On the interpretation of transcriptome-wide association studies

Christiaan de Leeuw[1], Josefin Werme[1], Jeanne E. Savage[1], Wouter J. 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, location VUmc, 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) aim to detect relationships between gene expression and a phenotype, and are commonly used for secondary analysis of genome-wide association study (GWAS) results. Results from TWAS analyses are often interpreted as indicating a genetic relationship between gene expression and a phenotype, but this interpretation is not consistent with the null hypothesis that is evaluated in the traditional TWAS framework. In this study we provide a mathematical outline of this TWAS framework, and elucidate what interpretations are warranted given the null hypothesis it actually tests. We then use both simulations and real data analysis to assess the implications of misinterpreting TWAS results as indicative of a genetic relationship between gene expression and the phenotype. Our simulation results show considerably inflated type 1 error rates for TWAS when interpreted this way, with 41% of significant TWAS associations detected in the real data analysis found to have insufficient statistical evidence to infer such a relationship. This demonstrates that in current implementations, TWAS cannot reliably be used to investigate genetic relationships between gene expression and a phenotype, but that local genetic correlation analysis can serve as a potential alternative.

Introduction

Transcriptome-wide association studies (TWAS)[1–8] constitute a statistical framework commonly used to study relationships between gene expression and a phenotype. Combining expression quantitative trait loci (eQTL) and genome-wide association study (GWAS) results, a TWAS analysis estimates the genetic component of the gene expression and then tests whether this is associated with the phenotype. Since TWAS does not require that gene expression and the phenotype are measured in the same sample, it can be used with any phenotype for which GWAS data is available, giving it a broad scope of application.

From a significant TWAS association, it would be tempting to conclude that there is a genetic relationship between the expression of the tested gene and the phenotype of interest, and this interpretation is indeed commonly found in applied TWAS studies[9–13]. Such an interpretation is also suggested by various TWAS method papers such as Gamazon (2015)[7] and Mancuso (2019)[8], with for example the latter stating that "[TWAS] can be viewed as a test for non-zero local genetic correlation between expression and trait." However, the traditional TWAS framework does not directly test the relationship of the phenotype with the genetic component of the gene expression. Rather, it only tests the relationship with the *predicted* genetic component of the gene expression. This is crucial, as traditional TWAS methods do not account for the uncertainty in that prediction[6,7], meaning they cannot be accurately used to infer a genetic relationship between gene expression and the phenotype on a population level.

In this study, we show that this distinction has important implications for the interpretation and utility of TWAS results. First, we dissect the mathematical structure of the TWAS framework, explicating the null hypothesis that is tested in TWAS analysis, and then demonstrating how this relates to the genetic relationship between gene expression and the phenotype. Second, we use large-scale simulations to demonstrate that if we were to incorrectly interpret TWAS as testing such genetic relationships, this can result in potentially highly inflated type 1 error rates. Finally, we use a real data application with five well-powered phenotypes to illustrate what these error rate inflations can look like in practice, showing that TWAS cannot be used as a reliable tool for detecting genetic relationships between gene expression and a phenotype of interest.

Results

The TWAS framework

The traditional TWAS framework is implemented as a two-stage procedure, as follows[1,6,7]. First, a model is specified for the expression level 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 a residual component independent of X. We can further define $G_E = X\alpha_E$, which reflects the true local genetic component of the gene expression captured by the SNPs in X. This model is fitted to a sample with genotype and gene expression data to obtain an estimated genetic effect vector $\hat{\alpha}_E$. With this, the estimated genetic component of the gene expression is then computed in the GWAS sample for the outcome phenotype Y as $\hat{G}_E = X\hat{\alpha}_E$. Finally, 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 the estimated genetic component $\hat{G}_E$ and the phenotype Y, evaluating the null hypothesis of no association $H_0: \beta = 0$.

Note that our presentation here is simplified for the sake of brevity, which includes assuming a continuous phenotype and omitting covariates. The second stage can also be rewritten to require only pre-existing GWAS summary statistics rather than raw genotype and phenotype data as input. See *Methods - Outline of TWAS framework* for a description of the more general case.

In practice, most existing TWAS methods have this mathematical structure[6–8,14–24], including all the 'linear' models listed in Table 1 aside from CoMM[25]. Where they differ is in their implementation, particularly in how $\hat{\alpha}_E$ is estimated[6–8,14–25]. Since the number of SNPs in X is usually much larger than the sample size of the gene expression data, and are also highly collinear, it is not possible to fit the model in (1) using standard procedures such as Ordinary Least Squares. Most TWAS implementations therefore use a penalized regression approach such as Elastic Net, LASSO or a spike-and-slab model to resolve this issue (see Table 1). Although different kinds of penalized regression procedures will also yield differ different estimates $\hat{\alpha}_E$, all these methods are still conceptually equivalent, each providing an estimate $\hat{G}_E$ of the genetic component of the gene expression for use in the second stage.

Not all existing methods adopt this structure, such as the methods listed as 'non-linear' in Table 1, but these methods run into the same issues that will be discussed below for the traditional 'linear' TWAS framework (see *Supplemental Information - Non-linear TWAS models* for more details). Throughout this paper, unless otherwise specified, we will use TWAS to refer specifically to the linear TWAS framework as described above.

Structure of the TWAS null hypothesis

TWAS evaluate the null hypothesis $H_0: \beta = 0$ for the model presented in equation (2), which be rewritten to mathematically equivalent formulations to provide a clearer understanding of what is being tested. Firstly, we can observe that $\beta = \frac{\mathrm{cov}(\hat{G}_E, Y)}{\mathrm{var}(\hat{G}_E)}$, and therefore β will be zero if and only if the covariance $\mathrm{cov}(\hat{G}_E, Y)$ is zero. As such, the TWAS null hypothesis is equivalent to $H_0: \mathrm{cov}(\hat{G}_E, Y) = 0$.

Secondly, we can partition the phenotype Y in the same way as we did with E in equation (1), such that

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

Analogous to G_E, this gives us a local genetic component G_Y for the phenotype that is based on the SNPs in X, and a residual ξ_Y that is independent of X (that is, $\mathrm{cov}(X_j, \xi_Y) = 0$ for every SNP j). Substituting $Y = G_Y + \xi_Y$, it follows that $\mathrm{cov}(\hat{G}_E, Y) = \mathrm{cov}(\hat{G}_E, G_Y + \xi_Y) = \mathrm{cov}(\hat{G}_E, G_Y) + \mathrm{cov}(\hat{G}_E, \xi_Y)$, with the last step following from the distributive property of covariance.

Because $\hat{G}_E = X\hat{\alpha}_E = \sum_j X_j \hat{\alpha}_{Ej}$ (with j indexing the individual SNPs), and as the residual ξ_Y is independent of X, it follows that $\mathrm{cov}(\hat{G}_E, \xi_Y) = \mathrm{cov}(\sum_j X_j \hat{\alpha}_{Ej}, \xi_Y) = \sum_j \hat{\alpha}_{Ej} \mathrm{cov}(X_j, \xi_Y) = 0$, and with that we can conclude that $\mathrm{cov}(\hat{G}_E, Y) = \mathrm{cov}(\hat{G}_E, G_Y)$. In other words, since G_Y captures all the association that the SNPs in X have with the phenotype Y, and $\hat{G}_E$ is entirely based on X, the covariance between $\hat{G}_E$ and the residual term ξ_Y will be zero by definition. The result of this is that we can further rewrite the TWAS null hypothesis into $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$.

Relation to local genetic covariance

Local genetic correlation, analogous to genome-wide genetic correlations, quantifies the genetic relationship between two phenotypes within a particular genomic region, with gene expression, in this context, being one of those phenotypes[26,27]. Since the local genetic correlation can be defined as $\frac{\mathrm{cov}(G_E, G_Y)}{\sqrt{\mathrm{var}(G_E)\mathrm{var}(G_Y)}}$, and thus will be zero if and only if $\mathrm{cov}(G_E, G_Y)$ is zero, the genetic relationship between gene expression and the phenotype can be tested using the null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$.

