A method to adjust for the uncertainty of predicted expression in transcriptome-wide association study

Arunabha Majumdar*[1] and Tanushree Haldar[2]

[1]Department of Mathematics, Indian Institute of Technology Hyderabad, Kandi, Telangana 502285, India

[2]Institute for Human Genetics, University of California San Francisco, San Francisco, California 94143, USA

*Correspondence: arun.majum@math.iith.ac.in

Abstract

Transcriptome-wide association study (TWAS) is a powerful approach to identifying novel genes associated with complex phenotypes. Standard TWAS approaches build a prediction model for the genetic component of expression based on reference transcriptome data. Next, an outcome is regressed on the predicted expression in separate GWAS data. The traditional TWAS approach disregards the uncertainty of predicted expression, which can lead to unreliable inference on gene-phenotype associations. We propose a novel approach to adjust for the uncertainty of predicted expression in TWAS. We adapt techniques from measurement error theory and implement bootstrapping algorithms for penalized regression to explicitly obtain an adjustment factor that needs to be incorporated in the unadjusted TWAS. We base the framework on adaptive Lasso. We use extensive simulations to show that the traditional TWAS inflates the type I error rate, whereas the adjusted TWAS adequately controls it. At the expense of an inflated false positive rate, the unadjusted TWAS offers a limited increase in power than the adjusted TWAS, which is statistically unjustified. Our approach produces more accurate estimates of the gene's effect size than a traditional approach. We demonstrate the merits of the adjusted approach by conducting TWAS for height and lipid phenotypes, LDL, HDL, and triglycerides while integrating the Geuvadis transcriptome and UK Biobank GWAS data.

Keywords: Gene-phenotype associations, predicted expression, type I error rate, penalized regression, residual bootstrap, trimmed mean.

1 Introduction

Genome-wide association studies (GWAS) have successfully associated single nucleotide polymorphism (SNP) loci with various complex human phenotypes. Overwhelmingly, GWAS have reported more than 167,000 SNP associations with numerous diseases and phenotypes in the NHGRI-EBI GWAS catalog [1]. The tremendous success in discovering SNP associations has inspired translational medicine studies to strive for improving precision medicine for many diseases, bringing into service genetic discoveries [2]. However, most of the GWAS signals reside in the non-coding regions of the human genome [3], lacking a comprehensive understanding of molecular mechanisms underlying the associations. The missing link between the susceptibility SNPs and the genes impacting the phenotype inhibits the translation of GWAS findings to clinical settings [2]. Non-coding regions of the human genome, e.g., intergenic regions, can regulate genes' transcriptional or translational activities, termed regulatory elements of gene expression. An expression quantitative trait locus, eQTL, is a locus that explains a fraction of the variance of a gene's expression [4]. Thus, eQTLs comprise various categories of regulatory elements, e.g., enhancers, promoters, transcription factor binding sites, etc. Recent expression mapping studies [5,6] have shown that local eQTLs (located near the gene) are colocalized with GWAS signals. Another study [7] demonstrated that GWAS-identified SNPs are more likely to be eQTLs. Therefore, a genetic variant identified by GWAS can influence the phenotype by modulating the expression level of its neighboring gene. This crucial speculation motivated the development of the transcriptome-wide association study (TWAS) framework.

TWAS [8,9] aggregates regulatory effects of multiple eQTLs and tests for association between a gene and an outcome. It combines the genotype and outcome data from GWAS with reference transcriptome data. For example, consider the lipid phenotype, HDL, the outcome of interest, and liver tissue as the relevant tissue type. As an example of the reference transcriptome data, the GTEx data [10,11] has measurements of gene expressions in various tissue types and genotypes for genome-wide SNPs. In the first step of TWAS, a

gene and its expression in the liver tissue are considered. Next, consider a set of local SNPs surrounding the gene, for example, within one megabase region from the gene's transcription start and end sites. We fit a regression model of expression on the local SNPs' genotypes in GTEx data to learn a prediction model for the expression. In the second step, we have separate GWAS data. For example, the UK Biobank [12] has data for HDL and genotypes for genome-wide SNPs, but expression data is unavailable. We implement the prediction model of expression obtained in the GTEx data to predict the expression for the GWAS individuals in the UK Biobank. Finally, we regress HDL on the predicted genetically regulated expression component to test for an association between HDL and the gene. TWAS has discovered many novel gene-phenotype associations and reconfirmed associations between numerous genetic loci and phenotypes initially identified by GWAS [13,14].

A significant criticism of the traditional TWAS approach is that it disregards the uncertainty of the predicted expression in the second-step regression. The uncertainty arises due to the variance of the prediction model for expression estimated in the reference transcriptome data. A recent study [15] demonstrated that the TWAS inference on gene-phenotype association, more specifically, on the relationship between the genetic component of expression and the outcome, can be subject to an uncontrolled rate of false positives due to such non-adjustment of the uncertainty in the predicted expression. They highlighted that many of the significant TWAS associations found in their real data analyses did not have substantial evidence to be accurate [15]. The inflation of false positive rates can reduce the reliability of the detected gene-phenotype associations. Therefore, extensive effort and valuable resources to reveal the biological mechanisms underlying the TWAS associations can be unsuccessful in follow-up functional studies. Despite considerable benefits and advantages over GWAS, TWAS must overcome this limitation.

This article investigates this vital question under a frequentist framework. We term the commonly used methods for TWAS that ignore the uncertainty of predicted expression as unadjusted TWAS. We adapt measurement error theory [16] to derive the adjustment factor that must be multiplied by the test statistic obtained from an unadjusted TWAS to account

for the uncertainty of imputed expression. We apply bootstrapping techniques for penalized regression [17] to estimate the adjustment factor and devise the procedure of testing gene-phenotype association. We base the approach on the adaptive Lasso framework [18] and refer to it as penaTWAS (**p**redicted **e**xpression **n**oise **a**djusted TWAS). We develop the approach for TWAS based on both individual-level and summary-statistics data. The latter requires summary-level association data from GWAS instead of individual-level data [8]. Therefore, the method can easily be applied to complex phenotypes using publicly available summary-level GWAS data.

We modify an adjusted variance approach proposed by Xue et al. [19] to account for the uncertainty of predicted expression, which requires individual-level data. Using simulations for individual-level data, we first show that the unadjusted TWAS inflates the type I error rate, whereas the penaTWAS and the adjusted variance approach adequately control the type I error rate. Unadjusted TWAS yields a limited increase in power than the penaTWAS and the adjusted variance approach at the cost of an inflated false positive rate, which is statistically invalid. PenaTWAS yields marginally higher power than the adjusted variance approach except for a higher choice of gene expression heritability when the latter leads to a marginal power gain. In penaTWAS, we obtain the adjustment of the gene-outcome effect size estimate. We show that the adjusted estimates of effect sizes are more accurate than those obtained by the unadjusted approaches. Even though we develop the approach for a continuous phenotype, we also apply the same adjustment factor in TWAS of a binary case-control phenotype. The main benefit here is that the effect size estimates obtained by penaTWAS are more accurate.

Next, for a TWAS of continuous outcome based on summary-level data, we confirm that the unadjusted TWAS inflates the false positive rate while the summary-level penaTWAS controls it at the desired level. Due to the type I error rate inflation, the unadjusted TWAS leads to a limited power increase. To demonstrate the method for real data, we implement the summary-level penaTWAS and an unadjusted TWAS for height, HDL cholesterol, LDL cholesterol, and triglycerides, combining Geuvadis transcriptome and UK Biobank GWAS

data.

2 Materials and methods

We present a schematic diagram outlining the main steps of the proposed method in Figure 1. We use the following main notations to describe the methods sections.

2.1 Various notations used for the transcriptome data

n_1: number of individuals.

$\mathbf{E}$: vector of expression values for a gene for n_1 individuals in the specific tissue of interest.

E_i: expression for i^{th} individual, $i = 1, \ldots, n_1$.

p: number of local SNPs for the gene.

G_{ref}: $n_1 \times p$ genotype matrix at local SNPs for individuals in reference transcriptome data.

$\mathbf{g}_{\text{ref},i}$: genotype vector at local SNPs for i^{th} individual in transcriptome data, $i = 1, \ldots, n_1$.

$\boldsymbol{\alpha}$: effect size vector for p local SNPs on the expression of the gene.

2.2 Various notations used for the GWAS data

n: number of individuals.

$\mathbf{y}$: vector of outcome values for n individuals in GWAS.

G: genotype matrix of order $n \times p$ for GWAS individuals at p local SNPs.

x_i: true genetic component of expression for i^{th} individual in GWAS, $i = 1, \ldots, n$.

X_i: predicted genetic component of expression for i^{th} individual in GWAS, $i = 1, \ldots, n$.

β_1: effect of the actual genetic component of expression on the outcome in GWAS.

b_1: effect of the predicted genetic component of expression on the outcome.

$\mathbf{z}_{\text{gwas}}$: vector of Z (Wald) statistics for marginal SNP-phenotype association at the p local SNPs in GWAS.

Z_{unadj}: Z-statistic for the unadjusted TWAS.

Z_{pena}: Z-statistic for the penaTWAS.

V: linkage disequilibrium (LD) matrix for the p local SNPs in GWAS.

2.3 Regression of gene expression on local SNPs in transcriptome data

Consider a given gene and its expression in a tissue of interest in the reference transcriptome data. Select a set of local SNPs located near the gene. We regress the gene's expression on the genotypes of the local SNPs. Consider the following linear model:

$$\mathbf{E} = G_{\text{ref}}\boldsymbol{\alpha} + \boldsymbol{\eta} \tag{1}$$

In G_{ref}, genotypes for each SNP in a column are centered for mean and then divided by standard deviation. $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ denotes the effect size vector, and $\boldsymbol{\eta}$ denotes the vector of residuals. We assume η_is are independent and normally distributed with mean zero and variance σ_η^2. Since the number of local SNPs can be larger than the sample size, a penalized regression is implemented in practice to estimate $\boldsymbol{\alpha}$. Lasso [20] and Elastic net [21] are two commonly used methods. In Lasso, $\hat{\boldsymbol{\alpha}}$ is obtained by minimizing the penalized loss function:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \sum_{i=1}^{n_1}(E_i - \mathbf{g}'_{\text{ref},i}\boldsymbol{\alpha})^2 + \lambda\sum_{j=1}^{p}|\alpha_j| \tag{2}$$

Here, $\mathbf{g}_{\text{ref},i}$ denotes the genotype vector for i^{th} individual, $i = 1, \ldots, n_1$. $\lambda \geq 0$ denotes the weight associated with the penalty term, which determines the amount of shrinkage. In Elastic net, $\hat{\boldsymbol{\alpha}}$ is given by:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \sum_{i=1}^{n_1}(E_i - \mathbf{g}'_{\text{ref},i}\boldsymbol{\alpha})^2 + \lambda\{\delta\sum_{j=1}^{p}|\alpha_j| + (1-\delta)\sum_{j=1}^{p}\alpha_j^2\} \tag{3}$$

Here, $\delta \in (0, 1)$ determines the relative contribution of the Lasso and ridge regression penalty and is chosen as 0.5 in practice. An optimal λ is chosen by cross-validation. Let $\hat{\boldsymbol{\alpha}}$ denote the estimate of $\boldsymbol{\alpha}$ obtained by the preferred penalized regression. Let $\text{E}(\hat{\boldsymbol{\alpha}})$ and $\text{cov}(\hat{\boldsymbol{\alpha}})$ denote the expectation vector and covariance matrix of $\hat{\boldsymbol{\alpha}}$.

2.4 Regression of phenotype on the genetic component of expression in GWAS data

Consider GWAS data; the subjects are non-overlapping with the transcriptome data. In general, expression measurements are not available in GWAS data. Assume that the data has the genotypes for the set of p local SNPs considered for the gene in the transcriptome data. We assume that the transcriptome and GWAS data are from the same population such that the effect of local SNPs on the expression is the same between the two datasets. If $\mathbf{g}$ denotes the genotype vector for the local SNPs in GWAS, the true genetic component of expression is given by $x = \mathbf{g}'\boldsymbol{\alpha}$. Since $\boldsymbol{\alpha}$ is not known, we predict x by $X = \mathbf{g}'\hat{\boldsymbol{\alpha}}$, where $\hat{\boldsymbol{\alpha}}$ was obtained in the first-step regression. The main goal of TWAS is to test for an association between the outcome and the true genetic component of expression. Let Y denote the phenotype of interest and $\mathbf{y} = (y_1, \ldots, y_n)$ be the phenotype values for the GWAS individuals. For i^{th} individual, let $\mathbf{g}_i$ denote the genotype vector for the p local SNPs, hence, $x_i = \mathbf{g}_i'\boldsymbol{\alpha}$, and $X_i = \mathbf{g}_i'\hat{\boldsymbol{\alpha}}$, $i = 1, \ldots, n$. Thus, X_i estimates x_i with some measurement error.

2.5 Ideal and working regressions

We consider two regressions, one considering x as the regressor and the other considering X as the regressor, where x is unobserved, and X is observed. We specify the ideal regression as the following:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n \tag{4}$$

β_1 is the effect of the true genetic component of expression on the phenotype. Assume that $\{\epsilon_i, i = 1, \ldots, n\}$ are i.i.d. $N(0, \sigma_\epsilon^2)$. Next, the working regression is specified as follows:

$$y_i = b_0 + b_1 X_i + e_i, i = 1, \ldots, n \tag{5}$$

Here, b_1 is the effect of the noisy predicted expression on the phenotype, and $\{e_i, i = 1, \ldots, n\}$ are i.i.d. $N(0, \sigma_e^2)$. We note that the definitions of the two regressions go along similar

lines of linear regression with measurement error in the regressors [16]. TWAS is a two-stage, two-sample framework where the two samples from the transcriptome and GWAS data are non-overlapping and independent. Hence, ηs in equation 1 and ϵs in equation 4 are independent.

