Reconsidering the validity 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), which aim to detect relationships between gene expression and a phenotype, are commonly used for secondary analysis of genome-wide association study results. Results from 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 effect estimates, this interpretation is not justified. In this study we provide a mathematical outline of this issue, and demonstrate its implications using both simulations and real data analysis. Our simulation results show considerably inflated type 1 error rates for TWAS when testing the genetic relationship between gene expression and the phenotype, and 41% of significant TWAS associations detected in the real data analysis were found to be invalid. This demonstrates that in current implementations, TWAS is unsuitable for investigating genetic relationships between gene expression and a phenotype.

Introduction

Transcriptome-wide association studies (TWAS)[1–5] are commonly presented as an alternative to differential gene expression analysis[6], 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, which gives it a very broad scope of application.

Results from TWAS are often interpreted as indicating a genetically mediated relationship between the expression of specific genes and the phenotype of interest. However, while such knowledge would provide valuable insight in the genetic etiology of the phenotype, this interpretation of TWAS results may not be justified. This is because the traditional TWAS framework does not account for the uncertainty in the eQTL effect estimates[7,8], which effectively implies the assumption that the estimated genetic component of the gene expression represents the underlying true population value without error (an assumption known to be false), raising the question of what statistical implications this has for the TWAS framework.

In this study we address this question, using three complementary approaches. First, we dissect the mathematical structure of the TWAS framework, demonstrating that there is a discrepancy between the null hypothesis that is tested and the way TWAS results are generally interpreted. Second, we use large-scale simulations to examine the impact of this discrepancy on type 1 error rates when using TWAS to evaluate the genetic relationship between gene expression and phenotype. Finally, we show what this translates to in practice, using a real data application with five well-powered phenotypes.

Results

The mathematical structure of TWAS

TWAS is implemented as a two-stage procedure, as follows[1,7,8]. 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 more general description and details).

In practice, essentially all TWAS methods have this mathematical structure, though they differ in their implementation, particularly in how $\hat{\alpha}_E$ is estimated[7–21]. As the number of SNPs in X is usually much larger than the sample size of the gene expression data, 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 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.

The TWAS null hypothesis

The conventional interpretation of TWAS as testing a genetic relationship between gene expression and the phenotype is intuitively sensible, given that it appears to extract the genetic component of the gene expression in equation (1) and tests the association between this genetic component G_E and the phenotype in equation (2). This interpretation would be valid if TWAS used the true genetic component G_E as the predictor in equation (2), but since G_E itself is unknown, the analysis instead uses the estimate $\hat{G}_E$. Being an estimate, the value of $\hat{G}_E$ is subject to uncertainty, being distributed around G_E akin to a form of measurement error. Yet this is not accounted for in equation (2), since regression models define a conditional distribution of the outcome variable given the predictors, with the predictors themselves assumed to be fixed.

Mathematically, we can break this down as follows. Firstly, we can observe that phenotype Y can be partitioned into a local genetic component G_Y based on X and a residual ξ_Y, analogous to G_E in model (1), such that

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

with $G_Y = X\alpha_Y$, and with ξ_Y being independent of X. Secondly, we note that the true value of coefficient β in equation (2) is a direct function of the covariance $\mathrm{cov}(\hat{G}_E, Y)$, with $\beta = \frac{\mathrm{cov}(\hat{G}_E, Y)}{\mathrm{var}(\hat{G}_E)}$. From this it follows that testing the null hypothesis $H_0: \beta = 0$ is equivalent to testing $H_0: \mathrm{cov}(\hat{G}_E, Y) = 0$. Putting this together, the covariance $\mathrm{cov}(\hat{G}_E, Y)$ can be rewritten as $\mathrm{cov}(\hat{G}_E, Y) = \mathrm{cov}(\hat{G}_E, G_Y + \xi_Y) = \mathrm{cov}(\hat{G}_E, G_Y) + \mathrm{cov}(\hat{G}_E, \xi_Y)$. Since $\hat{G}_E$ is a linear transformation of X, and X is independent of ξ_Y, we have $\mathrm{cov}(\hat{G}_E, \xi_Y) = 0$, showing that the TWAS null hypothesis is further equivalent to $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$.

Crucially, this covariance term $\mathrm{cov}(\hat{G}_E, G_Y)$ is a sample-dependent quantity, and its interpretation is therefore intrinsically tied to the gene expression sample from which $\hat{\alpha}_E$ was originally estimated. The test of $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$ therefore does not allow for any inference on the fully population-level parameter $\mathrm{cov}(G_E, G_Y)$ in which we are actually interested. This true genetic covariance relates to $\mathrm{cov}(\hat{G}_E, G_Y)$ as $\mathrm{cov}(\hat{G}_E, G_Y) = \mathrm{cov}(G_E, G_Y) + \hat{C}$, with $\hat{C}$ an unknown, non-zero value specific to the gene expression sample. Evaluating the TWAS null hypothesis $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$ is thus equivalent to testing the true genetic covariance $\mathrm{cov}(G_E, G_Y)$ against an unknown value $-\hat{C}$, ie. $H_0: \mathrm{cov}(G_E, G_Y) = -\hat{C}$, with $\hat{C}$ non-zero unless none of the SNPs in X are associated with Y (see *Supplemental Information - Derivation of the TWAS null hypothesis*).

This can be further illustrated by analogy to the t-test. Suppose we are interested in testing whether the population means μ_A and μ_B of some variable for groups A and B differ, yet for B we simply compute the sample mean $\hat{\mu}_B$ and proceed to test $H_0: \mu_A = \hat{\mu}_B$ using a single sample t-test, ignoring the fact that $\hat{\mu}_B$ is merely a noisy estimate of μ_B. Further suppose that for the true means μ_A and μ_B are both 100, and in our sample for B the estimate $\hat{\mu}_B$ equals 110. In this case our null hypothesis implicitly equates to $H_0: \mu_A = \mu_B + 10$, and since in reality $\mu_A = \mu_B$, with a sufficiently large sample for A we are likely to get a significant result and reject $H_0: \mu_A = \hat{\mu}_B$. But clearly, no matter how small our p-value, we cannot validly conclude from this that the population means of A and B differ, since this is not what we tested. Only accounting for the noise in $\hat{\mu}_B$ and performing a test of $H_0: \mu_A = \mu_B$ directly, would allow for such a conclusion.