However, by contrast, TWAS performs a test of $\mathrm{cov}(\hat{G}_E, G_Y)$ rather than $\mathrm{cov}(G_E, G_Y)$, as shown above. To see how these two covariances relate to each other, we can partition the estimated genetic component $\hat{G}_E$ as $\hat{G}_E = G_E + \hat{\Delta}_E$, with $\hat{\Delta}_E$ the deviation of the sample estimate from the true genetic component G_E. Again using the distributive property of covariance, we can then conclude

138 $\mathrm{cov}(\hat{G}_E, G_Y) = \mathrm{cov}(G_E + \hat{\Delta}_E, G_Y) = \mathrm{cov}(G_E, G_Y) + \mathrm{cov}(\hat{\Delta}_E, G_Y)$. For ease of notation, we will denote
139 $\hat{C} = \mathrm{cov}(\hat{\Delta}_E, G_Y)$. We can thus see that the covariance term $\mathrm{cov}(\hat{G}_E, G_Y)$ that is tested by TWAS is
140 offset from the local genetic covariance $\mathrm{cov}(G_E, G_Y)$ by $\hat{C}$, and plugging this relationship into
141 $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$, we find that the TWAS null hypothesis is equivalent to $H_0: \mathrm{cov}(G_E, G_Y) = -\hat{C}$.
142 Crucially, the regression model in equation (2) conditions on its predictor $\hat{G}_E$, thus considering
143 the value of $\hat{G}_E$ to be fixed. It follows that the values of the deviation term $\hat{\Delta}_E$, and therefore that of
144 $\hat{C}$, are fixed as well, even though $\hat{\Delta}_E$ is the result of the random sampling of E in the gene expression
145 data. As such, the quantity $-\hat{C}$ that TWAS implicitly tests $\mathrm{cov}(G_E, G_Y)$ against, will have an unknown,
146 arbitrary value (different for each gene) that reflects the specific error in the estimation that was
147 realized when the gene expression E was generated, with the distribution from which this value is
148 drawn also determined by choice of model used for estimating α_E (for more details, see *Supplemental*
149 *Information – The TWAS null value*).
150 To further illustrate this, consider an analogy to the t-test. Suppose we are interested in testing
151 whether the population means μ_A and μ_B of some variable for groups A and B differ. However, instead
152 of performing a two-sample t-test with $H_0: \mu_A = \mu_B$, or equivalently $H_0: \mu_A - \mu_B = 0$, we compute
153 the sample mean $\hat{\mu}_B$ for group B and use a single sample t-test with $H_0: \mu_A = \hat{\mu}_B$. Let us assume that
154 in this case, in our sample from group B the realized error in the estimate $\hat{\mu}_B$ is 1, ie. $\hat{\mu}_B = \mu_B + 1$. This
155 means that with our single sample t-test, we would be testing $H_0: \mu_A - \mu_B = 1$. As with TWAS, because
156 we are disregarding the estimation uncertainty in $\hat{\mu}_B$ when performing our test, we end up testing the
157 quantity of interest, $\mu_A - \mu_B$, against an unknown null value, which isn't very informative.
158 *Interpretation of TWAS results*
159 As shown above, TWAS implicitly tests the local genetic covariance $\mathrm{cov}(G_E, G_Y)$ against $-\hat{C}$ (for an
160 unknown value of $\hat{C}$) rather than zero, and as such it cannot be considered a direct test of the genetic
161 relationship between gene expression and the phenotype. This raises the question of how TWAS
162 results actually *can* be interpreted, which we can address by determining the circumstances in which
163 the TWAS null hypothesis $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$ is true.
164 As shown above, the null hypothesis $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$ is equivalent to $H_0: \mathrm{cov}(G_E, G_Y) =$
165 $-\hat{C}$. Examining the structure of $\hat{C}$ more closely we can see that it decomposes into a weighted sum
166 $\hat{C} = \sum_j \hat{w}_j \mathrm{cov}(X_j, Y)$ of the (population-level) covariances $\mathrm{cov}(X_j, Y)$ of the SNPs in X with the
167 phenotype (see *Supplemental Information – The TWAS null value*). These weights $\hat{w}_j$ reflect the error
168 in the effect size estimate for each SNP j, ie. $\hat{\alpha}_{Ej} = \alpha_{Ej} + \hat{w}_j$, and as such their values are the
169 realization of a random process.
170 It follows that, if there is any genetic association between the SNPs in X and the phenotype Y,
171 the TWAS null hypothesis will only be true if by chance the values of $\hat{w}_j$ are such that $\sum_j \hat{w}_j \mathrm{cov}(X_j, Y)$

172 happens to equal $-\mathrm{cov}(G_E, G_Y)$. Because the value of $\sum_j \widehat{w}_j \mathrm{cov}(X_j, Y)$ is effectively a random draw
173 from a continuous space, the probability of this happening is infinitessimal. Realistically, the TWAS null
174 hypothesis will therefore only be true if there is no genetic association between X and Y, with
175 $\mathrm{cov}(X_j, Y) = 0$ for every SNP j. In this case, $\sum_j \widehat{w}_j \mathrm{cov}(X_j, Y)$ will be zero as well, and at the same time
176 the absence of such genetic signal also means that the variance of G_Y is zero, and therefore no
177 covariance with G_E can exist. As such, in this case both $\mathrm{cov}(G_E, G_Y)$ and $\hat{C}$ will be zero, and
178 $H_0: \mathrm{cov}(G_E, G_Y) = -\hat{C}$ is therefore true.

179 In practice, the TWAS null hypothesis will therefore be true if, and only if, the SNPs in X are all
180 independent of the phenotype, and a test of this null hypothesis therefore amounts to a joint genetic
181 association test, ie. a test of independence between X and Y akin to testing $H_0: \alpha_Y = 0$ in the general
182 regression model in equation (3), though with somewhat different power characteristics. Although as
183 shown, the TWAS null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = -\hat{C}$ being true does imply that the null hypothesis
184 $H_0: \mathrm{cov}(G_E, G_Y) = 0$ of no local genetic relationship is true as well, the reverse does not hold. If genetic
185 association does exist between X and Y, then $H_0: \mathrm{cov}(G_E, G_Y) = -\hat{C}$ will be false, but as long as this
186 genetic association is independent of the genetic signal captured by G_E, then $\mathrm{cov}(G_E, G_Y)$ will still be
187 zero. This is why it does not follow from the TWAS null hypothesis being false that a local genetic
188 relationship exists between gene the expression and the phenotype.

189 Despite this, the intuition may remain that even though $\hat{G}_E$ is uncertain, it is still based on the
190 true G_E and should therefore still contain *some* information relevant to the genetic relationship
191 between gene expression and the phenotype. And this is indeed the case, because as shown the tested
192 covariance $\mathrm{cov}(\hat{G}_E, G_Y) = \mathrm{cov}(G_E, G_Y) + \hat{C}$ is partly based on the true local genetic covariance
193 $\mathrm{cov}(G_E, G_Y)$. But although present, this information cannot meaningfully be extracted from the TWAS
194 results, since even when $\mathrm{cov}(\hat{G}_E, G_Y)$ is found to deviate significantly from zero, the TWAS cannot tell
195 us whether this is due to a strong local genetic relationship through $\mathrm{cov}(G_E, G_Y)$, or just through the
196 noise term $\hat{C}$.

197

198

199 *TWAS as a test of genetic relationship*

200 As established above, TWAS does not technically provide a direct test of the genetic relationship
201 between gene expression and the phenotype, since its null hypothesis is equivalent to
202 $H_0: \mathrm{cov}(G_E, G_Y) = -\hat{C}$, with the value of $\hat{C}$ unknown. However, it is possible that in the scenarios that
203 are likely to arise in practice, the value of $\hat{C}$ might be so small as to be negligible. If so, we could simply
204 treat TWAS as if it is testing the null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$ after all, and interpret its

205 significant results as indicative of genetic relationships between gene expression and the phenotype
206 without issue.

207 In the remainder of this paper, we will evaluate the viability in practice of treating TWAS as a
208 test of the null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$, first using large-scale simulations to determine the
209 resulting type 1 error rates across a range of different scenarios, and then using an application to real
210 data to illustrate the impact in practice. We used FUSION[6] for this evaluation, as it is a widely used tool,
211 and it implements a range of different models within the same tool that are representative of those
212 used in many other TWAS methods (see Table 1). We also included CoMM[25], since it uses a setup that
213 is somewhat different than that of the other TWAS methods, and closer to that of local genetic
214 correlation methods[26,27] (see *Supplemental Information - Comparison with the CoMM model*). Note
215 that reported type 1 error rates will be relative to $H_0: \mathrm{cov}(G_E, G_Y) = 0$, rather than the null hypotheses
216 these methods are designed to test, as the aim is to evaluate the viability of using these methods when
217 they are treated as a test the genetic relationship between gene expression and the phenotype (as
218 they are often interpreted in practice).

219 To serve as a reference, we included the local genetic correlation model implemented in LAVA
220 (denoted LAVA-rG), which is designed and calibrated to test the null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$
221 directly. Furthermore, we also implemented a TWAS model in the LAVA framework (denoted LAVA-
222 TWAS), which is identical to LAVA-rG in every way, except that like other traditional TWAS methods it
223 treats $\hat{G}_E$ as given (rather than modelling its distribution under sampling of the gene expression E, like
224 LAVA-rG does; see *Methods - LAVA implementation of TWAS*). This implementation allows for a direct
225 comparison between testing $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$ and testing $H_0: \mathrm{cov}(G_E, G_Y) = 0$, unconfounded by
226 other differences in model or implementation.

227

228

229 *Simulation study*

230 To evaluate the performance of TWAS when interpreted as a test of the genetic relationship between
231 gene expression and outcome phenotype, we performed a series of simulations, generating gene
232 expression and phenotype values under the null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$ and computed type
233 1 error rates. Genotype data from the UK Biobank[28,29] was used as a basis for the simulations, and we
234 varied the number of SNPs per simulated gene, as well as the sample size and local heritability of the
235 gene expression and outcome phenotype data (see *Methods - Simulation study*). Analysis was
236 performed using LAVA-TWAS and LAVA-rG, as the four different models in FUSION (BLUP, Bayesian
237 LMM, Elastic Net, LASSO), and CoMM.

