Reconsidering the validity of transcriptome-wide association studies

Christiaan de Leeuw[1], Josefin Werme[1], Jeanne Savage[1], Wouter Peyrot[1,2], Danielle Posthuma[1,3]*

[1] Department of Complex Trait Genetics, Centre for Neurogenomics and Cognitive Research, VU University, Amsterdam, The Netherlands

[2] Department of Psychiatry, Amsterdam UMC, Vrije Universiteit, Amsterdam, the Netherlands

[3] Department of Child and Adolescent Psychology and Psychiatry, section Complex Trait Genetics, Amsterdam Neuroscience, VU University Medical Centre, Amsterdam, The Netherlands

* Corresponding Author: Christiaan de Leeuw, c.a.de.leeuw@vu.nl

Abstract

Transcriptome-wide association studies (TWAS)[1–5], which aim to detect relationships between gene expression and a phenotype, are commonly used for secondary analysis of genome-wide association study (GWAS) results. Results of TWAS analyses are often interpreted as indicating a genetically mediated relationship between gene expression and the phenotype, but because the traditional TWAS framework does not model the uncertainty in the expression quantitative trait loci (eQTL) effect estimates[6,7], this interpretation is not justified. In this study we outline the implications of this issue. Using simulations, we show severely inflated type 1 error rates for TWAS when evaluating a null hypothesis of no genetic relationship between gene expression and the phenotype. Moreover, in our application to real data only 51% of the TWAS associations were confirmed with local genetic correlation[8] analysis, an approach which correctly evaluates the same null. Our results thus demonstrate that TWAS is unsuitable for investigating genetic relationships between gene expression and a phenotype.

Main text

TWAS is commonly presented as an alternative to differential gene expression analysis[9], to study relationships between gene expression and a phenotype. Since data containing both gene expression levels and the phenotype is often unavailable, TWAS uses a separate sample to estimate genetic associations of SNPs with gene expression (ie. eQTLs). It then imputes the gene expression in a GWAS sample, and tests the association between this imputed expression and the phenotype.

This two-stage procedure is implemented as follows[1,6,7]. First, a model is specified for the expression E of a particular gene

$$E = X\alpha_E + \xi_E, \tag{1}$$

where X denotes the genotype matrix of SNPs local to that gene and ξ_E the residual, while $G_E = X\alpha_E$ reflects the genetic component of its expression captured by those SNPs. This model is fitted to a sample with expression data to obtain an estimated weight vector $\hat{\alpha}_E$.

Second, the imputed genetic component $\hat{G}_E = X\hat{\alpha}_E$ is computed in the GWAS sample for the phenotype of interest Y. Then, a linear regression model of the form

$$Y = \hat{G}_E\beta + \varepsilon_Y, \tag{2}$$

with coefficient β and residual ε_Y, is used to test the relationship between $\hat{G}_E$ and the phenotype. Essentially all TWAS methods have this structure (though often requiring only GWAS summary statistics), but they differ in their implementation, particularly in how $\hat{\alpha}_E$ is estimated[6,7,10–22] (see Table 1). Note that the presentation here is simplified for the sake of brevity, see *Methods - Outline of TWAS framework* for details.

This TWAS framework is generally interpreted as testing the genetically mediated relationship between gene expression levels and the phenotype. We can mathematically quantify this relation as the covariance $\mathrm{cov}(G_E, G_Y)$ of the true genetic components of E and Y, where G_Y is defined analogous to G_E in equation (1), such that

$$Y = X\alpha_Y + \xi_Y = G_Y + \xi_Y, \tag{3}$$

with $G_Y = X\alpha_Y$. Estimate $\hat{G}_E$ is seen as imputing G_E, and since the analysis is entirely based on the SNPs in X, equation (2) must therefore specifically fit the genetic relation between G_E and Y. And

indeed, coefficient β is a direct function of the covariance $\mathrm{cov}(\hat{G}_E, Y)$, and plugging in equation (3) this yields $\mathrm{cov}(\hat{G}_E, Y) = \mathrm{cov}(\hat{G}_E, G_Y) + \mathrm{cov}(\hat{G}_E, \xi_Y) = \mathrm{cov}(\hat{G}_E, G_Y)$ (since ξ_Y is independent of X, and therefore of $\hat{G}_E$). In other words, testing $\beta = 0$ is equivalent to testing $\mathrm{cov}(\hat{G}_E, G_Y) = 0$.

However, whereas the true genetic covariance $\mathrm{cov}(G_E, G_Y)$ is a population-level parameter, the $\mathrm{cov}(\hat{G}_E, G_Y)$ tested by TWAS is a function of the sample-dependent estimate $\hat{G}_E$. TWAS does not model the uncertainty in $\hat{G}_E$, treating it as a fixed quantity. It therefore cannot be interpreted as testing the true genetic covariance $\mathrm{cov}(G_E, G_Y)$, for the following two reasons.

First, $\mathrm{cov}(\hat{G}_E, G_Y)$ is offset from $\mathrm{cov}(G_E, G_Y)$ by an error term Δ, ie. $\mathrm{cov}(\hat{G}_E, G_Y) = \mathrm{cov}(G_E, G_Y) + \Delta$. But because $\hat{G}_E$ is fixed, so is Δ, and therefore under the TWAS null hypothesis of $\mathrm{cov}(\hat{G}_E, G_Y) = 0$, the true genetic covariance $\mathrm{cov}(G_E, G_Y)$ equals $-\Delta$ rather than 0. Second, the failure to model the uncertainty in $\hat{G}_E$ means an important source of sampling variance is ignored, resulting in underestimation of the standard errors. Interpreted as a test of $\mathrm{cov}(G_E, G_Y)$, TWAS would thus test the wrong null value using an underdispersed sampling distribution (Figure 1), resulting in a downward bias in p-values and inflated type 1 error rates (see *Supplemental Information - Mathematical structure of TWAS*).