Although the test that TWAS performs is still statistically valid for certain applications, the extent to which biologically relevant conclusions from it is very limited, and does not permit inference about any genetic relationship between the gene expression and the outcome phenotype. Essentially, the estimate $\hat{G}_E$ merely represents a sum of the SNPs in X, weighted by the elements of $\hat{\alpha}_E$. If we substitute $\hat{G}_E = X\hat{\alpha}_E$ in equation (2) we obtain

$$Y = X\hat{\alpha}_E\beta + \varepsilon_Y, \tag{4}$$

which shows that TWAS reduces to a constrained version of the general multiple regression model shown in equation (3), with coefficient vector α_Y proportional to weight vector $\hat{\alpha}_E$, ie. $\alpha_Y = \hat{\alpha}_E\beta$. The TWAS null hypothesis $\beta = 0$ thus implies the multiple regression null hypothesis $\alpha_Y = 0$, and vice versa. Like multiple regression, TWAS therefore only provides a test of whether the SNPs in X are jointly associated with Y, which does not warrant any conclusions about a role of gene expression in those genetic associations (see *Supplemental Information - Interpreting the invalid TWAS associations*).

Evaluating genetic covariance

For TWAS results to be used as they are generally interpreted, requires that the TWAS framework is modified to test the null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$ of no local genetic covariance between gene expression and the outcome phenotype. Although it would in principle be possible to alter existing TWAS methods accordingly, this is likely to be challenging in practice given that the penalized regression models commonly used for fitting equation (1) cause the distribution of $\hat{\alpha}_E$, and therefore that of $\hat{G}_E$, to acquire a complex mathematical form. At present, the only TWAS method that does explicitly model the uncertainty in $\hat{G}_E$ is CoMM[13], yet, due to specific constraints in its model structure, it still suffers from very similar issues as other TWAS methods (see *Supplemental Information - Comparison with the CoMM model* and results below).

A simpler alternative to using penalized regression is to project the genotype matrix X onto a pruned set of principal components, then fit a regression of E onto those principal components instead using Ordinary Least Squares. Although this approach is not as statistically efficient as the penalized regression models, it yields a much more tractable multivariate normal distribution of $\hat{G}_E$. This approach is already implemented by local genetic correlation analysis methods such as LAVA[22] and rho-HESS[23], which allows them to fully account for the uncertainty in $\hat{G}_E$ and directly test $H_0: \mathrm{cov}(G_E, G_Y) = 0$ (see *Methods – local genetic correlation*).

Finally, a more straightforward solution would be to simply assume that $\hat{G}_E = G_E$ and treat existing TWAS methods as if they are testing the null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$. Although this assumption does not hold, and omitting a source of variance generally results in a downward bias in p-values and inflated type 1 error rates due to underestimation of the standard errors, it is possible that the uncertainty in $\hat{G}_E$ is small enough that these issues are negligible in practice. This would make the distinction between testing $H_0: \mathrm{cov}(\hat{G}_E, G_Y) = 0$ and $H_0: \mathrm{cov}(G_E, G_Y) = 0$ would be a merely theoretical concern.

In order to evaluate to what extent making this assumption is problematic, we implemented a traditional TWAS model into the LAVA framework (denoted LAVA-TWAS), based on the same pruned principal components approach used in LAVA. This LAVA-TWAS model only differs from the local genetic correlation analysis (denoted LAVA-rG) in ignoring the uncertainty in $\hat{G}_E$, in the same way as existing TWAS methods, but is otherwise identical in model and implementation (see *Methods - LAVA implementation of TWAS*). As such, both models compare the same test statistic to a normal sampling distribution, but with the LAVA-TWAS standard errors always smaller than those of LAVA-rG, and its p-values will therefore always be lower as well.

Since the models are otherwise identical, any difference in results between the two must therefore inherently be due to LAVA-TWAS failing to account for the uncertainty in $\hat{G}_E$. Moreover, as LAVA-rG has been validated for testing $H_0: \mathrm{cov}(G_E, G_Y) = 0$ and has well-controlled type 1 error rates when doing so (see Werme (2022)[22] for details), we can directly quantify the underestimation in the LAVA-TWAS standard errors relative to this null hypothesis. It also allows LAVA-rG to serve as a baseline for LAVA-TWAS in the real data analyses, to determine how many invalid associations LAVA-TWAS yields because of these deflated standard errors.

Simulation study

To evaluate the performance of TWAS when interpreted as a test of the genetic relationship between gene expression and outcome phenotype, we performed a series of simulations, generating gene expression and phenotype values under the corresponding null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$ and computed type 1 error rates. We used genotype data from the UK Biobank[24,25] as a basis for the simulations, varying the number of SNPs per simulated gene, and the sample size and local heritability of the gene expression and outcome phenotype data (see *Methods - Simulation study*). Analysis was performed using LAVA-TWAS and LAVA-rG, as well as four different models in FUSION (BLUP, Bayesian LMM, Elastic Net, LASSO) which are representative of the range of different penalized regression models found in existing TWAS methods (see Table 1), and CoMM.

The simulations showed considerable inflation of type 1 error rates ($\alpha = 0.05$) when using LAVA-TWAS to evaluate $H_0: \mathrm{cov}(G_E, G_Y) = 0$, depending on the simulation condition. The inflation strongly increased with larger local heritability or sample size for the outcome phenotype (Figure 1), with the local heritability also interacting with the number of SNPs in the gene (Supplemental Figure 1). In addition, some decrease in the type 1 error rate inflation was also found with greater local heritability or sample size for the gene expression, but this effect is much less pronounced. Moreover, the type 1 error rate inflation also became progressively worse when evaluating it at stricter significance thresholds than the α of 0.05 used so far (Figure 2). Finally, type 1 error rates also increased when filtering out genes that exhibit insufficient genetic signal for the gene expression, as is general practice in TWAS analysis (Supplemental Figure 2).

Results for the four FUSION penalized regression models are largely the same as for LAVA-TWAS (Figure 3), suggesting that the type of model and penalization used for the estimation of $\hat{\alpha}_E$ does not strongly affect the subsequent type 1 error rate inflation. Although the Elastic Net and LASSO models showed lower inflation than other models in the 1% gene expression heritability conditions, this can be largely explained by the fact that they frequently failed to converge in these conditions when estimating equation (1). This difference was no longer present in the 5% heritability conditions (see Supplemental Figure 3). Similar patterns of type 1 error rate inflation were found for CoMM as well and were generally more pronounced, though the influence of the gene expression heritability was more complex than for the other methods (Supplemental Figure 4). As in the validation simulations in Werme (2022)[22], type 1 error rates for LAVA-rG were found to be well-controlled, further underscoring the importance of modelling the uncertainty in $\hat{G}_E$ in order to assess genetic relationships between gene expression and the outcome phenotype (Figure 3).

