Supplementary Materials

A unified framework for joint-tissue transcriptome-wide association and Mendelian Randomization analysis

Dan Zhou[1, 2], Yi Jiang[2, 3], Xue Zhong[1, 2], Nancy J. Cox[1, 2, 4], Chunyu Liu[3,5], Eric R. Gamazon[1, 2, 4, 6,7]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[2]Vanderbit Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

[3]Center for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan, China

[4]Data Science Institute, Vanderbilt University Medical Center, Nashville, TN

[5]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA

[6]Clare Hall, University of Cambridge, Cambridge, United Kingdom

[7]MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom

Send correspondence to:

Dan Zhou <zdangm@gmail.com>

Eric R. Gamazon <ericgamazon@gmail.com>

Supplementary Note

Type I error and statistical power

We performed extensive simulations to evaluate the type I and type II error rates of the association test based on the PrediXcan and JTI models.

Given the shared samples between tissues, it is important to evaluate whether JTI would show inflated false-positive findings. The generative model for the trait is given by:

$$y = \alpha_g(X\beta) + \varepsilon$$

where y is the phenotype and X encapsulates the genotype matrix for the variants in the multi-tissue imputation model (with effect-size vector β) of a gene g. Here, $\varepsilon \sim \mathrm{N}(0, \sigma_g^2)$ and varies with g. To evaluate the type I error, we set $\alpha_g = 0$ to simulate the null.

Compared to PrediXcan, JTI showed higher prediction performance across all tissues (Extended Data Fig. 3). Higher prediction quality will automatically imply higher power for the association test, assuming concordant gene-level effect for PrediXcan and JTI.

For power analysis, the true expression level (g_i) of each of m (m =100) randomly sampled causal genes was simulated ($g_i \sim N(0, 1)$). The effect size (α_i) of a gene expression trait on the phenotype (y) was simulated across the m genes, i.e.,

$$y = \sum_{i=1}^{m} \alpha_i \, g_i + \varepsilon$$

where $\alpha_i \sim N\left(0, \frac{h^2_{exp}}{m}\right)$, $\varepsilon \sim N\left(0, 1 - h^2_{exp}\right)$ is the residual, and h^2_{exp} is the overall variance explained by the gene expression traits ($h^2_{exp} = 0.50$). In this model, each gene, on average, contributes $E(\alpha_i) = \frac{h^2_{exp}}{m}$ (i.e., 0.005) to the phenotypic variance. For each gene, the predicted (i.e., genetically determined) expression level was generated according to the proportion of variance explained (PVE) for each of the three imputation approaches (PrediXcan, UTMOST, and JTI) based on actual prediction performance (R^2) in an external dataset (PsychENCODE). We note that, by design, the correlation between the true expression and the predicted expression is higher if the prediction performance is higher. We performed 100 simulations and then tested the association between the predicted expression and the phenotype. Power was estimated as the proportion of simulations that attain significance (defined as Bonferroni adjusted *P* < 0.05).

Summary-statistics based approach for association analysis with imputed expression

Given the imputed expression $\hat{g} = X\hat{\beta}$, where $\hat{\beta}$ is the *m*-dimensional effect-size vector and X is the standardized genotype matrix from the JTI imputation model, the correlation of $\hat{g}$ with standardized phenotype y and therefore the z-score Δ for the effect of g on this phenotype is given by the following expressions respectively, as we previously showed[1]:

$$\mathrm{corr}(X\hat{\beta},\mathrm{y}) = \hat{\beta}^T X^T y \Big/ \sqrt{\hat{\beta}^T X^T X \hat{\beta} y^T y}$$

$$\Delta = \hat{\beta}^T z \Big/ \sqrt{\hat{\beta}^T X^T X \hat{\beta}}$$

Here z is the vector of standardized effect sizes on the phenotype (from GWAS summary statistics) for the variants in the imputation model. The SNP-SNP covariance matrix $X^T X$ can be calculated using LD data from an ancestry-matched reference panel. We calculated the covariance matrix for each gene with an imputation model using LD information from the GTEx sample data. Under the null, $\Delta \sim N(0,1)$. In the case of uncorrelated variants, we can, as we previously noted for PrediXcan[1], estimate the effect size $\hat{\alpha}$ and corresponding standard error $SE(\widehat{\alpha})$ of the JTI-imputed genetically determined component of gene expression on trait as follows:

$$\hat{\alpha} = \frac{\sum_{j=1}^{m} \hat{\theta}_j \hat{\beta}_j S_j^{-2}}{\sum_{j=1}^{m} \hat{\beta}_j{}^2 S_j^{-2}}$$

$$SE(\widehat{\alpha}) = \sqrt{\frac{1}{\sum_{j=1}^{m} \hat{\beta}_j{}^2 S_j^{-2}}}$$