By way of analogy, this can be likened to using a single sample t-test to compare the means of a variable between two groups, testing the null hypothesis that the true population mean of one group is equal to the sample mean of the other group, rather than testing equality of the two true population means. This disregards that the sample mean is subject to uncertainty, treating the other group as a population of interest rather than merely a sample drawn from one, and therefore cannot be interpreted as a valid test of population-level differences. Similarly, like this sample mean, the $\mathrm{cov}(\hat{G}_E, G_Y)$ tested by TWAS is an inherently sample-dependent quantity, meaning that we cannot draw any population-level conclusions by testing its value.

This raises the question whether there are any valid and informative interpretation of significant TWAS results. And indeed there are considerable limits on which biologically relevant conclusions can be drawn from TWAS, because the estimate $\hat{G}_E$ merely represents a weighted sum of the SNPs in X. If we substitute $\hat{G}_E = X\hat{\alpha}_E$ in equation (2) we obtain

$$Y = X\hat{\alpha}_E\beta + \varepsilon_Y, \qquad (4)$$

which shows that TWAS reduces to a constrained version of the general multiple regression model shown in (3), with coefficient vector α_Y proportional to weight vector $\hat{\alpha}_E$, ie. $\alpha_Y = \hat{\alpha}_E\beta$. The TWAS null hypothesis $\beta = 0$ implies the multiple regression null hypothesis $\alpha_Y = \vec{0}$, and vice versa. Thus, like multiple regression, TWAS ultimately only provides a test of whether the SNPs in X are jointly

associated with Y, and therefore does not warrant any conclusions about a role of gene expression in those genetic associations (see *Supplemental Information - Relation to joint association testing*).

The modelling of uncertainty in eQTL estimates has occasionally been mentioned in TWAS literature[1,14,23,24], but implications for the validity of the TWAS framework have received little scrutiny. Although the CoMM[14] method does explicitly model this uncertainty, due to the model structure it still suffers from very similar issues as other TWAS methods (see *Supplemental Information - Comparison with CoMM*). Note that other methods like colocalization[25,26] and Mendelian Randomization[27–29] may address this issue within the context of their own respective frameworks, but as these have different aims and make different assumptions their evaluation is beyond the scope of this paper.

To evaluate the severity of the issues that arise when interpreting TWAS as testing genetically mediated relationships between gene expression and phenotype, we performed extensive simulations and applied TWAS to real data. To serve as a reference, we used the local genetic correlation analysis in LAVA[8], which directly tests the true genetic covariance $\mathrm{cov}(G_E, G_Y)$. To simplify comparison, we implemented TWAS analysis inside the LAVA framework, using the same preprocessing and test statistic as for the local genetic correlation analysis, ensuring that the only difference between the two analyses is the null model being evaluated (see *Methods - LAVA implementation of TWAS*).

Gene expression and phenotype values were simulated under a null model of no genetic covariance ($\mathrm{cov}(G_E, G_Y) = 0$), separately varying the levels of local genetic signal for each (see *Methods - Primary simulations*). In these simulations we found strongly inflated type 1 error rates for TWAS (Figure 2), with the inflation decreasing with greater eQTL signal strength or sample size, but increasing with the phenotype's genetic signal strength. The inflation also gets progressively worse at lower significance thresholds (Figure 3). Running two other TWAS implementations, FUSION[7] and CoMM[14], through these simulation, we observed largely the same pattern of results (Supplemental Figures 1 and 2). Error rates for local genetic correlation were well-controlled (Supplemental Figure 1). See *Supplemental Information - Simulation results* for further discussion.

To gauge the impact of this issue for real data, we applied both TWAS and local genetic correlation analysis to GWAS of five well-powered phenotypes[30–33] (see *Methods - Data* and *Methods - Real data analysis*), with eQTL data for 49 different tissues from GTEx[34] (v8). Only 51% of significant TWAS associations were confirmed by the local genetic correlation analyses (Table 2), showing that when used to detect genetic relations with gene expression, TWAS yields a very high rate of uncertain and likely spurious associations.

Additional empirical simulations were performed to gain insight into the type 1 error rates of TWAS in a real data context. For each phenotype, individual null simulation were run for each previously analysed gene-tissue pair, using the same SNPs and levels of genetic association observed in the real data (see *Methods - Empirical error rate simulations*). These simulations show that the type

1 error rate inflation is highly variable across genes and tissues (Table 3). Consistent with the earlier simulations, the level of inflation decreases as the eQTL signal gets stronger, but increases with stronger genetic associations for the phenotype (Supplemental Figure 3). See *Supplemental Information - Simulation results* for additional discussion.

Finally, to evaluate the impact of the eQTL-specific weighting in TWAS, we ran additional analyses on the real data for genes and tissues where no genetic association with gene expression was present (see *Methods - Real data analysis*). Despite the resulting $\hat{\alpha}_E$ thus essentially being random noise, these analyses still yielded large numbers of significant associations (Supplemental Table 1), further illustrating that significance in a TWAS analysis is not inherently related to eQTL-related information contained in the weights.

As we have shown, TWAS is unsuitable for testing genetic relations between gene expression and phenotypes, due to its failure to account for the uncertainty in the estimated gene expression, yielding severely inflated false positive rates when used for this purpose. Because of this, and since the null hypothesis evaluated by TWAS depends on a sample-specific quantity, the extent to which informative conclusions can be drawn from significant TWAS results are very limited. Investigating genetic expression-phenotype relationships thus requires more robust methods that can account for all the uncertainty in the data and can provide meaningful effect size estimates, such as local genetic correlation analysis methods like LAVA.

Methods

Outline of TWAS framework

The general TWAS framework consists of a two-stage procedure, based on the equations

$$E = X\alpha_E + \xi_E, \tag{1}$$

and

$$Y = \hat{G}_E\beta + \varepsilon_Y \tag{2}$$

also given in the main text. For ease of notation, we have omitted model intercepts and covariates from these equations, but in practice these will usually be included. An alternative model may also be used rather than the linear regression in equation (2), such as a logistic regression if the phenotype is dichotomous.