In the unadjusted TWAS, we implement the working regression (equation 5), ignoring the uncertainty or measurement error in X observations. Note that $\mathrm{E}(X_i) = \mathbf{g}_i^{'}\mathrm{E}(\hat{\boldsymbol{\alpha}})$ and $\mathrm{var}(X_i) = \mathbf{g}_i^{'}\mathrm{cov}(\hat{\boldsymbol{\alpha}})\mathbf{g}_i$. Our main goal is to test whether the true genetic component of expression affects the phenotype, i.e., $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. However, in unadjusted TWAS, we test for $H_0 : b_1 = 0$ vs $b_1 \neq 0$ using the least squares estimator of $b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$.

From the ideal regression model (equation 4), $\mathrm{E}(y_i) = \beta_0+\beta_1 x_i$ and $\mathrm{var}(y_i) = \mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$. We make a reasonable assumption that $\boldsymbol{\alpha}$, hence $x_i = \mathbf{g}_i^{'}\boldsymbol{\alpha}$ is non-random. Under the ideal regression model, $\mathrm{cov}(X_i, y_i) = \mathrm{cov}(\mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}}, \beta_0 + \beta_1 x_i + \epsilon_i) = \mathrm{cov}(\mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}}, \beta_1 x_i) + \mathrm{cov}(\mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}}, \epsilon_i)$. Both β_1 and x_i are unobserved constants, hence $\mathrm{cov}(\mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}}, \beta_1 x_i) = 0$. If we consider $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \hat{\boldsymbol{\psi}}$, $\mathrm{cov}(\mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}}, \epsilon_i) = \mathrm{cov}(\mathbf{g}_i^{'}\boldsymbol{\alpha}, \epsilon_i) + \mathrm{cov}(\mathbf{g}_i^{'}\hat{\boldsymbol{\psi}}, \epsilon_i)$. Here, $\hat{\boldsymbol{\psi}}$ is the noise part in $\hat{\boldsymbol{\alpha}}$ based on the transcriptome data. Since $\mathbf{g}_i^{'}\boldsymbol{\alpha} = x_i$ is non-random, $\mathrm{cov}(\mathbf{g}_i^{'}\boldsymbol{\alpha}, \epsilon_i) = 0$. Because $\mathbf{g}_i$ is a constant vector and $\hat{\boldsymbol{\psi}}$ is the error part of $\hat{\boldsymbol{\alpha}}$, which is separately obtained based on the reference transcriptome data, we assume that $\mathrm{cov}(\mathbf{g}_i^{'}\hat{\boldsymbol{\psi}}, \epsilon_i) = 0$, where ϵ_i is the random residual in the ideal regression of the outcome in the GWAS. Thus, $\mathrm{cov}(X_i, y_i) = 0$, under the ideal regression model (equation 4). We assume that the transcriptome data sample size is sufficiently large such that $\hat{\boldsymbol{\alpha}}$ asymptotically follows multivariate normal [18], hence, X_i is distributed as normal. Thus, X_i and y_i are assumed to be independent under the ideal regression model, $i = 1, \ldots, n$.

2.5.1 An adjusted estimator and related test

The least squares estimator of b_1 obtained from the working regression (equation 5) is $\hat{b}_1 = \frac{S_{Xy}}{S_{XX}} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(y_i-\bar{y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$. Note that $\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}} = (\sum_{i=1}^{n} \mathbf{g}_i)^{'}\hat{\boldsymbol{\alpha}} = 0$. Here, $\sum_{i=1}^{n} \mathbf{g}_i = \mathbf{0}$, because each local SNP's genotype data are normalized to have a mean zero and variance

one. Thus, $\bar{X} = 0$ and $\hat{b}_1 = \frac{\sum_{i=1}^n X_i(y_i - \bar{y})}{\sum_{i=1}^n X_i^2}$. We note that $\hat{b}_1$ is an unbiased estimator of b_1. Our main goal is to obtain an estimator of β_1 and test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

We obtain an estimator of β_1 in the ideal regression based on $\hat{b}_1$ in the following. Note that $\mathrm{E}(\hat{b}_1) = \mathrm{E}\left(\frac{\sum_{i=1}^n X_i(y_i - \bar{y})}{\sum_{j=1}^n X_j^2}\right) = \sum_{i=1}^n \mathrm{E}\left(\frac{X_i}{\sum_{j=1}^n X_j^2}(y_i - \bar{y})\right) = \sum_{i=1}^n \mathrm{E}\left(\frac{X_i}{\sum_{j=1}^n X_j^2}\right)\mathrm{E}(y_i - \bar{y})$, since $(X_1, \ldots, X_n)$ and $(y_1, \ldots, y_n)$ are independent under the ideal regression model (equation 4). $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$ and $\mathrm{E}(\bar{y}) = \frac{1}{n}\sum_{i=1}^n E(y_i) = \frac{1}{n}\sum_{i=1}^n(\beta_0 + \beta_1 x_i) = \beta_0 + \frac{\beta_1}{n}\sum_{i=1}^n x_i = \beta_0$. Because, $\sum_{i=1}^n x_i = \sum_{i=1}^n \mathbf{g}_i^{'}\boldsymbol{\alpha} = (\sum_{i=1}^n \mathbf{g}_i)^{'}\boldsymbol{\alpha} = 0$, since $\sum_{i=1}^n \mathbf{g}_i = \mathbf{0}$. So, $E(y_i - \bar{y}) = \beta_1 x_i$.

Next, $X_i = \mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}}$. $\sum_{j=1}^n X_j^2 = \sum_{j=1}^n X_j^{'}X_j = \sum_{j=1}^n \hat{\boldsymbol{\alpha}}^{'}\mathbf{g}_j\mathbf{g}_j^{'}\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}^{'}\left(\sum_{j=1}^n \mathbf{g}_j\mathbf{g}_j^{'}\right)\hat{\boldsymbol{\alpha}} = n\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}$. Here, $V = \frac{1}{n}\sum_{j=1}^n \mathbf{g}_j\mathbf{g}_j^{'}$ denotes the linkage disequilibrium (LD) matrix of the p local SNPs based on the GWAS data. Thus, $\mathrm{E}\left(\frac{X_i}{\sum_{j=1}^n X_j^2}\right) = \mathrm{E}\left(\frac{\mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}}}{n\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}\right) = \frac{1}{n}\mathbf{g}_i^{'}\mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}\right)$. Thus, $\mathrm{E}(\hat{b}_1) = \sum_{i=1}^n \mathrm{E}\left(\frac{X_i}{\sum_{j=1}^n X_j^2}\right)\mathrm{E}(y_i - \bar{y}) = \frac{\beta_1}{n}\sum_{i=1}^n x_i\mathbf{g}_i^{'}\mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}\right) = \frac{\beta_1}{n}\sum_{i=1}^n(\boldsymbol{\alpha}^{'}\mathbf{g}_i)\mathbf{g}_i^{'}\mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}\right) = \frac{\beta_1}{n}\boldsymbol{\alpha}^{'}(\sum_{i=1}^n \mathbf{g}_i\mathbf{g}_i^{'})\mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}\right) = \beta_1\boldsymbol{\alpha}^{'}V\mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}\right) = B\beta_1$, where $B = \boldsymbol{\alpha}^{'}V\mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}\right)$ is a constant. Note that B involves unknown terms. Therefore,

$$\mathrm{E}(\hat{b}_1) = B \times \beta_1, \text{ where } B = \boldsymbol{\alpha}^{'}V\mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}\right) \tag{6}$$

Hence, $\mathrm{E}\left(\frac{\hat{b}_1}{B}\right) = \beta_1$. Consider $\hat{\beta}_1 = \frac{\hat{b}_1}{\hat{B}}$. From the above, $\hat{b}_1 = \frac{\sum_{i=1}^n(X_i - \bar{X})(y_i - \bar{y})}{\sum_{j=1}^n(X_j - \bar{X})^2}$. Note that $X_i = \mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}}$, and the randomness of X_i comes through $\hat{\boldsymbol{\alpha}}$. Hence, $\mathbf{X}$ is fixed for a given $\hat{\boldsymbol{\alpha}}$. By the law of total variance, $\mathrm{var}(\hat{b}_1) = \mathrm{E}(\mathrm{var}(\hat{b}_1|\mathbf{X})) + \mathrm{var}(\mathrm{E}(\hat{b}_1|\mathbf{X}))$.

$\mathrm{var}(\hat{b}_1|\mathbf{X}) = \mathrm{var}\left(\frac{\sum_{i=1}^n(X_i - \bar{X})(y_i - \bar{y})}{\sum_{j=1}^n(X_j - \bar{X})^2}\middle|\mathbf{X}\right) = \frac{1}{\{\sum_{j=1}^n(X_j - \bar{X})^2\}^2}\mathrm{var}\{\sum_{i=1}^n(X_i - \bar{X})(y_i - \bar{y})|\mathbf{X}\}$ $= \frac{1}{\{\sum_{j=1}^n(X_j - \bar{X})^2\}^2}\mathrm{var}\{\sum_{i=1}^n(X_i - \bar{X})y_i|\mathbf{X}\}$, since $\sum_{i=1}^n(X_i - \bar{X})|\mathbf{X} = 0$. $\mathrm{var}\{\sum_{i=1}^n(X_i - \bar{X})y_i|\mathbf{X}\} = \sum_{i=1}^n(X_i - \bar{X})^2\mathrm{var}(y_i|X_i) = \sigma_e^2\sum_{i=1}^n(X_i - \bar{X})^2$. Based on the working regression model (5), $y_i|X_i$, $i = 1, \ldots, n$, are independently distributed with $\mathrm{var}(y_i|X_i) = \sigma_e^2$; hence, the covariance terms become zero. Thus, $\mathrm{var}(\hat{b}_1|\mathbf{X}) = \frac{\sigma_e^2}{\sum_{i=1}^n(X_i - \bar{X})^2}$. Furthermore, $\bar{X} = \mathbf{0}$, conditioned on $\mathbf{X}$. Hence, $\mathrm{var}(\hat{b}_1|\mathbf{X}) = \frac{\sigma_e^2}{\sum_{i=1}^n X_i^2} = \frac{\sigma_e^2}{n\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}$. Thus, $\mathrm{E}(\mathrm{var}(\hat{b}_1|\mathbf{X})) = \frac{\sigma_e^2}{n}\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}^{'}V\hat{\boldsymbol{\alpha}}}\right)$. The expectation concerns $\hat{\boldsymbol{\alpha}}$ based on the transcriptome data. The other part of $\mathrm{var}(\hat{b}_1)$ is $\mathrm{var}(\mathrm{E}(\hat{b}_1|\mathbf{X}))$. Again, we use that $\bar{X}|\mathbf{X} = 0$. Thus, $\mathrm{E}(\hat{b}_1|\mathbf{X}) = \mathrm{E}\left(\frac{\sum_{i=1}^n X_i(y_i - \bar{y})}{\sum_{j=1}^n X_j^2}\middle|\mathbf{X}\right)$ $= \sum_{i=1}^n \frac{X_i}{\sum_{j=1}^n X_j^2}\mathrm{E}\left((y_i - \bar{y})|\mathbf{X}\right)$. From the working regression model (5), $\mathrm{E}(y_i|X_i) = b_0 + b_1 X_i$. So $\mathrm{E}(\bar{y}|\mathbf{X}) = b_0$ and $\mathrm{E}\left((y_i - \bar{y})|\mathbf{X}\right) = b_1 X_i$. Thus, $\mathrm{E}(\hat{b}_1|\mathbf{X}) = b_1$ and $\mathrm{var}(\mathrm{E}(\hat{b}_1|\mathbf{X})) = 0$. Note

that conditioned on $\mathbf{X}$, $\hat{b}_1$ is an unbiased estimator of b_1. Thus, putting together all parts of the derivations,

$$\mathrm{var}(\hat{b}_1) = \frac{\sigma_e^2}{n}\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right) \tag{7}$$

If B was known in the equation 6, an unbiased estimator of β_1 would be $\frac{\hat{b}_1}{B}$ with the corresponding variance $\mathrm{var}\left(\frac{\hat{b}_1}{B}\right) = \frac{\sigma_e^2}{nB^2}\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$. However, B is unknown. To test for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, denote the Z (Wald) statistic in the penaTWAS as Z_{pena}. If we have $\hat{B}, \hat{\sigma}_e$, and $\hat{\mathrm{E}}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$ as the estimates of B, σ_e, and $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$, $Z_{\text{pena}} = \frac{\frac{\hat{b}_1}{\hat{B}}}{\frac{\hat{\sigma}_e}{\sqrt{n}\hat{B}}\sqrt{\hat{\mathrm{E}}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)}}$ $= \frac{\sqrt{n}\hat{b}_1}{\hat{\sigma}_e\sqrt{\hat{\mathrm{E}}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)}}$. The Z statistic for the unadjusted TWAS based on the working regression model (equation 5) is $\frac{\hat{b}_1}{se(\hat{b}_1)}$, where $\hat{b}_1 = \frac{\sum_{i=1}^n X_i(y_i - \bar{y})}{\sum_{j=1}^n X_j^2}$ and $\mathrm{var}(\hat{b}_1|\mathbf{X}) = \frac{\sigma_e^2}{n\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}$. Hence, the Z statistic for unadjusted TWAS is $Z_{\text{unadj}} = \frac{\sqrt{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\sqrt{n}\hat{b}_1}{\hat{\sigma}_e}$. Thus, we obtain the relation between Z_{pena} and Z_{unadj} as:

$$Z_{\text{pena}} = \hat{A} \times Z_{\text{unadj}}, \text{ where } A = \frac{\frac{1}{\sqrt{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}}}{\sqrt{\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)}} \tag{8}$$

To adjust for the uncertainty of the predicted expression, we need to multiply the Z statistic used in the unadjusted TWAS by an estimate of the adjustment factor A. Intuitively, the unadjusted TWAS ignores the expectation of the quantity $\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}$ in the denominator of A, in which case, A becomes one. In the following, we discuss an approach to estimating the adjustment factor A. To obtain an estimate of β_1, we use equation 6: $\hat{\beta}_1 = \frac{\hat{b}_1}{\hat{B}}$, where $B = \boldsymbol{\alpha}' V \mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$. Given an estimate of $\mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}'}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$, we plug in the estimate of $\boldsymbol{\alpha}$ based on full data, $\hat{\boldsymbol{\alpha}}$, and employ $\hat{B} = \hat{\boldsymbol{\alpha}}' V \hat{\mathrm{E}}\left(\frac{\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$.