Here $\hat{\beta}_j$ is the effect size of the *j*-th variant from the JTI model while $\hat{\theta}_j$ and S_j are the effect size and the corresponding standard error of the effect size from the GWAS summary statistics data. We note that this estimator corresponds to the estimator for the summary statistic (inverse-variance weighted) method[2,3] for Mendelian Randomization with multiple instrumental variables. In the section "Causal effect

inference and the calculus of MR-JTI" (Methods), we make the connection of JTI with Mendelian Randomization more explicit.

Weak instrument bias

In addition to the statistical challenge presented by invalid instruments due to horizontal pleiotropy, we considered the impact of weak instruments[4] arising from the fact that the genetic instruments may explain only a modest proportion of the gene expression variance. The strength of the genetic instruments is given by the F-statistic[4]:

$$F(R, n, J) = {R^2(n - J - 1)}\Big/{((1 - R^2)J)}$$

The instrument strength $F(R, n, J)$ depends on the sample size n, the number of genetic instruments J, and the expression variance explained by the model R^2. (An unmeasured confounder is another source of bias, but aside from the well-studied population stratification, such a confounder is likely to be independent of the genetic instruments because of random assortment of genetic variation at gamete formation (i.e., Mendel's second law).) The Wald estimator $\hat{\alpha} = \frac{\widehat{\theta_J}}{\widehat{\beta_J}}$ is highly biased when the effect β_j is small, which results in weak instrument bias[5]. However, our approach aims to substantially improve the R^2 by borrowing regulatory information across tissues and to capture the effect of weak instrument bias in our estimate of heterogeneity, which we explicitly model.

Causal inference using other MR approaches

We ran four additional widely-used MR methods for performance comparison with MR-JTI. JTI significant genes for LDL-C were tested using MR-Egger (R package 'MendelianRandomization'), weighted median (R package 'MendelianRandomization'), and MR-PRESSO ('rondolab/MR-PRESSO' on github). SMR-HEIDI (version 1.03), a variant-based approach, was performed for each gene with at least one eQTL. Significant genes were identified after Bonferroni correction ($P_{Bonferroni} < 0.05$). SMR-HEIDI also required $P_{HEIDI} > 0.05$ from the HEIDI approach to test against the null hypothesis that the association identified by SMR was due to pleiotropy. MR-Egger estimates an intercept in Egger regression as a measure of the average pleiotropic effect across the variants. The median estimator has been shown to be consistent when less than half of the instruments are invalid, with the consistency not dependent on the strength of the invalid instruments relative to valid instruments or their correlation structure[6]. MR-PRESSO[7] removes, in turn, genetic variants that have outlying Wald ratio estimates based on a decrease in a residual sum of squares (as a heterogeneity measure) relative to a simulation-derived expected distribution.

For additional functional support for the MR-implicated genes, we identified their overlap with, and tested their enrichment for, the genes in the "silver standard" list of well-known genes and a conserved cholesterol biosynthetic process module in mice (the full list of genes from which can be found in Li and colleagues' paper[8]). Briefly, the null distribution for the overlap count with the cholesterol module was generated by randomly drawing (1000 times) genes of the same count as the number of significant

genes from all the tested genes (imputable genes in liver). The empirical p value was calculated as the proportion of times the overlap count from a randomly generated set was at least as extreme as the actual overlap count from the set of significant genes.

Modifications to UTMOST

We elaborate on the reason for (and the details of) the modifications made to the original UTMOST below. Briefly, the original UTMOST generates a different pair of hyperparameters (λ_1 and λ_2) for each 5-fold. The best lambda pair is found for each fold (i.e., the 5-fold cross-validation generates 5 best lambda pairs). Then the final lambda pair is determined based on the prediction performance by retraining the model in the entire data across the 5 best lambda pairs. i.e., the performance is estimated in the final training *and* full dataset. Thus, the in-sample estimation at this final stage results in overestimation of the performance (Extended Data Fig. 6 and 7). Here, we describe our modifications to the UTMOST code. We addressed the overestimation due to the "double dipping" by using uniform hyperparameter pairs across the 5 folds, which makes the hyperparameter pairs' performance comparable across the 5 folds in the cross-validation step. This approach avoids conducting the performance estimation using the retrained (final) model. The modifications substantially reduced the inflated performance as observed in an external dataset (Extended Data Fig. 6 and 7).