For each gene and tissue, equation (1) is first fitted to the eQTL data to obtain the estimated weight vector $\hat{\alpha}_E$. This is then used to compute $\hat{G}_E$ in the target GWAS sample in the second stage, and plugged into the linear model in equation (2). A p-value is then obtained by performing a test on the coefficient β. Note that usually the second stage is only performed for genes and tissues that exhibit sufficient genetic association in the eQTL data. This second stage can also be rewritten in terms of GWAS summary statistics, allowing TWAS to be performed without having direct access to the GWAS sample (see also *Supplemental Information - Mathematical structure of TWAS*). In this case a genotype matrix X obtained from a separate reference sample is used to estimate LD.

Which SNPs are included in X varies, but a common choice is to use all available SNPs within one megabase of the transcription region of the gene. Although for simplicity the same genotype matrix X is used in equation (1) and (2), there will be separate X genotype matrices for each sample. The analysis is therefore restricted to using only those SNPs that are available in both samples, as well as in the LD reference sample when using summary statistics as input.

In practice equation (1) cannot be fitted with a traditional multiple linear regression model, due to the high LD between SNPs (leading to extreme collinearity), and the number of SNPs typically exceeding the sample sizes of eQTL data. Some form of regularization in the regression model is therefore required to obtain $\hat{\alpha}_E$, and consequently one of the main discrepancies between TWAS implementations is the specific regularization used (see Table 1). In some cases, rather than fitting equation (1), the elements of $\hat{\alpha}_E$ are simply set to the marginal SNP effect estimates instead.

Note that some methods[20–22] diverge from this linear model structure (Table 1, *non-linear models*). Statistically these can be seen as generalizations of the TWAS framework, though conceptually they can no longer be interpreted as imputing the genetic component of gene expression. See *Supplemental Information - Non-linear TWAS models* for more details.

Local genetic correlation

The LAVA implementation of local genetic correlation analysis has been described in detail in Werme et al. (2021)[8]. In brief, LAVA uses summary statistics and a reference genotype sample to fit equations (1) and (3), obtaining estimates of $\hat{\alpha}_E$ and $\hat{\alpha}_Y$ as well as a corresponding sampling covariance matrix for each (a logistic regression equivalent is used for binary phenotypes). To do so, a singular value decomposition for X is computed, pruning away excess principal components to attain regularization of the models and allowing them to be fitted. To accommodate the small sample sizes in the eQTL data, the pruning procedure from the original LAVA approach was adapted by capping the maximum number of principal components to be retained at 75% of the eQTL sample size for each tissue.

With the pruned and standardized principal component matrix $W = XR$ (with R the transformation matrix projecting the genotypes onto the principal components), we can write $G_E = W\gamma_E$ and $G_Y = W\gamma_Y$, where γ_E and γ_Y are the genetic effect size vectors for these principal components. Their estimates $\hat{\gamma}_E$ and $\hat{\gamma}_Y$ can be used to obtain $\hat{\alpha}_E$ and $\hat{\alpha}_Y$ by reversing the transformation through R, such that $\hat{\alpha}_E = R\hat{\gamma}_E$ and $\hat{\alpha}_Y = R\hat{\gamma}_Y$, with the projection to the principal components effectively providing a form of regularization in the estimation of α_E and α_Y. In practice however, LAVA is defined and implemented directly in terms of γ_E and γ_Y and its estimates, rather than working with α_E and α_Y explicitly.

For ease of notation, we define the combined matrix $G = (G_E, G_Y) = W\gamma$ for the genetic components, with combined effect size matrix $\gamma = (\gamma_E, \gamma_Y)$. We denote the effect sizes for a single principal component j as γ_j, corrresponding to the jth row of γ, and denote the number of principal components as K.

The estimates $\hat{\gamma}_E$ and $\hat{\gamma}_Y$ are obtained by reconstructing multiple linear regressions from the input summary statistics (see Werme et al. (2021) for details). This uses two separate equations of the form $E = W\gamma_E + \zeta_E$ and $Y = W\gamma_Y + \zeta_Y$, analogous to equations (1) and (3) but regressing on W rather than X, with residual variances η_E^2 and η_Y^2 for ζ_E and ζ_Y. From these models we have estimates of the form $\hat{\gamma}_E = (W^TW)^{-1}W^TE = \frac{W^TE}{N-1}$ (since $W^TW = I_K(N-1)$, with I_K the size K identity matrix) and similarly $\hat{\gamma}_Y = \frac{W^TY}{N-1}$, with corresponding sampling distributions $\hat{\gamma}_E \sim \mathrm{MVN}(\gamma_E, \sigma_E^2 I)$ and

$\hat{\gamma}_Y \sim \mathrm{MVN}(\gamma_Y, \sigma_Y^2 I)$, where $\sigma_E^2 = \frac{\eta_E^2}{N-1}$ and $\sigma_Y^2 = \frac{\eta_Y^2}{N-1}$ are the sampling variances (ie. squared standard errors).

For principal component j we therefore have $\hat{\gamma}_j \sim \mathrm{MVN}(\gamma_j, \Sigma)$, where the diagonal elements of Σ are σ_E^2 and σ_Y^2 and the off-diagonal elements are 0 (in the general case the off-diagonal elements represent the sampling covariance resulting from sample overlap, but this is not present in the analyses in this study). Since for the covariance matrix of G we have $\mathrm{cov}(G) = \frac{G^T G}{N-1} = \frac{\gamma^T W^T W \gamma}{N-1} = \gamma^T \gamma$, it follows that inference on $\mathrm{cov}(G)$ can be performed using the sampling distributions for $\hat{\gamma}_E$ and $\hat{\gamma}_Y$ directly.

Using this model, separate univariate tests of joint association of the SNPs in X with E and Y can be performed, testing the null hypotheses $\gamma_E = \vec{0}$ and $\gamma_Y = \vec{0}$ respectively (using standard linear regression F-test for continuous phenotypes (such as gene expression), or a χ^2 test for binary phenotypes). This is equivalent to testing the local genetic variances, a prerequisite for the analysis since genetic covariance can only exist in a genomic region where there both phenotypes exhibit some degree of genetic variance.

