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

1 Abstract

Transcriptome-wide association study (TWAS) is a promising approach to identifying novel genes associated with complex phenotypes. It offers better biological insights into the gene-phenotype associations than a genome-wide association study (GWAS). Standard TWAS approaches build a prediction model for the genetic component of expression based on reference transcriptome data. Next, an outcome phenotype is regressed on the predicted expression in separate GWAS data. The traditional unadjusted TWAS approach ignores the uncertainty of predicted expression, which can lead to unreliable discovery of 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 obtain an adjustment factor

that needs to be incorporated in the unadjusted TWAS. We use extensive simulations to show that the unadjusted TWAS severely inflates the type 1 error rate, whereas the adjusted TWAS adequately controls it. At the expense of an alarmingly inflated false positive rate, the unadjusted TWAS offers a higher power than the adjusted TWAS, which is statistically unjustified. We demonstrate the merits of adjusted TWAS by conducting TWAS for height and lipid phenotypes while integrating the Geuvadis transcriptome and UK Biobank GWAS data.

Keywords: Gene-phenotype associations, Predicted expression, Type 1 error rate, Penalized regression, Residual bootstrap, Trimmed mean.

2 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, promot-

ers, 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 phenotype. It combines the genotype and outcome phenotype data from GWAS with reference transcriptome data. For example, consider the lipid phenotype, HDL, the outcome phenotype 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, consider a gene and its expression in the liver tissue. 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 equation 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 phenotype, 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 41% of significant TWAS associations found in 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, it can increase the wastage of extensive effort and valuable resources to reveal the biological mechanisms underlying the associations 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 refer to this approach as adjusted TWAS. We develop the approach for the summary-statistics version of TWAS, which requires summary-level association data from GWAS instead of individual-level data [8]. The method can easily be applied to complex phenotypes using publicly available summary-level GWAS data. However, the procedure utilizes individual-level reference transcriptome data.

We use simulations to show that adjusted TWAS adequately controls the type 1 error rate, whereas unadjusted TWAS drastically inflates it. At the cost of an inflated false positive rate, unadjusted TWAS produces a higher power than adjusted TWAS. The difference in power decreases due to increased sample sizes of the transcriptome and GWAS data.

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. Contrary to the unadjusted approach, the method correctly controls the false positive rate in this setting. For a multi-phenotype TWAS, we compare the performance of the subset-based meta-analysis method ASSET [18], using the test statistics obtained by adjusted and unadjusted TWAS. ASSET sufficiently controls the type 1 error rate based on adjusted TWAS statistics but leads to an inflated type 1 error rate based on unadjusted TWAS statistics. Analogous to the univariate scenario, the multi-phenotype unadjusted TWAS power is higher than the adjusted approach due to an excessive type 1 error rate. We implement adjusted TWAS for height, combining Geuvadis transcriptome and UK Biobank GWAS data. We also conduct a multi-phenotype adjusted TWAS for lipid phenotypes to demonstrate the method.

3 Materials and methods

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

3.1 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 n_1 individuals in the transcriptome data.

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

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

3.2 Various notations used for the GWAS data

n: number of individuals.

$\mathbf{y}$: vector of outcome phenotype 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 phenotype in GWAS.

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

$\mathbf{z}_{\text{gwas}}$: 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_{adj}: Z-statistic for the adjusted TWAS.

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

3.3 Regression of gene expression on local SNPs in transcriptome data

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

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

Let $\mathbf{E}$ denote the vector of mean-centered expression values of the gene. G_{ref} denotes the genotype matrix for n_1 individuals and p local SNPs of the gene, genotypes for each SNP

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. 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 [19] and Elastic net [20] 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. Let $\hat{\boldsymbol{\alpha}}$ denote the estimate of $\boldsymbol{\alpha}$ obtained by the preferred penalized regression. Let $\mathrm{E}(\hat{\boldsymbol{\alpha}})$ and $\mathrm{cov}(\hat{\boldsymbol{\alpha}})$ denote the expectation vector and covariance matrix of $\hat{\boldsymbol{\alpha}}$.

3.4 Regression of phenotype on the genetic component of expression in GWAS

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 genotype measurements for the set of p local SNPs considered for the gene in the transcriptome data. The genetic component of expression is predicted using $\hat{\boldsymbol{\alpha}}$, obtained in the first-step regression. 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 $X = \mathbf{g}^{'}\hat{\boldsymbol{\alpha}}$. The main goal of TWAS is to test for an association between the outcome phenotype 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.

3.5 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. $N(0, \sigma_\epsilon^2)$. 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 with similar lines of measurement error in the regressors in linear regression [16]. 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 un-

adjusted 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 $\mathrm{var}(y_i) = \mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$. Under the ideal regression model, $\mathrm{cov}(X_i, y_i) = \mathrm{cov}(\mathbf{g}_i^{'}\hat{\boldsymbol{\alpha}}, \beta_0 + \beta_1 x_i + \epsilon_i) = 0$. We make reasonable assumptions that $\boldsymbol{\alpha}$, hence x_i, is an unknown non-random quantity, and the random error in ideal regression ϵ_i is independent of $\hat{\boldsymbol{\alpha}}$. Thus, $\mathrm{cov}(X_i, y_i) = 0$, and if the transcriptome data sample size is sufficiently large such that $\hat{\boldsymbol{\alpha}}$ asymptotically follows multivariate normal [21], X_i and y_i are independently distributed under the ideal regression model. In the following sections, we assume that $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ are independently distributed under the ideal regression model (equation 4).