How the original UTMOST script works for model building[9]:

The authors split the entire data into five parts. For convenience, we denote the five parts as A, B, C, D, and E.

1. Hyperparameter tuning (5-fold)

1.1 In the first fold, ABC and D were used as the training and the tuning set, respectively. E was not used at this stage.

1.2 Single tissue prediction model was trained by elastic net for each tissue independently (using 5-fold cross-validation in ABC).

1.3 A joint-tissue group-LASSO prediction model was built for each hyperparameter pair (λ_1 and λ_2, five values for each lambda. i.e., 25 combinations in total). The range of each lambda was learned from the single-tissue model training for each fold (i.e., the lambdas are fold-specific). In the optimization step, the parameters (betas) were initialized from the weights of the single-tissue model with the lowest cross-validation error. The optimization would stop if the new training error (error in ABC) or the new tuning error (error in D) was higher than the value from the previous step. The best lambda pair (for this fold) was chosen according to the average tuning error across all the tissues.

1.4 The same procedure was then conducted by taking BCD, CDE, DEA, EAB as the training set and taking E, A, B, C as tuning set, respectively. After the 5-fold training, 5 best lambda pairs were generated.

2. Training the model using the entire data

2.1 Single-tissue elastic net model was trained by 5-fold cross-validation using the entire data.

2.2 The joint-tissue group-LASSO was performed by applying each of the five best lambda pairs. The optimization was initialized by the parameters (betas) from the single-tissue model. The iteration would stop when the new training error was greater than the old value, or the difference in error was very small between the two steps (no "early stop" here). The final lambda pair was chosen by evaluating the average training error across all the tissues in the entire data. The model with parameters (betas) from the second to the last iteration was considered as the final model for downstream analysis.

3. Prediction accuracy evaluation

The prediction quality (i.e., correlation *r*) was calculated by applying the final model to the entire data.

The original script URL: https://github.com/yiminghu/CTIMP/blob/master/main.R (10/28/2018)

How we modified the original UTMOST to fix the inflated performance due to the double dipping:

We split the entire data into five parts, ABCDE.

1. Hyperparameter initialization in single-tissue elastic net

Prediction models were trained by 5-fold cross-validation elastic net using the entire data for each tissue independently. The range of lambdas for joint-tissue prediction was learned from the range of single-tissue lambdas. In total, 25 lambda pairs were generated, which were then uniformly applied across the 5-fold training.

2. Hyperparameter tuning and model training (5-fold)

2.1 In the first fold, ABCD and E were used as the training and the tuning set, respectively.

2.2 Single-tissue prediction model was trained by elastic net for each tissue independently (using 5-fold cross-validation in ABCD).

2.3 The joint model was trained using each of the 25 lambda pairs in the training set (ABCD). The optimization was initialized by single-tissue weights generated in the current fold. The optimization would stop if the new training error (error in ABCD) or the new tuning error (error in E) was higher than the value from the previous step.

2.4 The same procedure was then conducted by taking BCDE, CDEA, DEAB, EABC as training set and taking A, B, C, D as tuning set, respectively. After the 5-fold training, one of the 25 lambda pairs was chosen as the best lambda pair according to the average tuning error across the 5 folds.