Overall, the results of our simulations demonstrate that the failure to account for the uncertainty in $\hat{G}_E$ can cause considerable inflation of the type 1 error rates in many conditions, with the strong effect of the sample size and heritability of the outcome phenotypes, which is in line with the interpretation of TWAS as a constrained joint association test as in equation (4). Of particular concern is that the inflation increases both when filtering on the univariate signal of the gene expression as well as when using lower significance thresholds, both of which are likely to apply in practice and often at more stringent levels than shown in our simulations as well. See *Supplemental Information - Discussion of simulation results* for more discussion of the findings from these simulations.

235 *Real data analysis*

236 To gauge the magnitude of the impact this issue has in a real data context, we applied TWAS to GWAS
237 of five well-powered phenotypes[26–29] (see Table 2), with eQTL data for blood gene expression from
238 GTEx[30] (v8). A total of 14,584 genes were analysed, performing TWAS analysis for those genes for which
239 the univariate gene expression signal was significant at $\alpha_{\mathrm{BONF}} = 0.05/14{,}584 = 3.43 \times 10^{-6}$ (see
240 *Methods - Data* and *Methods - Real data analysis*). Here, the data was analysed with LAVA-TWAS, as
241 well as the Elastic Net and LASSO models in FUSION. To serve as comparison, the data was analysed
242 using LAVA-rG as well.

243 A summary of the results is given in Table 2 and as shown, the number of significant LAVA-
244 TWAS associations is considerably higher than those for LAVA-rG, with on average only about 59% of
245 LAVA-TWAS associations confirmed by LAVA-rG. Because LAVA-rG is calibrated for testing
246 $H_0\colon \mathrm{cov}(G_E, G_Y) = 0$, it follows that the remaining 41% of LAVA-TWAS associations cannot validly be
247 declared to be significant if the TWAS analysis is interpreted as providing a test of the genetic
248 relationship between gene expression and phenotype. The number of associations found when using
249 FUSION are at a comparable level as LAVA-TWAS (see also Supplemental Table 1), and taken together
250 with their very similar type 1 error rate inflation levels in the simulations as well, this confirms that
251 TWAS using penalized regression yields similar rates of invalid associations. Note that the invalid
252 associations found in these analyses do not all necessarily reflect false positives, and some of these
253 associations will likely become genuinely significant if sample sizes are increased such that sufficient
254 statistical evidence for their association becomes available.

255 We also evaluated results for genes with a univariate p-value for the gene expression greater
256 than 0.05, to illustrate that TWAS still yields significant associations even when the expression for a
257 gene has no discernible genetic component at all (Supplemental Table 2). As shown in this table,
258 virtually no local genetic correlations are found with LAVA-rG for these genes, yet results for the TWAS
259 analyses are at a similar level as in the primary analysis in Table 2. Although these genes would typically
260 not be included in a TWAS analysis, the fact that TWAS yields a considerable number of significant
261 associations in this context, when $\hat{\alpha}_E$ reflects essentially just noise, further illustrates how TWAS
262 functions as a form of joint association test as expressed in equation (4), and its results do not
263 necessarily reflect any genetic relationship between gene expression and the outcome phenotype (see
264 also *Supplemental Information - Interpreting the invalid TWAS associations*).

265 As previously mentioned, the failure to account for the uncertainty in $\hat{G}_E$ results in a deflation
266 of the standard errors. This deflation is the cause of the inflated type 1 error rates observed in the
267 simulations, but more generally results in a downward bias of the p-values for all genes, regardless of
268 their true association status. Because for LAVA-TWAS the corrected standard errors are known, we can
269 compute this standard error deflation for each gene in our analysis. These are summarized in Table 3

270 and, as shown, they vary considerably across genes, with negligible deflation for some genes but severe
271 deflation for others. Figure 4 illustrates the relation between this standard error deflation and the
272 resulting bias in p-values, using the BMI results as an example, with Supplemental Figure 5 further
273 showing the level of realized bias. This high variability of standard error deflation across genes implies
274 that this issue must be addressed at the level of the individual gene, and cannot be resolved by applying
275 a general post-hoc correction to the gene p-values.
276
277
278 **Discussion**
279
280 TWAS has commonly been used with the aim of evaluating genetic relationships between gene
281 expression and a phenotype of interest. But in this paper we have shown that the TWAS framework in
282 its existing form is not well-suited to this aim. Because TWAS does not account for the uncertainty in
283 the estimated genetic component $\hat{G}_E$ of the gene expression, the standard errors are underestimated.
284 This results in potentially severe downward bias in p-values and inflation of type 1 error rates when
285 used for this purpose, the situation is complicated further by the fact that different genes are affected
286 to different degrees by this issue.
287 What this means is that TWAS p-values cannot be relied upon to draw any conclusions about
288 a genetic relationship between gene expression and the phenotype, because as demonstrated these
289 p-values do not accurately reflect the level of statistical evidence in the data for such a relationship.
290 Moreover, our results demonstrate that this is issue is not merely theoretical and considerably affects
291 analysis in a wide range of conditions, making results of TWAS studies difficult to interpret and use in
292 practice . This is further illustrated by the fact that in our real data analysis, TWAS still yielded a
293 considerable number of significant associations even when genetic signal for the gene expression was
294 absent altogether.
295 Although the issue of uncertainty in eQTL estimates has occasionally been mentioned in TWAS
296 literature[1,13,31,32], its implications for the validity of the TWAS framework have received little scrutiny
297 thus far. Moreover, while various TWAS methods have included simulations to evaluate type 1 error
298 rates, these all either treated $\hat{G}_E$ as fixed in the simulations just as in the TWAS model itself[11,20], or only
299 simulated under a null scenario of Y having no genetic component at all (ie. $G_Y = 0$)[13,19,21]. This explains
300 why the issues demonstrated in this paper have not been uncovered before.
301 Addressing this issue with TWAS requires fully accounting for the uncertainty of the estimate
302 $\hat{G}_E$ in relation to the true genetic component G_E. While local genetic correlation analysis offers one
303 way to do so, it is in principle possible to incorporate this into existing TWAS models as well. In practice,

304 the feasibility of implementing this will strongly depend on the approach used to fit the model in
305 equation (1), as the distribution of the noise component induced by penalized regression models can
306 take mathematically very complex forms. An alternative is to employ other methods like
307 colocalization[33,34] and Mendelian Randomization (MR)[35–37], which, although they use quite different
308 model assumptions, still test null hypotheses that can answer similar research questions as TWAS is
309 intended to do. Of note as well is that the issue with TWAS described in this paper has a clear analogy
310 to the issue of the "no measurement error" (NOME) assumption in MR analysis[38], and strategies used
311 to remove this NOME assumption in the MR context may also help inform solutions for addressing the
312 issue in TWAS methods.