3.5.1 Adjusted estimator of β_1 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 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 unbiased 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 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. Thus,

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

Hence, $\mathrm{E}\left(\frac{\hat{b}_1}{B}\right) = \beta_1$. Consider $\hat{\beta}_1 = \frac{\hat{b}_1}{\hat{B}}$. $\mathrm{var}(\frac{\hat{b}_1}{B}) = \frac{\mathrm{var}(\hat{b}_1)}{B^2}$. 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 adjusted TWAS as Z_{adj}. 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 $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$, $Z_{\text{adj}} = \frac{\frac{\hat{b}_1}{\hat{B}}}{\frac{\hat{\sigma}_e}{\sqrt{n}\hat{B}}\sqrt{\hat{\mathrm{E}}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)}}$ $= \frac{\sqrt{n}\hat{b}_1}{\hat{\sigma}_e\sqrt{\hat{\mathrm{E}}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)}}$. 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 $\mathrm{var}(\hat{b}_1|\mathbf{X}) = \frac{\sigma_e^2}{n\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}$. Thus, 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_{adj} and Z_{unadj} as:

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

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.

3.6 Estimation of the adjustment factor based on bootstrapping

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 transcriptome 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 in the analytical form of A. 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, assuming 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 [21] 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 adjusted TWAS approach under the framework of adaptive Lasso [21]. The adaptive Lasso (ALasso) estimator for the local SNPs' effects on the expression is given by:

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

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^{'}$, and $\hat{\boldsymbol{\alpha}}_{ols} = \{\hat{\alpha}_{j,ols}, j = 1, \ldots, p\}$ denotes the ordinary least squares estimate (OLS) of $\boldsymbol{\alpha}$ based on the full sample. Here, p denotes the number of local SNPs, λ denotes the weight associated with the penalty, and $\gamma > 0$ allows for varying the influence of the OLS 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 five-fold cross-validation to choose the optimal λ. Adaptive Lasso enjoys the oracle property, i.e., the probability of identifying the true subset of local SNPs having non-zero effects converges to one as the sample size increases, and the estimator asymptotically follows multivariate normal under some mild regularity conditions [21]. 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.

3.6.1 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: $\hat{\epsilon}_i = E_i - \mathbf{g}_{\text{ref},i}^{'}\hat{\boldsymbol{\alpha}}$, 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 $(r_1^+, \ldots, r_{n_1}^+)$ denotes a random sample of size n_1 drawn from $(r_1, \ldots, r_{n_1})$ with replacement. Define $E_i^+ = \mathbf{g}_{\text{ref},i}^{'}\hat{\boldsymbol{\alpha}} + r_i^+$, $i = 1, \ldots, n_1$. Consider $(E_i^+, \mathbf{g}_{\text{ref},i}), i = 1, \ldots, n_1$ to implement ALasso:

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

where $\tilde{\boldsymbol{\alpha}}_{ols} = \{\tilde{\alpha}_{j,ols}, j = 1, \ldots, p\}$ is the OLS estimate of $\boldsymbol{\alpha}$ based on the residual bootstrap sample $\{(E_i^+, \mathbf{g}_{\text{ref},i}), i = 1, \ldots, n_1\}$. Thus, $\tilde{\boldsymbol{\alpha}}$ provides the bootstrapped ALasso estimate of $\boldsymbol{\alpha}$ for a given residual bootstrap sample. We can use five-fold cross-validation to choose optimal λ for each bootstrap sample, which is computationally intensive. Or, we can use the optimal λ obtained from the full data above (equation 9) to save computing time. We consider 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 $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$.

3.7 Summary statistics version of unadjusted TWAS

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 phenotype 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 phenotype. 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:

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

Without individual-level data from GWAS, we can use 1000 genome data to estimate V. To adjust for the uncertainty of the predicted expression in the unadjusted summary TWAS, we consider $Z_{\mathrm{adj}} = \hat{A} \times Z_{\mathrm{unadj}}$ (equation 8), where $\hat{A}$ is the estimated adjustment factor described in previous sections. Note that the adjusted summary TWAS requires individual-level transcriptome data and summary-level GWAS association data.

3.8 Expectation and variance of $Z_{\mathbf{adj}}$ under H_0

To test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, we require the expectation and variance of Z_{adj} under H_0. For the summary TWAS, $Z_{\text{adj}} = \hat{A} \times Z_{\text{unadj}}$, where $A = \frac{\frac{1}{\sqrt{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}}}{\sqrt{\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)}}$ 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, $\mathrm{E}_{H_0}(Z_{\text{adj}}) = \mathrm{E}(\frac{\hat{A}}{\sqrt{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}} \hat{\boldsymbol{\alpha}}')\mathrm{E}_{H_0}(\mathbf{z}_{\text{gwas}})$. Under the assumption that no TWAS association for the gene implies no GWAS association for the local SNPs, $\mathrm{E}_{H_0}(\mathbf{z}_{\text{gwas}}) = \mathbf{0}$, hence $\mathrm{E}(Z_{\text{adj}}) = 0$. Next, we obtain the variance term under H_0, making the same assumption.

