumes a linear relationship between GE and the trait, which however may not hold, leading to loss of power. In this study, we extend the standard TWAS by considering a quadratic effect of GE, in addition to the usual linear effect. We train imputation models for both linear and quadratic gene expression levels in Stage 1, then include both the imputed linear and quadratic expression levels in Stage 2. We applied both the standard TWAS and our approach first to the ADNI gene expression data and the IGAP Alzheimer's disease GWAS summary data, then to the GTEx (V8) gene expression data and the UK Biobank individual-level GWAS data for lipids, followed by validation with different GWAS data, suitable model checking and more robust TWAS methods. In all these applications, the new TWAS approach was able to identify additional genes associated with Alzheimer's disease, LDL and HDL cholesterol levels, suggesting its likely power gains and thus the need to account for potentially nonlinear effects of gene expression on complex traits.

Introduction

Although genome-wide association studies (GWAS) have identified thousands of genetic loci associated with many complex traits and diseases, a mechanistic understanding of the biological function of these loci remains largely elusive. It is hypothesized that a substantial proportion of GWAS risk variants influence complex traits through their regulatory roles on the expression levels of their target genes (1–3). Transcriptome-wide association studies (TWAS), also called PrediXcan, have become increasingly popular and important in identifying (putative) causal genes and thus underlying regulatory mechanisms associated with diseases and complex traits (4,5). TWAS leverages an independent expression quantitative trait locus (eQTL) dataset to discover gene–trait associations for a GWAS (summary) dataset. Specifically, first, in Stage 1, the eQTL data are used to build a prediction model for the genetically regulated component of each gene's expression level (GReX), usually using only its *cis*-acting genotypes/single-nucleotide polymorphisms (SNPs) around the gene. Then in Stage 2, the predictive model is used to impute gene expression using the GWAS genotypic data, which is then associated with the GWAS trait; the genes associated with the trait are claimed to be causal under suitable theoretical conditions.

In spite of many successes (6–8), the current standard practice with TWAS only imputes the mean expression level of each gene and associates it linearly with a GWAS trait. In other words, only a linear relationship between gene expression (GE) and a trait is considered. However, there is no reason, except perhaps for simplicity, to exclusively assume only a linear relationship; if there is a nonlinear relationship between GE and the trait, the standard TWAS is expected to lose statistical power. In this paper, we empirically confirm this point. It is noted that the quadratic effects of GE on a trait can be regarded as the influence of the GE variability on the trait. In addition to eQTL (or mean QTL), variable expression QTL (veQTL) or more generally variance QTL (vQTL) have been studied in the literature (9–11). In particular, the presence of vQTL may be due to omitted SNP–SNP or SNP–environment interactions. Thus, the quadratic effects of GE on a trait can capture the mediating effects of veQTL. Motivated by these considerations, we first extend the standard TWAS (called TWAS-L) by considering a quadratic effect, in addition to the usual linear effect, of GE on the trait; the two new versions including only a quadratic effect and both a linear and a quadratic effects are called TWAS-Q and TWAS-LQ, respectively. Then we apply the three TWAS methods to the (individual-level) Alzheimer's Disease Neuroimaging

Initiative (ADNI) eQTL data and the International Genomics of Alzheimer's Project (IGAP) Alzheimer's disease (AD) GWAS (summary) data, and the (individual-level) genotype-tissue expression (GTEx) v8 eQTL data and UK Biobank (UKB) GWAS data for lipids, showcasing that the extended TWAS identified additional genes associated with AD, low-density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol levels. These results were confirmed by (partial) validation with different GWAS data, suitable model checking and application of some more robust TWAS methods. Given the increasing importance of TWAS and its standard approach being applied, our findings suggest great potential with the application of our simple TWAS extension in practice.

Results

TWAS-LQ identified additional genes associated with AD

We applied the standard TWAS-L and its two new extensions, TWAS-Q and TWAS-LQ, to the discovery sample with the ADNI GE data in Stage 1 and the IGAP GWAS summary data in Stage 2. We used the common SNPs present in both the ADNI data and IGAP data to fit the GE imputation models in Stage 1. To avoid weak IV bias, we applied TWAS-L and TWAS-Q to the genes that had their F-statistics >10 in Stage 1 model Eqs. (1) and (2), respectively. We applied TWAS-LQ to the genes with an F-statistic >10 in either Eq. (1) or (2). For TWAS-L and TWAS-Q, we further removed the genes with only one SNP imputing GE and GE^2 , respectively, and for TWAS-LQ, we kept the genes with more than one SNP for both GE and GE^2 . After screening in Stage 1, we had 1278, 235 and 1279 genes left in Stage 2 for TWAS-L, TWAS-Q and TWAS-LQ, respectively. We identified significant genes on the basis of the corresponding Bonferroni-adjusted P -values.

Table 1 shows the significant genes identified by at least one of the three methods with the discovery sample. After the Bonferroni adjustment, TWAS-L, TWAS-Q and TWAS-LQ identified two, one and four genes, respectively. In particular, genes *HLA-DQA1* and *HLA-DQB1* were missed by the standard TWAS-L but identified by the new methods TWAS-Q and/or TWAS-LQ; *HLA-DQA1* and *HLA-DQB1* genes are part of the human leukocyte antigen (HLA) complex. They both belong to a group of major histocompatibility complex (MHC) genes called MHC class II, which play an important role in immune system. Their contributions to AD risk have been previously discussed in the literature (12–16). For example, a genome-wide pathway analysis has suggested that both *HLA-DQA1* and *HLA-DQB1* were associated with AD risk (15). Fine mapping of HLA region including *HLA-DQA1* and *HLA-DQB1* also suggested a central role of these two genes in late-onset AD (12,14).

