The American Journal of Human Genetics, Volume *109*

Supplemental information

Analyzing and reconciling colocalization and transcriptome-wide association studies from the perspective of inferential reproducibility

Abhay Hukku, Matthew G. Sampson, Francesca Luca, Roger Pique-Regi, and Xiaoquan Wen

Supplemental Figures

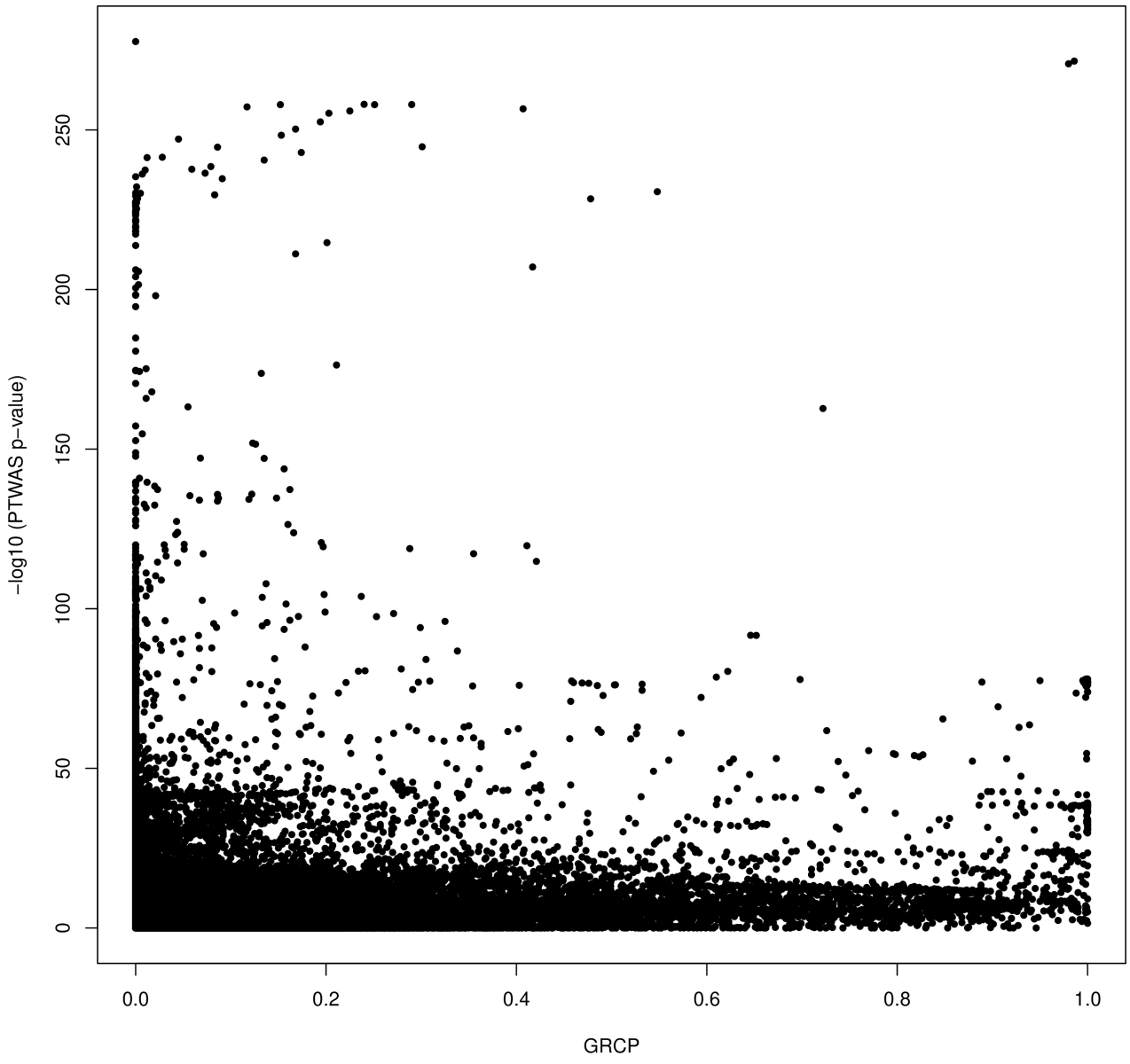

Figure S1: **Scatter plot of PTWAS -log10 p-values and GRCP values** Each data point in the plot represents a gene-tissue-trait combination. The plot indicates modest correlations between TWAS and colocalization results.