From the above distributions it follows that the expected value $E[\hat{\gamma}^T \hat{\gamma}] = \gamma^T \gamma + K\Sigma$, and we can therefore use the method of moments to estimate $\mathrm{cov}(G)$ as $\widehat{\mathrm{cov}}(G) = \hat{\gamma}^T \hat{\gamma} - K\Sigma$. Since in the present analyses there is assumed to be no sample overlap, the off-diagonal elements of Σ are 0, and for the estimate for the genetic covariance therefore reduces to $\widehat{\mathrm{cov}}(G_E, G_Y) = \hat{\gamma}_E^T \hat{\gamma}_Y$. The matrix $\hat{\gamma}^T \hat{\gamma}$ has a non-central Wishart sampling distribution, which is used to obtain p-values to test $\mathrm{cov}(G_E, G_Y) = 0$ using a simulation procedure (see Werme et al. (2021) for details).

LAVA implementation of TWAS

To construct a TWAS model within the LAVA framework, we note that in a linear regression for equation (2) we have $\hat{\beta} = \frac{\widehat{\mathrm{cov}}(\hat{G}_E, Y)}{\widehat{\mathrm{var}}(\hat{G}_E)}$. Since $\widehat{\mathrm{var}}(\hat{G}_E)$ is considered fixed in TWAS the sampling distribution of $\hat{\beta}$ directly proportional to the distribution of $\widehat{\mathrm{cov}}(\hat{G}_E, Y)$ (which is the sample estimate of $\mathrm{cov}(\hat{G}_E, Y)$). As both $\hat{G}_E$ and Y have means of zero, and since $\hat{G}_E = W\hat{\gamma}_E$, we have $\widehat{\mathrm{cov}}(\hat{G}_E, Y) = \frac{\hat{G}_E^T Y}{N-1} = \frac{\hat{\gamma}_E^T W^T Y}{N-1} = \hat{\gamma}_E^T \frac{W^T Y}{N-1}$. As previously derived $\hat{\gamma}_Y = \frac{W^T Y}{N-1}$, and it therefore follows that $\widehat{\mathrm{cov}}(\hat{G}_E, Y) = \hat{\gamma}_E^T \hat{\gamma}_Y$, the distribution of which depends entirely on the sampling distribution of $\hat{\gamma}_Y$ since $\hat{\gamma}_E$ is considered fixed. The distribution of $\widehat{\mathrm{cov}}(\hat{G}_E, Y)$ thus has the form $\mathrm{N}(\hat{\gamma}_E^T \gamma_Y, \hat{\gamma}_E^T \hat{\gamma}_E \sigma_Y^2)$. We note that analogous to its estimate, $\mathrm{cov}(\hat{G}_E, Y) = \hat{\gamma}_E^T \gamma_Y$, and we therefore see that the distribution of the estimate $\widehat{\mathrm{cov}}(\hat{G}_E, Y)$ centers on the parameter $\mathrm{cov}(\hat{G}_E, Y)$ it is intended to estimate.

What this shows is that we can use the same test statistic $\hat{\gamma}_E^T \hat{\gamma}_Y$ that is used in the local genetic correlation analysis to perform testing for the TWAS analysis as well, with the only difference being the sampling distribution against which this $\hat{\gamma}_E^T \hat{\gamma}_Y$ is compared to obtain the p-value. For the TWAS analysis, under the TWAS null hypothesis of $\mathrm{cov}\left(\hat{G}_E, G_Y\right) = 0$, this is a normal distribution with mean of 0 and a variance of $\hat{\gamma}_E^T \hat{\gamma}_E \sigma_Y^2$, which is used to compute the p-value. Note that this $\hat{\gamma}_E^T \hat{\gamma}_Y$ equals a weighted sum $\sum_j \hat{\gamma}_{Ej} \hat{\gamma}_{Yj}$ of the estimated genetic associations with $\hat{\gamma}_{Yj}$. This implementation is therefore essentially equivalent to how TWAS is performed using GWAS summary statistics in other TWAS methods (eg. Gusev (2016)[7]), except defined in terms of the estimated genetic associations of principal component matrix W rather than the original SNP genotype matrix X.

Data

The European panel of the 1,000 Genomes[35] data (N = 503, as downloaded from https://ctg.cncr.nl/software/magma) was used as genotype reference data to estimate LD. For eQTL data we used the published cis-eQTL summary statistics from GTEx[34] (v8, European subset), for 49 different tissues. For every analysed gene, this covers all SNPs in the data within one megabase of the transcription start site. Genes were filtered to include only autosomal protein-coding and RNA genes, for a total 24,836 different genes across all tissues (note that not all genes were available for all tissues).

GWAS summary statistics were selected for five well-powered phenotypes, chosen to reflect a range of different domains. These were BMI (GIANT)[30] (no waist-hip ratio adjustment), educational attainment (SSGAC)[31], schizophrenia (PGC, wave 3)[32], diastolic blood pressure (GWAS Atlas)[33] and type 2 diabetes (GWAS Atlas)[33]. Sample size and number of SNPs for each sample can be found in Table 2.

Primary simulations

Genotype data from the 1,000 Genomes data was used to perform the primary simulations, selecting ten blocks of 5,000 consecutive SNPs, each from a different chromosome. The sample size was scaled up by a factor 20 to obtain a sample size of 10,060 for use in the simulations. Per block, genotype data was projected onto standardized principal components W, pruning away redundant components based on the cumulative genotypic variance explained by the components (retaining those that jointly explain 99% of the total variance). The local heritability of the gene expression and the phenotype were both independently varied, at values of 1%, 2%, 5% and 10% for a total of 16 conditions. Each

condition was repeated for 10,000 iterations per block, and type 1 error rates were computed per condition across the blocks, ie. 100,000 iterations per condition in total.

For each iteration, true genetic effect sizes γ_E and γ_Y for the principal components under the null hypothesis of $\mathrm{cov}(G_E, G_Y) = 0$ were generated by drawing values from a normal distribution from a normal distribution for each, then regressing one vector on the other and retaining only the residuals for the outcome vector to ensure that γ_E and γ_Y were exactly independent. Simulated gene expression and phenotype values were then generated as $E = W\gamma_E + \xi_E$ and $Y = W\gamma_Y + \xi_Y$, drawing the residuals ξ_E and ξ_Y from normal distributions with variance parameters such that the expected explained variance equalled the desired local heritability for that condition.