$Z_{\text{adj}} = \frac{1}{\sqrt{\hat{C}}} \frac{\hat{\boldsymbol{\alpha}}' \mathbf{z}_{\text{gwas}}}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}$, where $C = \mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}}\right)$. $\mathrm{var}(Z_{\text{adj}}) = \mathrm{var}(\frac{\hat{\boldsymbol{\alpha}}' \mathbf{z}_{\text{gwas}}}{\sqrt{\hat{C}} \hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}})$. 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, $\mathrm{var}(Z_{\text{adj}}) = \mathrm{var}(\mathbf{u}'\mathbf{w})$. Note that $\mathrm{var}(\mathbf{u}'\mathbf{w}) = \mathrm{E}((\mathbf{u}'\mathbf{w})^2) - \mathrm{E}^2(\mathbf{u}'\mathbf{w})$. Since $\mathbf{u}$ and $\mathbf{w}$ are independent, $\mathrm{E}(\mathbf{u}'\mathbf{w}) = \mathrm{E}(\mathbf{u}')\mathrm{E}(\mathbf{w})$. Again, $\mathrm{E}(\mathbf{w}) = \mathbf{0}$ under the above assumption. Thus, $\mathrm{var}(\mathbf{u}'\mathbf{w}) = \mathrm{E}((\mathbf{u}'\mathbf{w})^2) = \mathrm{E}((\mathbf{u}'\mathbf{w}) \times (\mathbf{w}'\mathbf{u})) = \mathrm{tr}(\mathrm{E}(\mathbf{u}'\mathbf{w}\mathbf{w}'\mathbf{u}))$ $= \mathrm{E}(\mathrm{tr}(\mathbf{u}'\mathbf{w}\mathbf{w}'\mathbf{u})) = \mathrm{E}(\mathrm{tr}(\mathbf{u}\mathbf{u}'\mathbf{w}\mathbf{w}')) = \mathrm{tr}(\mathrm{E}(\mathbf{u}\mathbf{u}'\mathbf{w}\mathbf{w}')) = \mathrm{tr}(\mathrm{E}(\mathbf{u}\mathbf{u}')\mathrm{E}(\mathbf{w}\mathbf{w}'))$. 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. Under the assumption that all local SNPs are not associated with the outcome phenotype, $\mathrm{cov}(\mathbf{z}_{\text{gwas}})$ $= \mathrm{cov}(\frac{1}{\sqrt{n}} G'\mathbf{y}) = \frac{1}{n} G' \mathrm{cov}(\mathbf{y}) G = \frac{1}{n} G'G = V$. Thus, $\mathrm{E}(\mathbf{w}\mathbf{w}') = \mathrm{cov}(\mathbf{w}) = V$, where V is the LD matrix of local SNPs in the GWAS data. $\mathrm{var}(Z_{\text{adj}}) = \mathrm{var}(\mathbf{u}'\mathbf{w}) = \mathrm{tr}(\mathrm{E}(\mathbf{u}\mathbf{u}')V)$ $= \mathrm{tr}(V\mathrm{E}(\mathbf{u}\mathbf{u}')) = \mathrm{tr}(\mathrm{E}(V\mathbf{u}\mathbf{u}')) = \mathrm{E}(\mathrm{tr}(V\mathbf{u}\mathbf{u}')) = \mathrm{E}(\mathrm{tr}(\mathbf{u}'V\mathbf{u})) = \mathrm{tr}(\mathrm{E}(\mathbf{u}'V\mathbf{u})) = \mathrm{E}(\mathbf{u}'V\mathbf{u})$ $= \mathrm{E}(\frac{1}{\hat{C}\hat{\boldsymbol{\alpha}}' V \hat{\boldsymbol{\alpha}}})$. Thus, the estimate of $\mathrm{var}(Z_{\text{adj}})$ under H_0 is one.

The unadjusted TWAS test statistic Z_{unadj} follows $N(0, 1)$ under H_0. Based on Slutsky's lemma, if the adjustment factor $\hat{A}$ converges in probability to one, the adjusted test statistic Z_{adj} should converge to $N(0, 1)$ 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_{adj} 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 very well with $N(0,1)$ (results described in the simulation study section).

3.9 Multi-phenotype TWAS

We implement a multi-phenotype TWAS approach based on the adjusted TWAS statistics. Without individual-level data, multi-phenotype association analyses are conducted based on summary-level data. We extend one such popular subset-based meta-analysis approach, ASSET [18], from SNP-level to gene-level under the TWAS framework. Suppose $\mathbf{Y} = (Y_1, \ldots, Y_k)'$ denotes a multivariate phenotype comprising k outcome phenotypes, each of which can have a null or non-null effect due to the true genetic component of expression: $\mathbf{Y} = x\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_k)'$. $\boldsymbol{\beta}$ denotes the effect sizes of the true genetic component of expression on $\mathbf{Y}$. The random error vector $\boldsymbol{\epsilon}$ follows multivariate normal with zero mean vector and a dispersion matrix. To perform a multi-phenotype analysis using summary-level association data, we provide the TWAS Z statistics for individual phenotypes and their covariance matrix to ASSET as the inputs. We compute the phenotype-specific Z_{adj} for $Y_j = \beta_j x + \epsilon_j$, for each j at a time, $j = 1, \ldots, k$. ASSET tests for the global null hypothesis $H_0 : \beta_1 = \ldots = \beta_k = 0$ versus $H_1 : \cup_{j=1}^{k} \{\beta_j \neq 0\}$. We obtain the formula for covariances between the Z statistics across phenotypes under H_0 and discuss their estimation procedure in the supplementary materials.

4 Simulation study

We performed extensive simulations to evaluate the adjusted TWAS approach for univariate and multivariate phenotypes. We compared the type 1 error rates and powers estimated

by the unadjusted and adjusted TWAS. We implemented the *hapgen* [25] 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 outcome phenotype in the GWAS data.

4.1 Setting

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. In most simulations, we chose the transcriptome data sample size of $n_1 = 500$ and the GWAS data sample size of $n = 10,000$.

After generating the genotype data applying *hapgen*, 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 $h_E^2 = 20\%$. We simulated four choices

of $\boldsymbol{\alpha}$ for a continuous outcome phenotype. We considered one choice of $\boldsymbol{\alpha}$ for a case-control and multivariate outcome phenotype.

We next simulate the outcome phenotype 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 five choices of β_1 as $0.02, 0.04, 0.06, 0.08, 0.1$ for power study and $\beta_1 = 0$ for evaluating the type 1 error rate. To simulate a case-control phenotype, we employ the logistic model to obtain the probability of a GWAS individual being a case: $P(\text{case}|\beta_1, \mathbf{g}, \boldsymbol{\alpha}) = \frac{\exp(\beta_1 x)}{1+\exp(\beta_1 x)}, x = \mathbf{g}'\boldsymbol{\alpha}$. To compare the power between the unadjusted and adjusted TWAS when the sample size increases, we chose $n_1 = 750$ and $n = 20,000$ for a univariate continuous phenotype. Here, we studied more refined choices of β_1 as $0.01, 0.02, 0.03, \ldots, 0.1$.