2.6 Estimation of the adjustment factor

We need to estimate the adjustment factor since it is unknown for a given gene. $A = \frac{\frac{1}{\sqrt{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}}}{\sqrt{\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)}}$, where $\hat{\boldsymbol{\alpha}}$ is the estimated effects of local SNPs on expression based on the tran-

scriptome data, and V is the LD-matrix of the local SNPs based on the GWAS data. The numerator of A is already calculated from complete data. Deriving a closed form formula for $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$ and then estimating it is very challenging. Hence, we propose an estimation approach based on bootstrapping. Note that $\hat{\boldsymbol{\alpha}}$ but not V involves variability. We perform bootstrapping for the penalized regression of expression on local SNPs in the transcriptome data (equations 1, 2, 3). We employ the bootstrap expectation based on the bootstrapped estimates of $\boldsymbol{\alpha}$ to estimate $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$. There is a considerable amount of literature on bootstrapping for penalized regressions. We mainly focus on the residual bootstrap technique instead of a pair-wise bootstrap due to the former's better performance, given that the linear model assumption in equation 1 is correct.

The bootstrapped estimates of $\boldsymbol{\alpha}$ should correctly approximate the distribution of $\hat{\boldsymbol{\alpha}}$, the full-sample estimate. Chatterjee and Lahiri [22] showed that the estimates of $\boldsymbol{\alpha}$ based on the standard residual bootstrap technique for Lasso regression may not converge in distribution to $\hat{\boldsymbol{\alpha}}$. Chatterjee and Lahiri [17] proved a significant result: the residual bootstrap for adaptive Lasso [18] converges in distribution to the entire sample estimate. In the same paper, they proposed a version of residual bootstrap for Lasso, which converges. However, the algorithm is computationally intensive, rendering it infeasible for many genes (in the order of thousands). Hence, we develop the penaTWAS approach under the framework of adaptive Lasso [18]. The adaptive Lasso (ALasso) estimator for the local SNPs' effects on the expression is given by:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \sum_{i=1}^{n_1} (E_i - \mathbf{g}'_{\mathrm{ref},i} \boldsymbol{\alpha})^2 + \lambda \sum_{j=1}^{p} \frac{|\alpha_j|}{|\hat{\alpha}_{j,init}|^{\gamma}}, \tag{9}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$, and $\hat{\boldsymbol{\alpha}}_{init} = \{\hat{\alpha}_{j,init}, j = 1, \ldots, p\}$ denotes the initial estimate of $\boldsymbol{\alpha}$ based on the full sample. Here, p denotes the number of local SNPs, and λ denotes the weight associated with the penalty. The initial estimator of $\boldsymbol{\alpha}$ must be $\sqrt{n}$ consistent [18]. The ordinary least squares (OLS) estimator satisfies this condition and is commonly used to obtain the initial estimate. In the context of TWAS, the number of local SNPs for a

gene can be larger than the sample size of the transcriptome data. Some local SNPs can also be strongly correlated, i.e., in strong linkage disequilibrium (LD). Consequently, the OLS estimator may not be stable for many genes. On the other hand, the ridge regression estimator has the consistency property and can perform robustly in these scenarios. Hence, we implement the ridge regression to obtain the initial estimates of $\boldsymbol{\alpha}$ instead of the OLS approach.

A positive γ allows for varying the influence of the initial estimate on the penalty term. We consider $\gamma = 1$ for simplicity in this work. The adaptive Lasso allows differential shrinkage in the effects of the local SNPs on the expression. We use ten-fold cross-validation to choose the optimal λ. In ridge regression, we use ten-fold cross-validation, too. Next, we outline the residual bootstrap for adaptive Lasso to estimate $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$, where $\hat{\boldsymbol{\alpha}}$ is the full-sample ALasso estimate.

2.6.1 Residual bootstrap for adaptive Lasso

As defined above, let $\hat{\boldsymbol{\alpha}}$ denote the full sample ALasso estimate of $\boldsymbol{\alpha}$. Define the ALasso residuals as: $\hat{\epsilon}_i = E_i - \mathbf{g}'_{\mathrm{ref},i}\hat{\boldsymbol{\alpha}}$, and the mean-centered residuals as $r_i = \hat{\epsilon}_i - \frac{1}{n_1}\sum_{j=1}^{n_1}\hat{\epsilon}_j$, $i = 1, \ldots, n_1$. Let $(r_1^+, \ldots, r_{n_1}^+)$ denote a random sample of size n_1 drawn from $(r_1, \ldots, r_{n_1})$ with replacement. Define $E_i^+ = \mathbf{g}'_{\mathrm{ref},i}\hat{\boldsymbol{\alpha}} + r_i^+$, $i = 1, \ldots, n_1$. Consider $(E_i^+, \mathbf{g}_{\mathrm{ref},i}), i = 1, \ldots, n_1$ to implement ALasso:

$$\tilde{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \sum_{i=1}^{n_1}(E_i^+ - \mathbf{g}'_{\mathrm{ref},i}\boldsymbol{\alpha})^2 + \lambda \sum_{j=1}^{p} \frac{|\alpha_j|}{|\tilde{\alpha}_{j,ridge}|}, \tag{10}$$

where $\tilde{\boldsymbol{\alpha}}_{ridge} = \{\tilde{\alpha}_{j,ridge}, j = 1, \ldots, p\}$ is the ridge regression estimate of $\boldsymbol{\alpha}$ based on the residual bootstrap sample $\{(E_i^+, \mathbf{g}_{\mathrm{ref},i}), i = 1, \ldots, n_1\}$. We used the same choice of the penalty parameter in the ridge regression, which was obtained from the original complete data. We used five-fold cross-validation to choose the optimal penalty parameter in ALasso for each bootstrap sample. Thus, $\tilde{\boldsymbol{\alpha}}$ provides the bootstrapped ALasso estimate of $\boldsymbol{\alpha}$ for a given residual bootstrap sample. We regard T residual bootstrap samples and obtain T

bootstrap estimates of $\boldsymbol{\alpha}$: $\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_T$. We compute $\frac{1}{\tilde{\boldsymbol{\alpha}}_j' V \tilde{\boldsymbol{\alpha}}_j}, j = 1, \ldots, T$, where V is the LD matrix of local SNPs estimated based on the GWAS data. We use the sample mean of $\{\frac{1}{\tilde{\boldsymbol{\alpha}}_j' V \tilde{\boldsymbol{\alpha}}_j}, j = 1, \ldots, T\}$ to estimate $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$. Assuming V is positive-definite and $\tilde{\boldsymbol{\alpha}}_j \neq \mathbf{0}$, $\tilde{\boldsymbol{\alpha}}_j' V \tilde{\boldsymbol{\alpha}}_j > 0$, $\forall j$. We also consider trimmed mean of $\{\frac{1}{\tilde{\boldsymbol{\alpha}}_j' V \tilde{\boldsymbol{\alpha}}_j}, j = 1, \ldots, T\}$ as a robust estimate of $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$.

We devise another approach to estimate $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$ based on the bootstrap sample. Denote the multi-variable function as $f(\hat{\boldsymbol{\alpha}}) = \frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}$. Using Taylor's theorem for a multi-variable function, we can approximate $f(\hat{\boldsymbol{\alpha}}) \approx f(\hat{\boldsymbol{\alpha}}_{\text{obs}}) + (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{\text{obs}})' Df(\hat{\boldsymbol{\alpha}}_{\text{obs}}) + \frac{1}{2}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{\text{obs}})' Hf(\hat{\boldsymbol{\alpha}}_{\text{obs}})(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{\text{obs}})$. Here, $Df(\hat{\boldsymbol{\alpha}}_{\text{obs}})$ and $Hf(\hat{\boldsymbol{\alpha}}_{\text{obs}}) = D^2 f(\hat{\boldsymbol{\alpha}}_{\text{obs}})$ denote the gradient vector and the Hessian matrix of f evaluated at $\hat{\boldsymbol{\alpha}}_{\text{obs}}$, which is the original estimate of $\boldsymbol{\alpha}$ obtained by the adaptive Lasso based on the full data. Taking an expectation of both sides, we obtain: $\mathrm{E}(f(\hat{\boldsymbol{\alpha}})) \approx f(\hat{\boldsymbol{\alpha}}_{\text{obs}}) + \{\mathrm{E}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{\text{obs}})\}' Df(\hat{\boldsymbol{\alpha}}_{\text{obs}}) + \frac{1}{2}\mathrm{E}\{(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{\text{obs}})' Hf(\hat{\boldsymbol{\alpha}}_{\text{obs}})(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{\text{obs}})\}$. We estimate the bias term, $\mathrm{E}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{\text{obs}})$, using the bootstrap sample as $\frac{1}{T}\sum_{j=1}^{T}(\tilde{\boldsymbol{\alpha}}_j - \hat{\boldsymbol{\alpha}}_{\text{obs}})$, where $\tilde{\boldsymbol{\alpha}}_j$ is j^{th} bootstrap sample of $\boldsymbol{\alpha}$. Similarly, we estimate the expectation of the quardratic form $(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{\text{obs}})' Hf(\hat{\boldsymbol{\alpha}}_{\text{obs}})(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{\text{obs}})$ as: $\frac{1}{T}\sum_{j=1}^{T}\{(\tilde{\boldsymbol{\alpha}}_j - \hat{\boldsymbol{\alpha}}_{\text{obs}})' Hf(\hat{\boldsymbol{\alpha}}_{\text{obs}})(\tilde{\boldsymbol{\alpha}}_j - \hat{\boldsymbol{\alpha}}_{\text{obs}})\}$. Here, $Df(\hat{\boldsymbol{\alpha}}) = -\frac{2}{(\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}})^2} V \hat{\boldsymbol{\alpha}}$ and $Hf(\hat{\boldsymbol{\alpha}}) = -\frac{2}{(\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}})^2} V + \frac{8}{(\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}})^3} V \hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}' V$, which are evaluated at $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{\text{obs}}$.

For estimating the effect size β_1, we require an estimate of B which includes $\mathrm{E}\left(\frac{\hat{\boldsymbol{\alpha}}'}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$. We estimate this term based on the bootstrap sample as $\frac{1}{T}\sum_{j=1}^{T} \frac{1}{\tilde{\boldsymbol{\alpha}}_j' V \tilde{\boldsymbol{\alpha}}_j} \tilde{\boldsymbol{\alpha}}_j$. We note that the ALasso estimator $\hat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$. Altogether, $\hat{\beta}_1 = \frac{\hat{b}_1}{\hat{B}}$ (equation 6) is a sensible estimator of β_1.

2.7 An adjusted variance approach

Xue et al. [19] proposed an adjusted TWAS approach based on individual-level data. To tune for the uncertainty of predicted expression, they considered an adjusted variance of the estimated TWAS effect size [19]. Let $\hat{b}$ stand for the OLS estimate of b based on the working regression (Equation 5). Suppose $\mathrm{var}(\hat{b})$ denotes the variance of $\hat{b}$ based on the unadjusted TWAS. The calibrated variance [19] takes the following form: $\mathrm{var}_c(\hat{b}) = \mathrm{var}(\hat{b})\big(1 + \hat{b}^2 \frac{n}{n_1} \frac{\sigma_\eta^2}{\sigma_e^2}\big)$.

The adjusted variance $\text{var}_c(\hat{b})$ is inflated than the initial variance $\text{var}(\hat{b})$. Here, n_1 and n are the transcriptome and GWAS data sample sizes; σ_η^2 and σ_e^2 are the variances of the residual terms in the linear models (Equations 1 and 5) based on the transcriptome and GWAS data, respectively. Xue et al. [19] considered a less realistic scenario in the first stage of TWAS, where the number of local SNPs for a gene is much smaller than the sample size, allowing the simple multiple linear regression to predict the expression. On the contrary, the contemporary TWAS approaches consider penalized regression in the first stage of TWAS to incorporate many SNPs. However, the proposed adjusted variance can also be implemented for a penalized regression in the first stage. To modify the adjusted variance approach proposed by Xue et al., we consider the Lasso while regressing the expression on the genotypes of local SNPs (Equations 1). We follow Reid et al. [23] to estimate the error variance in a penalized regression. We devise the estimator of σ_η^2 as $\frac{1}{n_1-p_1}\sum_{i=1}^{n_1}\hat{\eta}_i^2$, where p_1 is the number of local SNPs selected in the Lasso variable selection, i.e., the number of non-zero estimated coefficients in $\hat{\boldsymbol{\alpha}}$ obtained by Lasso based on the complete transcriptome data. Here, $\hat{\eta}_i$, $i = 1, \ldots, n_1$, denote the estimated residuals from Lasso. σ_e^2 is estimated from the linear working regression in Equation 5. We note that the adjusted variance approach requires individual-level GWAS data for implementation.