3. Training the model using entire data

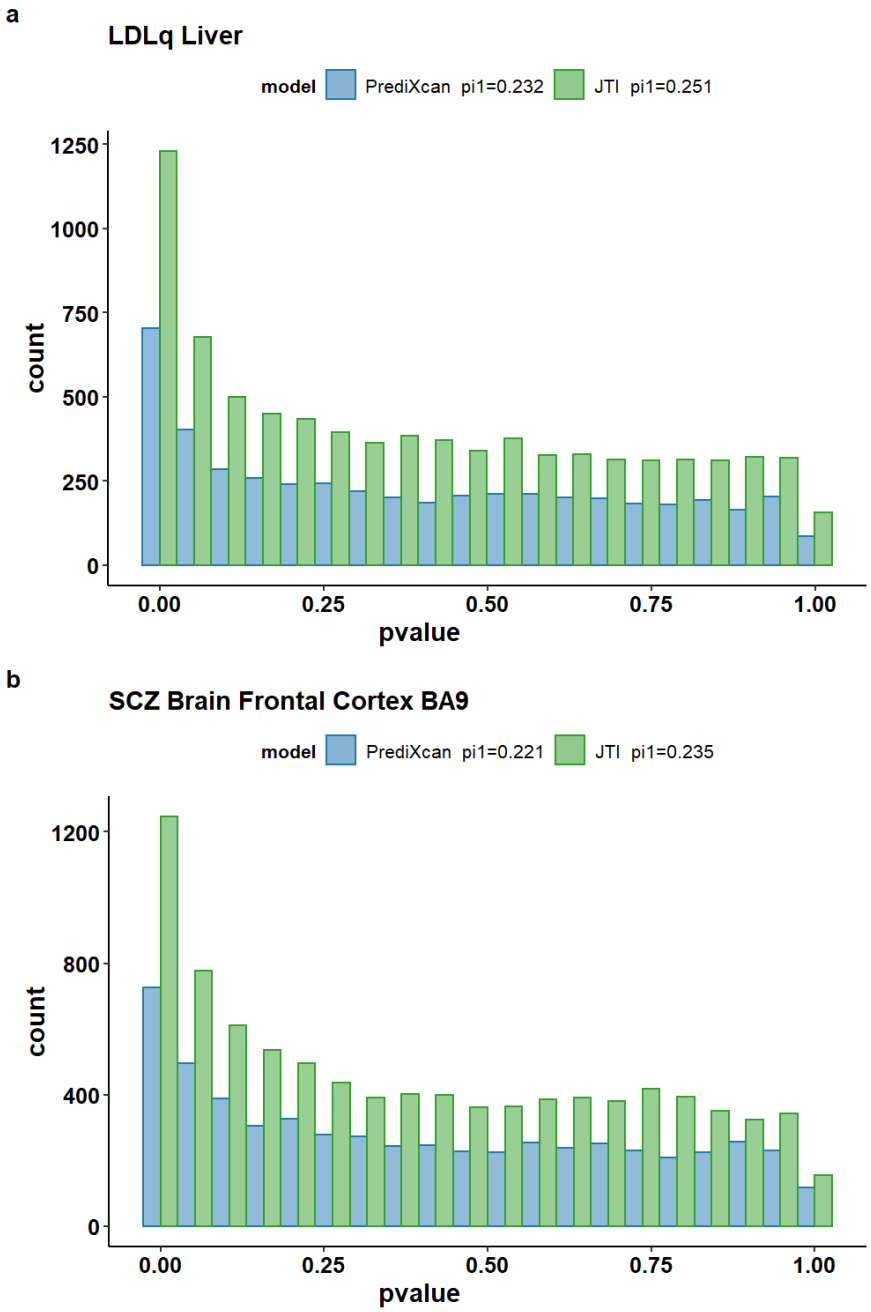
The joint-tissue training was performed by applying the best lambda combination. The optimization was initialized by the parameters (betas) from the single-tissue model

using the entire data. The iteration would stop if the new training error was greater than the previous one, or the difference in error was very small between the two steps.

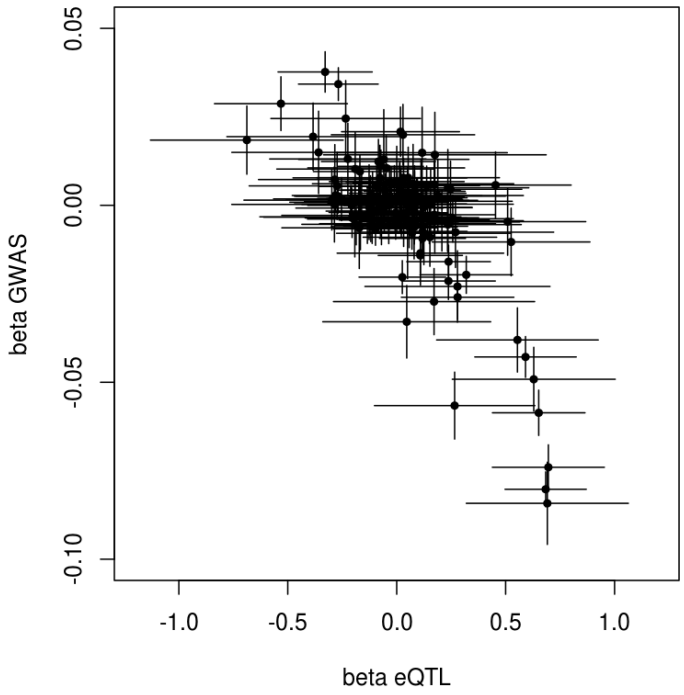
4. Prediction accuracy evaluation

The imputation accuracy was estimated using the correlation of the observed expression level and the predicted expression level calculated in the tuning set.