Effect size estimates $\hat{\gamma}_E$ and $\hat{\gamma}_Y$, as well as estimates of the residual variance parameters, were then obtained by multiple regression of E and Y on W. These were analysed with both the LAVA local genetic correlation model as well as the TWAS model implemented in the LAVA framework to obtain p-values. To evaluate the impact of a smaller sample size for the eQTL data, a second estimate $\hat{\gamma}_E^{(1K)}$ of the eQTL effects using only 10% of the full simulation sample was obtained, and also analyzed using the TWAS model.

To validate the implementation of the TWAS model, for each iteration an additional TWAS analysis was performed under the actual TWAS null model of $\mathrm{cov}(\hat{G}_E, G_Y) = 0$. This was accomplished by generating a new vector γ_Y that was exactly independent of the $\hat{\gamma}_E$ already estimated for the iteration, then generating a new Y and estimating $\hat{\gamma}_Y^{(TWAS)}$ as before and analysing this together with $\hat{\gamma}_E$ using the TWAS model.

Simulations were also performed for FUSION[7] and CoMM[14], using the same procedure but running only 1,000 iterations per condition (100 per block). For the eQTL weight estimation step in FUSION, heritability values were set to their true value for the condition rather than estimating them from the data, and both the elastic net and LASSO models were used. For CoMM, due to computational constraints only 1,000 SNPs could be used in the simulations. As an additional reference, the TWAS simulations using the implementation in LAVA were therefore repeated with only 1,000 SNPs as well.

Real data analysis

TWAS and local genetic correlation analyses were performed on the GWAS data as follows, separately for each of the five phenotypes. In all the analyses, SNP filtering was applied to remove all SNPs with a minor allele frequency below 0.5%, and only SNPs available in the 1,000 Genomes data, the GTEx data, and the GWAS sample for that phenotype were used.

For every gene-tissue pair, univariate analysis was first performed for that gene for both the gene expression and the phenotype, to determine the level of genetic association for each. Note that univariate p-values were only computed if the estimated genetic variances (that is, the diagonal elements of $\widehat{\mathrm{cov}}(G_E, G_Y)$) were both positive. Bivariate analyses were then performed if both univariate p-values were below 0.05/24,836 (ie. Bonferroni-corrected for the number of genes).

The significance threshold for the bivariate analyses was set separately for each phenotype, at a Bonferroni-correction for the total number of gene-tissue pairs for which bivariate analysis was performed for that phenotype (see Table 2). As the aim was to represent the level of TWAS and local genetic correlation result for a full analysis of a single phenotype, no further correction was applied across the phenotypes.

In a secondary analysis, to evaluate inflation in the absence of gene expression signal. TWAS analysis was also performed for all gene-tissue pairs for which the univariate p-value of the gene expression was greater than 0.05. Filtering on the univariate p-value for the phenotype was maintained at the same 0.05/24,836 level.

Empirical error rate simulations

In addition to the primary simulations, we ran additional empirical simulations to estimate type 1 error rates per phenotype and per gene-tissue pair, at genetic association levels observed in the real data. This procedure is conceptually equivalent to the primary simulations as described above, but was optimized for computational feasibility as follows.

As noted above in the section *Local genetic correlation*, the matrix $\hat{\gamma}^T\hat{\gamma}$ on which the TWAS test statistic is based has a non-central Wishart sampling distribution when accounting for the uncertainty in the eQTL estimates, with its parameters dependent on $\mathrm{cov}(G)$, Σ and K. Under the null hypothesis of $\mathrm{cov}(G_E, G_Y) = 0$ the off-diagonal elements of $\mathrm{cov}(G)$ are 0 (as are those of Σ, as there is no sample overlap in the present analyses), and the diagonal elements of both $\mathrm{cov}(G)$ and Σ are set to their corresponding estimates obtained from the local genetic correlation analysis output in the real data application.

These diagonal elements of $\mathrm{cov}(G)$ and Σ represent the genetic variance and residual variance, which together determine the relative level of detectable genetic association (separately for gene expression and the outcome phenotype), filling the role of the local heritability parameter in the primary simulations. As such, simulating based on these values provides type 1 error estimates at realistic levels of genetic association.

The type 1 error rates are computed by generating 20 million draws of $\hat{\gamma}^T\hat{\gamma}$ for each gene-tissue pair for each phenotype, then computing the corresponding TWAS p-values for each draw. Type

1 error rates for that gene-tissue pair are then computed as the proportion of iterations with p-value below the Bonferroni-corrected threshold used in the bivariate analyses for that phenotype (see Table 2).

By obtaining draws of $\hat{\gamma}^T\hat{\gamma}$ directly from this non-central Wishart distribution with parameters as specified above, the need to explicitly generate and analyse simulated E and Y is removed, considerably reducing the computational burden. Moreover, the simulation process was set up to allow the random draws to be shared across different gene-tissue pairs with different values for $\mathrm{cov}(G)$ and Σ but the same K, making it feasible to obtain separate simulations for each individual gene-tissue pair.

Acknowledgements

This work was funded by The Netherlands Organization for Scientific Research (Grant No. NWO VICI 435–14–005 [to DP]) and NWO Gravitation: BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (Grant No. 024.004.012 [to DP]), and a European Research Council advanced grant (Grant No, ERC-2018-AdG GWAS2FUNC 834057 [to DP]). CdL was funded by F. Hoffman-La Roche AG. WJP was funded by an NWO Veni grant (NWO: 916-19-152). The analyses were carried out on the Genetic Cluster Computer, which is financed by the Netherlands Organization for Scientific Research (NWO: 480-05-003), by the VU University (Amsterdam, The Netherlands) and the Dutch Brain Foundation, hosted by the Dutch National Computing and Networking Services SurfSARA.

Author contributions

CdL and JW conceived of the study. CdL performed the analyses, simulations, and wrote the manuscript. All authors participated in the interpretation of the results and revision of the manuscript, and provided meaningful contributions at each stage of the project.

Competing interests

The authors declare no competing financial interests.