2.8 Summary statistics version of penaTWAS

It may often be challenging to get access to individual-level GWAS data. Gusev et al. [8] proposed a summary statistics version of TWAS (summary TWAS), which only uses GWAS summary statistics and the prediction model for expression $\hat{\boldsymbol{\alpha}}$. As a result, it can easily be implemented for various phenotypes based on publicly available GWAS summary statistics. This work focuses on adjusting the Z statistic from the unadjusted summary TWAS. In the following, we outline the derivation of the unadjusted summary TWAS Z statistic. Consider the estimated vector of local SNPs' effects on expression, $\hat{\boldsymbol{\alpha}}$, obtained from the regression in equation 1. Next, consider the phenotype regression on the predicted genetic component of

expression in GWAS employing the working regression (equation 5).

$$\hat{b}_1 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{n}(\hat{\boldsymbol{\alpha}}'V\hat{\boldsymbol{\alpha}})^{-1}\hat{\boldsymbol{\alpha}}'G'\mathbf{y} = \frac{\hat{\boldsymbol{\alpha}}'G'\mathbf{y}}{n\hat{\boldsymbol{\alpha}}'V\hat{\boldsymbol{\alpha}}} \tag{11}$$

G denotes the genotype data matrix of the local SNPs in GWAS. Suppose the outcome Y is also individually regressed on the genotype of jth local SNP in GWAS: $\mathrm{E}(Y) = \gamma_j g_j$, where γ_j is the marginal effect of the SNP on the phenotype. If $\mathbf{g}_j$ denotes the normalized genotype vector of jth SNP, $\hat{\gamma}_j = (\mathbf{g}_j'\mathbf{g}_j)^{-1}\mathbf{g}_j'\mathbf{y} = \frac{1}{n}\mathbf{g}_j'\mathbf{y}$. Let $\boldsymbol{\gamma}$ denote the vector of marginal effects of the local SNPs on the outcome. Since, $\mathbf{g}_j'\mathbf{y} = n\hat{\gamma}_j$, $G'\mathbf{y} = n\hat{\boldsymbol{\gamma}}$. Thus, $\hat{b}_1 = \frac{\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\gamma}}}{\hat{\boldsymbol{\alpha}}'V\hat{\boldsymbol{\alpha}}}$. Note that $\hat{\boldsymbol{\alpha}}$ is considered non-random in the unadjusted TWAS. $\mathrm{var}(\hat{b}_1) = \frac{1}{(\hat{\boldsymbol{\alpha}}'V\hat{\boldsymbol{\alpha}})^2}\hat{\boldsymbol{\alpha}}'\mathrm{cov}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\alpha}}$. Here, $\mathrm{cov}(\hat{\boldsymbol{\gamma}}) = \frac{\sigma_y^2}{n}V$, where V is the LD matrix of the SNPs. Thus, $\mathrm{var}(\hat{b}_1) = \frac{\sigma_y^2}{n\hat{\boldsymbol{\alpha}}'V\hat{\boldsymbol{\alpha}}}$ and the unadjusted TWAS statistic $Z_{\mathrm{unadj}} = \frac{\sqrt{n}\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\gamma}}}{\hat{\sigma}_y\sqrt{\hat{\boldsymbol{\alpha}}'V\hat{\boldsymbol{\alpha}}}}$. Note that the Z statistic for marginally testing $\gamma_j = 0$ in the GWAS is given by: $\frac{\hat{\gamma}_j}{s.e.(\hat{\gamma}_j)} = \frac{\sqrt{n}\hat{\gamma}_j}{\hat{\sigma}_y}$. Hence, the vector of Z statistics for testing marginal effect vector $\boldsymbol{\gamma}$ obtained from the GWAS is: $\mathbf{z}_{\mathrm{gwas}} = \frac{\sqrt{n}}{\hat{\sigma}_y}\hat{\boldsymbol{\gamma}}$. Thus, the Z statistic for testing $b_1 = 0$ versus $b_1 \neq 0$ in the unadjusted TWAS is given by:

$$Z_{\mathrm{unadj}} = \frac{\hat{\boldsymbol{\alpha}}'\mathbf{z}_{\mathrm{gwas}}}{\sqrt{\hat{\boldsymbol{\alpha}}'V\hat{\boldsymbol{\alpha}}}} \tag{12}$$

To adjust for the uncertainty of the predicted expression in the unadjusted summary TWAS, we consider $Z_{\mathrm{pena}} = \hat{A} \times Z_{\mathrm{unadj}}$ (equation 8), where $\hat{A}$ is the estimated adjustment factor described in previous sections and Z_{unadj} is the unajusted summary-level TWAS Z statistic. Note that the adjusted summary TWAS, i.e., summary-level penaTWAS, requires summary-level GWAS association data. The approach can be implemented using publicly available GWAS summary statistics and the prediction model for expression based on ALasso and the adjustment factor. Without individual-level data from GWAS, we can use 1000 genome data to estimate V. Due to a relatively lower sample size of reference panel genotype data, regularization of the calculated LD matrix is preferred. We implement a simple regularization strategy following truncated SVD. After the eigenvalue decomposition of the matrix, we discard the negative eigenvalues if present and the corresponding eigenvectors. Next, we

retain the top eigenvalues (and the corresponding eigenvectors) explaining 99% of the total variance, i.e., the sum of the eigenvalues. Finally, we retrieve the LD matrix based on the retained eigenvalues and eigenvectors.

2.9 Expectation and variance under null hypothesis

To test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, we require the expectation and variance of Z_{pena} under H_0. For the summary TWAS, $Z_{\text{pena}} = \hat{A} \times Z_{\text{unadj}}$, where $A = \frac{\frac{1}{\sqrt{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}}}{\sqrt{\text{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)}}$ and $Z_{\text{unadj}} = \frac{\hat{\boldsymbol{\alpha}}' \mathbf{z}_{\text{gwas}}}{\sqrt{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}}$. Since the transcriptome and GWAS data have separate groups of unrelated individuals, $\hat{\boldsymbol{\alpha}}$ and $\mathbf{z}_{\text{gwas}}$ are independent. Also, V is a fixed LD matrix based on the GWAS or external data. Thus, $\text{E}_{H_0}(Z_{\text{pena}}) = \text{E}(\frac{\hat{A}}{\sqrt{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}} \hat{\boldsymbol{\alpha}}')\text{E}_{H_0}(\mathbf{z}_{\text{gwas}})$. Since the genotypes for each SNP and the phenotype data are normalized to have mean zero and variance one, $\mathbf{z}_{\text{gwas}}$ can be approximated as $\frac{1}{\sqrt{n}} G' \mathbf{y}$. Here it is also assumed that the heritability of Y due to a single SNP is negligible. Under the ideal regression model (Equation 4), $\text{E}(y_i) = \beta_1 x_i$; $\beta_0 = 0$ because ys are mean-centered. Thus, $\text{E}_{H_0}(\mathbf{z}_{\text{gwas}}) = \frac{\beta_1}{\sqrt{n}} G' \mathbf{x} = \mathbf{0}$, since $\beta_1 = 0$ under H_0. Hence, $\text{E}_{H_0}(Z_{\text{pena}}) = 0$. Next, we obtain the variance term under H_0.

$Z_{\text{pena}} = \frac{1}{\sqrt{\hat{C}}} \frac{\hat{\boldsymbol{\alpha}}' \mathbf{z}_{\text{gwas}}}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}$, where $C = \text{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$. $\text{var}(Z_{\text{pena}}) = \text{var}(\frac{\hat{\boldsymbol{\alpha}}' \mathbf{z}_{\text{gwas}}}{\sqrt{\hat{C}} \hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}})$. Denote $\mathbf{u} = \frac{1}{\sqrt{\hat{C}} \hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}} \hat{\boldsymbol{\alpha}}$ and $\mathbf{w} = \mathbf{z}_{\text{gwas}}$. Thus, $\text{var}(Z_{\text{pena}}) = \text{var}(\mathbf{u}'\mathbf{w})$. Note that $\text{var}(\mathbf{u}'\mathbf{w}) = \text{E}((\mathbf{u}'\mathbf{w})^2) - \text{E}^2(\mathbf{u}'\mathbf{w})$. Since $\mathbf{u}$ and $\mathbf{w}$ are independent, $\text{E}(\mathbf{u}'\mathbf{w}) = \text{E}(\mathbf{u}')\text{E}(\mathbf{w})$. Again, $\text{E}(\mathbf{w}) = \mathbf{0}$ under H_0. Thus, $\text{var}(\mathbf{u}'\mathbf{w}) = \text{E}((\mathbf{u}'\mathbf{w})^2) = \text{E}((\mathbf{u}'\mathbf{w}) \times (\mathbf{w}'\mathbf{u})) = \text{tr}(\text{E}(\mathbf{u}'\mathbf{w}\mathbf{w}'\mathbf{u})) = \text{E}(\text{tr}(\mathbf{u}'\mathbf{w}\mathbf{w}'\mathbf{u})) = \text{E}(\text{tr}(\mathbf{u}\mathbf{u}'\mathbf{w}\mathbf{w}')) = \text{tr}(\text{E}(\mathbf{u}\mathbf{u}'\mathbf{w}\mathbf{w}')) = \text{tr}(\text{E}(\mathbf{u}\mathbf{u}')\text{E}(\mathbf{w}\mathbf{w}'))$. Note that $\mathbf{w} = \mathbf{z}_{\text{gwas}} = \frac{1}{\sqrt{n}} G' \mathbf{y}$, where G is the genotype matrix for the local SNPs in GWAS. The genotype and phenotype data are normalized to have mean zero and variance one. $\text{cov}(\mathbf{z}_{\text{gwas}}) = \text{cov}(\frac{1}{\sqrt{n}} G' \mathbf{y}) = \frac{1}{n} G' \text{cov}(\mathbf{y}) G = \frac{\sigma_\epsilon^2}{n} G' G = V$. Under the assumption that a single gene has a negligible heritability for the phenotype, which is normalized to have variance one, $\text{cov}(\mathbf{z}_{\text{gwas}}) = \text{cov}(\mathbf{w}) = V$. $\text{var}(Z_{\text{pena}}) = \text{var}(\mathbf{u}'\mathbf{w}) = \text{tr}(\text{E}(\mathbf{u}\mathbf{u}')V) = \text{tr}(V\text{E}(\mathbf{u}\mathbf{u}')) = \text{tr}(\text{E}(V\mathbf{u}\mathbf{u}')) = \text{E}(\text{tr}(V\mathbf{u}\mathbf{u}')) = \text{E}(\text{tr}(\mathbf{u}'V\mathbf{u})) = \text{tr}(\text{E}(\mathbf{u}'V\mathbf{u})) = \text{E}(\mathbf{u}'V\mathbf{u}) = \text{E}(\frac{1}{\hat{C}\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}})$. Thus, $\text{var}(Z_{\text{pena}}) = 1$ under H_0.

The LSE of b_1, $\hat{b}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$, follows normal under the ideal regression model (Equation 4), since y_i follows normal. For large sample sizes of contemporary GWAS, the

unadjusted TWAS test statistic Z_{unadj} follows normal under H_0. Based on Slutsky's theorem, if the adjustment factor $\hat{A}$ converges in probability to a constant, the adjusted test statistic Z_{pena} should converge to a normal random variable in distribution under H_0. Since it is very challenging to show this condition analytically, we explore the null distribution using simulations. We assume the null distribution of Z_{pena} is $N(0,1)$ since the expectation and variance are shown to be zero and one under H_0. Using empirical density plots and QQ plots, we demonstrate that the null distribution matches excellently with $N(0,1)$ (results described in the simulation study section).

3 Simulation study

We performed extensive simulations to evaluate the penaTWAS approach compared to the unadjusted approaches and another adjusted approach. We compared the type I error rates and powers estimated by the different methods. We contrasted the accuracy in estimating the effect sizes. We implemented the *hapgen* [24] software to simulate genotype data in the transcriptome and GWAS data. We used simulation models to generate the gene expression in the transcriptome data and the outcome in the GWAS data.

Gusev et al. [8] analyzed the transcriptome data from the Young Finish Sequencing (YFS) study to identify all the locally heritable genes in the whole blood tissue. They provided the results from the analysis in the software package Fusion. Since we test one gene at a time in traditional TWAS, we selected a locally heritable gene, DGCR8, on chromosome 22 from the YFS data analysis for simplicity in computation. To run *hapgen*, we considered a Finnish (European) population's LD structure for the region neighboring DGCR8. We considered 300 local SNPs, with MAF > 0.05, randomly selected from the 1MB local region surrounding DGCR8. To mimic the TWAS setup, we generate the transcriptome and GWAS data independently with non-overlapping individuals from the same population. We chose the transcriptome data sample size of $n_1 = 500$ and the GWAS data sample size of $n = 10,000$. *Hapgen* is a widely used method to generate genotype data mimicking real-life populations. Using *hapgen* allows the creation of samples for genotype data of subjects

independent across the iterations in a given simulation scenario.

After generating the genotype data, we simulate the expressions of DGCR8 in the transcriptome data based on the local SNPs' genotypes for the n_1 individuals. Genotype values for each local SNP are normalized to have mean zero and variance one. We consider a linear model to simulate the expression: $E = \mathbf{g}_{ref}^{'}\boldsymbol{\alpha} + \eta$. E denotes the expression, $\mathbf{g}_{ref}$ denotes the genotype values for the local SNPs, and $\boldsymbol{\alpha}$ denotes the effect size vector for the local SNPs on expression. Assuming that $\text{var}(E) = 1$, we consider $\eta \sim N(0, 1 - h_E^2)$, where h_E^2 is the local heritability of the expression due to the SNPs. We assume that 5% of the local SNPs affect the expression. If m_c denotes the number of such SNPs, each element of $\boldsymbol{\alpha}$ follows $N(0, \frac{h_E^2}{m_c})$. We chose $h_E^2 = 10\%, 20\%$. We simulated a few different choices of $\boldsymbol{\alpha}$.