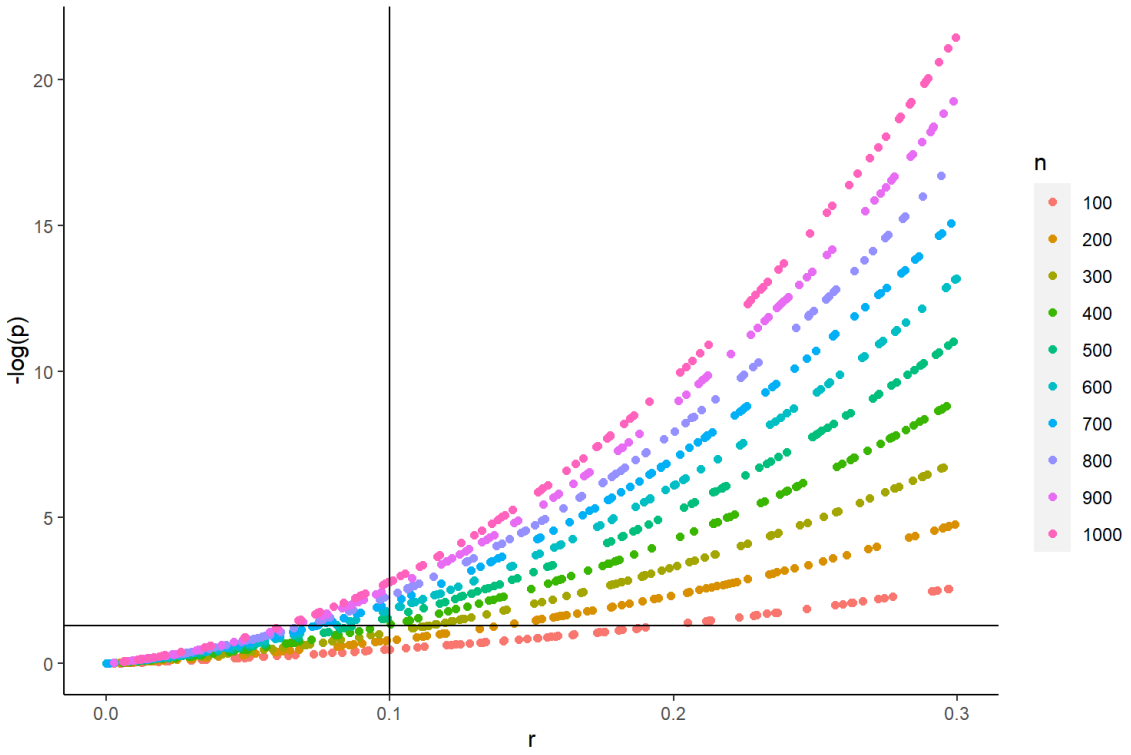
Supplementary Figures



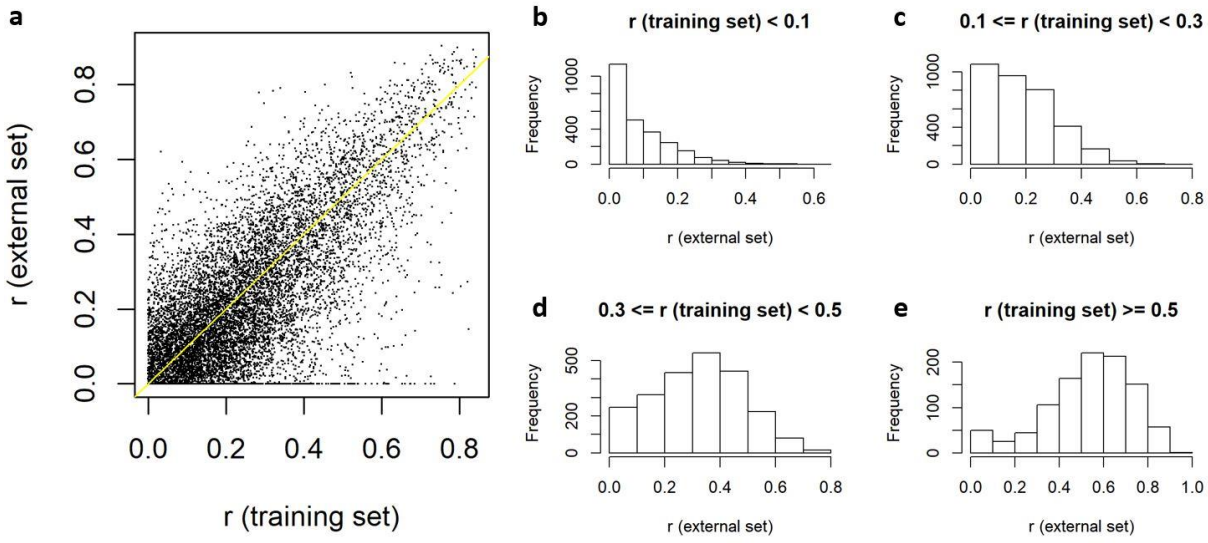
Supplementary Fig. 1: TWAS *P*-value distribution from PrediXcan (blue) and JTI (green) prediction models. a, Application to LDL-C using predicted expression in liver. **b**, Application to schizophrenia (SCZ) using predicted expression in brain frontal cortex BA9. The proportion of true positives π_1 (see Methods) was compared.



Supplementary Fig. 2: MR-JTI supports a causal role for the expression of *SORT1* (in liver) in determining LDL-C levels. X-axis and y-axis show the SNP effect sizes for *SORT1* expression level (beta eQTL) and LDL-C level (beta GWAS). The error bars indicate 1.96 times the standard error of the effect size estimate.



Supplementary Fig. 3: Simulations show that *r* > 0.1 and *P* < 0.05 could be considered a reasonable nominal significance threshold for an imputable gene (iGene). We simulated the observed and predicted expression level with the correlation *r* uniformly drawn from 0 to 0.3 while varying the sample size for the training set. The x-axis and y-axis denote the Pearson correlation *r* and the minus log *P* value of the correlation test, respectively. Given the scale of the GTEx v8 training dataset (mean tissue sample size = 310), when *P* = 0.05, then *r* is close to 0.1. Using *r* > 0.1 and *P* < 0.05 as the threshold, an iGene from tissues with limited samples (e.g. *n* = 100) will have a reasonable significance level (statistically meaningful) and an iGene from tissues with big sample sizes (e.g. n = 1000) will have a reasonable proportion of trait variance explained by the genetically regulated expression (biologically meaningful).