We applied the same analysis pipeline as aforementioned to the validation data with the ADNI GE data

in Stage 1 and Jansen's GWAS summary data in Stage 2 to (partially) validate the findings from the discovery sample. We focused on the four significant genes i.e. two genes identified by TWAS-L, one gene by TWAS-Q and four genes by TWAS-LQ, and used the corresponding Bonferroni adjustment for each method. As shown in Table 1, all the previously significant genes were confirmed by the validation data.

We also performed the analysis on other genes with the validation data. The results are given in the [Supplementary Material](#).

TWAS-LQ identified additional genes associated with lipids

We applied TWAS-L, TWAS-Q and TWAS-LQ using the GTEx whole blood GE data in Stage 1 and UKB LDL and HDL GWAS data in Stage 2. As before, we used the common SNPs present in both the GTEx and UKB data to train Stage 1 models, and we screened out the genes with possible weak IV issues using the F-statistic threshold >10 in Stage 1 and excluded the genes with only one SNP. After screening, there were 4685, 161, 3815 genes left in Stage 2 for TWAS-L, TWAS-Q and TWAS-LQ, respectively. The results using different screening criteria are shown in [Supplementary Material](#).

Figure 1 shows the numbers of the significant genes identified by the three TWAS methods for HDL (left) and LDL (right). TWAS-LQ identified 67 (out of 3815) genes, TWAS-L identified 87 (out of 4685) genes and TWAS-Q only identified 4 (out of 161) genes associated with HDL. Although the standard TWAS-L identified the largest number of the significant genes, it is noted that TWAS-Q and TWAS-LQ identified many additional genes other than the ones by TWAS-L, which may bring in some new insights. For example, there were 15 significant genes associated with HDL only identified by TWAS-LQ, and one was uniquely identified by TWAS-Q. In particular, two genes, *CDK2AP1* and *SPDYC*, were identified by TWAS-LQ and TWAS-Q, but not by TWAS-L. *CDK2AP1* has been shown to interact with cyclin-dependent kinase 2 (*CDK2*) (17) and *SPDYC* was also found to be an activator of *CDK2* (18), whereas *CDK2* plays a critical role in cell cycle progression, and it is related to many liver diseases, such as hepatocellular carcinoma (HCC) and liver cancer (19,20). It is also noted that, among the 37 genes only identified by TWAS-L, 31 were not in the analysis by TWAS-LQ (because there was no SNP left in imputing quadratic GE). For LDL, again TWAS-Q and TWAS-LQ could identify some additional genes, and TWAS-LQ identified most genes among the three methods.

We further examined the unique genes identified by TWAS-Q (but not by TWAS-L and TWAS-LQ). For HDL it was gene *RHD*. Its P -values from the three methods were actually close at $4.95e-05$, $3.92e-05$ and $2.12e-04$ for TWAS-L, TWAS-Q and TWAS-LQ, respectively. Because of the different numbers of the genes being tested and thus

Table 1. P-values of four significant genes associated with AD from the discovery (or validation) data are outside (or inside) parentheses

Gene	chr	P-values in Stage 2		
		TWAS-L	TWAS-Q	TWAS-LQ
HLA-DQA1	6	2.45e-01	4.27e-01	2.09e-09 (1.34e-05)
HLA-DQB1	6	1.05e-03	8.45e-06 (6.26e-10)	2.73e-05 (4.93e-09)
HLA-DRB5	6	4.43e-06 (1.21e-03)	2.41e-04	8.12e-08 (1.67e-08)
MS4A6A	11	2.65e-11 (5.75e-12)	1.68e-06	9.22e-11 (4.73e-11)

The significant P-values (after the Bonferroni adjustment) are bold faced.

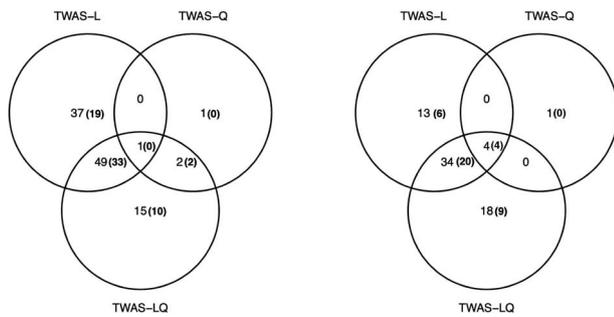


Figure 1. Venn diagrams for the significant genes for HDL (left) or LDL (right) identified by TWAS-L, TWAS-Q and TWAS-LQ. The numbers of the significant genes using the discovery (UKB) or validation (GLGC) GWAS data are shown outside or inside parentheses, respectively.

different Bonferroni adjustments for the three methods, only the one from TWAS-Q was significant after the Bonferroni adjustment. In addition, the P-value from TWAS-LQ was less significant than those from the other two methods because the imputed GE and imputed GE² were highly correlated for this gene: the set of the three eSNPs used for GE² were all included in the set of the six eSNPs for GE, and their corresponding weights were also correlated. For LDL, the unique gene only identified by TWAS-Q was SPDYC. The P-values for TWAS-L, TWAS-Q and TWAS-LQ were 2.13e-03, 8.28e-05 and 5.44e-05, respectively. Again, due to the different Bonferroni adjustments, although TWAS-LQ gave a more significant P-value, only that of TWAS-Q was significant after the Bonferroni adjustment.