313 Investigating genetic relationships between gene expression and phenotypes, such as TWAS is
314 generally intended to do, has the potential to provide valuable insights into the genetic etiology of our
315 phenotypes. But this requires more robust models than those currently implemented in existing TWAS
316 methods, that is, models which can fully account for all the uncertainty in the data. Further
317 development of existing TWAS methods to account for this issue, or adopting the use local genetic
318 correlation analysis methods, will therefore be essential to support future investigations of the
319 relationship between gene expression and phenotypes.

Methods

Ethics statement

This study relied on simulated data based on the UK Biobank genotype data[24], as well as secondary analysis of publicly available summary statistics. Ethical approval for this data was obtained by the primary researchers.

Outline of TWAS framework

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

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

and

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

with the estimated genetic component $\hat{G}_E = X\hat{\alpha}_E$, 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. 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 for 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[19–21] 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* for more details.

Local genetic correlation

The LAVA implementation of local genetic correlation analysis has been described in detail in Werme et al. (2022)[22]. 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 $G_E = W\delta_E$ and $G_Y = W\delta_Y$, where δ_E and δ_Y are the genetic effect size vectors for these principal 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 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 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 δ_E and δ_Y and its estimates, rather than working with α_E and α_Y explicitly.

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

The estimates $\hat{\delta}_E$ and $\hat{\delta}_Y$ are obtained by reconstructing multiple linear regressions from the input summary statistics (see Werme et al. (2022)[22] for details). This uses two separate equations of 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 $\widehat{\omega}_{EY}$ has an expected value of zero and a variance of $V_{\mathrm{rG}} = K\hat{\sigma}_E^2\hat{\sigma}_Y^2 + \widehat{\omega}_E^2\hat{\sigma}_Y^2 + \hat{\sigma}_E^2\widehat{\omega}_Y^2$. 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 normal approximation of the sampling distribution of $\widehat{\omega}_{EY}$. For the simulations and real data analysis we computed LAVA-rG p-values in both ways, and as shown in Supplemental Figure 6 these were found to be virtually identical. As such, all LAVA-rG results reported in this paper are based on p-values computed using the distribution $\widehat{\omega}_{EY} \sim \mathrm{N}(0, V_{\mathrm{rG}})$, to provide a mathematically more direct comparison for the LAVA-TWAS implementation.

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 $\widehat{\mathrm{cov}}(\hat{G}_E, Y)$, the sample estimate of $\mathrm{cov}(\hat{G}_E, Y)$. As both $\hat{G}_E$ and Y have means of zero, and since $\hat{G}_E = W\hat{\delta}_E$, we have $\widehat{\mathrm{cov}}(\hat{G}_E, Y) = \frac{\hat{G}_E^T Y}{N-1} = \frac{\hat{\delta}_E^T W^T Y}{N-1} = \hat{\delta}_E^T \frac{W^T Y}{N-1}$. As previously derived $\hat{\delta}_Y = \frac{W^T Y}{N-1}$, and it therefore follows that $\widehat{\mathrm{cov}}(\hat{G}_E, Y) = \hat{\delta}_E^T \hat{\delta}_Y$. This $\widehat{\mathrm{cov}}(\hat{G}_E, Y)$ is therefore equal to the genetic covariance estimate $\widehat{\omega}_{EY}$ used by LAVA-rG, and thusboth LAVA-TWAS and LAVA-rG are shown to use the same test statistic $\hat{\delta}_E^T \hat{\delta}_Y = \widehat{\omega}_{EY}$.

Under the TWAS null hypothesis of H_0: $\mathrm{cov}(\hat{G}_E, Y) = \mathrm{cov}(\hat{G}_E, G_Y) = 0$, the distribution of $\widehat{\omega}_{EY}$ depends entirely on the sampling distribution of $\hat{\delta}_Y$, since $\hat{G}_E$ and therefore $\hat{\delta}_E$ is considered fixed. The sampling distribution of $\widehat{\omega}_{EY}$ under this null hypothesis thus has the form $\mathrm{N}(0, V_{\mathrm{TWAS}})$, with $V_{\mathrm{TWAS}} = \hat{\delta}_E^T \hat{\delta}_E \hat{\sigma}_Y^2$. From the LAVA-rG model we have $\widehat{\omega}_E^2 = \hat{\delta}_E^T \hat{\delta}_E - K\hat{\sigma}_E^2$ and therefore $\hat{\delta}_E^T \hat{\delta}_E = K\hat{\sigma}_E^2 + \widehat{\omega}_E^2$, from which it follows that $V_{\mathrm{TWAS}} = (K\hat{\sigma}_E^2 + \widehat{\omega}_E^2)\hat{\sigma}_Y^2 = K\hat{\sigma}_E^2\hat{\sigma}_Y^2 + \widehat{\omega}_E^2\hat{\sigma}_Y^2$ and thus $V_{\mathrm{rG}} = V_{\mathrm{TWAS}} + \hat{\sigma}_E^2\widehat{\omega}_Y^2$. Because $\hat{\sigma}_E^2\widehat{\omega}_Y^2$ is the product of two variance terms and therefore by definition positive, V_{TWAS} is always smaller than V_{rG} and will there yield inherently lower p-values as well.

As such, the only difference between LAVA-TWAS and LAVA-rG is the variance of the sampling distribution against which the test statistic $\widehat{\omega}_{EY}$ is compared to obtain the p-value. Moreover, because V_{rG} has a known mathematical form we can also directly compute the deflation of the LAVA-TWAS standard errors relative to their correct value for testing H_0: $\mathrm{cov}(G_E, G_Y) = 0$, as $\sqrt{V_{\mathrm{TWAS}}/V_{\mathrm{rG}}} = \sqrt{1 - \hat{\sigma}_E^2\widehat{\omega}_Y^2/(K\hat{\sigma}_E^2\hat{\sigma}_Y^2 + \widehat{\omega}_E^2\hat{\sigma}_Y^2 + \hat{\sigma}_E^2\widehat{\omega}_Y^2)}$.