238 The simulations showed considerable inflation of type 1 error rates ($\alpha = 0.05$) when using
239 LAVA-TWAS to evaluate $H_0: \mathrm{cov}(G_E, G_Y) = 0$, depending on the simulation condition. The inflation
240 strongly increased with larger local heritability or sample size for the outcome phenotype (Figure 1),
241 with the local heritability also interacting with the number of SNPs in the gene (Supplemental Figure
242 1). In addition, some decrease in the type 1 error rate inflation was also found with greater local
243 heritability or sample size for the gene expression, but this effect is much less pronounced. Moreover,
244 the type 1 error rate inflation also became progressively worse when evaluating it at stricter
245 significance thresholds than the α of 0.05 used so far (Figure 2). Type 1 error rates also increased when
246 filtering out genes that exhibit insufficient genetic signal for the gene expression, as is general practice
247 in TWAS analysis (Supplemental Figure 2). To validate the LAVA-TWAS model, type 1 error rates for
248 LAVA-TWAS under its actual null hypothesis of $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$ were also assessed. These were
249 found to be well controlled in all conditions (Supplemental Figure 3), further emphasizing the
250 disconnect between this null hypothesis and the null hypothesis of no genetic relationship.
251 Results for the four FUSION penalized regression models are largely the same as for LAVA-
252 TWAS (Figure 3), suggesting that the type of model and penalization used for the estimation of $\hat{\alpha}_E$ does
253 not strongly affect the subsequent type 1 error rate inflation. Although the Elastic Net and LASSO
254 models showed lower inflation than other models in the 1% gene expression heritability conditions,
255 this can be explained by the fact that they frequently failed to converge in these conditions when
256 estimating equation (1). This difference was no longer present in the 5% heritability conditions (see
257 Supplemental Figure 4).
258 Similar patterns of type 1 error rate inflation were found for CoMM as well and were generally
259 more pronounced, though the influence of the gene expression heritability was more complex than
260 for the other methods (Supplemental Figure 5). This is due to specific constraints in the model, which
261 are analogous to the $\alpha_Y = \hat{\alpha}_E \beta$ constraint the traditional TWAS models impose the multiple regression
262 in equation (4) (see *Supplemental Information - Comparison with the CoMM model*). As in the
263 validation simulations in Werme (2022)[26], type 1 error rates for LAVA-rG were found to be well-
264 controlled (Figure 3).
265 Overall, the results of our simulations demonstrate that treating TWAS analysis as if they were
266 a test of the null hypothesis of no genetic relationship $H_0: \mathrm{cov}(G_E, G_Y) = 0$ can cause considerable
267 inflation of the type 1 error rates in many conditions. A strong effect was observed for the sample size
268 and heritability of the outcome phenotypes, in line with the interpretation of TWAS as a joint
269 association test. Of particular concern is that the inflation increases when filtering on the univariate
270 signal of the gene expression, which is commonly done in TWAS studies. The inflation also increases
271 when using lower significance thresholds, which is likely to apply in practice to correct for multiple
272 testing, and often at more stringent levels than shown in our simulations. See *Supplemental*

Information - Discussion of simulation results for more discussion of the findings from these simulations.

Real data analysis

To gauge the magnitude of the impact this issue has in a real data context, we applied TWAS to GWAS results of five well-powered phenotypes[30–33] (see Table 2), with eQTL data for blood gene expression from GTEx[34] (v8). A total of 14,584 genes were analysed, performing TWAS analysis for those genes for which the univariate gene expression signal was significant at $\alpha_{\mathrm{BONF}} = 0.05/14{,}584 = 3.43 \times 10^{-6}$ (see *Methods - Data* and *Methods - Real data analysis*). The data was analysed with the Elastic Net and LASSO models in FUSION and LAVA-TWAS. We analyzed the data with LAVA-rG as well to provide a baseline for the LAVA-TWAS results, and determine for how many associations the $H_0: \mathrm{cov}(G_E, G_Y) = 0$ null hypothesis can be validly rejected.

A summary of the results is given in Table 2 and as shown, the number of significant LAVA-TWAS associations is considerably higher than those for LAVA-rG, with on average only about 59% of LAVA-TWAS associations confirmed by LAVA-rG. From this it follows that if we were to interpret LAVA-TWAS as a test of genetic relationship between gene expression and the phenotype, 41% of the associations found would not be valid. This is because LAVA-rG is calibrated for testing $H_0: \mathrm{cov}(G_E, G_Y) = 0$, and LAVA-TWAS only differs from LAVA-rG in the fact that it conditions on $\hat{G}_E$ whereas LAVA-rG models the distribution of the estimation uncertainty in $\hat{G}_E$ that we know exists. This single difference is the cause of the inflated type 1 error rates observed for LAVA-TWAS (but not LAVA-rG) in the simulations. Consequently, because there is no other difference between the two, the 41% of significant LAVA-TWAS associations not shared with LAVA-rG cannot be due to a legitimate increase in power, meaning that they are not valid if interpreting LAVA-TWAS as a test of $H_0: \mathrm{cov}(G_E, G_Y) = 0$ (see also *Supplemental Information – Underlying distributions*).

Although no equivalent reference model directly testing $H_0: \mathrm{cov}(G_E, G_Y) = 0$ is available for the FUSION models, the FUSION results paint a similar picture. The number of associations found when using FUSION are at a comparable level to LAVA-TWAS (Table 2, Supplemental Table 1), and the FUSION models had very similar type 1 error rate inflation levels as LAVA-TWAS in the simulations as well. Taken together, it can reasonably be concluded that the proportion of FUSION results in the real data analysis that are not supported by sufficient statistical evidence, when interpreted as a test of the genetic relationship between gene expression and the phenotype, is likely at a very similar level as LAVA-TWAS

We also evaluated results for genes with a univariate p-value for the gene expression greater than 0.05. As shown in Supplemental Table 2, virtually no local genetic correlations are found with

LAVA-rG for these genes, yet results for the TWAS analyses remain at a similar level as in the primary analysis in Table 2. Because these genes have no detectable genetic component for the gene expression, they would typically not be included in a TWAS analysis. But the fact that TWAS still yields a considerable number of significant associations in this context, when $\hat{\alpha}_E$ reflects essentially just noise, serves to further illustrate how TWAS simply functions as a form of joint association test, with results not necessarily reflecting any genetic relationship between gene expression and the outcome phenotype.

Deflation of standard errors

Fundamentally, the statistical reason behind the results in our simulations and real data analysis is that the TWAS model omits a source of variance that we know is present. The estimated genetic effect vector $\hat{\alpha}_E$ is subject to estimation uncertainty due to the sampling of the gene expression E, but this uncertainty is not accounted for in the TWAS test. If the aim is to evaluate the genetic relationship with the phenotype, by testing the null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$, the resulting standard errors will therefore only reflect the uncertainty due to sampling of Y, and not that of E. As such, the standard errors will be too small, creating a downward bias in the p-values (see *Supplemental Information – Underlying distributions* for more details). For LAVA-TWAS the corrected standard errors are known, as these are the LAVA-rG standard errors, and we can therefore use these to illustrate the issue. We computed this standard error deflation individually for each gene in our analysis, as the ratio of the LAVA-TWAS standard errors to their corrected value. These are summarized in Table 3 and, as shown, they vary considerably across genes, with negligible deflation for some genes but severe deflation for others. Figure 4 illustrates the relation between this standard error deflation and the resulting bias in p-values, using the BMI results as an example, with Supplemental Figure 6 further showing the level of realized bias. This high variability of standard error deflation across genes also demonstrates that this issue must be addressed at the level of the individual gene, and cannot be resolved by applying a simple post-hoc correction to the gene p-values.

Discussion

In this paper, our aim was to elucidate the interpretation of the TWAS null hypothesis and results. Although TWAS results are often interpreted as indicating a genetic relationship between gene expression and the phenotype, we showed that this is not what the TWAS null hypothesis allows us to test. Instead, the null hypothesis that the traditional TWAS models actually evaluate effectively makes

343 them a test of joint genetic association between the SNPs included in the analysis and the phenotype,
344 irrespective of any genetic relationship with the gene expression. However, because it is not
345 uncommon for TWAS results to be misinterpreted as a test of such a genetic relationship, we aimed to
346 determine to what extent this interpretation was viable in practice. To this end, we performed
347 simulations to determine the degree of type 1 error rate inflation that occurs when treating TWAS
348 results as a test of genetic relationship, and performed real data analysis to demonstrate what this
349 translated to in practice.
350 Our results showed that using TWAS to infer genetic relationships between gene expression and the
351 phenotype can lead to strongly inflated type 1 error rates, as well as a substantial proportion of invalid
352 significant results, for which such a genetic relationship is insufficiently supported by the statistical
353 evidence in the data. This was further emphasized by the fact that in our real data analysis, TWAS still
354 yielded a considerable number of significant associations even when genetic signal for the gene
355 expression was absent altogether, clearly illustrating how TWAS functions simply as a form of joint
356 association test. The issue is further complicated by the fact that the degree to which standard errors
357 are deflated varies across genes, depending on factors such as the gene size and level of genetic signal
358 for both gene expression and phenotype. This therefore rules out the possibility of a simple post-hoc
359 correction on existing TWAS results, making them difficult to interpret and use in practice. Although
360 TWAS results can still be validly interpreted as a form of joint association test for the genetic
361 relationship between the SNPs and the phenotype, as out results show, it does not allow for any
362 conclusions to be drawn about the relationship of the phenotype with the gene expression, and
363 therefore does not address the research questions that most researchers are looking to answer with
364 such data. While the issue of uncertainty in eQTL estimates has occasionally been mentioned in
365 TWAS literature[1,25,35,36], its implications for the validity of the TWAS framework have received little
366 scrutiny thus far. Moreover, although various TWAS methods have included simulations to evaluate
367 type 1 error rates, these all either treated $\hat{G}_E$ as fixed in the simulations just as in the TWAS model
368 itself[16,23], or only simulated under a null scenario of the phenotype having no genetic component at all
369 (ie. $G_Y = 0$)[22,24,25]. This explains why the issue demonstrated in this paper has not yet been widely
370 discussed.
371 Addressing this issue with TWAS generally requires fully accounting for the uncertainty of the
372 estimate $\hat{G}_E$ in relation to the true genetic component G_E. While it is in principle possible to
373 incorporate this into existing TWAS models, in practice, the feasibility of implementing this will strongly
374 depend on the approach used to fit the model in equation (1), as the distribution of the noise
375 component induced by penalized regression models can take mathematically very complex forms.
376 Another option is to use alternative methods to detect genetic relationships between gene
377 expression and a phenotype. As demonstrated, local genetic correlation analysis methods could