We next simulate the outcome in GWAS data using the linear model: $y = \beta_1 x + \epsilon$. Here, y denotes the phenotype, $x = \mathbf{g}^{'}\boldsymbol{\alpha}$ denotes the true genetic component of expression where $\mathbf{g}$ is the genotype vector for the local SNPs for a GWAS individual, and $\boldsymbol{\alpha}$ is the same as in the transcriptome data. The genetic component of expression for one gene should not explain a significant proportion of the variance of a complex phenotype. Hence, we assume that $\text{var}(y) = \text{var}(\epsilon) = 1$. We selected ten choices of β_1 as $0.02, 0.04, \ldots, 0.18, 0.2$ to study power and estimation of effect sizes, and $\beta_1 = 0$ for evaluating the type I error rate. To simulate a case-control phenotype, we employ the logistic model to obtain the probability of a GWAS individual being a case: $P(\text{case}|\beta_1, \mathbf{g}, \boldsymbol{\alpha}) = \frac{\exp(\beta_1 x)}{1+\exp(\beta_1 x)}, x = \mathbf{g}^{'}\boldsymbol{\alpha}$.

3.1 Analysis pipeline

We implemented the GCTA software [25] to compute the p-value of testing the null hypothesis that the local heritability of the gene is zero versus the alternate hypothesis that it is positive. We applied the BH FDR controlling procedure [26] to the p-values obtained from 2,000 iterations under a given simulation scenario. In a simulation scenario, we regarded an iteration for downstream analysis if the gene was heritable in the iteration based on the FDR correction. In the simulation settings considered, we always found more than 1800 iterations to pass the correction, based on which we evaluated the various measures for the inference.

For unadjusted TWAS, we implemented penalized regression using the Lasso, adaptive Lasso, and Elastic net penalty to estimate the genetically regulated expression component prediction model based on the transcriptome data. To perform a TWAS based on individual-level data, we imputed the expression using the prediction model in the GWAS data. Then, we tested for an association between the predicted expression and the outcome using linear or logistic regression. We obtained the adjustment factor for penaTWAS by applying the adaptive Lasso and residual bootstrapping. Next, we computed the summary-level association data for the local SNPs in GWAS and calculated the test statistic for an unadjusted TWAS based on summary-level data. In an iteration, we utilized the LD matrix of local SNPs estimated from the transcriptome data in the next iteration. This strategy imitates the TWAS procedure based on summary-level data, where LD structure is calculated from external reference data, such as 1000 genome data. The latter has a sample size of 503 for the European population. To assess power and type I error rate, we considered the standard choice of significance level as 0.05.

We examined $0\%, 5\%, 10\%$ trimmed means of the bootstrapped estimates while estimating the denominator of the adjustment factor in penaTWAS. We denote the penaTWAS that utilizes $0\%, 5\%, 10\%$ trimmed means as penaTWAS0, penaTWAS5, and penaTWAS10, respectively. We used $T = 300$ residual bootstrap samples in the penaTWAS. To explore the normality of the null distribution of penaTWAS statistics, we constructed the QQ and empirical density plots for the test statistic values under the null hypothesis of no association.

3.2 Results

We first discuss the results for a TWAS based on individual-level data. To select a representative adjusted TWAS approach, we first compare the performances of the approaches that calibrate the predicted expression's uncertainty. We denote the Taylor expanded version of penaTWAS as PENAtaylor. All the adjusted approaches control the type I error rate below the threshold of 0.05 (Figure 2). While PENAtaylor produced the most conservative estimate of 0.021, the adjusted variance approach also led to a lower estimate of 0.033. On the

other hand, penaTWAS0, penaTWAS5, and penaTWAS10 resulted in an estimate of 0.038, 0.051, and 0.056, respectively (Figure 2). Thus, penaTWAS5 produced a desirable type I error rate. In this simulation setting, the gene expression heritability was 10%.

Concerning power, PENAtaylor led to a lower power than the other alternatives (Figure 2). The adjusted variance approach results in a marginally higher power than penaTWAS0 except for higher choices of β as 0.16, 0.18, and 0.2, for which penaTWAS0 produces a marginally higher power (Figure 2). We find that penaTWAS5 and penaTWAS10 offer consistently a marginally higher power than the adjusted variance approach. For example, penaTWAS5 produced $2-5\%$ power gain than the adjusted variance approach (Figure 2). penaTWAS10 results in a slight power gain of 1% for most choices of β than penaTWAS5. However, we chose penaTWAS5 as the representative adjusted TWAS approach because it better controlled the type I error rate than penaTWAS10 (Figure 2). We refer to penaTWAS5 as penaTWAS in the following sections. We observed similar comparative performances of the various adjusted TWAS approaches in the other simulation settings. In the following description for individual-level TWAS, we skip presenting the results obtained by the other adjusted TWAS approaches for brevity and clarity.

Next, we compare the performances of the unadjusted TWAS approaches and penaTWAS and contrast them with those of Gold TWAS. In the latter, we plug in the unknown true local SNPs' effects on expression while conducting TWAS. We refer to the unadjusted TWAS approaches as Lasso, ALasso (adaptive Lasso), and Enet (elastic net) TWAS, respectively. For the current simulation setting with gene expression heritability 10%, the unadjusted TWASs produced an inflated type I error rate (Figure 3). The Lasso, ALasso, and Enet TWAS led to an estimate of 0.061, 0.064, and 0.063, respectively. The Gold TWAS produced a type I error rate of 0.056, which is lower than that of the unadjusted TWASs (Figure 3). The type I error rate estimated by penaTWAS was 0.051. Thus, the unadjusted TWASs produced inflated type I error rates, whereas the penaTWAS adequately controls it, a little lower than the Gold TWAS. We note that the three choices of the unadjusted TWASs performed comparably concerning the control of the false positive rate.

As expected, the Gold TWAS produced the maximum power, which no method could achieve (Figure 3). Lasso, ALasso, and Enet TWAS offered higher powers than the penaTWAS at the expense of inflated type I error rates. The Lasso and Enet TWAS yielded almost the same power, resulting in a marginal power increase of $1-5\%$ than ALasso TWAS. However, the ALasso TWAS produced a slightly lower false positive rate than Lasso and Enet. The nearly comparable results of these three types of unadjusted TWAS justify the development of penaTWAS based on the adaptive Lasso framework. For simplicity, we pick the Enet TWAS to contrast with the adjusted TWAS approach penaTWAS. We note that the Enet TWAS for individual-level data is the same method as PrediXcan [9]. Enet TWAS led to a higher power of $1-10\%$ than penaTWAS at the expense of an inflated false positive rate, and hence, the power gain is not statistically justified. Given its calibrated type I error rate, the bearable extent of power loss by penaTWAS is preferable.

We now describe the results for a summary-level TWAS. Among the adjusted TWAS approaches, penaTWAS5 provides a tuned estimate of the type I error rate as 0.054 (Figure 4). penaTWAS10 yielded a more inflated estimate of 0.057, whereas penaTWAS0 and PENAtaylor produced estimates of 0.042 and 0.023, respectively. Regarding power, penaTWAS10 offered a marginal increase of 1% than penaTWAS5 (Figure 4). Due to better control of the type I error rate and good power, we again selected penaTWAS5 as the representative adjusted TWAS approach. Note that the adjusted variance approach could only be applied for TWAS based on individual-level data.

For the current simulation setting with gene expression heritability 10%, the unadjusted summary-level TWASs yielded an inflated type I error rate (Figure 5). The Lasso, ALasso, and Enet TWAS estimated a type I error rate of 0.063, 0.064, and 0.066, respectively. The three types of unadjusted TWASs performed comparably regarding the false positive rate. The Gold TWAS led to a type I error rate of 0.055, lower than the unadjusted TWASs. The type I error rate estimated by penaTWAS was 0.054 (Figure 5). Again, the unadjusted summary-level TWASs produced inflated type I error rates, whereas the summary-level penaTWAS adequately controls it. The Gold TWAS based on summary-level data produced

the maximum power, as expected (Figure 5). Lasso, ALasso, and Enet TWAS offered higher powers than the penaTWAS at the expense of inflated type I error rates. The Lasso and Enet TWAS produced almost the same power (Figure 5). They offered a marginal power gain of $1-5\%$ than ALasso TWAS. Enet TWAS led to a higher power of $1-10\%$ than penaTWAS at the expense of an inflated false positive rate, and hence, the power gain is not justified. Again, given its tuned type I error rate, penaTWAS is preferable for a summary-level TWAS despite the tolerable extent of power loss.

Next, we describe the accuracy of penaTWAS and the unadjusted TWAS approaches based on individual-level data while estimating the effect size of the genetic component of expression on the outcome. PenaTWAS performs well overall in estimating the effect size (Table 1). The mean of the estimated effects is the same as the true effect when $\beta_1 = 0.02, 0.04, 0.06$. For the higher choices of β_1, the mean falls very close to the true value, 0.01 larger than the actual value (Table 1). We utilize the relative bias, defined as $\frac{\hat{\beta}_1-\beta_1}{\beta_1}$, and the square root of the mean squared error (MSE) to compare the estimation accuracy of the different approaches. ALasso TWAS produced a downward bias with a mean relative bias of -32% (Table S1). Lasso and Enet TWAS resulted in an upward bias with the mean relative bias in the range of $14\% - 17\%$ and $17-23\%$, respectively (Table S1). PenaTWAS yielded the lowest upward bias with $6-7\%$ mean relative bias. The s.d. of the relative bias is the lowest for ALasso TWAS, whereas it is comparable for the other approaches. The mean bias for a given method excluding ALasso decreases marginally for increasing values of β_1. However, the s.d. of the relative bias decreases rapidly as the actual value of β_1 increases (Table S1). Thus, penaTWAS offered the most accurate estimates of the effect sizes concerning the relative bias. However, ALasso TWAS provided the smallest MSE of the β_1 estimates, and the other approaches yielded comparable MSEs (Table S2).

We regarded another choice of the true eQTL effects with expression heritability as 10% and repeated the above simulations. First, we summarize the results for the TWAS using individual-level data. We chose the penaTWAS5, referred to as penaTWAS in the following, as the preferred adjusted TWAS approach based on the relative performance of

the alternatives (Figure S1). Among the unadjusted approaches, Enet TWAS inflated the type I error rate to 0.06; Lasso TWAS marginally overestimated it as 0.055; ALasso TWAS estimated it as 0.053 with adequate control. On the other hand, penaTWAS correctly estimated the error rate as 0.045 below the desired level, which is very close to the estimate of 0.044 by the Gold TWAS (Figure S2). Regarding power, Enet TWAS offered a $1-8\%$ increase in power than penaTWAS at the expenditure of a higher type I error rate (Figure S2). For summary-level TWAS, we observe the same pattern, i.e., penaTWAS provides better control of the false positive rate at the cost of $2-7\%$ power loss compared to the unadjusted approaches (Figure S4). Concerning the accuracy of the effect size estimation, penaTWAS produced zero mean relative bias and hence performed substantially better than the unadjusted approaches (Table S3). The mean estimate is the same as the true effect sizes (Table 2). PenaTWAS yielded a marginally lower MSE in the estimates than the Lasso and Enet TWAS but larger than the ALasso TWAS (Table S4). Overall, penaTWAS provides more accurate estimates of the effect sizes.

We explored whether out-of-sample LD and GWAS summary statistics used in the summary-level TWAS affect the type I error rate and power compared to the TWAS based on individual-level data. Interestingly, we found no consistent pattern regarding the type I error rate. For the first choice of the actual eQTL effects with expression heritability 10%, the summary-level penaTWAS yielded a marginally higher type I error rate of 0.054 (Figure 4) compared to 0.051 by its individual-level data-based version (Figure 2). Conversely, for the second choice of the actual eQTL effects above, the estimated rate was 0.042 (Figure S3) versus 0.045 for the summary-level and individual-level penaTWAS (Figure S1), respectively. Regarding power, both strategies performed comparably.

We next consider a higher choice of the expression heritability as 20%. We find that the adjusted variance approach yields a slight power gain of $1-5\%$ than penaTWAS for the individual-level data-based analyses (Figure S5). For brevity, we briefly discuss the accuracy of effect size estimation using individual-level TWAS approaches and the results from hypothesis testing obtained using summary-level TWAS approaches in the following

sections. In general, we observed that penaTWAS controlled the type I error rate better than the unadjusted TWAS approaches at the cost of a limited loss in power (Figure S8). PenaTWAS estimated the effect sizes more accurately than the unadjusted approaches (Table S5) concerning lower bias (Table S6). The mean squared error estimates were comparable among the approaches (Table S7).

For a binary case-control phenotype, the unadjusted TWASs control the false positive rates nearly adequately with marginal inflation. The Lasso and Enet TWAS lead to estimated type I error rates of 0.052 and 0.055, respectively. PenaTWAS curbs it below the threshold with an estimate of 0.04, and as a consequence, it loses a limited power of $1 - 8\%$ compared to the unadjusted methods. On the other hand, penaTWAS estimates the effect sizes more accurately (Table 3). For each choice of β_1, the mean of the estimates coincides with the true choice. The s.d. of the estimates is marginally smaller than the Lasso and Enet TWAS and a little larger than the ALasso TWAS (Table 3). The latter produces a substantial downward bias, and the formers lead to significant upward bias (Table S8). PenaTWAS yields a limited downward bias, which is expected because a binary phenotype contains less information than a continuous phenotype for a finite sample (Table S8). PenaTWAS has a marginally lower MSE than Lasso and Enet TWAS and a marginally higher MSE than ALasso TWAS (Table S9). Overall, penaTWAS offers more accurate effect size estimates than the unadjusted approaches.

We find that the median of the adjustment factors to obtain the penaTWAS test statistics from the unadjusted ALasso TWAS test statistics is 0.94, 0.95, and 0.94 in the three different simulation scenarios considered for a continuous phenotype above (Figure 6). We note that the adjustment factors can sometimes be higher than one, which, however, happens less frequently, e.g., in 14%, 10%, and 2% of all iterations in the above three simulation scenarios. An adjustment factor lower than one in most simulation iterations partly explains the better control of the type I error rates by penaTWAS. The empirical density plots for the penaTWAS test statistics obtained from an individual-level TWAS appear bell-shaped and symmetric around the mean zero under the null hypothesis (Figure S9). QQ plots

demonstrate an excellent match between the null distribution and $N(0,1)$ (Figure S10). The Kolmogorov Smirnov (KS) test for deviation from Normal distribution indicated an excellent match with normality for the test statistics under null when the adjustment factor in penaTWAS was calculated using 5% and 10% trimmed means of the boostrap estimates to estimate the denominator (Equation 8). The KS test p-values were 0.7 and 0.44, respectively. However, without trimming, the test statistics suggest deviation from normality at a nominal level with the KS test p-value of 0.02. These results motivated using the penaTWAS version employing 5% trimmed mean. We obtain a similar finding for summary-level TWAS (Figure S11, S12). The p-values from the KS test are 0.7, 0.46, and 0.03, respectively.