References

1. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
2. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, (2020).
3. Zhao, B. *et al.* Transcriptome-wide association analysis of brain structures yields insights into pleiotropy with complex neuropsychiatric traits. *Nat. Commun.* **12**, 2878 (2021).
4. Strunz, T., Lauwen, S., Kiel, C., Hollander, A. den & Weber, B. H. F. A transcriptome-wide association study based on 27 tissues identifies 106 genes potentially relevant for disease pathology in age-related macular degeneration. *Sci. Rep.* **10**, 1584 (2020).
5. Derks, E. M. & Gamazon, E. R. Transcriptome-wide association analysis offers novel opportunities for clinical translation of genetic discoveries on mental disorders. *World Psychiatry* **19**, 113–114 (2020).
6. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
7. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
8. Werme, J., Sluis, S. van der, Posthuma, D. & Leeuw, C. A. de. LAVA: An integrated framework for local genetic correlation analysis. *bioRxiv* 2020.12.31.424652 (2021) doi:10.1101/2020.12.31.424652.
9. McDermaid, A., Monier, B., Zhao, J., Liu, B. & Ma, Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. *Brief. Bioinform.* **20**, 2044–2054 (2019).
10. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).
11. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
12. Su, Y.-R. *et al.* A Mixed-Effects Model for Powerful Association Tests in Integrative Functional Genomics. *Am. J. Hum. Genet.* **102**, 904–919 (2018).
13. Hu, Y. *et al.* A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* **51**, 568–576 (2019).
14. Yang, C. *et al.* CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics* **35**, 1644–1652 (2019).
15. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. 22 (2019).

16. Nagpal, S. *et al.* TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *Am. J. Hum. Genet.* **105**, 258–266 (2019).
17. Liu, W. *et al.* Leveraging functional annotation to identify genes associated with complex diseases. *PLOS Comput. Biol.* **16**, e1008315 (2020).
18. Luningham, J. M. *et al.* Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *Am. J. Hum. Genet.* **107**, 714–726 (2020).
19. Bhattacharya, A., Li, Y. & Love, M. I. MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. *PLOS Genet.* **17**, e1009398 (2021).
20. Xu, Z., Wu, C., Wei, P. & Pan, W. A Powerful Framework for Integrating eQTL and GWAS Summary Data. *Genetics* **207**, 893–902 (2017).
21. Zhang, J., Xie, S., Gonzales, S., Liu, J. & Wang, X. A fast and powerful eQTL weighted method to detect genes associated with complex trait using GWAS summary data. *Genet. Epidemiol.* **44**, 550–563 (2020).
22. Tang, S. *et al.* Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. *PLOS Genet.* **17**, e1009482 (2021).
23. Xue, H. & Pan, W. Some statistical consideration in transcriptome-wide association studies. *Genet. Epidemiol.* **44**, 221–232 (2020).
24. Zhu, H. & Zhou, X. Transcriptome-wide association studies: a view from Mendelian randomization. *Quant. Biol.* (2020) doi:10.1007/s40484-020-0207-4.
25. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).
26. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
27. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
28. Yuan, Z. *et al.* Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat. Commun.* **11**, 3861 (2020).
29. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 3300 (2019).
30. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).
31. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

32. The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, S., Walters, J. T. & O'Donovan, M. C. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv* 2020.09.12.20192922 (2020) doi:10.1101/2020.09.12.20192922.
33. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
34. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
35. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

Table 1. Overview of available TWAS analysis methods.

Method	Weight estimation[a]	Base model extensions
Linear models		
Gamazon (2015)[6] - PrediXcan	Marginal LASSO Elastic net	-
Gusev (2016)[7] - FUSION[b]	Top eQTL BLUP Bayesian LMM LASSO Elastic net	-
Mancuso (2017)[10] - RhoGE	BLUP	-
Barbeira (2018)[11] - MetaXcan	Marginal LASSO Elastic net	-
Su (2018)[12] - MiST	External	Models additional variance component for genetic effects not mediated by predicted expression
Hu (2019)[13] - UTMOST	Multivariate LASSO	Simultaneously models multiple tissues during weight estimation
Yang (2019)[14] - CoMM	Collaborative mixed model	Estimates weights and associations with phenotype simultaneously in single model
Mancuso (2019)[15] - FOCUS	External	Models multiple genes at once, as well as additional pleiotropic genetic effects on phenotype
Nagpal (2019)[16] - TIGAR	Dirichlet process regression	Multivariate model with multiple outcome phenotypes
Liu (2020)[17] - T-GEN	Spike & Slab	Incorporates epigenetic information into weight estimation process
Luningham (2020)[18] - BGW-TWAS	Spike & Slab	Models additional trans-eQTL component
Bhattacharya (2021)[19] - MOSTWAS	Elastic net BLUP	Models additional components for trans-eQTL or other molecular phenotypes
Non-linear models		
Xu (2017)[20] - ASPU	External	Uses adaptive test combining sums of powers of score statistics for different powers (includes linear model)
Zhang (2020)[21]	External	Uses adaptive test combining linear model with sum of squared score statistics
Tang (2021)[22] - VC-TWAS	External	Uses sum of powers of score statistics instead of linear model

LASSO: least absolute shrinkage and selection operator; BLUP: best linear unbiased predictor; LMM: linear mixed model

[a] Multiple entries for a method denote different options; 'marginal' refers to marginal SNP effect sizes being used as weights, 'external' means the method requires precomputed weights from an external source

[b] The name 'FUSION' and the LASSO and elastic net options for this method were added after publication of the Gusev (2016) paper

Table 2. Summary of results of TWAS and local genetic correlation analyses of five phenotypes.

Phenotype	Sample size[a]	Number of SNPs[b]	Number of tests	Significance threshold	Significant associations			Significant genes[c]		
					TWAS	LAVA r_G	Both	TWAS	LAVA r_G	Both
BMI[30]	807K	6.28M	84,567	5.91×10^{-7}	2,227	1,098	1,094	1,400	695	693
Blood pressure[33]	361K	5.94M	54,622	9.15×10^{-7}	760	437	436	533	293	292
Diabetes[33]	18.5K/366K	5.94M	18,967	2.63×10^{-6}	320	144	142	209	114	112
Educational attainment[31]	766K	6.18M	45,160	1.11×10^{-6}	846	499	499	514	292	292
Schizophrenia[32]	67.4K/94.0K	6.08M	61,137	8.18×10^{-7}	655	302	301	472	228	228
Total					4,808	2,480	2,472	3,128	1,622	1,617
% of TWAS						51.6%	51.4%		51.8%	51.7%

Results were Bonferroni corrected per phenotype for the number gene-tissue pairs for which both the gene expression as well as the phenotype showed significant univariate genetic association at $p < 0.05/24{,}836$ (see *Methods - Real data analysis*). r_G denotes the local genetic correlation.