378 potentially fill this role, as these are explicitly designed to evaluate local genetic relationships between
379 phenotypes. However, methods like colocalization[37,38] or Mendelian Randomization (MR)[39–41] could
380 potentially be used for this as well. These methods use quite different model assumptions, but test
381 null hypotheses that can answer similar research questions pertaining to local genetic relationships
382 between gene expression and phenotypes. Of note as well is that the issue with TWAS described in
383 this paper has a clear analogy to the issue of the "no measurement error" (NOME) assumption in MR
384 analysis[42]. As such, strategies used to remove this NOME assumption in the MR context may also help
385 inform solutions for addressing the issue in TWAS methods.

386 Investigating genetic relationships between gene expression and phenotypes, as TWAS is often
387 intended to do, has the potential to provide valuable insights into the genetic etiology of many
388 phenotypes. But although traditional TWAS models are statistically valid relative to their own null
389 hypothesis, ultimately they only perform a test of joint association of the phenotype with the SNPs in
390 a region, and thus cannot be used to reliably draw conclusions about any genetic relationship with
391 gene expression. Further development of existing TWAS methods to account for this issue, or adoption
392 of alternative methods such as local genetic correlation analysis, will therefore be essential to support
393 future investigations of the relationship between gene expression and phenotypes.

394 **Methods**

395

396 *Ethics statement*

397 This study relied on simulated data based on the UK Biobank genotype data[28], as well as secondary
398 analysis of publicly available summary statistics. Ethical approval for this data was obtained by the
399 primary researchers.

400

401 *Outline of TWAS framework*

402 For a standardized genotype matrix X containing SNPs for a particular gene, and with E and Y the
403 centered gene expression for that gene and outcome phenotype respectively, the general TWAS
404 framework consists of a two-stage procedure, based on the equations

405

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

406

407 and

408

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

409

410 with the estimated genetic component $\hat{G}_E = X\hat{\alpha}_E$, also given in the main text. For ease of notation,
411 we have omitted model intercepts and covariates from these equations, but in practice these will
412 usually be included. An alternative model may also be used rather than the linear regression in
413 equation (2), such as a logistic regression if the phenotype is binary.

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

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

427 In practice, equation (1) cannot be fitted with a traditional multiple linear regression model,
428 due to the high LD between SNPs (leading to extreme collinearity), and the number of SNPs typically
429 exceeding the sample sizes of eQTL data. Some form of regularization in the regression model is
430 therefore required to obtain $\hat{\alpha}_E$, and consequently one of the main discrepancies between TWAS
431 implementations is the specific regularization used (see Table 1). In some cases, rather than fitting
432 equation (1), the elements of $\hat{\alpha}_E$ are simply set to the marginal SNP effect estimates instead.
433 Note that some methods[22–24] diverge from this linear model structure (Table 1, *non-linear*
434 *models*). Statistically these can be seen as generalizations of the TWAS framework, though
435 conceptually they can no longer be interpreted as imputing the genetic component of gene expression.
436 See *Supplemental Information - Non-linear TWAS models* for more details.
437
438
439 *Local genetic correlation*
440 The LAVA implementation of local genetic correlation analysis has been described in detail in Werme
441 et al. (2022)[26]. In brief, LAVA uses summary statistics and a reference genotype sample to fit equations
442 (1) and (3), obtaining estimates of $\hat{\alpha}_E$ and $\hat{\alpha}_Y$ as well as a corresponding sampling covariance matrix
443 for each (a logistic regression equivalent is used for binary phenotypes). To do so, a singular value
444 decomposition for X is computed, pruning away excess principal components to attain regularization
445 of the models and allowing them to be fitted.
446 With the pruned and standardized principal component matrix $W = XR$ (with R the
447 transformation matrix projecting the genotypes onto the principal components), we can write $G_E =$
448 $W\delta_E$ and $G_Y = W\delta_Y$, where δ_E and δ_Y are the genetic effect size vectors for these principal
449 components. Their estimates $\hat{\delta}_E$ and $\hat{\delta}_Y$ can be used to obtain $\hat{\alpha}_E$ and $\hat{\alpha}_Y$ by reversing the
450 transformation through R, such that $\hat{\alpha}_E = R\hat{\delta}_E$ and $\hat{\alpha}_Y = R\hat{\delta}_Y$, with the projection to the principal
451 components effectively providing a form of regularization in the estimation of α_E and α_Y. In practice
452 however, LAVA is defined and implemented directly in terms of δ_E and δ_Y and its estimates, rather
453 than working with α_E and α_Y explicitly.
454 For ease of notation, we define the combined matrix $G = (G_E, G_Y) = W\delta$ for the genetic
455 components, with combined effect size matrix $\delta = (\delta_E, \delta_Y)$. We denote the effect sizes for a single
456 principal component j as δ_j, corrresponding to the jth row of γ, and denote the number of principal
457 components as K.
458 The estimates $\hat{\delta}_E$ and $\hat{\delta}_Y$ are obtained by reconstructing multiple linear regressions from the
459 input summary statistics (see Werme et al. (2022)[26] for details). This uses two separate equations of
460 the form $E = W\delta_E + \zeta_E$ and $Y = W\delta_Y + \zeta_Y$, 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 $\hat{\delta}_E = (W^T W)^{-1} W^T E = \frac{W^T E}{N-1}$ (since $W^T W = I_K(N-1)$, with I_K the size K identity matrix) and similarly $\hat{\delta}_Y = \frac{W^T Y}{N-1}$, with corresponding sampling distributions $\hat{\delta}_E \sim \mathrm{MVN}(\delta_E, \sigma_E^2 I)$ and $\hat{\delta}_Y \sim \mathrm{MVN}(\delta_Y, \sigma_Y^2 I)$, 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). In practice, we obtain these by estimating η_E^2 and η_Y^2 and plugging these in to get estimates $\hat{\sigma}_E^2$ and $\hat{\sigma}_Y^2$.

For principal component j we therefore have $\hat{\delta}_j \sim \mathrm{MVN}(\delta_j, \hat{\Sigma})$, where the diagonal elements of $\hat{\Sigma}$ are $\hat{\sigma}_E^2$ and $\hat{\sigma}_Y^2$ 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, denoted Ω, we have $\Omega = \mathrm{cov}(G) = \frac{G^T G}{N-1} = \frac{\delta^T W^T W \delta}{N-1} = \delta^T \delta$, it follows that inference on Ω can be performed using the sampling distributions for $\hat{\delta}_E$ and $\hat{\delta}_Y$ 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 $\delta_E = \vec{0}$ and $\delta_Y = \vec{0}$ 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 ω_E^2 and ω_Y^2, the diagonal elements of Ω. In general, it is recommended in LAVA to test both genetic variances before performing the bivariate analysis, since genetic covariance can only exist in a genomic region where there both phenotypes exhibit some degree of genetic variance, though analogous to standard practice in TWAS literature in this paper we only test on the genetic variance for the gene expression.

From the above distributions it follows that the expected value $E[\hat{\delta}^T \hat{\delta}] = \delta^T \delta + K\hat{\Sigma}$, and we can therefore use the method of moments to estimate Ω as $\hat{\Omega} = \hat{\delta}^T \hat{\delta} - K\hat{\Sigma}$. Since in the present analyses there is assumed to be no sample overlap, the off-diagonal elements of Σ are 0, and the estimate for the genetic covariance therefore reduces to $\hat{\omega}_{EY} = \hat{\delta}_E^T \hat{\delta}_Y$. The matrix $\hat{\delta}^T \hat{\delta}$ has a non-central Wishart sampling distribution, which in the current implementation of LAVA is used to obtain p-values to test $H_0: \omega_{EY} = \mathrm{cov}(G_E, G_Y) = 0$ using a simulation procedure (see Werme et al. (2022) for details). Note that the LAVA model as described in Werme et al. (2022) is technically defined in terms of testing $H_0: \mathrm{cov}(\delta_E, \delta_Y) = 0$, but since $\mathrm{cov}(G_E, G_Y) = \frac{G_E^T G_Y}{N-1} = \frac{\delta_E^T W^T W \delta_Y}{N-1} = \delta_E^T \delta_Y = K\mathrm{cov}(\delta_E, \delta_Y)$, this is equivalent to testing $H_0: \mathrm{cov}(G_E, G_Y) = 0$.