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 genetic association $\hat{\delta}_{Yj}$ by the corresponding eQTL estimate $\hat{\delta}_{Ej}$. This implementation is therefore analogous to how TWAS is performed using GWAS summary statistics in other TWAS methods (eg.

451 Gusev (2016)[8]), except defined in terms of the estimated genetic associations of principal component
452 matrix W rather than the original SNP genotype matrix X.
453
454
455 *Data*
456 Genotype data for simulations came from the UK Biobank [24], a national genomic study of over 500,000
457 volunteer participants in the UK. The study was approved by the National Research Ethics Service
458 Committee North West–Haydock (reference 11/NW/0382) and data were accessed under application
459 #16406. Data collection, primary quality control, and imputation of the genotype data were performed
460 by the UK Biobank[25]. Data was converted to hardcalled genotypes, and then additionally filtered, using
461 only SNPs with an info score of at least 0.9 and missingness no greater than 5%, and restricting the
462 sample to unrelated, European-ancestry individuals with concordant sex. Remaining missing genotype
463 values were then mean-imputed to ensure consistency of the input data across different methods.
464 The European panel of the 1,000 Genomes[39] data (N = 503, as downloaded from
465 https://ctg.cncr.nl/software/magma) was used as genotype reference data to estimate LD for the real
466 data analysis for both LAVA and FUSION, removing SNP with a minor allele frequency (MAF) below
467 0.5%. For eQTL data we used the GTEx[30] data (v8, European subset), for whole blood gene expression.
468 The published cis-eQTL summary statistics (European ancestry) from GTEx Portal were used for the
469 LAVA analysis, pre-computed FUSION model files were used for the FUSION analysis. For every
470 analysed gene, we included all SNPs in the data within one megabase of the transcription start site.
471 Genes were filtered to include only autosomal protein-coding and RNA genes expressed in blood, and
472 only genes also available in the FUSION model files were used, resulting in a total of 14,584 genes
473 available for analysis.
474 GWAS summary statistics were selected for five well-powered phenotypes, chosen to reflect
475 a range of different domains. These were BMI (GIANT)[26] (no waist-hip ratio adjustment), educational
476 attainment (SSGAC)[27], schizophrenia (PGC, wave 3)[28], diastolic blood pressure (GWAS Atlas)[29] and type
477 2 diabetes (GWAS Atlas)[29]. Sample size and number of SNPs for each sample can be found in Table 2.
478
479
480 *Simulation study*
481 Simulation studies were performed to evaluate the type 1 error rates of TWAS when used to evaluate
482 the null hypothesis $H_0: \mathrm{cov}(G_E, G_Y) = 0$. The simulations were based on the UK Biobank data,
483 generating gene expression and outcome phenotypes while varying their sample sizes and the local
484 heritabilities, as well as the number of SNPs per gene. The number of SNPs K in the genes was varied
485 across values of 100, 500, 1,000 and 2,500 SNPs; sample sizes of either 250 or 1,000 were used for the

gene expression (denoted N_E), and 10,000, 50,000 or 100,000 for the outcome phenotype (denoted N_Y); and local heritabilities, defined as the percentage of variance explained by the SNPs in the gene, were set at either 1%, 5% or 10% for the gene expression (denoted h_E^2)[40], and 0.001%, 0.005%, 0.01%, 0.05%, 0.1%, 0.5% or 1% for the outcome phenotype (denoted h_Y^2). The number of SNPs K was only varied at sample size settings of $N_E = 1{,}000$ and $N_Y = 10{,}000$, and the sample sizes were only varied at $K = 1{,}000$. All combinations of the two local heritabilities were simulated for each setting of K and sample sizes. The simulations were analysed using LAVA-TWAS, LAVA-rG, the four FUSION models and CoMM.

To create the samples to use, we first drew a random subset of 100,000 individuals from the UK Biobank data. For each sample size N used in the simulations, the first N individuals from this subset were used for that sample, to ensure that samples of different sizes are all nested. The MAF for all SNPs was computed based on the $N = 100$ subsample, discarding all SNPs with MAF below 5%. Since the subsamples are nested, this ensures that all SNPs have a sufficient minor allele count at every sample size used. Ten blocks of 2,500 consecutive SNPs were then selected to represent simulated genes, one block each from around the middle of the first ten chromosomes (avoiding the centromere region). The number of SNPs K was varied across conditions by using only a subset of the these 2,500 SNPs, always selecting the first K SNPs from each block. For each simulation condition, the total number of iterations for that condition was evenly distributed over these ten blocks, and results were aggregated over the blocks. Type 1 error rates per condition were computed as the proportion of iterations for that condition for which the p-value was smaller than the significance threshold, which was set to 0.05 unless otherwise specified.

To simulate gene expression E and outcome phenotype Y for the standardized genotypes X of the SNPs in a gene, X 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). For each iteration, true genetic effect sizes δ_E and δ_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, then regressing one vector on the other and retaining only the residuals for the outcome vector to ensure that δ_E and δ_Y were exactly independent. The two genetic components were then computed as $G_E = W\delta_E$ and $G_Y = W\delta_Y$, and both were subsequently standardized to a variance of one.

Simulated gene expression and phenotype values were then generated as $E = W\delta_E + \xi_E$ and $Y = W\delta_Y + \xi_Y$, drawing the residuals ξ_E and ξ_Y from normal distributions with variance parameters 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