We used the GTEx GE data in Stage 1 and the Global Lipids Genetics Consortium (GLGC) lipid GWAS summary data in Stage 2 to validate the previous findings. For HDL, there were 87 genes identified by TWAS-L, 4 genes by TWAS-Q and 67 genes by TWAS-LQ. After we refitted the Stage 1 model, there were a few with only one SNP; after removing these genes, we had 85, 4 and 61 genes left for TWAS-L, TWAS-Q and TWAS-LQ, respectively. Similarly, for LDL, there were 50, 5 and 54 genes left for the three methods, respectively. We used the corresponding Bonferroni adjustments to identify significant genes with the validation data. Figure 1 compares the numbers of the significant genes from the discovery and validation data. Most of the significant genes were confirmed, including in particular the two genes associated with HDL (CDK2AP1 and SPDYC) uniquely identified by TWAS-LQ and TWAS-Q.

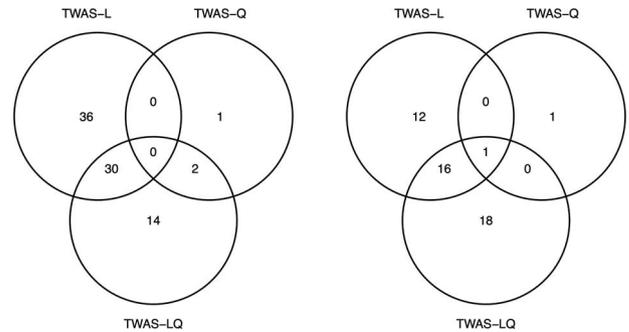


Figure 2. Venn diagrams for the significant genes for HDL (left) or LDL (right) after surviving the TEDE tests (i.e. without evidence for horizontal pleiotropy).

Model checking

As shown before, TWAS-LQ identified some additional genes. However, it is possible that some of the significant genes were due to horizontal pleiotropy of the SNPs being used, especially given that TWAS-LQ used more SNPs than the other two methods to impute both linear and quadratic GE levels. Thus, we applied the TESting Direct Effects (TEDE) test on the significant genes identified by TWAS-L, TWAS-Q and TWAS-LQ to detect possible horizontal pleiotropy with the GTEx and UKB lipids GWAS data. Specifically, we used the score test with a modified covariance estimate (21); the details are given in the [Supplementary Material](#). We used the Bonferroni correction for the TEDE tests as well. For example, TWAS-L identified 87 significant genes associated with HDL, we then applied the TEDE test to these 87 genes; we claimed that there was evidence for pleiotropy of a gene's SNPs (being used to impute its expression) if the TEDE test gave a P-value < 0.05/87.

Figure 2 shows the numbers of the significant genes without pleiotropy and identified by the three TWAS methods for HDL (left) and LDL (right). We can see that after excluding the genes with possible pleiotropic SNP effects, TWAS-Q and TWAS-LQ together identified 17 and 19 additional genes missed by TWAS-L for HDL and LDL, respectively. Again, this demonstrates that incorporating imputed quadratic GE could detect additional associated genes.

Robust TWAS accounting for horizontal pleiotropy

We applied the more robust linear mixed-effects (LME) model-based TWAS-L-LME, TWAS-Q-LME and TWAS-LQ-

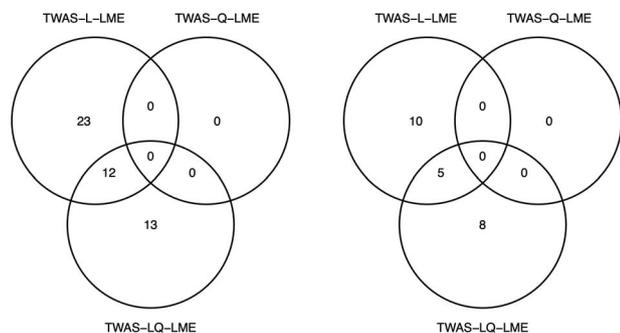


Figure 3. Venn diagrams for the significant genes for HDL (left) or LDL (right) identified by TWAS-L-LME, TWAS-Q-LME and TWAS-LQ-LME.

LME, as the counterparts of TWAS-L, TWAS-Q and TWAS-LQ, to take account of possible horizontal pleiotropy. Figure 3 shows the numbers of the significant genes by the three LME-based TWAS methods for HDL (left) and LDL (right) when applied to the GTEx eQTL and UKB lipids GWAS data. Although TWAS-Q-LME did not identify any significant genes associated with HDL or LDL, TWAS-LQ-LME still identified some additional genes missed by TWAS-L-LME. For example, as shown in the right of Figure 3, there were eight genes associated with LDL uniquely identified by TWAS-LQ-LME, including, for instance, *ITIH4*, which was found to be the most characteristic protein corresponding with nonalcoholic fatty liver disease (NAFLD) progression and HCC development in the NAFLD pigs in an HCC pig model study. A human serum samples analysis supported this observation (22).

Similar results were obtained with the three corresponding LME0 methods under the assumption of balanced pleiotropy, instead of directional pleiotropy with the above LME methods, as shown in the Supplementary Material.