However, under this null hypothesis $\hat{\omega}_{EY}$ has an expected value of zero and a variance of $V_{\mathrm{rG}} = K\hat{\sigma}_E^2 \hat{\sigma}_Y^2 + \hat{\omega}_E^2 \hat{\sigma}_Y^2 + \hat{\sigma}_E^2 \hat{\omega}_Y^2$. Although it is a sum of product-normal distributions, with large enough K its distribution converges on a normal distribution following the central limit theorem, allowing for a

493 normal approximation of the sampling distribution of $\widehat{\omega}_{EY}$. For the simulations and real data analysis
494 we computed LAVA-rG p-values in both ways, and as shown in Supplemental Figure 7 these were found
495 to be virtually identical. As such, all LAVA-rG results reported in this paper are based on p-values
496 computed using the distribution $\widehat{\omega}_{EY} \sim \mathrm{N}(0, V_{\mathrm{rG}})$, to provide a mathematically more direct comparison
497 for the LAVA-TWAS implementation.
498
499
500 *LAVA implementation of TWAS*
501 To construct a TWAS model within the LAVA framework, we note that in a linear regression for
502 equation (2) we have estimator $\hat{\beta} = \frac{\widehat{\mathrm{cov}}(\hat{G}_E, Y)}{\widehat{\mathrm{var}}(\hat{G}_E)}$. Since $\widehat{\mathrm{var}}(\hat{G}_E)$ is considered fixed in TWAS the sampling
503 distribution of $\hat{\beta}$ directly proportional to the distribution of $\widehat{\mathrm{cov}}(\hat{G}_E, Y)$, the sample estimate of
504 $\mathrm{cov}(\hat{G}_E, Y)$. As both $\hat{G}_E$ and Y have means of zero, and since $\hat{G}_E = W\hat{\delta}_E$, we have $\widehat{\mathrm{cov}}(\hat{G}_E, Y) = \frac{\hat{G}_E^T Y}{N-1} =$
505 $\frac{\hat{\delta}_E^T W^T Y}{N-1} = \hat{\delta}_E^T \frac{W^T Y}{N-1}$. As previously derived $\hat{\delta}_Y = \frac{W^T Y}{N-1}$, and it therefore follows that $\widehat{\mathrm{cov}}(\hat{G}_E, Y) = \hat{\delta}_E^T \hat{\delta}_Y$.
506 This $\widehat{\mathrm{cov}}(\hat{G}_E, Y)$ is therefore equal to the genetic covariance estimate $\widehat{\omega}_{EY}$ used by LAVA-rG, and thus
507 both LAVA-TWAS and LAVA-rG are shown to use the same test statistic $\hat{\delta}_E^T \hat{\delta}_Y = \widehat{\omega}_{EY}$.
508 We can also observe that this $\widehat{\omega}_{EY}$ can be rewritten as the sum $\sum_j \hat{\delta}_{Ej} \hat{\delta}_{Yj}$, weighting each
509 genetic association $\hat{\delta}_{Yj}$ by the corresponding eQTL estimate $\hat{\delta}_{Ej}$. This implementation is therefore
510 analogous to how TWAS is performed using GWAS summary statistics in other TWAS methods (eg.
511 Gusev (2016)[6]), except defined in terms of the estimated genetic associations of principal component
512 matrix W rather than the original SNP genotype matrix X.
513
514
515 *Data*
516 Genotype data for simulations came from the UK Biobank [28], a national genomic study of over 500,000
517 volunteer participants in the UK. The study was approved by the National Research Ethics Service
518 Committee North West–Haydock (reference 11/NW/0382) and data were accessed under application
519 #16406. Data collection, primary quality control, and imputation of the genotype data were performed
520 by the UK Biobank[29]. Data was converted to hardcalled genotypes, and then additionally filtered, using
521 only SNPs with an info score of at least 0.9 and missingness no greater than 5%, and restricting the
522 sample to unrelated, European-ancestry individuals with concordant sex. Remaining missing genotype
523 values were then mean-imputed to ensure consistency of the input data across different methods.
524 The European panel of the 1,000 Genomes[43] data (N = 503, as downloaded from
525 https://ctg.cncr.nl/software/magma) was used as genotype reference data to estimate LD for the real

526 data analysis for both LAVA and FUSION, removing SNP with a minor allele frequency (MAF) below
527 0.5%. For eQTL data we used the GTEx[34] data (v8, European subset), for whole blood gene expression.
528 The published cis-eQTL summary statistics (European ancestry) from GTEx Portal were used for the
529 LAVA analysis, pre-computed FUSION model files were used for the FUSION analysis. For every
530 analysed gene, we included all SNPs in the data within one megabase of the transcription start site.
531 Genes were filtered to include only autosomal protein-coding and RNA genes expressed in blood, and
532 only genes also available in the FUSION model files were used, resulting in a total of 14,584 genes
533 available for analysis.
534 GWAS summary statistics were selected for five well-powered phenotypes, chosen to reflect
535 a range of different domains. These were BMI (GIANT)[30] (no waist-hip ratio adjustment), educational
536 attainment (SSGAC)[31], schizophrenia (PGC, wave 3)[32], diastolic blood pressure (GWAS Atlas)[33] and type
537 2 diabetes (GWAS Atlas)[33]. Sample size and number of SNPs for each sample can be found in Table 2.
538
539
540 *Simulation study*
541 Simulation studies were performed to evaluate the type 1 error rates of TWAS when used to evaluate
542 the null hypothesis H_0:$\mathrm{cov}(G_E, G_Y) = 0$. The simulations were based on the UK Biobank data,
543 generating gene expression and outcome phenotypes while varying their sample sizes and the local
544 heritabilities, as well as the number of SNPs per gene. The number of SNPs K in the genes was varied
545 across values of 100, 500, 1,000 and 2,500 SNPs; sample sizes of either 250 or 1,000 were used for the
546 gene expression (denoted N_E), and 10,000, 50,000 or 100,000 for the outcome phenotype (denoted
547 N_Y); and local heritabilities, defined as the percentage of variance explained by the SNPs in the gene,
548 were set at either 1%, 5% or 10% for the gene expression (denoted h_E^2)[44], and 0.001%, 0.005%, 0.01%,
549 0.05%, 0.1%, 0.5% or 1% for the outcome phenotype (denoted h_Y^2). The number of SNPs K was only
550 varied at sample size settings of $N_E = 1{,}000$ and $N_Y = 10{,}000$, and the sample sizes were only varied
551 at $K = 1{,}000$. All combinations of the two local heritabilities were simulated for each setting of K and
552 sample sizes. The simulations were analysed using LAVA-TWAS, LAVA-rG, the four FUSION models and
553 CoMM.
554 To create the samples to use, we first drew a random subset of 100,000 individuals from the
555 UK Biobank data. For each sample size N used in the simulations, the first N individuals from this subset
556 were used for that sample, to ensure that samples of different sizes are all nested. The MAF for all
557 SNPs was computed based on the $N = 100$ subsample, discarding all SNPs with MAF below 5%. Since
558 the subsamples are nested, this ensures that all SNPs have a sufficient minor allele count at every
559 sample size used. Ten blocks of 2,500 consecutive SNPs were then selected to represent simulated
560 genes, one block each from around the middle of the first ten chromosomes (avoiding the centromere

561 region). The number of SNPs K was varied across conditions by using only a subset of these 2,500 SNPs,
562 always selecting the first K SNPs from each block. For each simulation condition, the total number of
563 iterations for that condition was evenly distributed over these ten blocks, and results were aggregated
564 over the blocks. Type 1 error rates per condition were computed as the proportion of iterations for
565 that condition for which the p-value was smaller than the significance threshold, which was set to 0.05
566 unless otherwise specified.

567 To simulate gene expression E and outcome phenotype Y for the standardized genotypes X
568 of the SNPs in a gene, X was projected onto standardized principal components W, pruning away
569 redundant components based on the cumulative genotypic variance explained by the components
570 (retaining those that jointly explain 99% of the total variance). For each iteration, true genetic effect
571 sizes δ_E and δ_Y for the principal components under the null hypothesis of $\mathrm{cov}(G_E, G_Y) = 0$ were
572 generated by drawing values from a normal distribution, then regressing one vector on the other and
573 retaining only the residuals for the outcome vector to ensure that δ_E and δ_Y were exactly independent.
574 The two genetic components were then computed as $G_E = W\delta_E$ and $G_Y = W\delta_Y$, and both were
575 subsequently standardized to a variance of one.

576 Simulated gene expression and phenotype values were then generated as $E = W\delta_E + \xi_E$ and
577 $Y = W\delta_Y + \xi_Y$, drawing the residuals ξ_E and ξ_Y from normal distributions with variance parameters
578 equal to $\frac{100-h_E^2}{h_E^2}$ and $\frac{100-h_Y^2}{h_Y^2}$ respectively, ensuring the desired level of explained variance. Both E and
579 Y were then regressed on each individual SNP in X using simple linear regression, and the resulting
580 test statistics as well as the simulated E and Y themselves were stored for subsequent analysis.