519 Y were then regressed on each individual SNP in X using simple linear regression, and the resulting
520 test statistics as well as the simulated E and Y themselves were stored for subsequent analysis.
521 To compare the different methods, a baseline set of simulation conditions with 1,000
522 iterations, with $K = 1{,}000$, $N_E = 1{,}000$ and the remaining parameters across their full range, was
523 generated and analysed with all methods: LAVA-TWAS, LAVA-rG, FUSION-BLUP, FUSION-BSLMM,
524 FUSION-ElastNet, FUSION-LASSO and CoMM. LAVA analyses were performed using the summary
525 statistics as input, FUSION analyses with the raw gene expression values for the first stage and outcome
526 phenotype summary statistics for input, and CoMM using the raw values for both gene expression and
527 outcome phenotype. For the first stage of FUSION, cross-validation was turned off and the true $R_E^2 =$
528 $\frac{h_E^2}{100}$ value was set using the --hsq_set option. In the CoMM analyses, the model was found to run into
529 convergence issues due to K being equal to N_E, and the simulations were rerun setting K to 100
530 instead to resolve this. Moreover, an additional set of simulations was performed with h_e^2 values of
531 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%.
532 For LAVA-TWAS, further of simulations with 10,000 iterations were performed across the full
533 range of parameter settings for analysis, to evaluate the impact of these parameters on type 1 error
534 rates. To determine the effect of filtering genes on the presence of eQTL signal, as is common practice
535 in TWAS analyses, for these simulations we also computed the LAVA univariate p-value for the gene
536 expression. In addition to the regular type 1 error rates, filtered type 1 error rates were also computed
537 based on only the iterations for which the univariate p-value was below either 0.05 or 0.0001. Finally,
538 to allow for reliable computation of type 1 error rates at lower significance thresholds, for the subset
539 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%
540 the number of iterations was increased to 100,000. Note that although for LAVA-TWAS the number of
541 iterations differs from the other methods and varies across conditions, for each plot in this paper the
542 type 1 error rate values within that plot are always based on the same number of iterations.
543
544
545
546 *Real data analysis*
547 TWAS and local genetic correlation analyses were performed on the GWAS data as follows, separately
548 for each of the five phenotypes. In all the analyses, SNP filtering was applied to remove all SNPs with
549 a minor allele frequency below 0.5%, and only SNPs available in the 1,000 Genomes data, the GTEx
550 data, and the GWAS sample for that phenotype were used.
551 For every gene, univariate LAVA analysis for the eQTL signal, LAVA-rG and LAVA-TWAS
552 analyses, and FUSION analyses with the Elastic Net and LASSO models was performed. For FUSION, the

553 BLUP and BSLMM models were not used as these were not available in the precomputed model files.
554 FUSION univariate p-values for the eQTL signal were also obtained from these model files. Note that
555 separate SNP filtering was applied prior to these models being computed, and as such the LAVA and
556 FUSION analyses are based on different sets of SNPs per gene.
557 For the primary analyses, the bivariate LAVA and FUSION analyses were only performed for
558 genes for which the LAVA univariate tests were significant after correcting for the total number of
559 genes, at a threshold of $0.05/14{,}584 = 3.43 \times 10^{-6}$ (see Supplemental Table 1 for results when using
560 the FUSION univariate p-values instead). The significance threshold for the bivariate analyses was set
561 separately for each phenotype, at a Bonferroni-correction for the total number of genes that was
562 univariate significant for that phenotype (see Table 2). In a secondary analysis, to evaluate TWAS
563 results in the absence of gene expression signal, bivariate analyses were performed for all genes for
564 which the LAVA univariate p-value was greater than 0.05, again setting the significance threshold per
565 phenotype for the number of genes with univariate p-value greater than 0.05 for that phenotype.
566

Data availability

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

Code availability

LAVA is implemented as an R package, which is publicly available at the LAVA website (https://ctg.cncr.nl/software/lava) and LAVA GitHub repository (https://github.com/josefin-werme/LAVA). GTEx v8 input files compatible with LAVA are available from the LAVA website as well. Version 0.1.0 of LAVA was used for the LAVA analyses in this paper. FUSION can be obtained from http://gusevlab.org/projects/fusion/ and CoMM from https://github.com/gordonliu810822/CoMM. Scripts used for the simulations and analyses performed for this paper are available from https://github.com/cadeleeuw/twas-validity-scripts2022.

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. Hoffmann-La Roche AG. WJP was funded by an NWO Veni grant (NWO: 916-19-152). JES was supported by an NWO Veni grant (NWO: 201G-064). The research has been conducted using the UK Biobank Resource (application no. 16406). 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 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. 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).
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. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
9. 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).
10. 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).
11. 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).
12. Hu, Y. *et al.* A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* **51**, 568–576 (2019).
13. 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).
14. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. 22 (2019).
15. 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).

16. Liu, W. *et al.* Leveraging functional annotation to identify genes associated with complex diseases. *PLOS Comput. Biol.* **16**, e1008315 (2020).
17. 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).
18. Bhattacharya, A., Li, Y. & Love, M. I. MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. *PLOS Genet.* **17**, e1009398 (2021).
19. Xu, Z., Wu, C., Wei, P. & Pan, W. A Powerful Framework for Integrating eQTL and GWAS Summary Data. *Genetics* **207**, 893–902 (2017).
20. 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).
21. 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).
22. 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).
23. 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).
24. 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).
25. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
26. 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).
27. 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).
28. 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.
29. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
30. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
31. Xue, H. & Pan, W. Some statistical consideration in transcriptome-wide association studies. *Genet. Epidemiol.* **44**, 221–232 (2020).
32. Zhu, H. & Zhou, X. Transcriptome-wide association studies: a view from Mendelian randomization. *Quant. Biol.* (2020) doi:10.1007/s40484-020-0207-4.

33. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).
34. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
35. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
36. Yuan, Z. *et al.* Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat. Commun.* **11**, 3861 (2020).
37. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 3300 (2019).
38. Bowden, J. *et al.* Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int. J. Epidemiol.* **48**, 728–742 (2019).
39. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
40. Lloyd-Jones, L. R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am. J. Hum. Genet.* **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)[8] - FUSION[b]	Top eQTL BLUP Bayesian LMM LASSO Elastic net	-
Mancuso (2017)[9] - RhoGE	BLUP	-
Barbeira (2018)[10] - MetaXcan	Marginal LASSO Elastic net	-
Su (2018)[11] - MiST	External	Models additional variance component for genetic effects not mediated by predicted expression
Hu (2019)[12] - UTMOST	Multivariate LASSO	Simultaneously models multiple tissues during weight estimation
Yang (2019)[13] - CoMM	Collaborative mixed model	Estimates weights and associations with phenotype simultaneously in single model
Mancuso (2019)[14] - FOCUS	External	Models multiple genes at once, as well as additional pleiotropic genetic effects on phenotype
Nagpal (2019)[15] - TIGAR	Dirichlet process regression	Multivariate model with multiple outcome phenotypes
Liu (2020)[16] - T-GEN	Spike & Slab	Incorporates epigenetic information into weight estimation process
Luningham (2020)[17] - BGW-TWAS	Spike & Slab	Models additional trans-eQTL component
Bhattacharya (2021)[18] - MOSTWAS	Elastic net BLUP	Models additional components for trans-eQTL or other molecular phenotypes
Non-linear models		
Xu (2017)[19] - ASPU	External	Uses adaptive test combining sums of powers of score statistics for different powers (includes linear model)
Zhang (2020)[20]	External	Uses adaptive test combining linear model with sum of squared score statistics
Tang (2021)[21] - 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 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[29]	361K	5.94M	6,686	7.48×10^{-6}	68	107	63.6	100	109
BMI[26]	807K	6.28M	5,453	9.17×10^{-6}	134	232	57.8	297	328
Type 2 diabetes[29]	18.5K/366K	5.94M	6,686	7.48×10^{-6}	27	45	60.0	26	33
Educational attainment[27]	766K	6.18M	5,532	9.04×10^{-6}	59	108	54.6	149	165
Schizophrenia[28]	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 $H_0: \mathrm{cov}(G_E, G_Y) = 0$).