We use a similar setup for a multi-phenotype TWAS. We consider a linear model for four phenotypes: $\boldsymbol{y} = x\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y} = (y_1, y_2, y_3, y_4)'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$. We presume that $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_4)' \sim N(\mathbf{0}, \Sigma_\epsilon)$, where Σ_ϵ is a correlation matrix in which off-diagonal elements are generated from Uniform$(-0.4, 0.4)$. For simplicity, we regard the same tissue type as causal for the component traits in the multivariate phenotype. We test for the global null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus the alternative hypothesis $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$. To estimate the power, we made five different choices of $\boldsymbol{\beta}$ as $(0, a, 0, a)'$, where $a = 0.02, 0.04, 0.06, 0.08, 0.1$.

4.2 Analysis pipeline

We applied the GCTA software [26] to determine if a gene is locally heritable based on a commonly used threshold of p-value: $\frac{0.05}{20000}$. Around 20,000 protein-coding genes are tested for local heritability, hence the Bonferroni correction. In a simulation scenario, we

regarded an iteration for downstream analysis if the gene was heritable in the iteration. For unadjusted TWAS, we implemented penalized regression using the Lasso and Elastic net penalty to estimate the genetically regulated expression component prediction model based on the transcriptome data. Next, we computed the summary-level association data for the local SNPs in GWAS and calculated the unadjusted TWAS statistic. We obtained the adjustment factor for adjusted TWAS by applying the adaptive Lasso and residual bootstrapping. In an iteration, we utilized the LD matrix of local SNPs estimated from the transcriptome data in a randomly selected iteration, excluding the current one. This strategy imitates the TWAS procedure based on summary-level data, where LD structure is estimated from external reference data, such as 1000 genome data. The latter has a sample size of around 500 for the European population. The only change in TWAS for a case-control phenotype is to compute the GWAS summary statistics for the local SNPs based on a logistic regression. For multi-phenotype TWAS, we calculated the unadjusted and adjusted TWAS statistics for each continuous phenotype separately. We applied ASSET using the phenotype covariance matrix as the estimated covariance matrix of TWAS Z statistics [24]. To assess power and type 1 error rate, we considered the level of significance 0.05 and 1500 iterations in a simulation scenario.

We examined $0\%, 5\%, 10\%$ trimmed means of the bootstrapped estimates while estimating the denominator of the adjustment factor in adjusted TWAS. We denote the adjusted TWAS that utilizes $0\%, 5\%, 10\%$ trimmed means as adjTWAS0, adjTWAS1, and adjTWAS2, respectively. To implement the adaptive Lasso in the residual bootstrap samples (equation 10) while conducting the univariate adjusted TWAS, we utilized the optimal λ, obtained from the adaptive Lasso for complete data (equation 9). For multivariate adjusted TWAS, we determined the optimal λ for each residual bootstrap sample using five-fold cross-validation. We used $T = 350$ residual bootstrap samples in the adjusted TWAS. Let TWAS.Lasso and TWAS.Enet denote the unadjusted TWAS implemented using the Lasso

and Elastic net penalty in the penalized regression for expression prediction.

To explore the normality of the null distribution of adjusted TWAS statistics, we constructed the QQ and empirical density plots for the test statistic values from the univariate TWAS of the continuous phenotype under the null hypothesis of no association.

4.3 Results

The unadjusted TWAS approach produced a drastically inflated type 1 error rate for univariate TWAS of a continuous phenotype (Figure 2). In some scenarios (choices 1 and 4 in Figure 2), the inflation is high at 5% level, reaching 10% type 1 error rate. Both TWAS.Lasso and TWAS.Enet resulted in inflated false positive rates due to ignoring the uncertainty of predicted expression. On the contrary, adjusted TWAS controls the type 1 error rate adequately. For choice 1, the adjusted TWAS estimated the type 1 error rate as 0.058 compared to 0.097 by TWAS.Lasso, and 0.1 by TWAS.Enet. For choices 2 and 3, the adjusted TWAS evaluated the type 1 error rate as 0.035 and 0.046. For these choices, adjTWAS0, adjTWAS1, and adjTWAS2 produced the same type 1 error rate estimates. For choice 4, adjTWAS0, adjTWAS1, and adjTWAS2 assessed the type 1 error rate as 0.051, 0.055, and 0.056, respectively. The unadjusted TWAS led to a substantially inflated type 1 error rate for all these choices (Figure 2).

We observed a similar pattern for a case-control phenotype and continuous multivariate phenotype (Figure 3,4). For a case-control phenotype, the adjusted TWAS estimated a type 1 error rate of 0.05, while the unadjusted TWAS using Lasso and Elastic net penalty generated a type 1 error rate of 0.07 and 0.1 (Figure 3). Thus, the adjusted TWAS approach controls a binary phenotype's TWAS type 1 error rate. For a multivariate phenotype, ASSET based on unadjusted TWAS statistics provided a 0.12 estimate of type 1 error rate (Figure 4). Based on adjTWAS0, adjTWAS1, and adjTWAS2 statistics, ASSET evaluated the type 1 error rate as 0.026, 0.031, and 0.033 (Figure 4). Thus, ASSET's type 1 error rate

using unadjusted TWAS is severely inflated, but the method achieves the desired control of type 1 errors based on adjusted TWAS. We note that the adjusted TWAS using the optimal λ, obtained from the adaptive Lasso for complete data (equation 9) while implementing the adaptive Lasso in the residual bootstrap samples (equation 10), does not control the type 1 error rate adequately for ASSET. Even though the procedure performed well for a univariate TWAS. Thus, the adjusted TWAS outperforms the unadjusted TWAS to control the type 1 error rate.