Supplementary Fig. 4: Prediction performance comparison between the training set and the external test set. a, an overview of the comparison. The horizontal axis denotes the correlation *r* in the training set (GTEx brain frontal cortex BA9) and the vertical axis denotes the correlation *r* in the replication dataset (PsychENCODE). **b**, **c**, **d**, and **e**, To visualize the prediction performance in the external dataset among different bins, we grouped genes into four bins (*r* < 0.1, 0.1 <= *r* < 0.3, 0.3 <= *r* < 0.5, and *r* >= 0.5) according to the cross-validation *r* in the training dataset (GTEx brain frontal cortex BA9). The distribution of correlation *r* between predicted and observed expression (on right panels b-e) indicated that 0.1 should be a reasonable cut-off for *r*.

Reference

1. Gamazon, E.R., Zwinderman, A.H., Cox, N.J., Denys, D. & Derks, E.M. Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. *Nature genetics* **51**, 933 (2019).
2. Johnson, T. Efficient calculation for multi-SNP genetic risk scores Technical Report, The Comprehensive R Archive Network. (2013).
3. Burgess, S., Butterworth, A. & Thompson, S.G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology* **37**, 658-665 (2013).
4. Burgess, S., Thompson, S.G. & CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *International journal of epidemiology* **40**, 755-764 (2011).
5. Bound, J., Jaeger, D.A. & Baker, R.M. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association* **90**, 443-450 (1995).
6. Windmeijer, F., Farbmacher, H., Davies, N. & Davey Smith, G. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 1-12 (2019).
7. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics* **50**, 693 (2018).
8. Li, Z. *et al.* Integrating mouse and human genetic data to move beyond GWAS and identify causal genes in cholesterol metabolism. *Cell Metabolism* (2020).
9. Hu, Y. *et al.* A statistical framework for cross-tissue transcriptome-wide association analysis. (Nature Publishing Group, 2019).