Other results

The Q-Q plots of the various methods for the real data analyses are shown in the Supplementary Material. In general, the Q-Q plots for TWAS-L and TWAS-LQ were similar, showing the enrichment of significant genes, whereas their LME counterparts showed less enrichment. There are several possible reasons for the enrichment of significant genes. First, according to the omnigenic model (23), there are indeed many genes associated with a complex trait such as HDL and LDL. Second, the imputed expression levels of physically nearby or co-expressed genes are correlated due to their shared eQTLs and/or co-expression, leading to inflated statistical significance of some genes (24). In addition, the non-independence of some genes' imputed expression levels and thus of their *P*-values violated the independence assumption in a Q-Q plot. Third, there is possible population structure in GWAS data. As a reviewer suggested, we performed analysis of the UKB individual-level data after adjusting for top genetic principal components to check whether such enrichment was due to population stratification. The results (shown in the Supplementary Material) were similar to those

presenting here. Finally, in each Q-Q plot, only a selected subset of the genes with their expression levels better imputed (i.e. surviving the screening procedure in stage 1 with large *F*-statistics) were included, hence any enrichment was only over this subset of the genes.

As shown in the Supplementary Material, we also did a simulation study to confirm that the proposed new TWAS methods could control type I errors satisfactorily. In addition, as expected, TWASL, TWAS-Q and TWAS-LQ were most powerful under the alternative hypotheses of a gene's expression having only a linear effect, only a quadratic effect and both a linear and a quadratic effects on the trait, respectively. Furthermore, as a more general omnibus test, TWAS-LQ could maintain high power across all the three scenarios, whereas the other two could lose substantial power in some cases.

Discussion

In this study, we have explored possibly a quadratic relationship between GE and complex diseases/traits, going beyond the current practice of considering only a linear relationship in the standard TWAS. We implemented two extensions of the standard TWAS by incorporating a term of imputed squared gene expression (GE^2), with or without the usual linear term, in the Stage 2 model of TWAS. We applied the extended TWAS methods, TWAS-Q and TWAS-LQ, to the ADNI eQTL data and the IGAP AD GWAS summary data, uncovering two genes that would be missed by the standard TWAS (TWAS-L). These two genes, namely *HLA-DQA1* and *HLA-DQB1*, have been shown to be associated with AD risk through a fine mapping study of the HLA locus (12). We also applied the methods to the GTEx eQTL data and UKB individual-level GWAS data for lipids. We observed a similar pattern that TWAS-Q and especially TWAS-LQ were able to identify a number of additional genes missed by TWAS-L, even after excluding those with potential pleiotropic SNPs as detected by model checking, or after accounting for pleiotropic effects in more robust LME-based TWAS methods. We also validated the findings using different AD and lipids GWAS data. These results suggest that there could be nonlinear relationships between GE and common diseases or complex traits, and accounting for nonlinear effects in TWAS would boost statistical power for new discovery that could give new insights into the underlying causal mechanism. Another example of possible application is to help increase the estimated heritability mediated through GE, which was previously estimated through only linear effects of GE and was suspected to be under estimated (25).

Although there is a large literature on nonlinear or nonadditive effects of genotypes from large-scale studies of model organisms (26,27), such evidence is known to be much harder to find in human studies, likely due to lack of power. Nevertheless, thanks to the increasing sample size, there is an emerging literature of human studies supporting nonlinear or nonconstant effects of genotypes on complex traits/diseases (28), in particular,

of nonlinear effects of ApoE on cognitive decline and AD risk (29). Perhaps even more surprisingly and thus more convincingly, the nonlinear effects of SNPs on GE in Stage 1 of TWAS were detected and a nonlinear machine learning method, random forest, for some genes performed better than linear model-based methods (e.g. Lasso) in predicting/imputing GE, as reported by Grinberg and Wallace (30). Similarly, although elastic net penalized linear regression is widely used e.g. in PrediXcan (nonlinear), random forest could sometimes outperform elastic net penalized linear regression in imputing/predicting GE in Stage 1 of TWAS/PrediXcan, as reported in Okoro *et al* (31). However, these results on nonlinear effects of genotypes on GE or traits are different from what we study here: nonlinear effects of genetically regulated components of GE on complex traits. To our best knowledge, we are not aware of any other study on nonlinear effects of GE on a trait in the context of TWAS, which is the main point of this paper.

There are a few limitations of this study. First, to impute nonlinear GE levels in Stage 1 of TWAS, we require the availability of individual-level eQTL data, whereas the standard TWAS-L could be implemented using eQTL summary data or even some pretrained Stage 1 model as available on the FUSION website (4). However, it is noted that, if eQTL summary data for quadratic GE are offered and available, our methods can be directly applied to such eQTL summary data and GWAS summary data. The computational time of TWAS-LQ is almost the same as that of TWAS-L. Second, we used a simple backward variable selection scheme to fit a Stage 1 regression model, whereas other possibly more efficient methods such as elastic net penalized linear regression, or a nonlinear method such as random forest, can be also applied and may further improve the performance. Third, we only studied the quadratic effect, though in theory there may be other functional forms of nonlinear effects, so other more sophisticated parametric or nonparametric models can be explored (though they will require the availability of individual-level GWAS data, in contrast to that of GWAS summary data required by our methods). Fourth, population structure in GWAS data may lead to false positives. With GWAS individual-level data, one can adjust for population stratification using genetic principal components or mixed models, as we did with the UKB data. With GWAS summary data, we did not consider this issue because presumably suitable adjustments for population structure should have already been applied in published GWAS summary data. If needed, we may perform genomic control (32,33) to adjust for population stratification with GWAS summary data, or treat population structure as a source of pleiotropy in TWAS or Mendelian randomization as in Hu *et al*. (34), which can be one of future research directions. Fifth, we only focused on TWAS with GE as the endophenotype; it is straightforward to apply the methods to other molecular traits such as methylation, proteomic and metabolomic QTL (xQTL) data to investigate possibly nonlinear relationships