The unadjusted TWAS offered a higher power than the adjusted TWAS at the expense of a severely inflated false positive rate, and hence, the power gain is not statistically justified. The power difference for univariate TWAS of continuous phenotype varies in the range of $3\% - 22\%$ for choice 1, $8\% - 25\%$ for choice 2, $8\% - 28\%$ for choice 3, and $10\% - 28\%$ for choice 4 (Figure 2). adjTWAS2 (10% trimmed mean) produced a marginally higher power than adjTWAS0 (no trimming) and adjTWAS1 (5% trimming). Increasing the trimming percentage did not improve the power noticeably further (results not reported for brevity). Unadjusted TWAS using the Elastic net penalty led to $9 - 28\%$ higher power than adjTWAS2 for a binary phenotype (Figure 3). Regarding the power of a multi-phenotype TWAS, we observe a similar comparative performance of the methods (Figure 4). The adjTWAS2 resulted in a substantially higher power than adjTWAS0 and marginally higher than adjTWAS1 (Figure 4). Hence, we use adjTWAS2 for adjusted TWAS in the remaining sections of the paper. ASSET employing unadjusted TWAS.Enet statistics led to maximum power with an increase of $15 - 39\%$ than adjTWAS2. Again, the power gain is caused by the acute hike in false positive rate by ASSET while using the unadjusted TWAS statistics (Figure 4).

We experimented with larger transcriptome and GWAS data sample sizes for univariate TWAS of a continuous phenotype, $(n_1 = 750, n = 20,000)$ vs $(n_1 = 500, n = 10,000)$. The maximum power difference between unadjusted and adjusted TWAS marginally decreases

due to increased sample sizes. For TWAS.Lasso, the power difference with adjTWAS2 varied in $7-25\%$ when $n_1 = 750$ and $n = 20,000$ compared to $6-28\%$ when $n_1 = 500$ and $n = 10,000$. For TWAS.Enet, the power difference varied in $7-24\%$ for the larger sample sizes compared to $5-28\%$.

The empirical density plots for the adjusted TWAS test statistics appear bell-shaped and symmetric around the mean zero under the null hypothesis (Figure S1). QQ plots demonstrate an excellent match between the null distribution and $N(0,1)$ (Figure S2).

5 Real data application

We applied the adjusted and unadjusted TWAS approaches to height for univariate analysis and lipids for multivariate analysis. We considered LDL, HDL, and triglycerides as the lipid phenotypes. 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, gene expressions were measured in the lymphoblastoid cell line in peripheral blood tissue. 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 outcome phenotypes. For example, liver tissue instead of blood tissue is one of the most pertinent tissue types for lipids [28]. The number of European individuals included in the Geuvadis data was 358, and the number of white-British individuals considered in UKB was 337,000.

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. 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 [26] to determine if a gene is locally heritable. We applied the BH FDR

controlling procedure at 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. We employed the penalized regression based on the Elastic net penalty for the unadjusted TWAS. We utilized 350 residual bootstrap samples for the adjusted TWAS. We applied the summary TWAS using the summary-level GWAS association data from UKB, obtained by adjusting for relevant covariates, such as age, sex, and PCs for genetic ancestry. 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., $\frac{0.05}{1515}$.

First, we discuss the results from the univariate analysis for height. Adjusted TWAS discovered 46 genes (Table 1), whereas the unadjusted TWAS detected 86 genes (Table S1). The number of significant genes identified by the approaches resembles the higher power of unadjusted TWAS at the expense of inflated false positive rates observed in simulations. The adjusted and unadjusted TWAS identified 33 common genes, leaving 13 genes discovered only by the adjusted approach and 53 genes found only by the unadjusted approach (Figure S3). In the multi-phenotype analysis for lipids, ASSET tested for a global association that at least one lipid phenotype is associated with the gene. ASSET identified 41 genes based on adjusted TWAS statistics (Table 2) and 63 genes using the unadjusted TWAS statistics (Table S2). Again, both approaches detected 31 common genes; 10 genes were discovered only by adjusted TWAS; 32 genes were found only by unadjusted TWAS (Figure S3).

Many of the genes found by the adjusted TWAS (Table 1 and 2) are reported to be associated with the analyzed phenotypes in the NHGRI-EBI GWAS catalog. For height, *INPP5B* gene in the chromosomal region 1p34.3 has been associated with body height in the Japanese population [32]. Similarly, *RPS20* gene in 8q12.1, *CLHC1* in 2p16.1, and

PACSIN2 in 22q13.2 are associated with height (Table 1). Few genes are not known to be directly associated with height but instead with other anthropometric phenotypes. For example, STOX1 in 10q22.1 is associated with BMI-adjusted hip circumference but not height. For lipid phenotypes, *CTSS* in 1q21.3 is communicated to be associated with HDL cholesterol, *CGREF1* in 2p23.3 with LDL cholesterol and triglyceride, *MRPL45P2* in 17q21.32 with LDL cholesterol (Table 2), etc. Thus, the adjusted TWAS effectively discovers gene-phenotype associations.

6 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 phenotype on the imputed expression. This shortcoming can lead to inadequate control of type 1 or 2 error rates 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 statistic. Then, we implement bootstrapping for penalized regression to devise an estimation procedure for the adjustment factor. We extend the adjusted TWAS approach for a multivariate phenotype. We implement ASSET for the multi-phenotype TWAS, but other suitable meta-analysis methods can also be applied. The adjusted TWAS is based on GWAS summary statistics and individual-level transcriptome data. Therefore, it can easily be implemented using publicly available GWAS summary statistics data for complex phenotypes. However, contrary to unadjusted TWAS, the adjusted approach requires individual-level transcriptome data.

The adjusted TWAS performs reliably to control the type 1 error rate. However, the unadjusted TWAS can drastically inflate the type 1 error rate, which resembles the results

communicated in Leeuw et al. [15]. The findings are consistent across continuous, case-control, and multivariate phenotypes. Thus, many gene-phenotype associations discovered by unadjusted TWAS can be false positive signals. The unadjusted TWAS offers higher power than the adjusted TWAS at the expense of an alarmingly inflated type 1 error rate. The power difference decreases marginally with sample size increases for transcriptome and GWAS data. As the sample sizes increase in contemporary studies, we expect the adjusted TWAS to provide higher power.