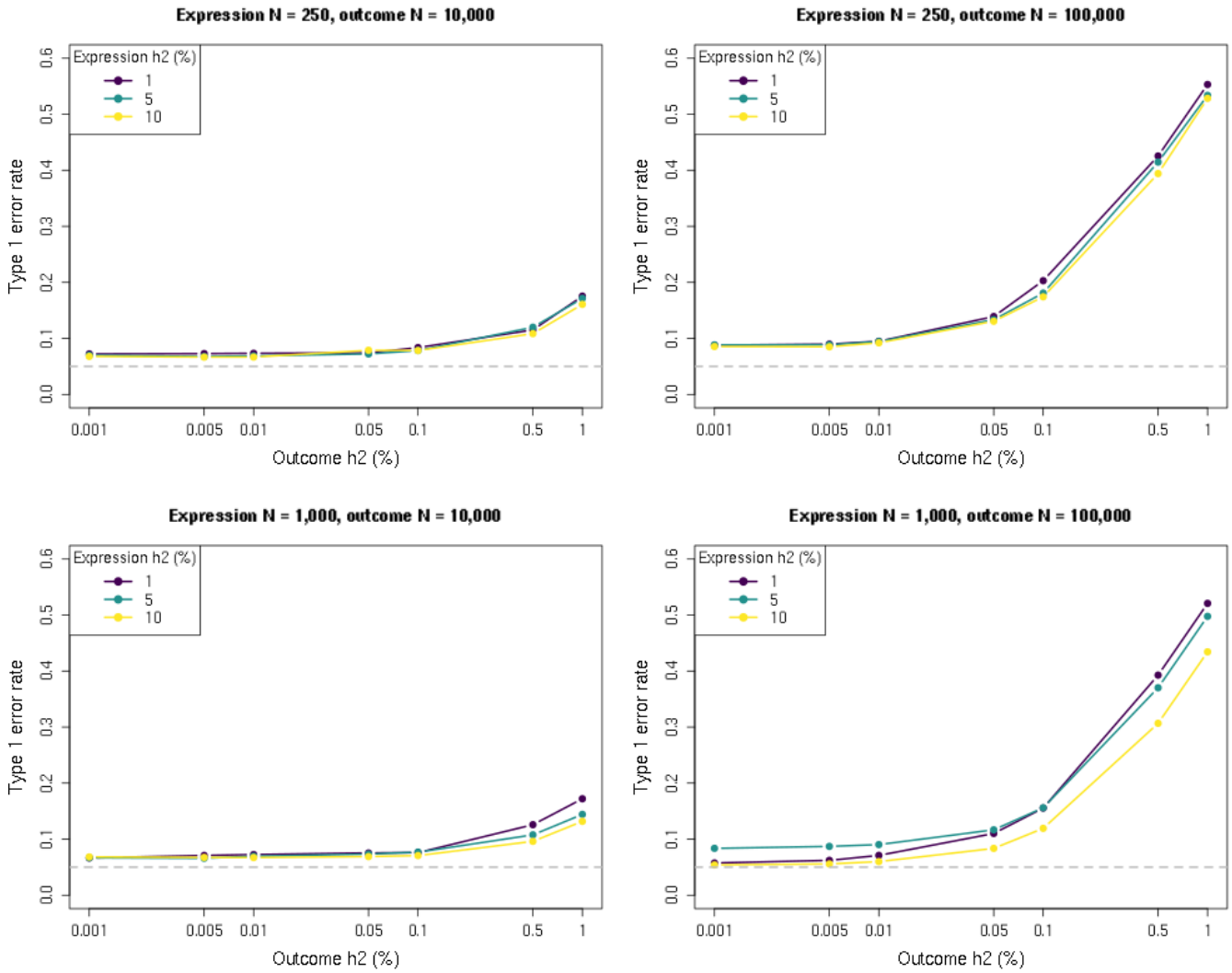

Figure 1. Type 1 error rate simulation results as a function of sample size and local heritability. *Shown is the type 1 error rate (at significance threshold of 0.05) of the LAVA-TWAS model for the null hypothesis of no genetic covariance ($cov(G_E, G_Y) = 0$), at different levels of local heritability (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.*