between their genetic components and other complex traits/diseases. Similarly, our proposed methods can be applied in the context of Mendelian randomization as well (35). Sixth, as for the standard TWAS, we have considered the use of only one eQTL dataset from one tissue (or cell type), but it may be extended to multiple tissues following several existing approaches (36–38). Lastly, with the ever increasing availability of large (individual-level) eQTL/xQTL data and GWAS (summary) data, it is both convenient and important to conduct empirical studies on more GWAS traits to investigate how wide-spread are nonlinear effects, and their specific functional forms, of GE and other molecular/imaging endophenotypes on complex traits, which will advance our understanding of the genetic architecture for complex traits.

Materials and Methods

Data

eQTL data

The ADNI GE data

The ADNI is a longitudinal multicenter study designed to develop clinical, imaging, genetic and biochemical biomarkers for the early detection and tracking of AD. To date the three phases of this study (ADNI-1, ADNI-GO and ADNI-2) have recruited over 1500 adults of ages 55–90 to participate in the research, consisting of cognitively normal older individuals, those with early or late minor cognitive impairment (MCI), and those with AD.

In this article, the ADNI data (39) were used in TWAS Stage 1, containing individuals' GE data, whole genome sequence (WGS) data and five covariates including age, gender, year of education, handedness and intracranial volume (ICV). After cleaning and merging, we had a sample size of 711. We first regressed (linear) GE on the five covariates, then used the residuals as the 'standardized' (linear) GE; the linear GE and its squared term (GE^2) of 17 256 genes on the autosomes were used in the standard and extended TWAS. For each gene, we defined its cis-region by expanding 100 kb upstream and downstream its coding region (i.e. from its transcription start site (TSS) and transcription end site (TES)), respectively. We excluded SNPs with minor allele frequency (MAF) ≤ 0.05 or with missing values, or failing the Hardy–Weinberg equilibrium test (P -value ≤ 0.001). We then matched the SNPs with the IGAP GWAS summary statistics and pruned the SNPs to ensure that any of their pairwise Pearson's absolute correlations was no > 0.8 . Finally, if there were > 50 SNPs left, we took the 50 SNPs with the largest absolute values of their marginal Pearson's correlations with GE as I_1 , and the 50 SNPs with the largest absolute values of correlations with GE^2 as I_2 and used the SNPs in the union $I_1 \cup I_2$ with no > 100 SNPs being used in TWAS Stage 1.

The GTEx GE data

The GTEx project is a comprehensive public resource to study tissue-specific GE and regulation. Samples were

collected from 54 nondiseased tissue sites across nearly 1000 individuals, primarily for molecular assays including GE by RNA-Seq and genotyping by WGS.

We used the GTEx v8 whole blood data in TWAS Stage 1, including 670 individuals' GE data, WGS data and covariates (first 5 genotype PCs, WGS sequencing platform, WGS library construction protocol, donor sex and PEER factors) (40). The standardized (linear) GE and squared expression (GE^2) of 19718 genes on the autosomes were used. The rest of the quality control procedures were similar to that for the ADNI data aforementioned except that we matched the GTEx SNPs with the UKB genotypic data.

GWAS data

The IGAP AD GWAS summary data

We used the imputed Stage 1 summary statistics (with imputation $R^2 \geq 0.3$) with 17 008 cases and 37 154 controls generated by the IGAP Consortium (41) as the discovery sample for our TWAS Stage 2 analysis.

Jansen's AD GWAS summary data

Jansen *et al.* (42) conducted a large genome-wide association study with 71 880 cases (defined either as clinically diagnosed AD or AD-by-proxy) and 383 378 controls, including the IGAP data. We used this dataset for (partial) validation of the findings from the IGAP GWAS dataset.

The UKB (individual-level) GWAS data

The UKB is a large-scale prospective cohort study with phenotypic and genetic data on about 500 000 subjects. We used the data from the individuals who were self-reported white British and did not have any (close) relatives among the set of the genotyped individuals. We took LDL and HDL cholesterol levels as the traits of interest. The number of SNPs is around 800 000. We used the UKB data as the discovery sample.

The GLGC lipid GWAS summary data

We used the HDL and LDL GWAS summary statistics generated by the GLGC (43) as the validation data for the findings from the UKB individual-level GWAS data. The number of SNPs is around 2 450 000.

Methods: TWAS and Extensions

GE imputation models

In Stage 1 of TWAS, we need to build a predictive model for GE. To incorporate quadratic effects of GE on a trait, in addition to the (linear) GE, we need to impute the GE^2 level as well. If there are covariates, we may need to first regress out their effects: for each gene, we regress the observed GE (and GE^2) on the covariates, then take the residuals as the adjusted GE (and GE^2) to be used. Next, for each gene with (adjusted) GE level X , we fit the following linear models

$$X = Z\beta_1 + \epsilon_1, \quad (1)$$

$$X^2 = Z\beta_2 + \epsilon_2, \quad (2)$$

where Z is the genotype matrix of the cis-SNPs for this gene. We estimated the regression coefficients β_1 and β_2 using backward step-wise variable regression with the AIC for variable selection, though other methods (such as elastic net penalized regression) can be equally applied.