[a] Showing case/control for binary phenotypes

[b] After filtering for overlap with 1,000 Genomes and GTEx SNPs

[c] Genes that showed significant association in at least one tissue

Table 3. Summary of type 1 error rate inflation estimates from empirical simulations for each individual gene-tissue pair, at Bonferroni-corrected significance threshold.

			Quantiles				
Phenotype	**Mean**	**Maximum**	**5%**	**25%**	**Median**	**75%**	**95%**
BMI	60.2	45,898	1.44	4.23	10.5	29.6	196.5
Blood pressure	12.5	1,317	1.26	2.95	6.23	12.5	41.0
Diabetes	41.0	6,676	1.19	2.54	4.89	8.55	59.8
Educational attainment	16.1	2,328	1.26	2.94	6.10	11.6	41.2
Schizophrenia	20.3	6,263	1.28	3.18	6.73	13.9	56.6

Type 1 error rate inflation is defined as the estimated error rate divided by the significance threshold, computed at the Bonferroni-corrected significance thresholds listed in Table 2.

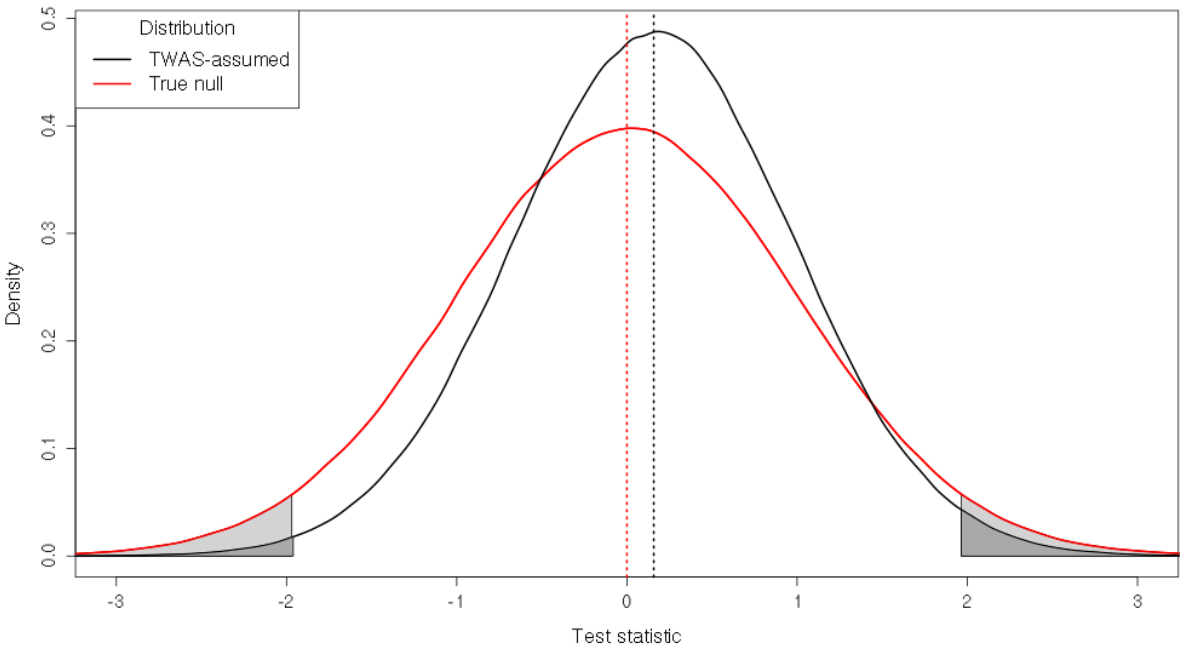

Figure 1. Illustration of distributions under the null hypothesis $\boldsymbol{H_0: cov(G_E, G_Y) = 0}$***.*** *Shown is an example of the distributions of the test statistic (for the LAVA TWAS implementation), scaled such that the true null distribution has a variance of one. The true null distribution (red) is the true distribution of the test statistic under this* H_0*, accounting for the uncertainty in eQTL estimates. The TWAS-assumed distribution is the sampling distribution that the TWAS model compares the same test statistic against to compute its p-value. As shown, the TWAS-assumed distribution has a smaller variance than the null distribution, resulting from the fact that it does not account for the uncertainty in* $\hat{G}_E$*. Unlike the true null distribution it also does not center on 0, reflecting the fact that under the TWAS-assumed distribution* $cov(G_E, G_Y)$ *equals the error term* $-\Delta$ *rather than 0 (see* Supplemental Information - Mathematical structure of TWAS*). The direction and degree to which this distribution is shifted away from 0 depends on the data, and will vary across genes and tissues. The areas corresponding to the p-value for a test statistic value of 1.96 have been shaded in, which gives a p-value of 0.05 for under the true null distribution but a p-value of 0.016 under the TWAS-assumed null. This shows that for the same observed value of the test statistic, the p-value computed by the TWAS model will be too low.*