581 To compare the different methods, a baseline set of simulation conditions with 1,000
582 iterations, with $K = 1{,}000$, $N_E = 1{,}000$ and the remaining parameters across their full range, was
583 generated and analysed with all methods: LAVA-TWAS, LAVA-rG, FUSION-BLUP, FUSION-BSLMM,
584 FUSION-ElastNet, FUSION-LASSO and CoMM. LAVA analyses were performed using the summary
585 statistics as input, FUSION analyses with the raw gene expression values for the first stage and outcome
586 phenotype summary statistics for input, and CoMM using the raw values for both gene expression and
587 outcome phenotype. For the first stage of FUSION, cross-validation was turned off and the true $R_E^2 =$
588 $\frac{h_E^2}{100}$ value was set using the --hsq_set option. In the CoMM analyses, the model was found to run into
589 convergence issues due to K being equal to N_E, and the simulations were rerun setting K to 100
590 instead to resolve this. Moreover, an additional set of simulations was performed with h_e^2 values of
591 1%, 2%, 4%, 5%, 6%, 8% and 10%, and h_Y^2 values of 0.01%, 0.02%, 0.05%, 0.1%, 0.2%, 0.5% and 1%.

592 For LAVA-TWAS, further of simulations with 10,000 iterations were performed across the full
593 range of parameter settings for analysis, to evaluate the impact of these parameters on type 1 error
594 rates. To determine the effect of filtering genes on the presence of eQTL signal, as is common practice

595 in TWAS analyses, for these simulations we also computed the LAVA univariate p-value for the gene
596 expression. In addition to the regular type 1 error rates, filtered type 1 error rates were also computed
597 based on only the iterations for which the univariate p-value was below either 0.05 or 0.0001. Finally,
598 to allow for reliable computation of type 1 error rates at lower significance thresholds, for the subset
599 of these conditions with $K = 1{,}000$, $N_E = 1{,}000$, $h_E^2 = 5\%$ and $h_Y^2 = 0.001\%, 0.01\%, 0.1\%$ or 1%
600 the number of iterations was increased to 100,000. Note that although for LAVA-TWAS the number of
601 iterations differs from the other methods and varies across conditions, for each plot in this paper the
602 type 1 error rate values within that plot are always based on the same number of iterations.
603 Additional simulations were also performed under the TWAS null hypothesis of
604 $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$ to validate the LAVA-TWAS model. To do so, the above procedure was modified
605 to first generate δ_E and the corresponding G_E, from which E was then simulated. This was used to
606 estimate SNP summary statistics for the associations between X and E, from which the estimate vector
607 $\hat{\delta}_E$ was then computed for the outcome data. Finally, δ_Y was then generated to be exactly
608 independent of this $\hat{\delta}_E$, with the rest of the simulation process proceeding as normal. For h_Y^2, the
609 parameter range was extended to also include $h_Y^2 = 0\%$.
610
611
612
613 *Real data analysis*
614 TWAS and local genetic correlation analyses were performed on the GWAS data as follows, separately
615 for each of the five phenotypes. In all the analyses, SNP filtering was applied to remove all SNPs with
616 a minor allele frequency below 0.5%, and only SNPs available in the 1,000 Genomes data, the GTEx
617 data, and the GWAS sample for that phenotype were used.
618 For every gene, univariate LAVA analysis for the eQTL signal, LAVA-rG and LAVA-TWAS
619 analyses, and FUSION analyses with the Elastic Net and LASSO models was performed. For FUSION, the
620 BLUP and BSLMM models were not used as these were not available in the precomputed model files.
621 FUSION univariate p-values for the eQTL signal were also obtained from these model files. Note that
622 separate SNP filtering was applied prior to these models being computed, and as such the LAVA and
623 FUSION analyses are based on different sets of SNPs per gene.
624 For the primary analyses, the bivariate LAVA and FUSION analyses were only performed for
625 genes for which the LAVA univariate tests were significant after correcting for the total number of
626 genes, at a threshold of $0.05/14{,}584 = 3.43 \times 10^{-6}$ (see Supplemental Table 1 for results when using
627 the FUSION univariate p-values instead). The significance threshold for the bivariate analyses was set
628 separately for each phenotype, at a Bonferroni-correction for the total number of genes that was

629 univariate significant for that phenotype (see Table 2). In a secondary analysis, to evaluate TWAS
630 results in the absence of gene expression signal, bivariate analyses were performed for all genes for
631 which the LAVA univariate p-value was greater than 0.05, again setting the significance threshold per
632 phenotype for the number of genes with univariate p-value greater than 0.05 for that phenotype.
633

634 **Data availability**

635 For the real data analysis, summary statistics were obtained for five different phenotypes: blood
636 pressure (https://atlas.ctglab.nl/traitDB/3379), BMI
637 (https://zenodo.org/record/1251813#.YBAfYBbvJhE, combined.23May2018), educational attainment
638 (https://www.thessgac.org/data), schizophrenia (https://www.med.unc.edu/pgc/download-results/,
639 scz2022), type 2 diabetes (https://atlas.ctglab.nl/traitDB/3328). GTEx v8 European ancestry whole
640 blood eQTL summary statistics and the gene definition file used for the LAVA-TWAS and LAVA-rG
641 analyses were obtained via https://www.gtexportal.org/home/datasets, the corresponding FUSION
642 model files from http://gusevlab.org/projects/fusion/#single-tissue-gene-expression (EUR Samples).
643 The European subset of the 1000 Genomes data used as LD reference for the real data analyses was
644 downloaded from https://ctg.cncr.nl/software/magma. UK Biobank genotype data was obtained from
645 https://www.ukbiobank.ac.uk/ under application #16406.
646
647

648 **Code availability**

649 LAVA is implemented as an R package, which is publicly available at the LAVA website
650 (https://ctg.cncr.nl/software/lava) and LAVA GitHub repository (https://github.com/josefin-
651 werme/LAVA). GTEx v8 input files compatible with LAVA are available from the LAVA website as well.
652 Version 0.1.0 of LAVA was used for the LAVA analyses in this paper; this version contains specific
653 functionality for handling eQTL summary statistics input files. FUSION can be obtained from
654 http://gusevlab.org/projects/fusion/ and CoMM from https://github.com/gordonliu810822/CoMM.
655 Scripts used for the simulations and analyses performed for this paper are available from
656 https://github.com/cadeleeuw/twas-validity-scripts2022.
657

658 **Acknowledgements**

659 This work was funded by The Netherlands Organization for Scientific Research (Grant No. NWO VICI
660 435–14–005 [to DP]) and NWO Gravitation: BRAINSCAPES: A Roadmap from Neurogenetics to
661 Neurobiology (Grant No. 024.004.012 [to DP]), and a European Research Council advanced grant
662 (Grant No, ERC-2018-AdG GWAS2FUNC 834057 [to DP, and funding CdL]). CdL was funded by F.
663 Hoffmann-La Roche AG. WJP was funded by an NWO Veni grant (NWO: 916-19-152). JES was
664 supported by an NWO Veni grant (NWO: 201G-064). The research has been conducted using the UK
665 Biobank Resource (application no. 16406). The analyses were carried out on the Genetic Cluster
666 Computer, which is financed by the Netherlands Organization for Scientific Research (NWO: 480-05-
667 003), by the VU University (Amsterdam, The Netherlands) and the Dutch Brain Foundation, hosted by
668 the Dutch National Computing and Networking Services SurfSARA.

Author contributions

CdL and JW conceived of the study. CdL performed the analyses and simulations. CdL wrote the manuscript, in collaboration with JW. 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. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245–252 (2016).
7. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015).
8. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. 22 (2019).
9. Cheng, W. *et al.* Shared genetic architecture between schizophrenia and subcortical brain volumes implicates early neurodevelopmental processes and brain development in childhood. *Mol Psychiatry* (2022) doi:10.1038/s41380-022-01751-z.
10. Chen, X. *et al.* A novel genetic variant potentially altering the expression of *MANBA* in the cerebellum associated with attention deficit hyperactivity disorder in Han Chinese children. *The World Journal of Biological Psychiatry* 1–12 (2022) doi:10.1080/15622975.2021.2014248.
11. Li, L. *et al.* Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes. *Basic Res Cardiol* **117**, 6 (2022).

12. Zhu, Q. *et al.* UACA locus is associated with breast cancer chemoresistance and survival. *npj Breast Cancer* **8**, 39 (2022).

13. Asefa, N. G. *et al.* Bioinformatic Prioritization and Functional Annotation of GWAS-Based Candidate Genes for Primary Open-Angle Glaucoma. *Genes* **13**, 1055 (2022).

14. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *The American Journal of Human Genetics* **100**, 473–487 (2017).

15. 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).

16. Su, Y.-R. *et al.* A Mixed-Effects Model for Powerful Association Tests in Integrative Functional Genomics. *The American Journal of Human Genetics* **102**, 904–919 (2018).

17. Hu, Y. *et al.* A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat Genet* **51**, 568–576 (2019).

18. Nagpal, S. *et al.* TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *The American Journal of Human Genetics* **105**, 258–266 (2019).

19. Liu, W. *et al.* Leveraging functional annotation to identify genes associated with complex diseases. *PLoS Comput Biol* **16**, e1008315 (2020).

20. Luningham, J. M. *et al.* Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *The American Journal of Human Genetics* **107**, 714–726 (2020).

21. Bhattacharya, A., Li, Y. & Love, M. I. MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. *PLoS Genet* **17**, e1009398 (2021).

22. Xu, Z., Wu, C., Wei, P. & Pan, W. A Powerful Framework for Integrating eQTL and GWAS Summary Data. *Genetics* **207**, 893–902 (2017).

23. 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. *Genetic Epidemiology* **44**, 550–563 (2020).

24. 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).

25. 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).

26. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C. A. An integrated framework for local genetic correlation analysis. *Nat Genet* **54**, 274–282 (2022).

27. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. *Am J Hum Genet* **101**, 737–751 (2017).

28. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* **12**, e1001779 (2015).

29. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

30. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Human Molecular Genetics* **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. Xue, H. & Pan, W. Some statistical consideration in transcriptome-wide association studies. *Genet. Epidemiol.* **44**, 221–232 (2020).

36. Zhu, H. & Zhou, X. Transcriptome-wide association studies: a view from Mendelian randomization. *Quant Biol* (2020) doi:10.1007/s40484-020-0207-4.

37. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet* **10**, e1004383 (2014).

38. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics* **99**, 1245–1260 (2016).

39. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481–487 (2016).

40. Yuan, Z. *et al.* Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat Commun* **11**, 3861 (2020).

41. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat Commun* **10**, 3300 (2019).

42. Bowden, J. *et al.* Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International Journal of Epidemiology* **48**, 728–742 (2019).

43. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

44. Lloyd-Jones, L. R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *The American Journal of Human Genetics* **100**, 228–237 (2017).

Table 1. Overview of available TWAS analysis methods.

Method	Weight estimation[a]	Base model extensions
Linear models		
Gamazon (2015)[7] - PrediXcan	Marginal LASSO Elastic net	-
Gusev (2016)[6] - FUSION[b]	Top eQTL BLUP Bayesian LMM LASSO Elastic net	-
Mancuso (2017)[14] - RhoGE	BLUP	-
Barbeira (2018)[15] - MetaXcan	Marginal LASSO Elastic net	-
Su (2018)[16] - MiST	External	Models additional variance component for genetic effects not mediated by predicted expression
Hu (2019)[17] - UTMOST	Multivariate LASSO	Simultaneously models multiple tissues during weight estimation
Yang (2019)[25] - CoMM	Collaborative mixed model	Estimates weights and associations with phenotype simultaneously in single model
Mancuso (2019)[8] - FOCUS	External	Models multiple genes at once, as well as additional pleiotropic genetic effects on phenotype
Nagpal (2019)[18] - TIGAR	Dirichlet process regression	Multivariate model with multiple outcome phenotypes
Liu (2020)[19] - T-GEN	Spike & Slab	Incorporates epigenetic information into weight estimation process
Luningham (2020)[20] - BGW-TWAS	Spike & Slab	Models additional trans-eQTL component
Bhattacharya (2021)[21] - MOSTWAS	Elastic net BLUP	Models additional components for trans-eQTL or other molecular phenotypes
Non-linear models		
Xu (2017)[22] - ASPU	External	Uses adaptive test combining sums of powers of score statistics for different powers (includes linear model)
Zhang (2020)[23]	External	Uses adaptive test combining linear model with sum of squared score statistics
Tang (2021)[24] - 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) pape r

Table 2. Summary of results of TWAS and local genetic correlation analyses of published summary statistics for five phenotypes.

		Number	Genes	Significance	LAVA signif. associations			FUSION signif. associations	
Phenotype	Sample size[a]	of SNPs[b]	tested	threshold	r_G	TWAS	Overlap %[c]	Elast. net	LASSO
Blood pressure[33]	361K	5.94M	6,686	7.48×10^{-6}	68	107	63.6	100	109
BMI[30]	807K	6.28M	5,453	9.17×10^{-6}	134	232	57.8	297	328
Type 2 diabetes[33]	18.5K/366K	5.94M	6,686	7.48×10^{-6}	27	45	60.0	26	33
Educational attainment[31]	766K	6.18M	5,532	9.04×10^{-6}	59	108	54.6	149	165
Schizophrenia[32]	67.4K/94.0K	6.08M	5,765	8.67×10^{-6}	53	92	57.6	109	114

Genes were included for bivariate testing if they exhibited significant univariate eQTL signal in the LAVA univariate test at $\alpha_{BONF} = 0.05/14{,}584 = 3.43 \times 10^{-6}$ (see Supplemental Table 1 for results filtered on FUSION univariate test). Results were Bonferroni corrected per phenotype for the number genes tested (see *Methods - Real data analysis*). Note that all significant LAVA-rG associations were also significant in LAVA-TWAS.

[a] Showing case/control for binary phenotypes

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

[c] Percentage of significant LAVA-TWAS associations that were also LAVA-rG associations

Table 3. Summary of TWAS standard error deflation values per gene.

			Quantiles				
Phenotype	**Mean**	**Minimum**	**5%**	**25%**	**Median**	**75%**	**95%**
Blood pressure	0.93	0.69	0.84	0.90	0.94	0.97	0.99
BMI	0.88	0.46	0.75	0.84	0.89	0.94	0.98
Type 2 diabetes	0.95	0.51	0.89	0.93	0.96	0.98	1.00
Educational attainment	0.93	0.62	0.84	0.90	0.94	0.97	0.99
Schizophrenia	0.92	0.65	0.83	0.89	0.92	0.96	0.99

Based on genes with significant univariate eQTL signal in the LAVA univariate test, see Table 2. The standard error deflation is calculated as the ratio of the standard error for genes in the LAVA-TWAS model (not correcting for uncertainty in the estimate) to the corresponding standard errors in the LAVA-rG model (calibrated for testing $H_0: \mathrm{cov}(G_E, G_Y) = 0$).

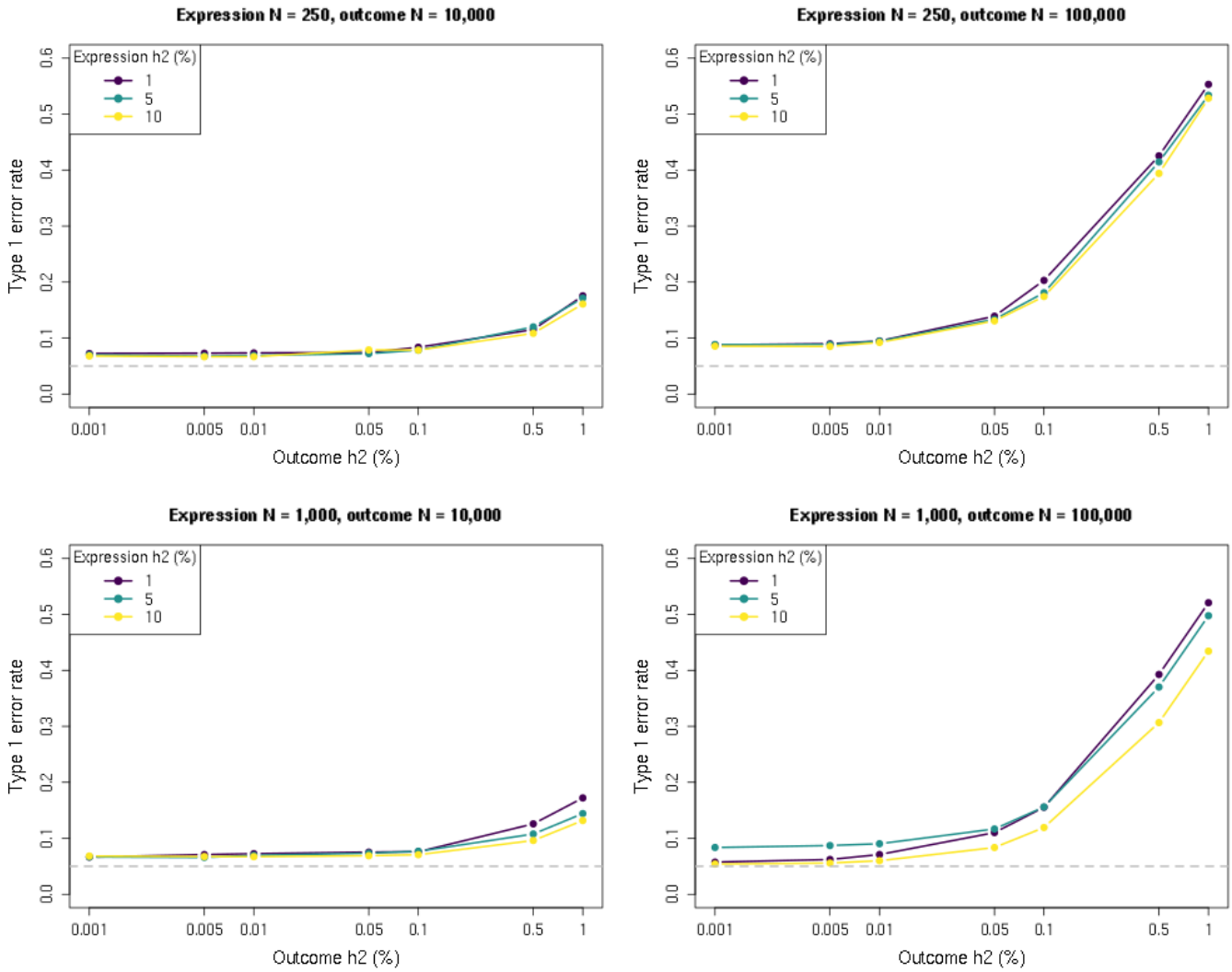

Figure 1. Type 1 error rate simulation results as a function of sample size and local heritability. *Shown is the type 1 error rate (at significance threshold of 0.05) of the LAVA-TWAS model for the null hypothesis of no genetic covariance ($cov(G_E, G_Y) = 0$), at different levels of local heritability (h2) for outcome phenotype (horizontal axis) and gene expression (separate lines). Results are shown for different sample sizes (N) for the expression (rows) and outcome (columns) data, and at 1,000 SNPs. As shown, the type 1 error rates get considerably larger with increasing sample size or local heritability for the outcome. Conversely, increasing these for the expression reduced the type 1 error rate, but the effect of this is much less pronounced.*