To avoid biases due to weak instrumental variables (IVs) (while ensuring that 1 of the 3 valid IV assumptions holds; that is, an IV is associated with GE or GE^2), we performed the F-test on each gene in Stage 1; only those genes with their F-statistics >10 were retained for Stage 2 analysis as to be described next.

Standard TWAS (TWAS-L) and extensions (TWAS-Q and TWAS-LQ)

In Stage 2 of TWAS, for a given gene, we may consider three linear models:

$$Y = \theta_L \hat{X} + e, \quad (3)$$

$$Y = \theta_Q \hat{X}^2 + e, \quad (4)$$

$$Y = \theta_{LQ,1} \hat{X} + \theta_{LQ,2} \hat{X}^2 + e, \quad (5)$$

where Y is the vector of GWAS trait, \hat{X} and \hat{X}^2 are vectors of imputed (mean or linear) GE and imputed GE^2 levels, respectively, and e is the vector of normal noises with mean 0. We call model Eq. (3) TWAS-L, Eq. (4) TWAS-Q and Eq. (5) TWAS-LQ for their modeling only linear, only quadratic and both linear and quadratic effects of GE on the trait, respectively. TWAS-L is the standard TWAS that has been exclusively used in practice.

With only GWAS summary data (i.e. no individual-level genotypic data) in Stage 2, as usual, we need a reference panel for genotypic data along with a pretrained Stage 1 model (i.e. with $\hat{\beta}_1$ and $\hat{\beta}_2$); then we perform the F-tests to obtain the P -values in the above three models with the corresponding null and alternative hypotheses: (i) $H_{0,L}$: $\theta_L = 0$ versus $\theta_L \neq 0$; (ii) $H_{0,Q}$: $\theta_Q = 0$ versus $\theta_Q \neq 0$; and (iii) $H_{0,LQ}$: $(\theta_{LQ,1}, \theta_{LQ,2}) = (0, 0)$ versus $(\theta_{LQ,1}, \theta_{LQ,2}) \neq (0, 0)$, for TWAS-L, TWAS-Q and TWAS-LQ, respectively. Details are given in the [Supplementary Material](#).

Model checking and robust TWAS

We also performed model checking for TWAS via the TEDE method (21) on the three linear models in TWAS-L, TWAS-Q and TWAS-LQ. The TEDE method tests whether there are any direct effects of the SNPs other than mediated through the gene. Since the original TEDE test applies to only the standard TWAS-L (or TWAS-Q) with only one single term of imputed GE, we extended it to TWAS-LQ with two (or more) imputed terms. Although more details are given in the [Supplementary Material](#), here we give a brief description. First, the direct effects of the SNPs are explicitly specified in the Stage 2 model of TWAS: for individual i ,

$$Y_i = \sum_{j=1}^p \alpha_j Z_{i,j} + W_i^T \theta + \epsilon_i, \quad (6)$$

where α_j is the direct effect of the j th SNP, $Z_{i,j}$ is the j th SNP for subject i , W_i is for one or more terms of imputed GE and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is the error term. To test for no direct effects, the TEDE method tests the null hypothesis $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$ (with a score test).

Alternatively, to account for possible horizontal (and directional) pleiotropy of the SNPs while estimating the causal effect of a gene i.e. θ , similar to MR-Egger (44), we can treat and fit model Eq. (6) as a LME model: we assume that $\alpha_j \sim N(\mu, \sigma_\alpha^2)$ iid are random effects, θ are fixed effects and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ iid are errors. Corresponding to Eqs. (3) to (5), the fixed effects W_i are $(1, \hat{X}_i)^T$, $(1, \hat{X}_i^2)^T$, $(1, \hat{X}_i, \hat{X}_i^2)^T$, respectively (with 1 for the intercept). We call these three corresponding LME models as TWAS-L-LME, TWAS-Q-LME and TWAS-LQ-LME. We can equivalently rewrite $\alpha_j \sim N(0, \sigma_\alpha^2)$ iid by including μ as a component of the fixed effects as follows:

$$Y_i = \sum_{j=1}^p \alpha_j Z_{i,j} + \mu \sum_{j=1}^p Z_{i,j} + W_i^T \theta + \epsilon_i, \quad (7)$$

where μ is the average pleiotropic effect. We fitted the mixed effects model Eq. (7) with R package *nlme* to individual-level GWAS data. A simplified version of each model, similar to TWAS-Egger (45), can be fitted with GWAS summary data, though we do not pursue it here.

A special case of each TWAS-L-LME, TWAS-Q-LME and TWAS-LQ-LME is to specify balanced pleiotropy with $\mu = 0$, as in MR-IVW (random-effects) (46) and RAPS (47); we call the corresponding models TWAS-L-LME0, TWAS-Q-LME0 and TWAS-LQ-LME0, respectively. Although such an LME0 model is more restrictive in not allowing directional pleiotropy, it may avoid the problem in LME (and MR-Egger) of the dependence on the orientations of the SNPs/IVs.

Remarks

It is noted that in the extended TWAS, or more generally in IV regression, to account for nonlinear effects of predictor or GE X on Y , in general one needs to impute the nonlinear effects explicitly: for the quadratic effect of X , we propose imputing or estimating $E(X^2|Z)$, which differs from $E(X|Z)^2$. However, in Eq. (1), under the usual assumption of $\text{var}(\epsilon_1) = \sigma_1^2$ being constant, we have

$$E(X^2|Z) = [E(X|Z)]^2 + \text{var}(X|Z) = [Z\beta_1]^2 + \sigma_1^2,$$

by which the quadratic GE X^2 can be imputed on the basis of the imputed X , $\hat{X} = E(\hat{X}|Z)$, instead of imputing X^2 directly in Eq. (2). In fact, by an argument of recursion, other higher moments of GE, such as the cubic GE, can be imputed as some higher order polynomials of \hat{X} . On the other hand, such an imputed quadratic (or other higher order) GE term is also a quadratic (or other higher order) function of genotypes Z , leading to its requirement of using individual-level GWAS data in Stage 2 of TWAS.