While computing the adjusted TWAS statistics for implementing ASSET, we used five-fold cross-validation to determine the optimal λ in adaptive Lasso for each residual bootstrap sample (equation 10). ASSET does not adequately control the type 1 error rate if we use the optimal λ obtained from the adaptive Lasso for complete data (equation 9) to implement the adaptive Lasso in the residual bootstrap samples (equation 10). We recommend choosing the optimal λ in each residual bootstrap sample for a univariate and multivariate TWAS. However, the complete data-based choice of λ worked effectively for the univariate TWAS in simulations and saved significant computing time.

We conducted univariate TWAS for height and multivariate TWAS for lipid phenotypes while integrating the Geuvadis transcriptome and UK Biobank GWAS data. The adjusted TWAS performed well; the number of genes discovered was smaller than the unadjusted TWAS approach, falling in line with the simulation results. Adjusted TWAS found some genes overlooked by the unadjusted TWAS and vice versa. However, a substantial proportion of genes were detected by both approaches, validating the use of adjusted TWAS. Thus, the adjusted TWAS identified many gene-phenotype associations, and the number will likely increase as the transcriptome and GWAS data sample sizes increase. We recommend using the adjusted TWAS rather than the unadjusted approach due to better control of the former concerning the type 1 error rate.

We acknowledge the limitation of our real data analyses: the tissue type considered may

not be the most relevant tissue type for the outcome phenotypes. In future work, we aim to apply the method to more phenotypes while integrating transcriptome data with larger sample sizes. We plan to extend this approach to incorporate multiple populations [33]. Another important direction is to devise the strategy to combine multiple tissue types available in transcriptome data together [34]. Multi-ethnic and multi-tissue type TWAS can substantially boost the power of adjusted TWAS.

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. The method is technically sound and computationally efficient. It can be implemented for complex phenotypes using publicly available GWAS summary statistics and individual-level transcriptome data.

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] Samsiddhi Bhattacharjee, Preetha Rajaraman, Kevin B Jacobs, William A Wheeler, Beatrice S Melin, Patricia Hartge, Meredith Yeager, Charles C Chung, Stephen J Chanock, and Nilanjan Chatterjee. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics*, 90(5):821–835, 2012.

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

[20] 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.

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

[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] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Laramie Duncan, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236–1241, 2015.

[24] Helian Feng, Nicholas Mancuso, Bogdan Pasaniuc, and Peter Kraft. Multitrait transcriptome-wide association study (twas) tests. *Genetic Epidemiology*, 45(6):563–576, 2021.

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

[26] 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.

[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] Masato Akiyama, Kazuyoshi Ishigaki, Saori Sakaue, Yukihide Momozawa, Momoko Horikoshi, Makoto Hirata, Koichi Matsuda, Shiro Ikegawa, Atsushi Takahashi, Masahiro Kanai, et al. Characterizing rare and low-frequency height-associated variants in the japanese population. *Nature communications*, 10(1):4393, 2019.

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

[34] 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. The yellow boxes represent the adjustment for the uncertainty of the predicted expression.

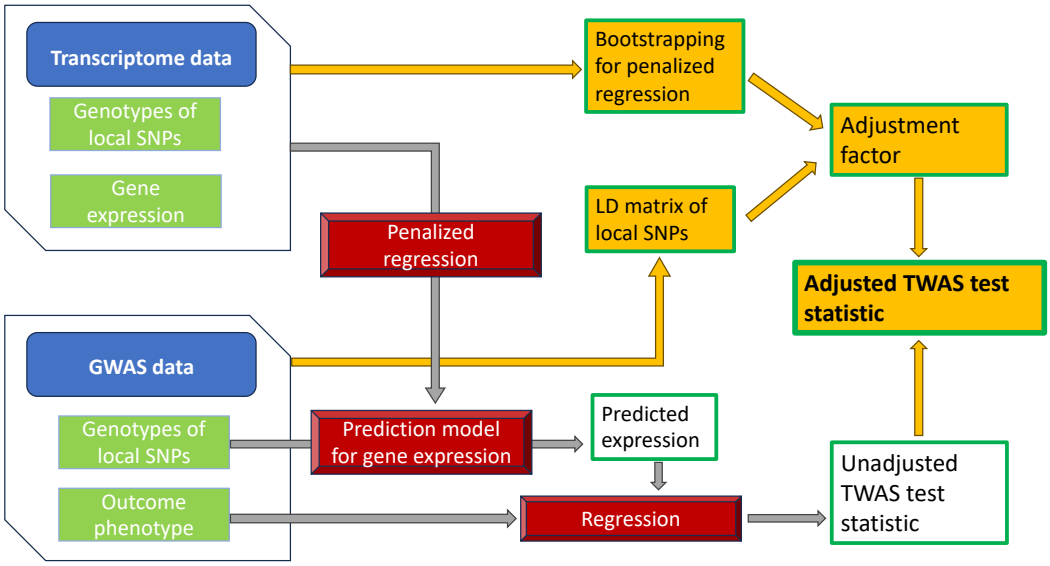

Figure 2: Comparison of type 1 error rate and power between the unadjusted TWAS and adjusted TWAS for the univariate continuous phenotype. adjTWAS0, adjTWAS1, adjTWAS2 denote the adjusted TWAS approaches that use $0\%, 5\%, 10\%$ trimmed means while estimating the adjustment factor. TWAS.Lasso and TWAS.Enet denote the unadjusted TWAS using the Elastic net and Lasso penalties in penalized regression.