4 Real data application

We applied the summary-level adjusted and unadjusted TWAS approaches to an anthropometric phenotype, height, and three lipid phenotypes: LDL cholesterol, HDL cholesterol, and triglycerides. We integrated the Geuvadis transcriptome data [27] and the UK Biobank (UKB) GWAS data [12]. We included the European individuals in Geuvadis and the white-British European individuals in UKB data. In Geuvadis data comprising 462 individuals from the 1000 Genomes Project, RNA sequencing was conducted for the lymphoblastoid cell line in peripheral blood tissue. Gene expression levels were measured for individuals who belonged to five different populations: CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI), and Yoruba (YRI). The individuals from CEU, FIN, GBR, and TSI were combined to form a sample from broader European (EUR) ancestry with a size of 358. Even though peripheral blood is one of the most studied tissue types, we acknowledge that it may not be the most relevant tissue type for the outcomes. For example, liver tissue instead of blood tissue is one of the most pertinent tissue types for lipids [28]. We implemented the following QC steps for the genotype data. We filtered out SNPs with MAF < 0.05. We tested for the Hardy-Weinberg equilibrium (HWE) and removed the SNPs with p-value $< 5 \times 10^{-8}$.

We analyzed $14,123$ protein-coding genes. We adjusted the expression of each gene for relevant covariates, e.g., sex and the top 10 principal components (PCs) of genetic ancestry.

The genetic PCs were included to adjust for the effect of population stratification on gene expression. For each gene, we regarded the SNPs located in the one megabase region from the transcription start and end sites of the gene as the local SNPs. We implemented the GCTA software [25] to determine if a gene is locally heritable. Using GCTA, we obtained the p-value for each gene, testing the non-null hypothesis that its local heritability is significantly positive. We applied the BH FDR controlling procedure to the p-values based on 5% FDR level to select the locally heritable genes. If a gene is locally heritable, we analyze it further. We found 1515 locally heritable genes; a limited number of genes were detected mainly due to the small sample size of Geuvadis data.

The same QC steps above were implemented to filter the genotype data in UKB. The MAF threshold was considered as 0.01 since the UKB data had an enormous sample size. After filtering for relatedness, 290,641 individuals were included in UKB. We obtained the summary-level GWAS association data by adjusting for relevant covariates, such as age, sex, and top 20 PCs for genetic ancestry. We employed the penalized regression based on the Elastic net penalty for the unadjusted TWAS, i.e., Enet TWAS. We utilized 300 residual bootstrap samples for the penaTWAS. Analogous to the subset testing approach [29–31], we apply the Bonferroni-corrected threshold for TWAS p-values in the second step regression for the 1515 genes selected in the first step, i.e., $\frac{0.05}{1515} = 3.3 \times 10^{-5}$.

First, we discuss the number of genes identified by the adjusted and unadjusted approaches for the phenotypes analyzed. PenaTWAS discovered 120 genes for height, whereas Enet TWAS detected 142 genes. The two approaches identified a common subset of 102 genes (Table S16), leaving 18 genes discovered only by penaTWAS (Figure 7, Table S17) and 40 genes found only by Enet TWAS (Figure 7, Table S17). PenaTWAS discovered 33 genes for HDL cholesterol, and Enet TWAS found 40 genes (Table S10, S11). Both approaches identified a common subset of 26 genes (Figure 7, Table S10). PenaTWAS spotted 25 genes for LDL cholesterol, and Enet TWAS found 26 genes. Both approaches detected a common set of 18 genes (Table S12, Figure 7). For triglycerides, penaTWAS detected 32 genes, and Enet TWAS found 30 genes. Both methods identified a common set of 20

genes (Table S14), Figure 7). Thus, except for triglycerides, Enet TWAS found a few more genes than penaTWAS, which resembles a limited higher power of unadjusted TWAS at the expense of inflated false positive rates observed in simulations. For each phenotype, nearly half or more of the discovered genes were identified by both approaches.

We recognize that it is very challenging to show inflation in the false positive rate of the unadjusted approach based on real data analyses. For a heuristic exploration, we calculated the genomic inflation factor (GIF) based on the test statistics calculated by penaTWAS and Enet TWAS for 1515 genes found locally heritable by GCTA. The GIF for penaTWAS and Enet TWAS were 3.81 and 4 for height, 1.92 and 2.29 for HDL cholesterol, 1.68 and 1.77 for LDL cholesterol, 1.7 and 1.9 for triglycerides. The GIFs are substantially higher than one, mainly due to the estimation of these based on a limited number of genes. Ideally, the GIF should be calculated based on all genes, which is at least ten-fold higher than the number of locally heritable genes in our analyses. However, relatively higher inflation for Enet TWAS than penaTWAS is consistently observed across phenotypes. Interestingly, penaTWAS found a marginally higher number of genes associated with triglycerides despite a lower GIF than Enet TWAS. As expected, height appeared most polygenic.

Among the genes uniquely identified by Enet TWAS, many genes were missed by penaTWAS with nominally significant association p-values. However, for some genes, the penaTWAS p-values are entirely insignificant. For height, the SLC25A16 gene located in the 10q21.3 region on chromosome 10 had a p-value of 2.6×10^{-10} by Enet TWAS, but a penaTWAS p-value of 0.7 (Table S17). For HDL cholesterol, the EEF2 gene located in 19p13.3 was detected by Enet TWAS but excluded by penaTWAS with a p-value of 0.5 (Table S11). We observe similar patterns for LDL cholesterol (Table S13) and triglycerides (Table S15). We assessed how many uniquely detected lipid phenotypes genes are reported in the NHGRI-EBI GWAS catalog. For HDL, 5 out of 14 genes exclusively found by Enet TWAS and 1 out of 7 genes spotted only by penaTWAS are reported (Table S11). 4 out of 8 genes for Enet TWAS, and 2 out of 7 genes for penaTWAS are reported for LDL (Table S13). For triglycerides, 2 out of 10 genes for Enet TWAS and 5 out of 12 genes for penaTWAS

are reported (Table S15). These exclusively identified genes that are not reported for lipids were communicated to be associated with many other biomedically important phenotypes in the catalog (Table S11, S13, S15).

Interestingly, the gene TMEM120A on chromosome 7 was uniquely identified by penaTWAS for all three lipid phenotypes, which suggests that it is a pleiotropic gene. Similarly, penaTWAS detected the SLC25A16 gene on chromosome 10 to be significantly associated with height, LDL cholesterol, and triglycerides; hence, it is likely to be a pleiotropic gene. Thus, penaTWAS performed well in discovering many gene-phenotype associations, including known and novel signals.

5 Discussion

The unadjusted TWAS approach suffers from the inherent methodological limitation that the uncertainty or variability of predicted expression is disregarded in the second step regression of the outcome in GWAS data. This shortcoming can lead to inference inaccuracies while identifying the gene-phenotype associations. We propose a novel approach to adjust for the uncertainty of predicted expression in TWAS. Adapting techniques from measurement error theory, we derive the adjustment factor to be multiplied with the unadjusted TWAS test statistic. Then, we implement bootstrapping for adaptive Lasso in the transcriptome data to devise an estimation procedure for the adjustment factor. The penaTWAS can be implemented both for individual-level and summary-level data.

The penaTWAS performs reliably to control the type I error rate. However, the unadjusted TWAS can inflate the type I error rate, which resembles the results communicated in Leeuw et al. [15]. We also found simulation scenarios where an unadjusted TWAS adequately controls the type I error rate. However, for real-life data, it is infeasible to assess whether an unadjusted approach would maintain the error rate for a gene or not. On the contrary, penaTWAS calibrates the uncertainty of predicted expression and is a more reliable alternative. The power loss of penaTWAS compared to an unadjusted TWAS due to better control of type I error rate is limited. In the simulation study, we allowed a gene in the

downstream analysis if the GCTA p-value, for testing whether the local heritability of the gene's expression is positive, passed the FDR correction. However, if the Bonferroni correction were applied instead to control FWER, the inflation of the type I error rates would be higher for an unadjusted TWAS [15]. The estimation of effect sizes is substantially more accurate for penaTWAS than an unadjusted TWAS, which is observed for both continuous and case-control phenotypes. The more precise estimation of effect sizes should improve the polygenic transcriptomic risk score (PTRS) calculation while predicting a quantitative or disease phenotype [32].

Under the null hypothesis $H_0 : \beta_1 = 0$, $\mathrm{E}(\hat{b}_1) = B \times \beta_1 = 0$, based on equation 6. However, an unadjusted TWAS uses $\mathrm{var}(\hat{b}_1) = \frac{\sigma_e^2}{n} \frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}$, which ignores the expectation in the variance formula calibrated for the uncertainty in the prediction model for expression: $\mathrm{var}(\hat{b}_1) = \frac{\sigma_e^2}{n} \mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$ (equation 7). Consequently, the type I error rate for an unadjusted TWAS may not be controlled. Leeuw et al. [15] argued that the TWAS framework tests for a non-zero local genetic correlation between the gene expression and the outcome. However, due to ignoring the uncertainty of predicted expression, the null hypothesis in standard TWAS boils down to the local genetic correlation being equal to a constant, which is not necessarily zero [15]. Following their derivation, it is straightforward to show that our null hypothesis $H_0 : \beta_1 = 0$ based on the ideal regression is equivalent to the correct null hypothesis of the local genetic correlation being equal to zero. We note that β_1 is proportional to the covariance between the outcome and true genetic component of expression based on the ideal regression (equation 4). We proposed a modification for an individual-level data-based adjusted variance approach [19], which also controls the false positive rate and provides comparable powers as penaTWAS. However, the former requires individual-level GWAS data. Moreover, penaTWAS presents an explicit adjustment both for the effect size estimation and hypothesis testing.

We conducted TWAS for height and lipid phenotypes while integrating the Geuvadis transcriptome and UK Biobank GWAS data. The penaTWAS performed well; the number of genes discovered was a little smaller than the unadjusted TWAS approach for height,

HDL, and LDL cholesterols, falling in line with the simulation results. However, the former detected a few more genes for triglycerides. PenaTWAS found some genes overlooked by the unadjusted TWAS and vice versa. However, for each phenotype, nearly half or more genes were detected by both approaches, validating the use of penaTWAS. Many of the identified gene-phenotype associations have already been communicated in the literature. However, evaluating the accuracy of the adjusted versus unadjusted approaches based on reported associations is challenging. Appropriate replication studies are crucial in this regard. The penaTWAS identified many gene-phenotype associations. We recommend using the penaTWAS rather than the unadjusted approaches due to better control of the former concerning the type I error rate and more precise estimation of effect sizes. PenaTWAS exclusively identified some genes for each phenotype, which unadjusted TWAS missed. The adjustment factor in penaTWAS is not bounded above by one (equation 8). Our simulation studies also demonstrate it. However, we note that relatively few genes were identified exclusively by penaTWAS.

Metro [33] is another approach that models the GWAS and transcriptome data jointly to account for the uncertainty of predicted expression. However, the method focuses on a multi-ancestry set-up and models the cross-ancestry GWAS data in a combined manner, whereas penaTWAS models the single ancestry scenario. Moreover, Metro considers a normal prior for the regression coefficients in the linear model in the transcriptome data, which is equivalent to a penalized ridge regression. However, our approach is based on the adaptive Lasso with desirable shrinkage and variable selection properties. We acknowledge the limitation of our real data analyses: the tissue type considered may not be the most relevant tissue type for the outcomes. In future work, we aim to apply the method to more phenotypes while integrating transcriptome data with larger sample sizes in various tissue or cell types. We plan to extend this approach to incorporate multiple populations [34]. Another important direction is to devise the strategy to combine multiple tissue types available in transcriptome data together [35].

In summary, we propose a TWAS approach that adjusts for the predicted expression's

uncertainty and is a statistically valid test for gene-phenotype association. It can be implemented for complex phenotypes using publicly available GWAS summary statistics data to obtain novel insights into their genetic architecture.

Acknowledgement

We acknowledge Dr. Arindam Chatterjee and Dr. Prajamitra Bhuyan for helpful discussions related to this work. This research used the UK Biobank resource under application 77327.

Data availability

The datasets that we have analyzed or used in our analyses in this work are available (either openly or via applications) from the following websites:

Geuvadis data: https://www.internationalgenome.org/data-portal/data-collection/geuvadis

UK Biobank: https://www.ukbiobank.ac.uk/

Fusion: http://gusevlab.org/projects/fusion/

References

[1] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[2] Binglan Li and Marylyn D Ritchie. From gwas to gene: transcriptome-wide association studies and other methods to functionally understand gwas discoveries. *Frontiers in Genetics*, 12:713230, 2021.

[3] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[4] Alexandra C Nica and Emmanouil T Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362, 2013.

[5] Jonathan Beesley, Haran Sivakumaran, Mahdi Moradi Marjaneh, Wei Shi, Kristine M Hillman, Susanne Kaufmann, Nehal Hussein, Siddhartha Kar, Luize G Lima, Sunyoung Ham, et al. eqtl colocalization analyses identify ntn4 as a candidate breast cancer risk gene. *The American Journal of Human Genetics*, 107(4):778–787, 2020.