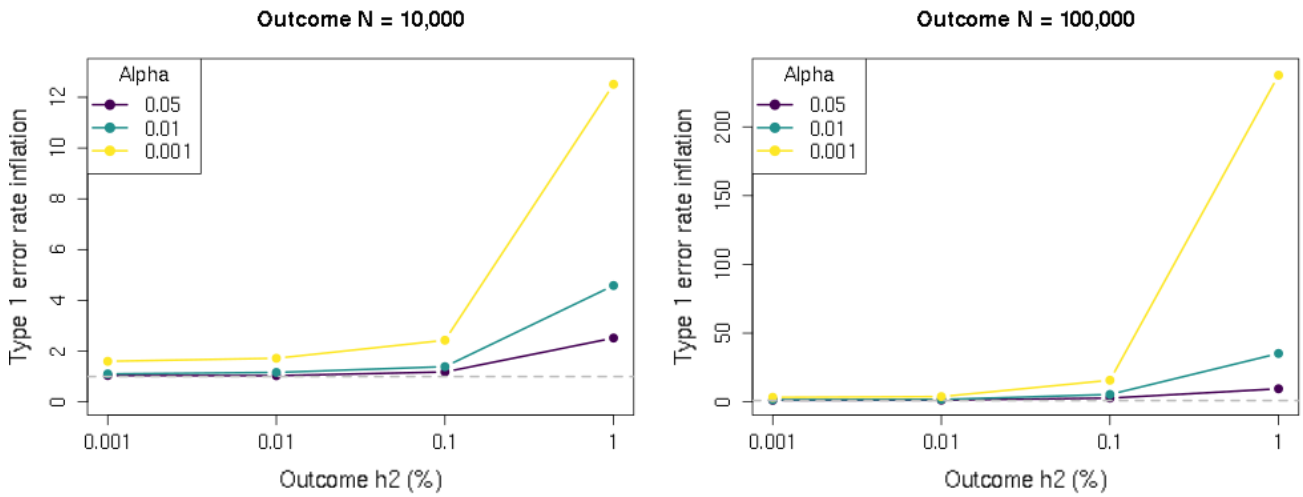

Figure 2. Type 1 error rate inflation simulations results as a function of significance threshold. *Shown is the type 1 error rate inflation of the LAVA-TWAS model for the null hypothesis of no genetic covariance ($cov(G_E, G_Y) = 0$), for different levels of significance threshold α. The type 1 error 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 shown for simulations with expression sample size of 1,000, expression local heritability of 5%, and 1,000 SNPs. As shown, the type 1 error rate becomes relatively more inflated the lower the significance threshold used (separate lines), with the difference becoming more pronounced at higher local heritability (h2) (horizontal axis) and sample size (N) (separate panels) for the outcome.*

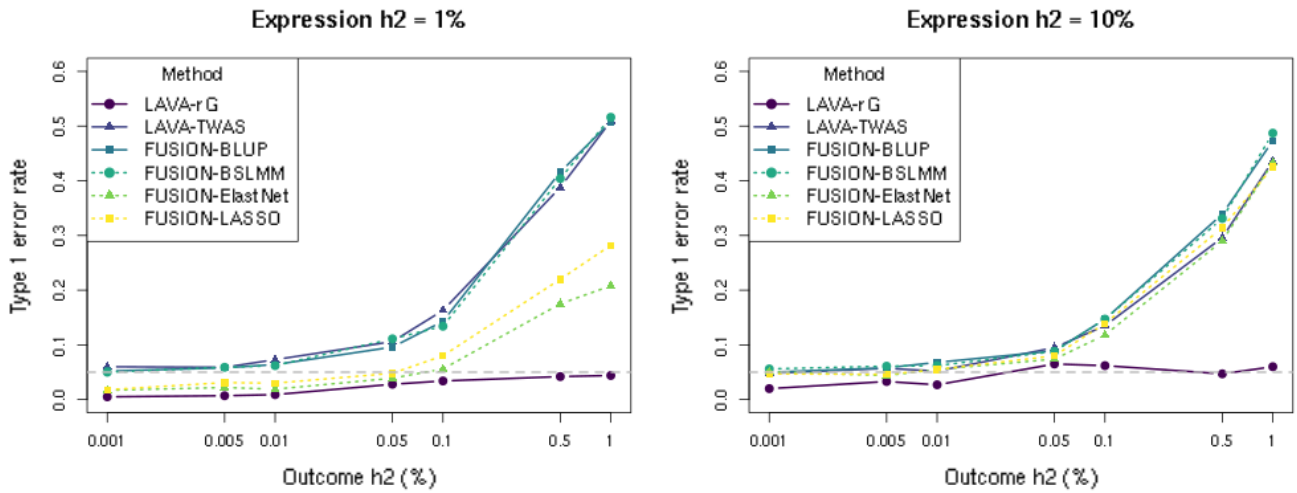

Figure 3. Comparison of type 1 error rate simulation results across models. *Shown is the type 1 error rate (at significance threshold of 0.05) of the five different TWAS models and LAVA-rG for the null hypothesis of no genetic covariance ($cov(G_E, G_Y) = 0$), at different levels of local heritability (h2) for outcome phenotype (horizontal axis) and gene expression (separate panels). Results are shown for sample sizes of 1,000 and 100,000 for the expression and outcome respectively, with 1,000 SNPs. As shown, the results are very similar across the different models. The lower type 1 error rate for Elastic Net and LASSO compared to the other TWAS models for the 1% expression heritability condition is due to these two models sometimes failing to converge at very low genetic signal for the expression (see also Supplemental Figure 4).*

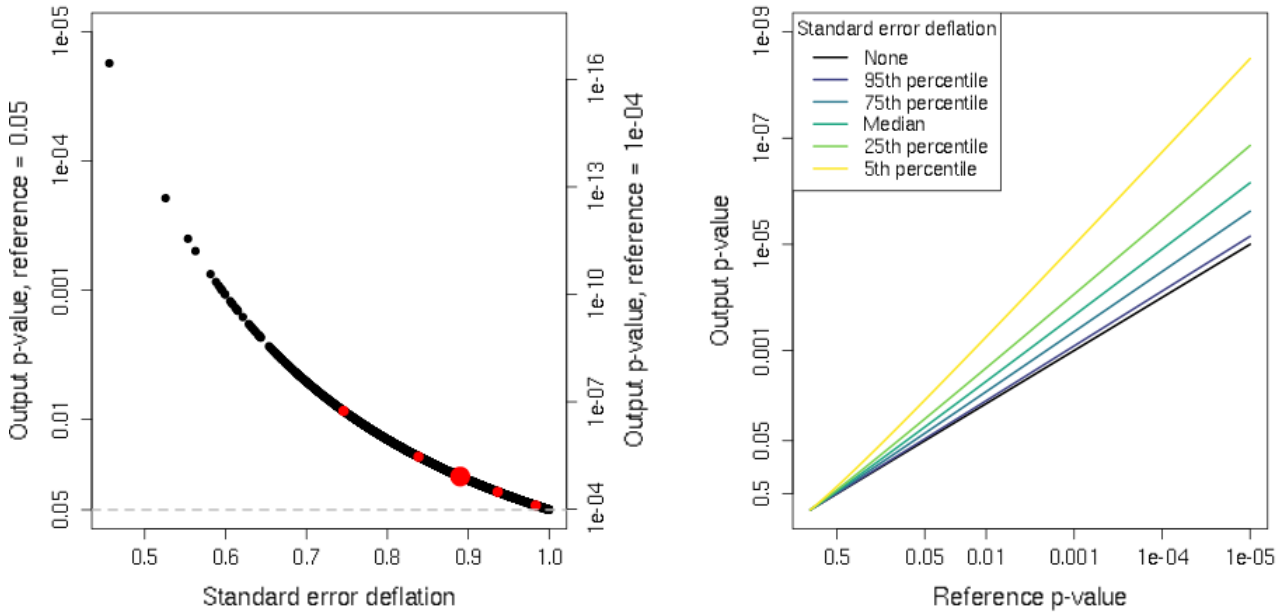

Figure 4. Illustration of p-value bias as a result of deflated standard errors. *Shown in the left panel are the p-values that will result from a LAVA-TWAS analysis at different levels of standard error deflation, when the true p-value computed with correct standard errors would either be 0.05 (left vertical axis) or 0.0001 (right vertical axis). The standard error deflation values are taken from the BMI analysis, with the large red point denoting the median and the other red points the 5th, 25th, 75th and 95th quantiles of their distribution (see Table 3). As shown, the relative bias is stronger the lower the p-value. For example, at the median standard error deflation of 0.89, at a reference p-value of 0.05 the biased p-value is about 0.028, about a 1.8-fold decrease in the p-value relative to its true value. At a reference p-value of 0.0001 however, the biased p-value is about* 1.2×10^{-5}*, an 8.1-fold decrease. In the right panel is the general relationship between the true reference p-value and the p-value that would result at different levels of standard error deflation, based on the percentiles for the BMI analysis. This further illustrates the relative increase in bias for lower p-values, with the gap with the reference p-value (black line) getting larger at lower reference p-values. Note that all axes with p-values use a -log10(p-value) scaling, but labels are expressed as p-values for ease of interpretation.*