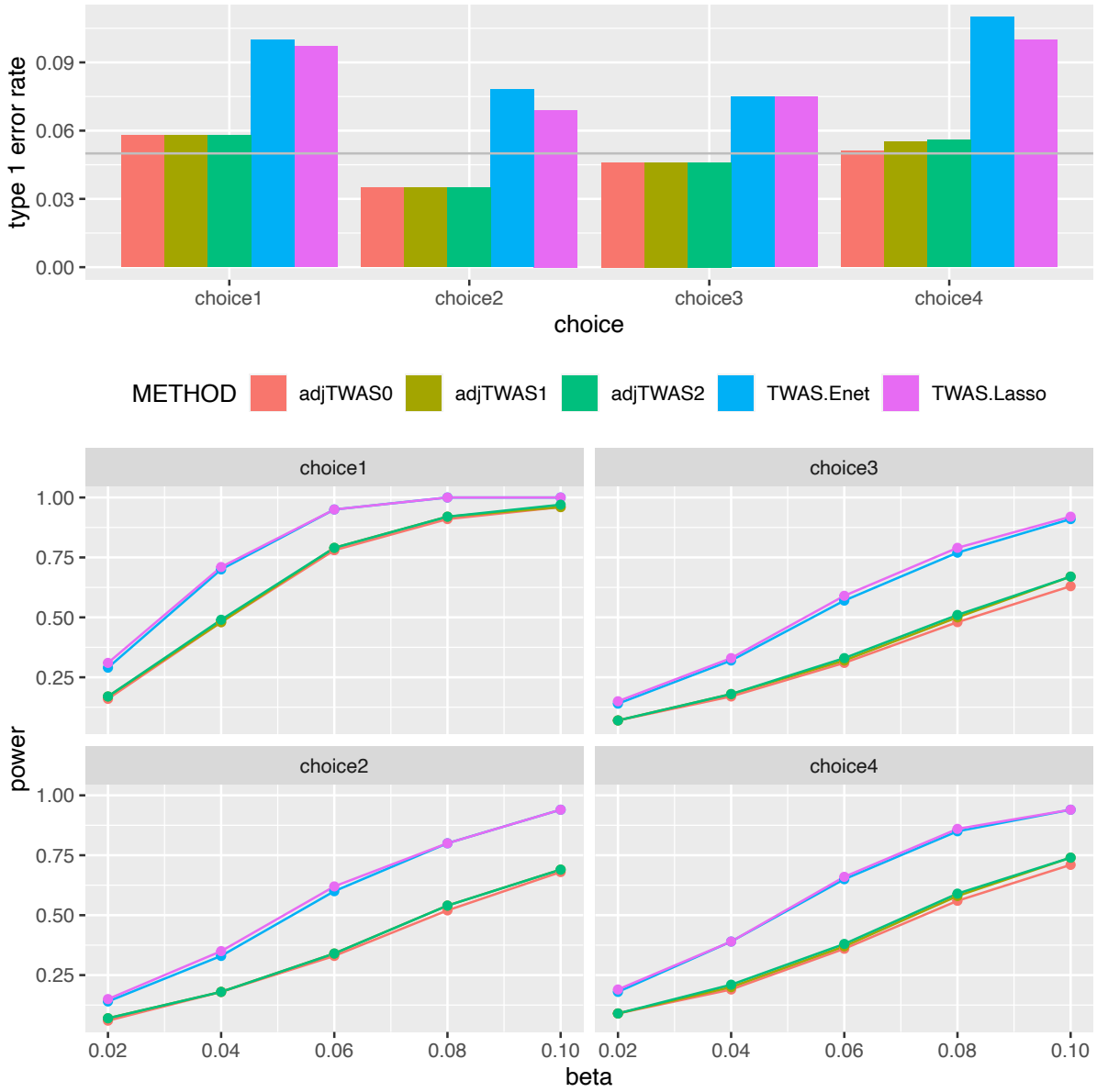

Figure 3: Comparison of type 1 error rate and power between the unadjusted TWAS and adjusted TWAS for the univariate binary phenotype. adjTWAS0, adjTWAS1, adjTWAS2 denote the adjusted TWAS approaches that use $0\%, 5\%, 10\%$ trimmed means while estimating the adjustment factor. TWAS.Lasso and TWAS.Enet denote the unadjusted TWAS using the Elastic net and Lasso penalties in penalized regression.