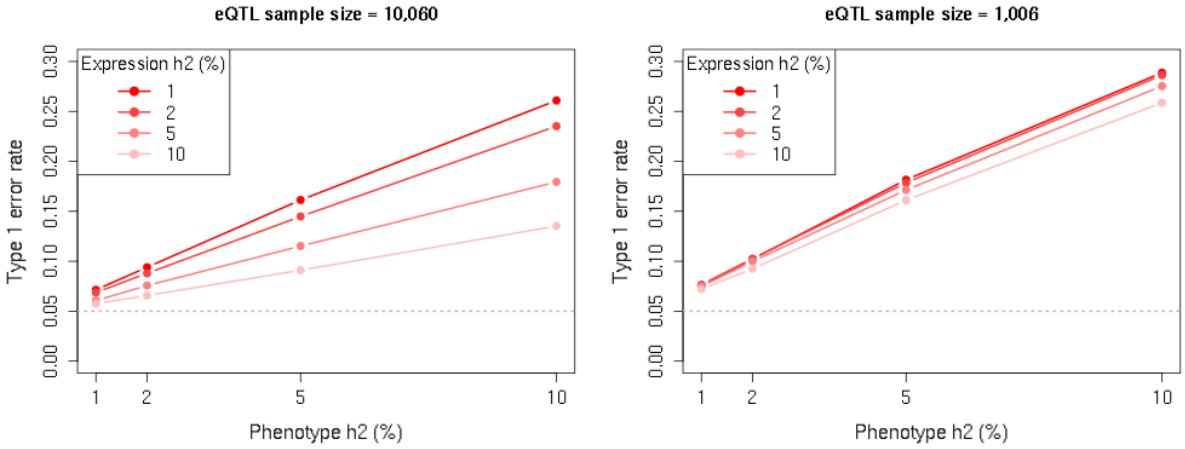

Figure 2. Results from primary simulations. *Shown is the type 1 error rate (at significance threshold of 0.05) of the TWAS model relative to the null hypothesis of no genetic covariance ($cov(G_E, G_Y) = 0$), at different levels of local heritability for outcome phenotype (horizontal axis) and gene expression (separate lines). Simulation sample size is 10,060 for the outcome phenotype, and either 10,060 (left) or 1,006 (right) for the eQTL data. As shown, the type 1 error rates become increasingly inflated at higher phenotype heritability as well as at lower gene expression heritability or sample size.*

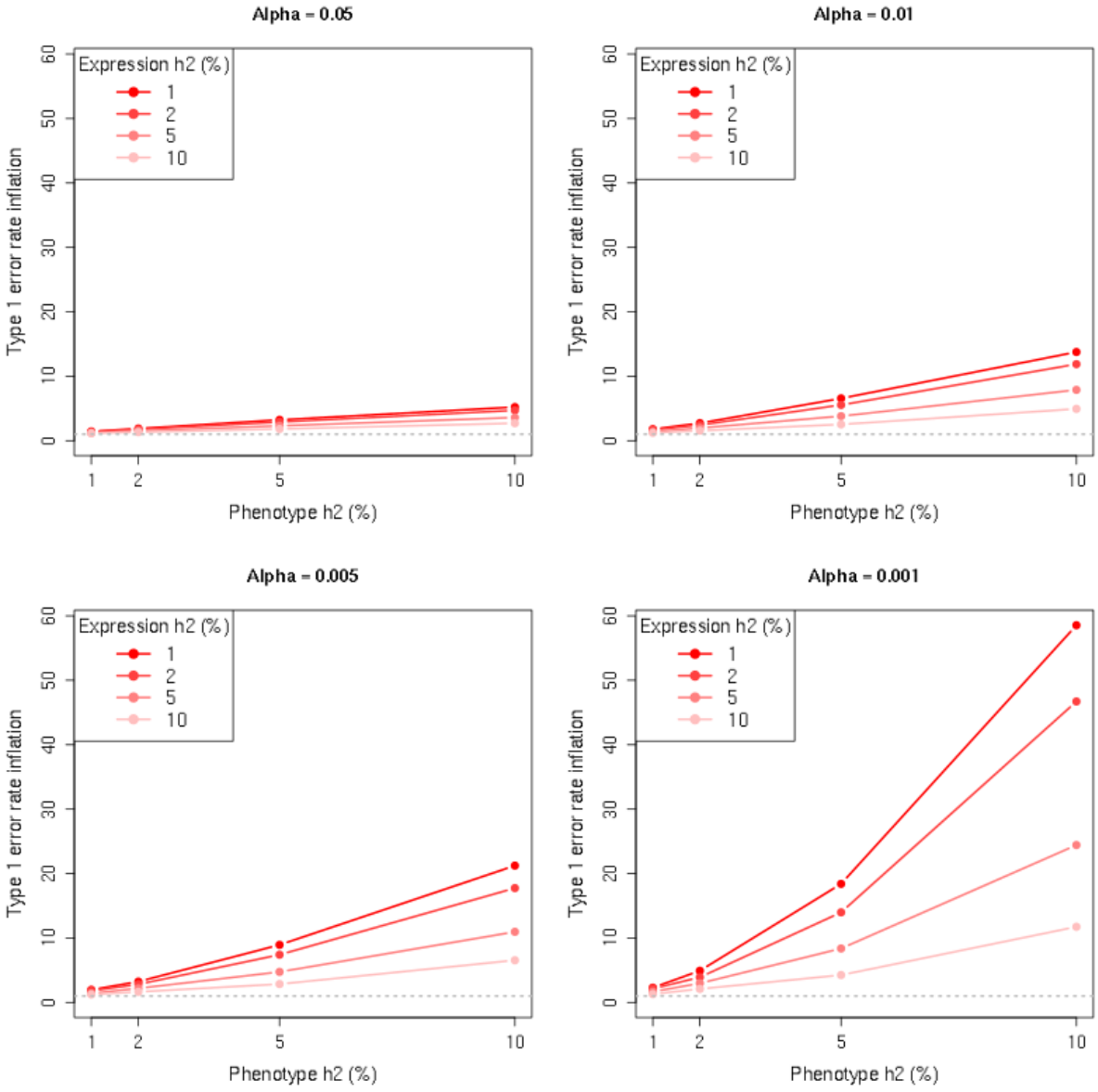

Figure 3. Type 1 error rate inflation as a function of significance threshold in primary simulations. *Shown is the type 1 error rate inflation relative to the null hypothesis of no genetic covariance ($cov(G_E, G_Y) = 0$), for different levels of α; the error rate inflation is defined as the type 1 error rate divided by the significance rate α, and equals 1 if the error rates are well-controlled. Results are for the same simulations as depicted in Figure 2 (left panel, with eQTL sample size of 10,060). As shown, the error rate inflation becomes progressively more pronounced as lower α are used.*