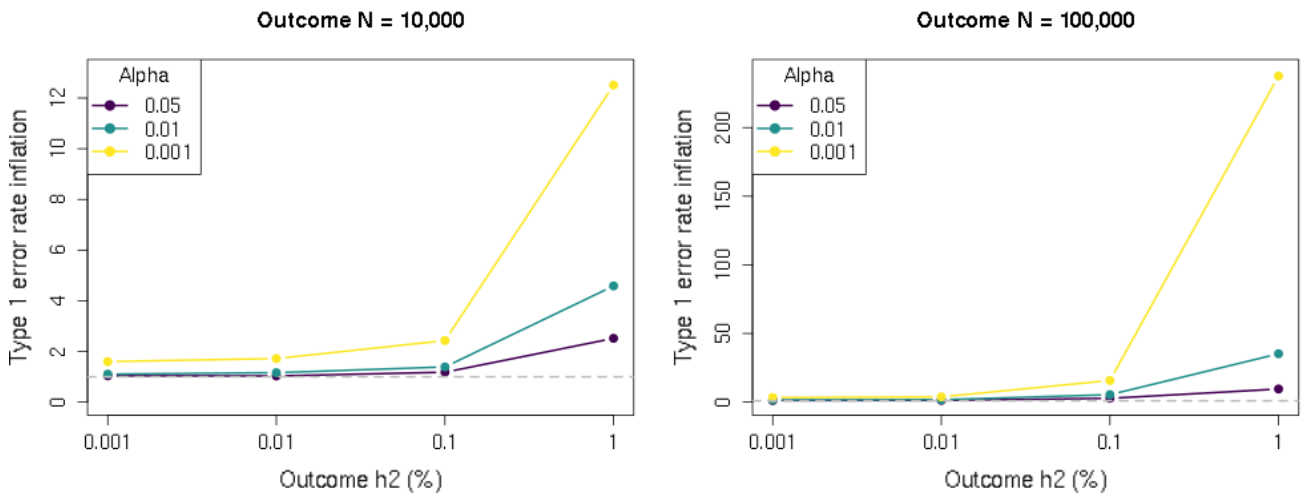

Figure 2. Type 1 error rate inflation simulations results as a function of significance threshold. *Shown is the type 1 error rate inflation of the LAVA-TWAS model for the null hypothesis of no genetic covariance ($cov(G_E, G_Y) = 0$), for different levels of significance threshold α. The type 1 error 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.*

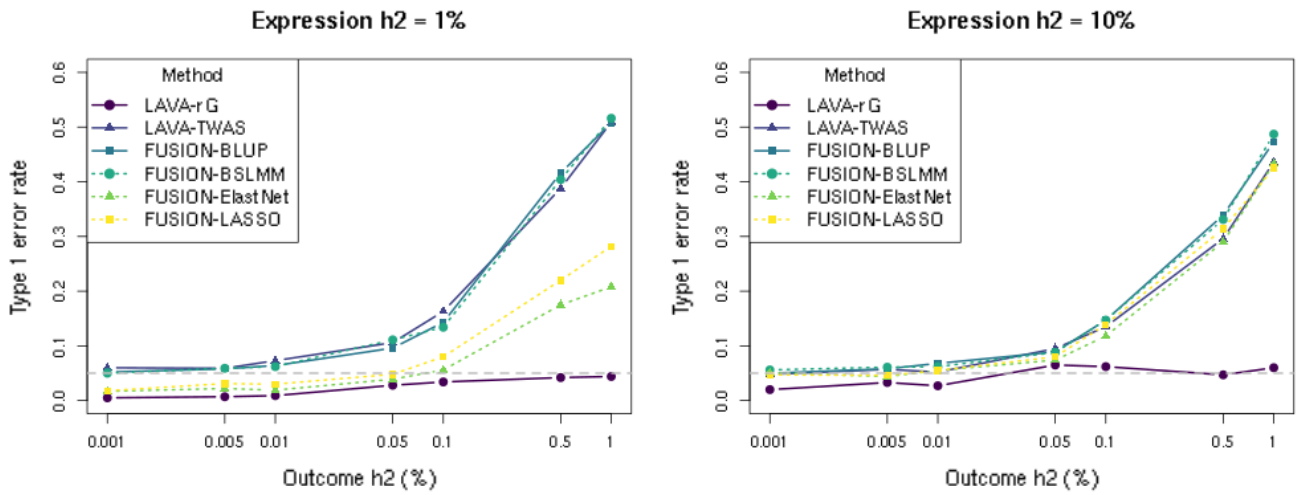

Figure 3. Comparison of type 1 error rate simulation results across models. *Shown is the type 1 error rate (at significance threshold of 0.05) of the five different TWAS models and LAVA-rG for the null hypothesis of no genetic covariance ($cov(G_E, G_Y) = 0$), at different levels of local heritability (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 3).*

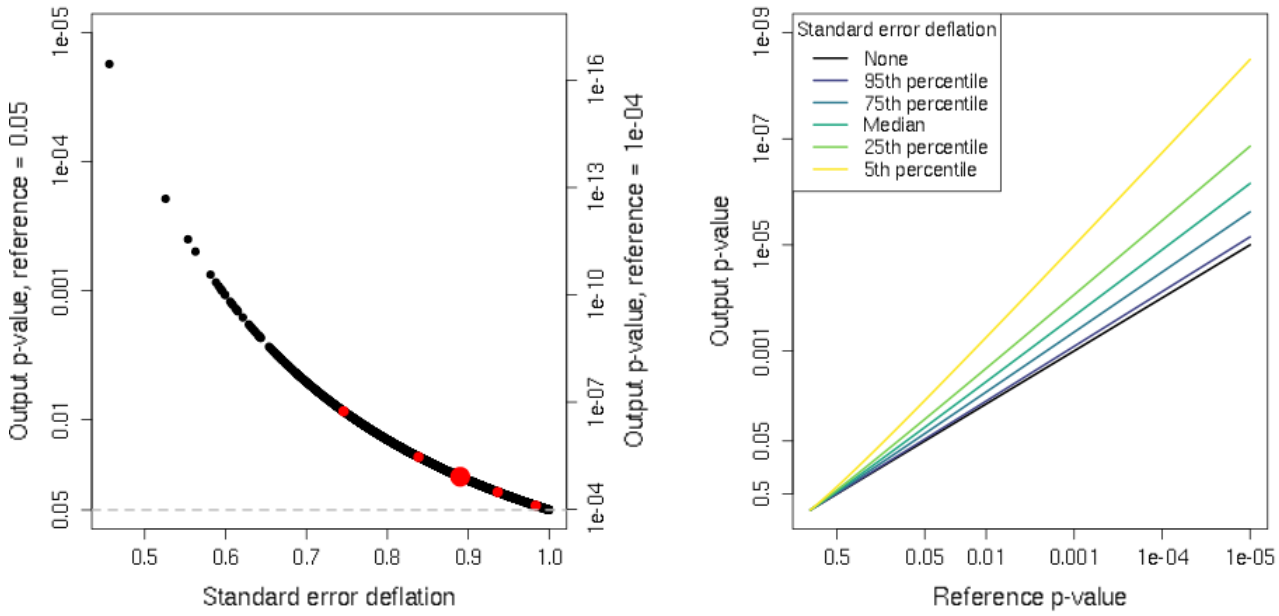

Figure 4. Illustration of p-value bias as a result of deflated standard errors. *Shown in the left panel are the p-values that will result from a LAVA-TWAS analysis at different levels of standard error deflation, when the true p-value computed with correct standard errors would either be 0.05 (left 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.*