It turned out that this alternative implementation did not perform as well as using Eq. (2), perhaps due to

the former's dependence on the sufficient adequacy of imputing X in Eq. (1). Given often relatively small sample sizes of an eQTL dataset, it is perhaps better to empirically impute X^2 directly in Eq. (2), instead of completely depending on using \hat{X} in Eq. (1). Alternatively, it may be also due to the nonconstant variance $\text{var}(\epsilon_1)$, which can be caused by omitted SNP-SNP or SNP-environment interactions, leading to the presence of veQTL for model (2). Importantly, because it is more likely (and convenient) to have individual-level (and smaller) eQTL data but only (larger) GWAS summary data, we recommend the use of Eq. (2) in the previous implementation and will skip the discussion of the alternative implementation.

Data availability

The ADNI data are available to the approved user at the ADNI site (<http://adni.loni.usc.edu>). The IGAP and Jansen's AD GWAS summary data can be downloaded at <https://www.ebi.ac.uk/gwas/studies/GCST002245> and https://ctg.cncr.nl/software/summary_statistics, respectively. The GTEx data and UKB individual-level eQTL/GWAS data are available at the dbGaP (<https://www.ncbi.nlm.nih.gov/gap/>) and UK Biobank (<https://www.ukbiobank.ac.uk/>), respectively, to the approved user.

Supplementary Material

Supplementary Material is available at HMG online.

Acknowledgement

The authors thank a reviewer for constructive comments.

Conflict of Interest statement. None declared.

Funding

This work was supported by the National Institutes of Health grants R01AG065636, RF1AG067924 and U01AG073079 and by the Minnesota Supercomputing Institute. The access to the GTEx and UKB data was approved through dbGaP Project #26511 and UKB Application #35107, respectively.

References

- Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E. and Cox, N.J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, **6**, e1000888.
- Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I. and Dermitzakis, E.T. (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.*, **6**, e1000895.
- Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S. et al. (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.

4. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., De Geus, E.J., Boomsma, D.I., Wright, F.A. et al. (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, **48**, 245–252.
5. 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. *Nat. Genet.*, **47**, 1091–1098.
6. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B.M. et al. (2018) Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.*, **50**, 538–548.
7. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A. and Pasaniuc, B. (2017) Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.*, **100**, 473–487.
8. Raj, T., Li, Y.I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.-C., Ng, B., Gupta, I., Haroutunian, V. et al. (2018) Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.*, **50**, 1584–1592.
9. Cao, Y., Wei, P., Bailey, M., Kauwe, J.S., Maxwell, T.J. and for the Alzheimer's Disease Neuroimaging Initiative (2014) A versatile omnibus test for detecting mean and variance heterogeneity. *Genet. Epidemiol.*, **38**, 51–59.
10. Sarkar, A.K., Tung, P.-Y., Blischak, J.D., Burnett, J.E., Li, Y.I., Stephens, M. and Gilad, Y. (2019) Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet.*, **15**, e1008045.
11. Wiggins, G.A., Black, M.A., Dunbier, A., Merriman, T.R., Pearson, J.F. and Walker, L.C. (2021) Variable expression quantitative trait loci analysis of breast cancer risk variants. *Sci. Rep.*, **11**, 1–10.
12. Steele, N.Z., Carr, J.S., Bonham, L.W., Geier, E.G., Damotte, V., Miller, Z.A., Desikan, R.S., Boehme, K.L., Mukherjee, S., Crane, P.K. et al. (2017) Fine-mapping of the human leukocyte antigen locus as a risk factor for Alzheimer disease: a case-control study. *PLoS Med.*, **14**, e1002272.
13. Rustenhoven, J., Smith, A.M., Smyth, L.C., Jansson, D., Scotter, E.L., Swanson, M.E., Aalderink, M., Coppeters, N., Narayan, P., Handley, R. et al. (2018) PU. 1 regulates Alzheimer's disease-associated genes in primary human microglia. *Mol. Neurodegener.*, **13**, 1–16.
14. 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 $A\beta$, tau, immunity and lipid processing. *Nat. Genet.*, **51**, 414–430.
15. Lambert, J.-C., Grenier-Boley, B., Chouraki, V., Heath, S., Zelenika, D., Fievet, N., Hannequin, D., Pasquier, F., Hanon, O., Brice, A. et al. (2010) Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis. *J. Alzheimers Dis.*, **20**, 1107–1118.
16. Mansouri, L., Klai, S., Gritli, N., Fekih-Mrissa, N., Messalmani, M., Bedoui, I., Derbali, H. and Mrissa, R. (2015) Association of HLA-DR/DQ polymorphism with Alzheimer's disease. *Am. J. Med. Sci.*, **349**, 334–337.
17. Shintani, S., Ohyama, H., Zhang, X., McBride, J., Matsuo, K., Tsuji, T., Hu, M.G., Hu, G., Kohno, Y., Lerman, M. et al. (2000) p12DOC-1 is a novel cyclin-dependent kinase 2-associated protein. *Mol. Cell. Biol.*, **20**, 6300–6307.
18. Cheng, A., Xiong, W., Ferrell, Jr, J.E. and Solomon, M.J. (2005) Identification and comparative analysis of multiple mammalian speedy/Ringo proteins. *Cell Cycle*, **4**, 155–165.
19. Li, K.K., Ng, I.O., Fan, S.T., Albrecht, J.H., Yamashita, K. and Poon, R.Y. (2002) Activation of cyclin-dependent kinases CDC2 and CDK2 in hepatocellular carcinoma. *Liver*, **22**, 259–268.
20. Otto, T. and Sicinski, P. (2017) Cell cycle proteins as promising targets in cancer therapy. *Nat. Rev. Cancer*, **17**, 93–115.
21. Deng, Y. and Pan, W. (2021) Model checking via testing for direct effects in Mendelian randomization and transcriptome-wide association studies. *PLoS Comput. Biol.*, **17**, e1009266.
22. Nakamura, N., Hatano, E., Iguchi, K., Sato, M., Kawaguchi, H., Ohtsu, I., Sakurai, T., Aizawa, N., Iijima, H., Nishiguchi, S. et al. (2019) Elevated levels of circulating ITIH4 are associated with hepatocellular carcinoma with nonalcoholic fatty liver disease: from pig model to human study. *BMC Cancer*, **19**, 1–14.
23. Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell*, **169**, 1177–1186.
24. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K. et al. (2019) Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.*, **51**, 592–599.
25. Yao, D.W., O'Connor, L.J., Price, A.L. and Gusev, A. (2020) Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.*, **52**, 626–633.
26. Celaj, A., Gebbia, M., Musa, L., Cote, A.G., Snider, J., Wong, V., Ko, M., Fong, T., Bansal, P., Mellor, J.C. et al. (2020) Highly combinatorial genetic interaction analysis reveals a multi-drug transporter influence network. *Cell Syst.*, **10**, 25–38.e10.
27. Campbell, R.F., McGrath, P.T. and Paaby, A.B. (2018) Analysis of epistasis in natural traits using model organisms. *Trends Genet.*, **34**, 883–898.
28. Jiang, X., Holmes, C. and McVean, G. (2021) The impact of age on genetic risk for common diseases. *PLoS Genet.*, **17**, e1009723.
29. Xiang, Q., Andersen, S.L., Perls, T.T. and Sebastiani, P. (2020) Studying the interplay between apolipoprotein E and education on cognitive decline in centenarians using Bayesian beta regression. *Front. Genet.*, **11**, 1673.
30. Grinberg, N.F. and Wallace, C. (2021) Multi-tissue transcriptome-wide association studies. *Genet. Epidemiol.*, **45**, 324–337.
31. Okoro, P.C., Schubert, R., Guo, X., Johnson, W.C., Rotter, J.I., Hoeschele, I., Liu, Y., Im, H.K., Luke, A., Dugas, L.R. et al. (2021) Transcriptome prediction performance across machine learning models and diverse ancestries. *Hum. Genet. Genom. Adv.*, **2**, 100019.
32. Reich, D.E. and Goldstein, D.B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.*, **20**, 4–16.
33. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.
34. Hu, X., Zhao, J., Lin, Z., Wang, Y., Peng, H., Zhao, H., Wan, X. and Yang, C. (2021) MR-APSS: a unified approach to Mendelian randomization accounting for pleiotropy and sample structure using genome-wide summary statistics. *bioRxiv*. doi: <https://doi.org/10.1101/2021.03.11.434915>
35. Sulc, J., Sjaarda, J. and Kutalik, Z. (2021) Polynomial Mendelian randomization reveals widespread non-linear causal effects in the UK biobank. *bioRxiv*. doi: <https://doi.org/10.1101/2021.12.08.471751>
36. Zhang, J., Xie, S., Gonzales, S., Liu, J. and Wang, X. (2020) A fast and powerful eQTL weighted method to detect genes associated

- with complex trait using GWAS summary data. *Genet. Epidemiol.*, **44**, 550–563.
37. 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. *Nat. Genet.*, **51**, 568–576.
 38. Feng, H., Mancuso, N., Gusev, A., Majumdar, A., Major, M., Pasaniuc, B. and Kraft, P. (2021) Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. *PLoS Genet.*, **17**, e1008973.
 39. Shen, L., Thompson, P.M., Potkin, S.G., Bertram, L., Farrer, L.A., Foroud, T.M., Green, R.C., Hu, X., Huentelman, M.J., Kim, S. et al. (2014) Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. *Brain Imaging Behav.*, **8**, 183–207.
 40. Consortium, G et al. (2020) The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
 41. Lambert, J.-C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A.L., Bis, J.C., Beecham, G.W. et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.*, **45**, 1452–1458.
 42. 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 metaanalysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.*, **51**, 404–413.
 43. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S. et al. (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274.
 44. Bowden, J., Davey Smith, G. and Burgess, S. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *Int. J. Epidemiol.*, **44**, 512–525.
 45. Knutson, K.A., Deng, Y. and Pan, W. (2020) Implicating causal brain imaging endophenotypes in Alzheimer's disease using multivariable IWAS and GWAS summary data. *Neuroimage*, **223**, 117347.
 46. Burgess, S., Butterworth, A. and Thompson, S.G. (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.*, **37**, 658–665.
 47. Zhao, Q., Wang, J., Hemani, G., Bowden, J. and Small, D.S. (2020) Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann. Stat.*, **48**, 1742–1769.