[6] Nora Franceschini, Claudia Giambartolomei, Paul S De Vries, Chris Finan, Joshua C Bis, Rachael P Huntley, Ruth C Lovering, Salman M Tajuddin, Thomas W Winkler, Misa Graff, et al. Gwas and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nature communications*, 9(1):5141, 2018.

[7] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4):e1000888, 2010.

[8] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245, 2016.

[9] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae,

Nancy J Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

[10] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[11] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[12] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018.

[13] Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Steven McCarroll, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature genetics*, 50(4):538–548, 2018.

[14] Nicholas Mancuso, Simon Gayther, Alexander Gusev, Wei Zheng, Kathryn L Penney, Zsofia Kote-Jarai, Rosalind Eeles, Matthew Freedman, Christopher Haiman, and Bogdan Pasaniuc. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nature communications*, 9(1):1–11, 2018.

[15] Christiaan de Leeuw, Josefin Werme, Jeanne E Savage, Wouter J Peyrot, and Danielle Posthuma. On the interpretation of transcriptome-wide association studies. *PLoS Genetics*, 19(9):e1010921, 2023.

[16] Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.

[17] Arindam Chatterjee and Soumendra Nath Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

[18] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[19] Haoran Xue, Wei Pan, and Alzheimer's Disease Neuroimaging Initiative. Some statistical consideration in transcriptome-wide association studies. *Genetic epidemiology*, 44(3):221–232, 2020.

[20] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[21] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[22] Arindam Chatterjee and S Lahiri. Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12):4497–4509, 2010.

[23] Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.

[24] Zhan Su, Jonathan Marchini, and Peter Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305, 2011.

[25] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[26] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[27] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel,

Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[28] Hilary K Finucane, Yakir A Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics*, 50(4):621–629, 2018.

[29] Charles Kooperberg and Michael LeBlanc. Increasing the power of identifying gene× gene interactions in genome-wide association studies. *Genetic Epidemiology*, 32(3):255–263, 2008.

[30] W James Gauderman, Pingye Zhang, John L Morrison, and Juan Pablo Lewinger. Finding novel genes by testing g× e interactions in a genome-wide association study. *Genetic Epidemiology*, 37(6):603–613, 2013.

[31] Arunabha Majumdar, Kathryn S Burch, Tanushree Haldar, Sriram Sankararaman, Bogdan Pasaniuc, W James Gauderman, and John S Witte. A two-step approach to testing overall effect of gene-environment interaction for multiple phenotypes. *Bioinformatics*, 36(24):5640–5648, 2020.

[32] Yanyu Liang, Milton Pividori, Ani Manichaikul, Abraham A Palmer, Nancy J Cox, Heather E Wheeler, and Hae Kyung Im. Polygenic transcriptome risk scores (ptrs) can improve portability of polygenic risk scores across ancestries. *Genome biology*, 23:1–18, 2022.

[33] Zheng Li, Wei Zhao, Lulu Shang, Thomas H Mosley, Sharon LR Kardia, Jennifer A Smith, and Xiang Zhou. Metro: Multi-ancestry transcriptome-wide association studies for powerful gene-trait association detection. *The American Journal of Human Genetics*, 109(5):783–801, 2022.

[34] Fang Chen, Xingyan Wang, Seon-Kyeong Jang, Bryan C Quach, J Dylan Weissenkampen, Chachrit Khunsriraksakul, Lina Yang, Renan Sauteraud, Christine M Albert, Nic-

holette DD Allred, et al. Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. *Nature genetics*, 55(2):291–300, 2023.

[35] Helian Feng, Nicholas Mancuso, Alexander Gusev, Arunabha Majumdar, Megan Major, Bogdan Pasaniuc, and Peter Kraft. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. *PLoS genetics*, 17(4):e1008973, 2021.

Figure 1: Flow chart of different steps in the adjusted TWAS method penaTWAS. The yellow boxes represent the adjustment for the uncertainty of the predicted expression.

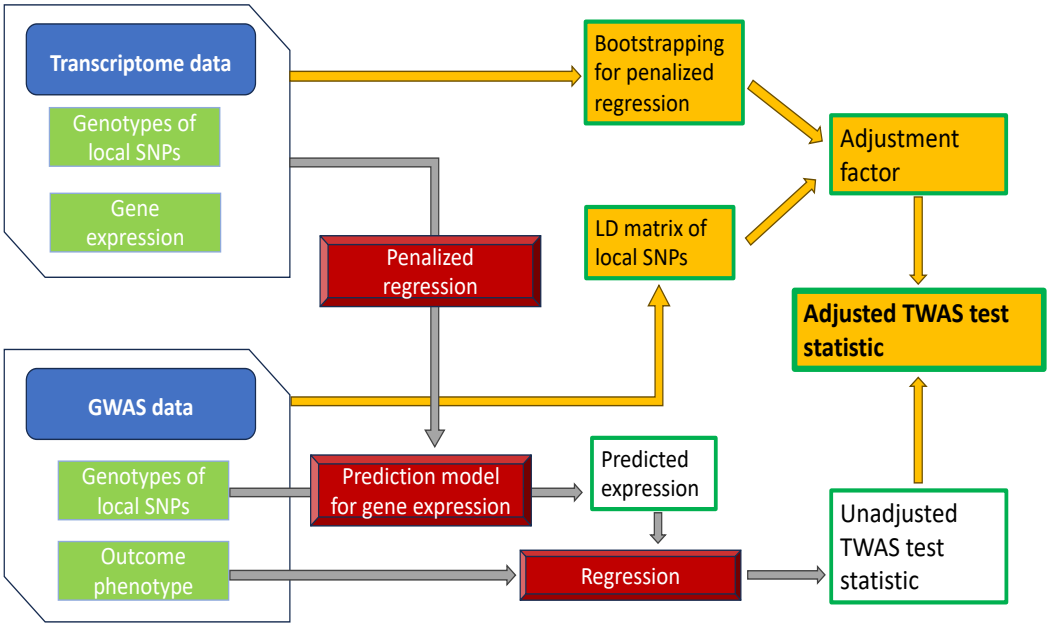

Figure 2: Comparison of type I error rate and power between the adjusted TWAS approaches based on individual-level data. Here, penaTWAS0, penaTWAS5, penaTWAS10 denote the penaTWAS approaches that use $0\%, 5\%, 10\%$ trimmed means based on the bootstrap sample while estimating the adjustment factor. PENAtaylor stands for the Taylor series expanded version of estimating the adjustment factor in penaTWAS. AdjVar represents the modified adjusted variance approach. The heritability of gene expression is 10%. The first choice of true eQTL effects was considered.

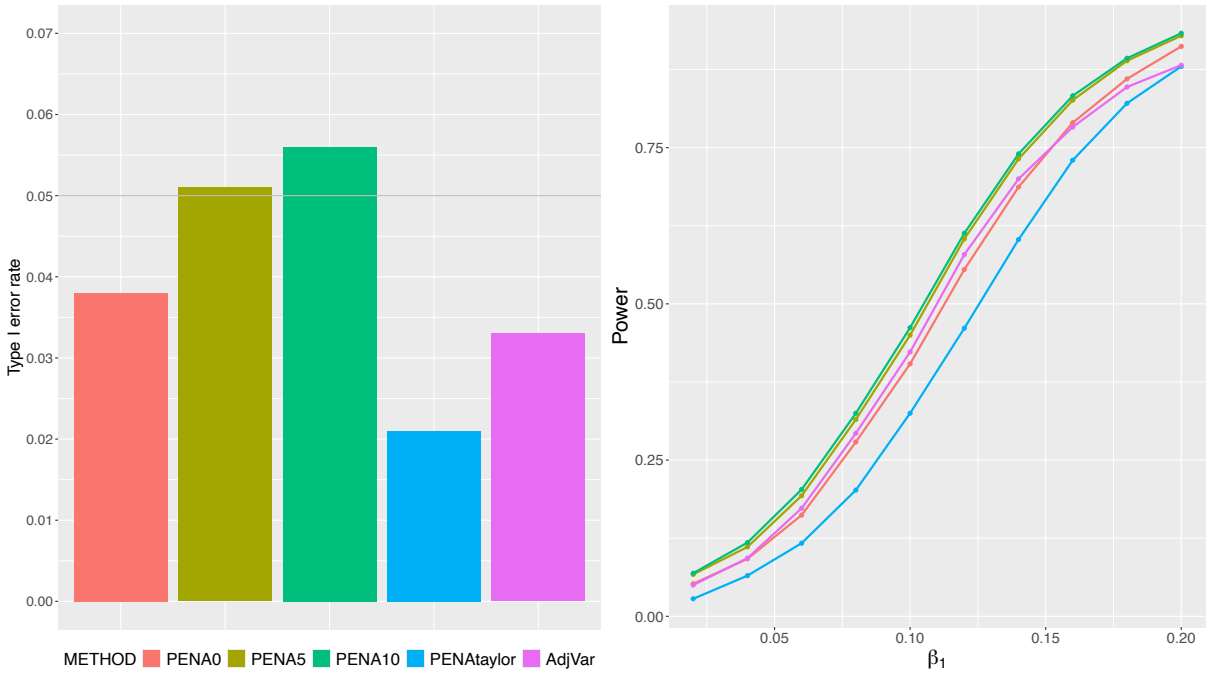

Figure 3: Comparison of type I error rate and power between the unadjusted TWAS approaches and the adjusted TWAS approach penaTWAS based on individual-level data. Results for the gold standard TWAS in which true eQTL effects are plugged in are also presented. The unadjusted TWAS approaches are referred to as Lasso, ALasso (adaptive Lasso), and Enet (elastic net) TWAS, depending on the type of penalty function used in the penalized regression for transcriptome data. The heritability of gene expression is 10%. The first choice of true eQTL effects was considered.

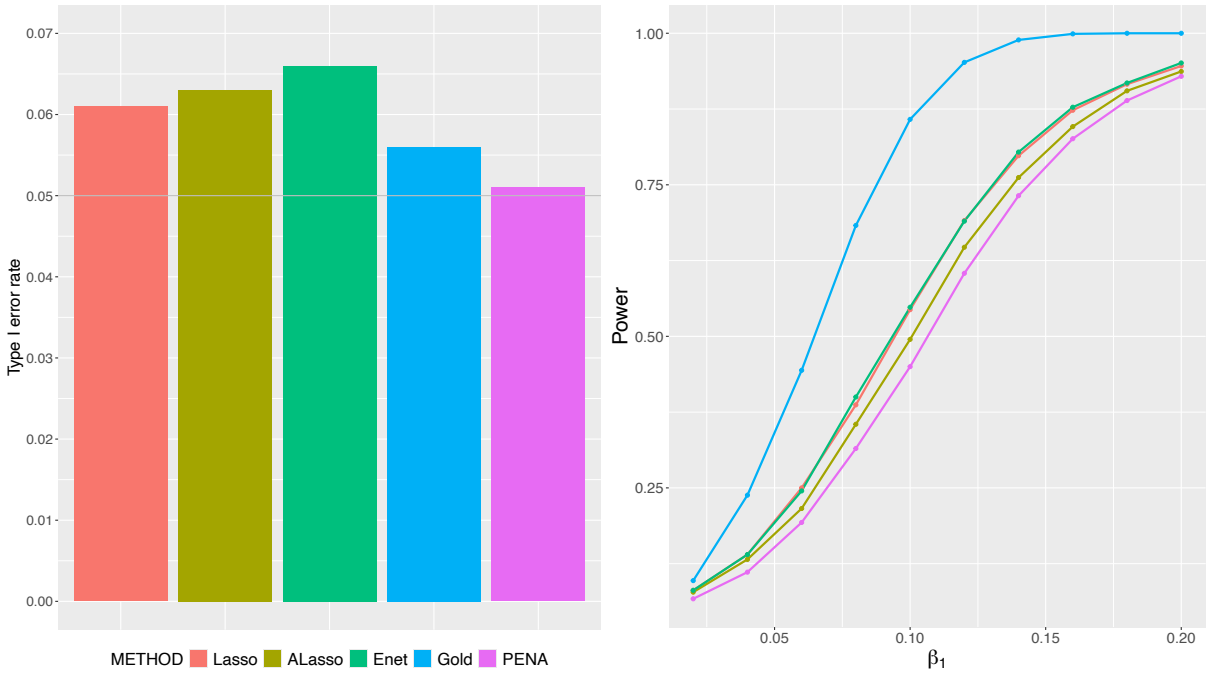

Figure 4: Comparison of type I error rate and power between the adjusted TWAS approaches based on summary-level data. Here, penaTWAS0, penaTWAS5, penaTWAS10 denote the penaTWAS approaches that use $0\%, 5\%, 10\%$ trimmed means based on the bootstrap sample while estimating the adjustment factor. PENAtaylor stands for the Taylor series expanded version of estimating the adjustment factor in penaTWAS. The heritability of gene expression is 10%. The first choice of true eQTL effects was considered.