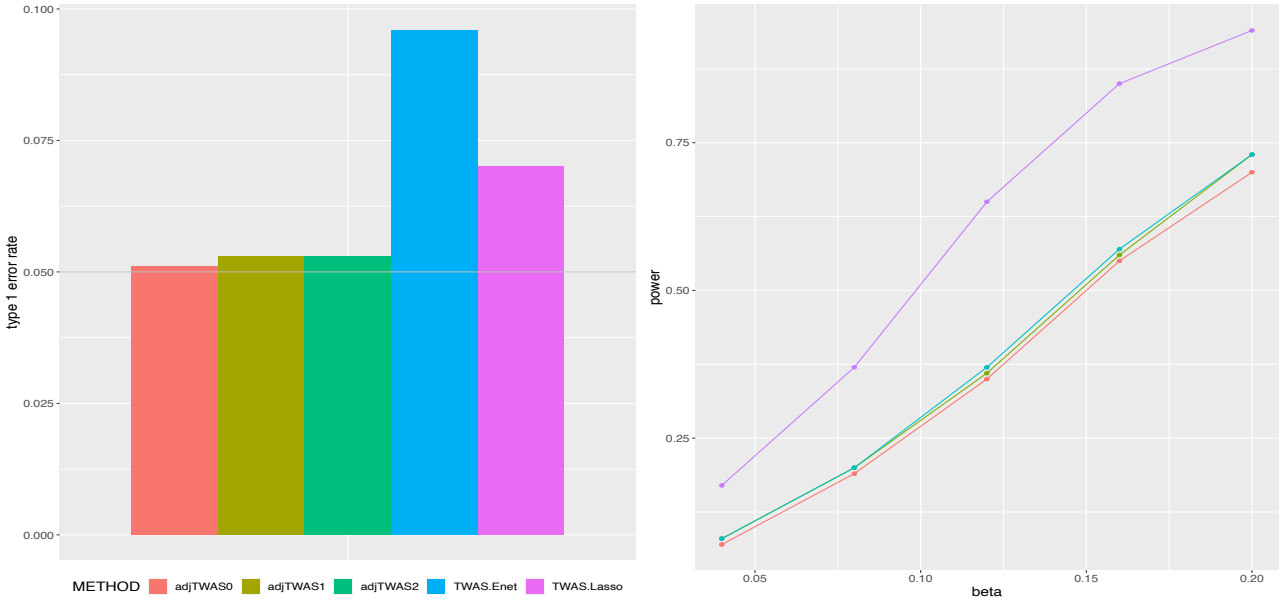

Figure 4: Comparison of type 1 error rate and power produced by ASSET using the unadjusted TWAS and adjusted TWAS statistics for a multivariate phenotype. adjTWAS0, adjTWAS1, adjTWAS2 denote the adjusted TWAS approaches that use $0\%, 5\%, 10\%$ trimmed means when estimating the adjustment factor. TWAS.Lasso and TWAS.Enet denote the unadjusted TWAS using the Elastic net and Lasso penalties in penalized regression.

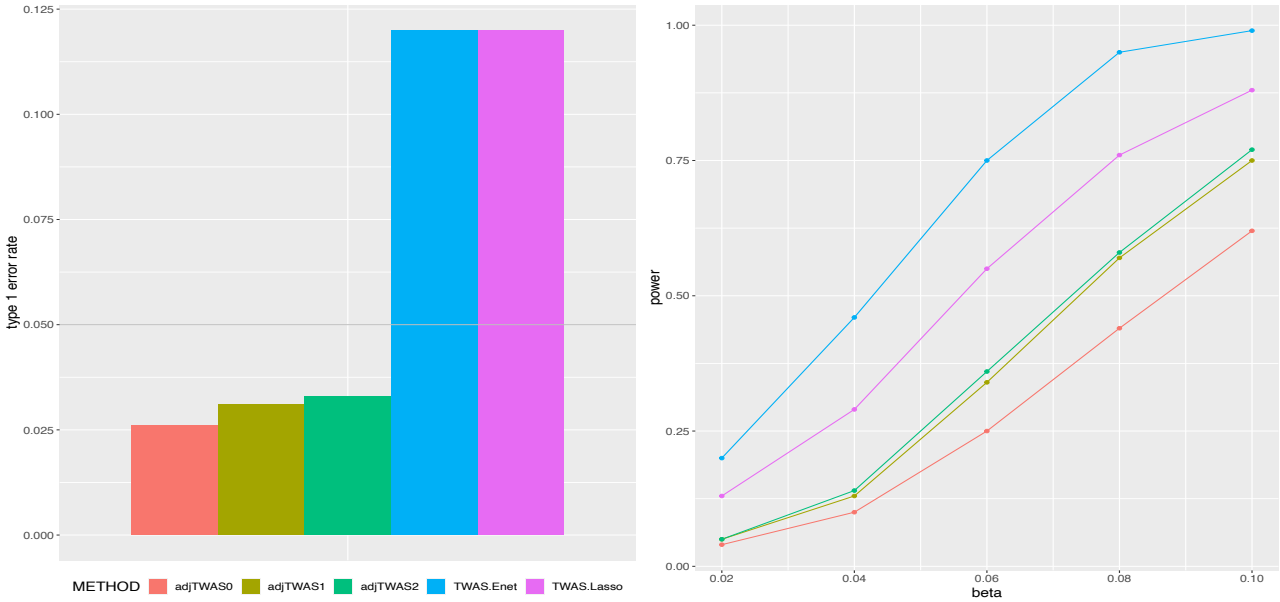

Table 1: Genes associated with height identified by adjusted TWAS. Genes colored blue were discovered by adjusted TWAS, not by unadjusted TWAS.

Chrom	Gene	Chrom	Gene	Chrom	Gene	Chrom	Gene
1	INPP5B	6	SMPD2	12	CDK2AP1	16	MARVELD3
1	FHL3	6	AK9	12	DIP2B	17	GJC1
1	KIF1B	6	TCP11	12	IRAK3	17	SHMT1
1	ECHDC2	7	GNA12	13	EBPL	17	TBKBP1
2	LINC00471	8	RPS20	13	ARL11	17	MRPL45P2
2	TMBIM1	9	LINC00476	13	CYSLTR2	17	LRRC37A
2	CGREF1	9	HABP4	13	LINC00462	17	DDX42
2	VAMP5	9	ERCC6L2	13	CDC16	17	EVI2A
2	CLHC1	10	STOX1	15	NDUFAF1	20	CPNE1
3	ANAPC13	10	WAC-AS1	16	TBX6	20	GSS
3	KLHDC8B	10	SLC25A16	16	SYNGR3	22	PACSIN2
5	SETD9	11	XRRA1				

Table 2: Genes associated with lipid phenotypes identified by ASSET using adjusted TWAS statistics for the multi-phenotype TWAS. Genes colored blue were discovered by adjusted TWAS, not by unadjusted TWAS.

Chrom	Gene	Chrom	Gene	Chrom	Gene	Chrom	Gene
1	CTSS	8	FAM167A	15	NDUFAF1	17	MRPS7
1	GSTM1	10	STOX1	15	C15orf39	17	ALKBH5
1	GSTM4	10	OR13A1	15	SERF2	17	TRIM65
2	TGOLN2	11	MRPL21	15	ELL3	17	EVI2A
2	CGREF1	11	CDHR5	17	GJC1	19	ZNF554
3	NT5DC2	11	C1QTNF4	17	ORMDL3	19	PEPD
3	KLHDC8B	12	RPS26	17	TBKBP1	19	ZNF575
4	DGKQ	12	CDK2AP1	17	MRPL45P2	20	CPNE1
5	ENC1	12	MMAB	17	TMEM97	20	GSS
6	PEX6	12	DIP2B	17	RAPGEFL1	22	MEI1
6	SNHG5						