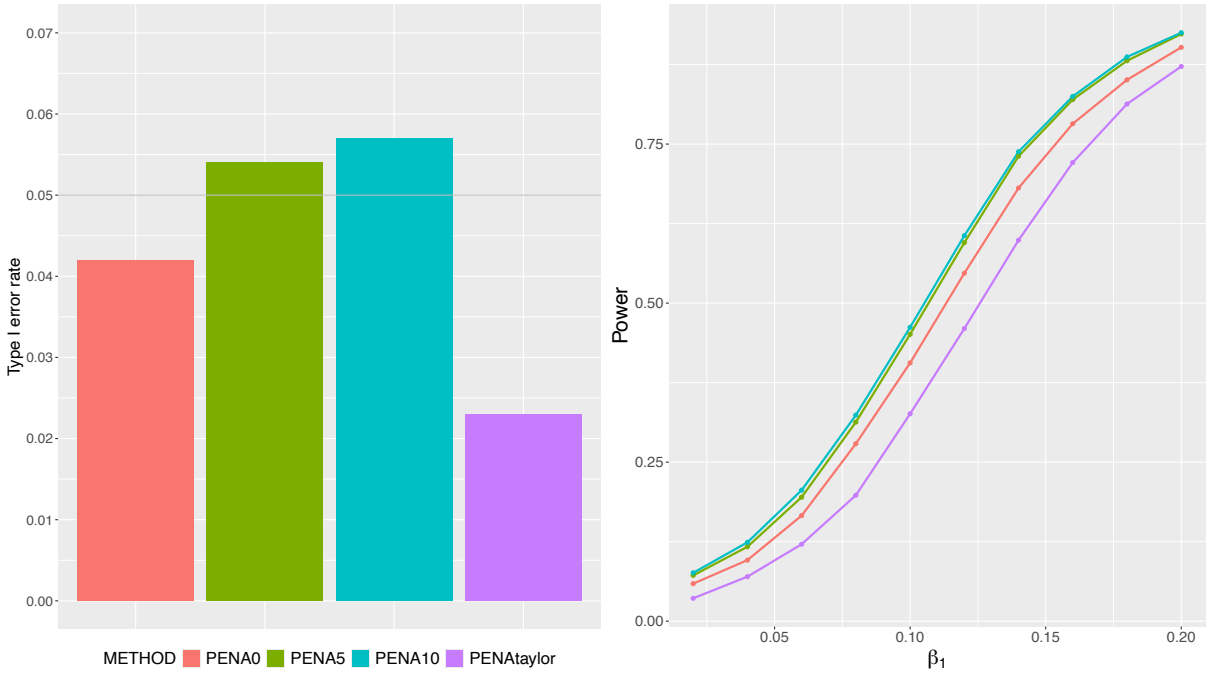

Figure 5: Comparison of type I error rate and power between the unadjusted TWAS approaches and the adjusted TWAS approach penaTWAS based on summary-level data. Results for the gold standard TWAS in which true eQTL effects are plugged in are also presented. The unadjusted TWAS approaches are referred to as Lasso, ALasso (adaptive Lasso), and Enet (elastic net) TWAS, depending on the type of penalty function used. The heritability of gene expression is 10%. The first choice of true eQTL effects was considered.

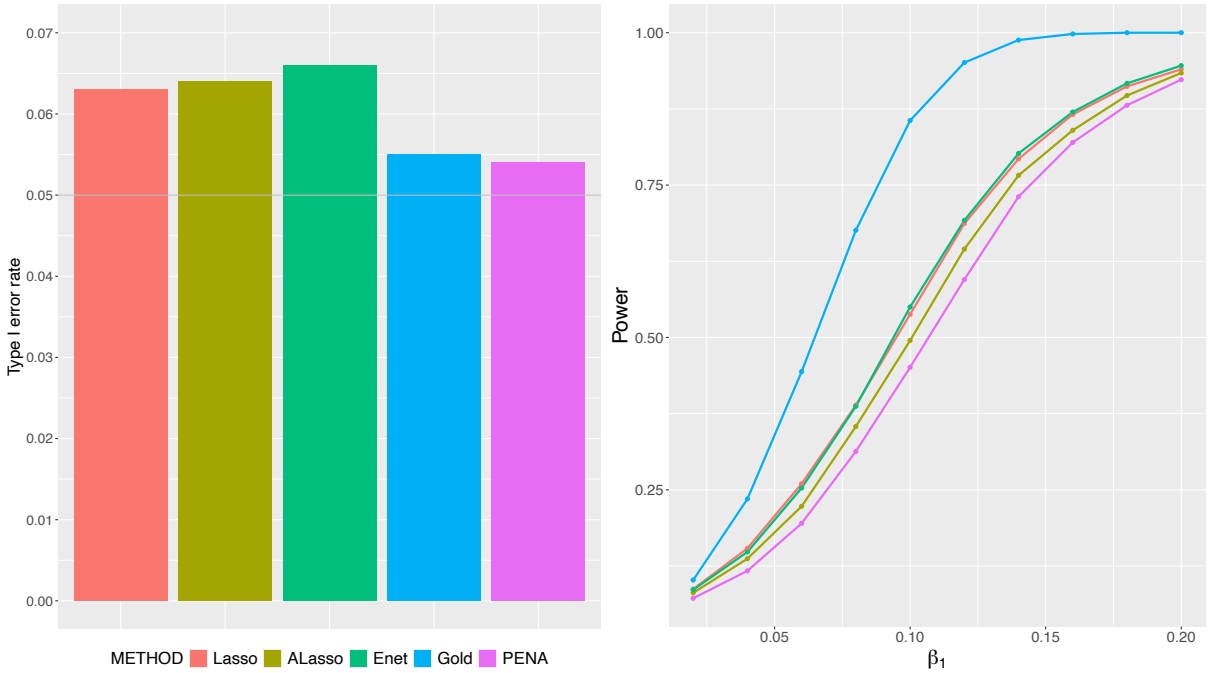

Figure 6: The adjustment factors to obtain the Z statistics for penaTWAS based on individual-level data in the three simulation scenarios considered for a continuous phenotype. In the first two scenarios, the gene expression local heritability is 10%, while in the third scenario, it is 20%.

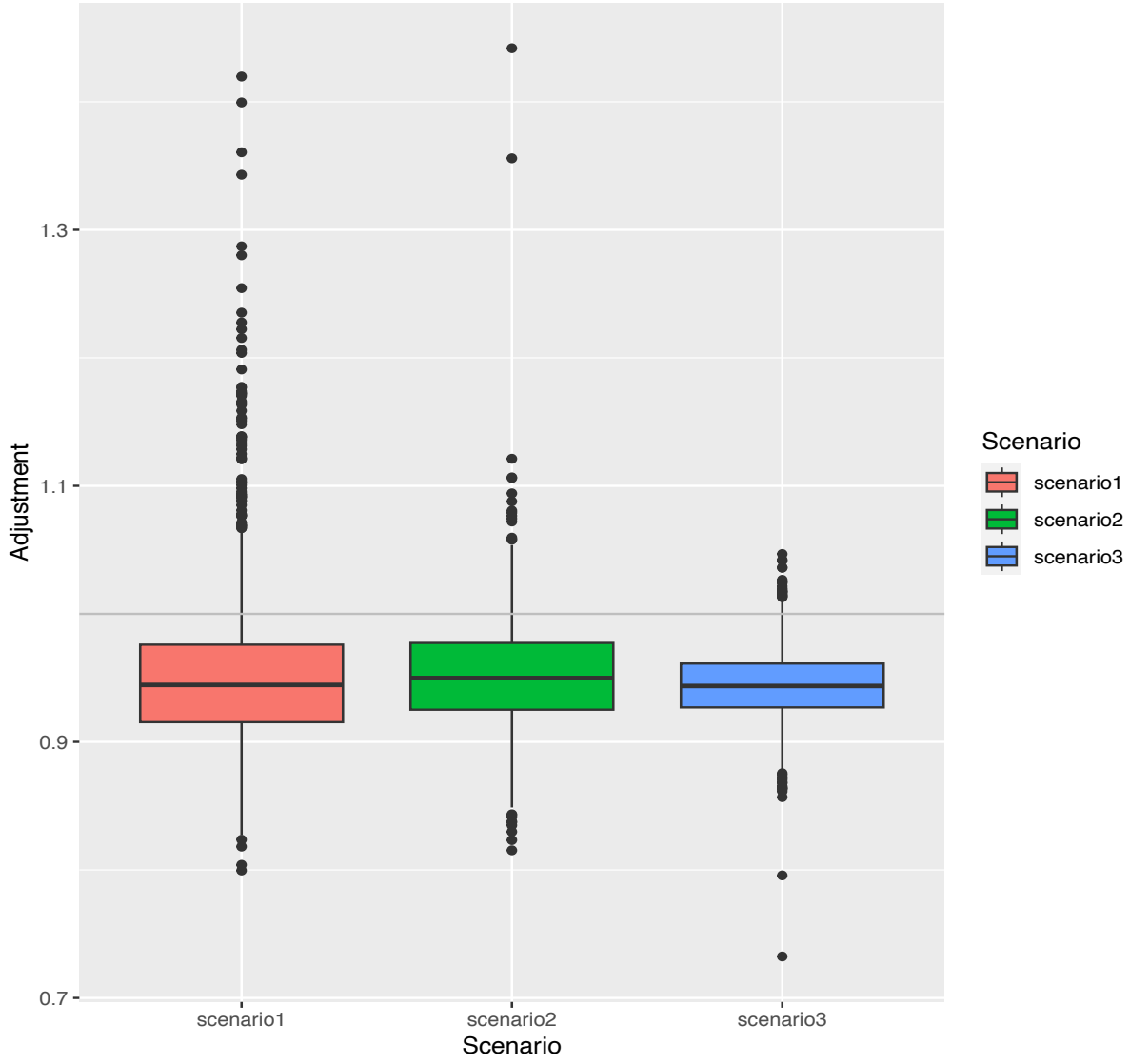

Table 1: Summary of estimates of the effect size of genetic component of expression on the outcome. The unadjusted TWAS approaches are referred to as Lasso, ALasso (adaptive Lasso), and Enet (elastic net) TWAS, depending on the type of penalty function used. PENA denotes the adjusted TWAS approach penaTWAS. The heritability of gene expression is 10%. The first choice of true eQTL effects was considered. The mean and s.d. of the estimates obtained across simulation iterations are provided for each method.

β_1	Lasso	Enet	ALasso	PENA
0.02	0.02 (0.07)	0.02 (0.07)	0.01 (0.04)	0.02 (0.07)
0.04	0.05 (0.07)	0.05 (0.07)	0.03 (0.04)	0.04 (0.07)
0.06	0.07 (0.08)	0.07 (0.08)	0.04 (0.04)	0.06 (0.08)
0.08	0.09 (0.08)	0.09 (0.08)	0.05 (0.04)	0.09 (0.08)
0.1	0.11 (0.09)	0.12 (0.09)	0.07 (0.04)	0.11 (0.09)
0.12	0.14 (0.09)	0.14 (0.1)	0.08 (0.05)	0.13 (0.1)
0.14	0.16 (0.1)	0.16 (0.11)	0.1 (0.05)	0.15 (0.11)
0.16	0.18 (0.11)	0.19 (0.11)	0.11 (0.05)	0.17 (0.11)
0.18	0.21 (0.12)	0.21 (0.12)	0.12 (0.05)	0.19 (0.12)
0.2	0.23 (0.13)	0.23 (0.13)	0.14 (0.06)	0.21 (0.13)

Table 2: Summary of estimates of the effect size of genetic component of expression on the outcome. The unadjusted TWAS approaches are referred to as Lasso, ALasso (adaptive Lasso), and Enet (elastic net) TWAS, depending on the type of penalty function used. PENA denotes the adjusted TWAS approach penaTWAS. The heritability of gene expression is 10%. The second choice of true eQTL effects was considered. The mean and s.d. of the estimates obtained across simulation iterations are provided for each method.

β_1	Lasso	Enet	ALasso	PENA
0.02	0.02 (0.06)	0.02 (0.06)	0.01 (0.03)	0.02 (0.05)
0.04	0.05 (0.06)	0.05 (0.06)	0.03 (0.04)	0.04 (0.05)
0.06	0.07 (0.07)	0.07 (0.07)	0.04 (0.04)	0.06 (0.06)
0.08	0.09 (0.07)	0.1 (0.07)	0.06 (0.04)	0.08 (0.06)
0.1	0.12 (0.08)	0.12 (0.08)	0.07 (0.04)	0.1 (0.07)
0.12	0.14 (0.08)	0.14 (0.08)	0.08 (0.04)	0.12 (0.08)
0.14	0.16 (0.09)	0.17 (0.09)	0.1 (0.04)	0.14 (0.09)
0.16	0.19 (0.1)	0.19 (0.1)	0.11 (0.05)	0.16 (0.1)
0.18	0.21 (0.1)	0.21 (0.11)	0.13 (0.05)	0.18 (0.1)
0.2	0.23 (0.11)	0.24 (0.12)	0.14 (0.05)	0.2 (0.11)

Table 3: Summary of estimates of the effect size of genetic component of expression on a binary case-control phenotype. The unadjusted TWAS approaches are referred to as Lasso, ALasso (adaptive Lasso), and Enet (elastic net) TWAS, depending on the type of penalty function used. PENA denotes the adjusted TWAS approach penaTWAS. The heritability of gene expression is 20%. The mean and s.d. of the estimates obtained across simulation iterations are provided for each method.

β_1	Lasso	Enet	ALasso	PENA
0.02	0.02 (0.08)	0.02 (0.08)	0.01 (0.06)	0.02 (0.07)
0.04	0.05 (0.08)	0.05 (0.08)	0.03 (0.06)	0.04 (0.07)
0.06	0.07 (0.08)	0.07 (0.08)	0.05 (0.06)	0.06 (0.07)
0.08	0.09 (0.08)	0.09 (0.09)	0.06 (0.06)	0.08 (0.07)
0.1	0.12 (0.08)	0.12 (0.09)	0.08 (0.06)	0.1 (0.08)
0.12	0.14 (0.09)	0.14 (0.09)	0.1 (0.06)	0.12 (0.08)
0.14	0.16 (0.09)	0.17 (0.09)	0.11 (0.06)	0.14 (0.08)
0.16	0.19 (0.09)	0.19 (0.1)	0.13 (0.06)	0.16 (0.08)
0.18	0.21 (0.1)	0.21 (0.1)	0.14 (0.07)	0.18 (0.08)
0.2	0.23 (0.1)	0.24 (0.1)	0.16 (0.07)	0.2 (0.09)

Figure 7: Number of genes associated with a phenotype identified by both penaTWAS and Enet TWAS, only by penaTWAS, and only by Enet TWAS. The subfigure (a) presents the results for height and HDL cholesterol, and the subfigure (b) presents the results for LDL cholesterol and triglycerides. PenaTWAS represents the adjusted TWAS approach, and Enet TWAS represents the unadjusted TWAS approach.

(a)

(b)