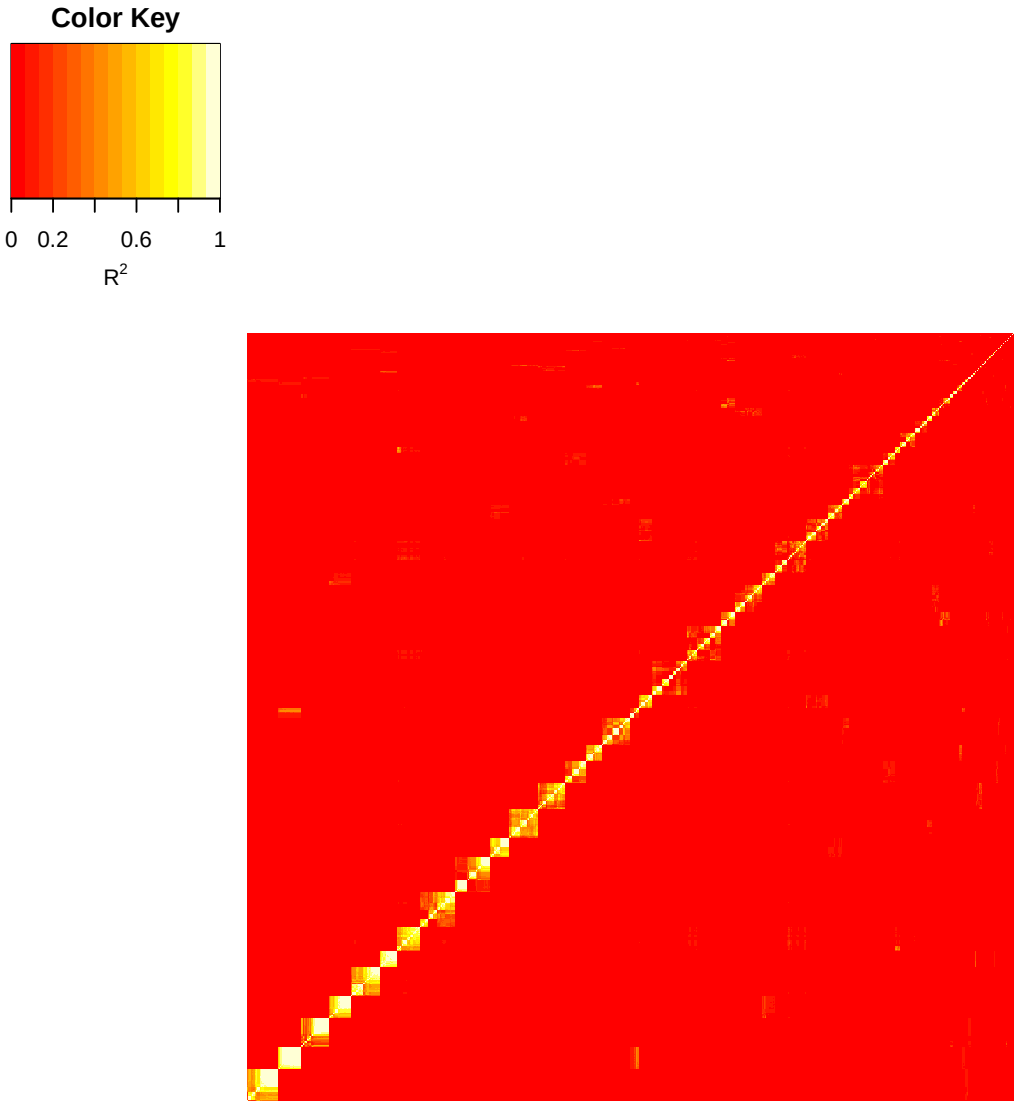

Figure S2: **LD structure of assembled genome used in the simulation study - overall view** LD is measured by pair-wise r^2 between SNPs. The assembled genome contains 22 LD blocks. Each block is selected from a different chromosome and consists of 50 consecutive common SNPs (MAF $\geq$ 0.1). The genotype data for the selected SNPs are taken directly from 838 GTEx muscle skeletal samples.

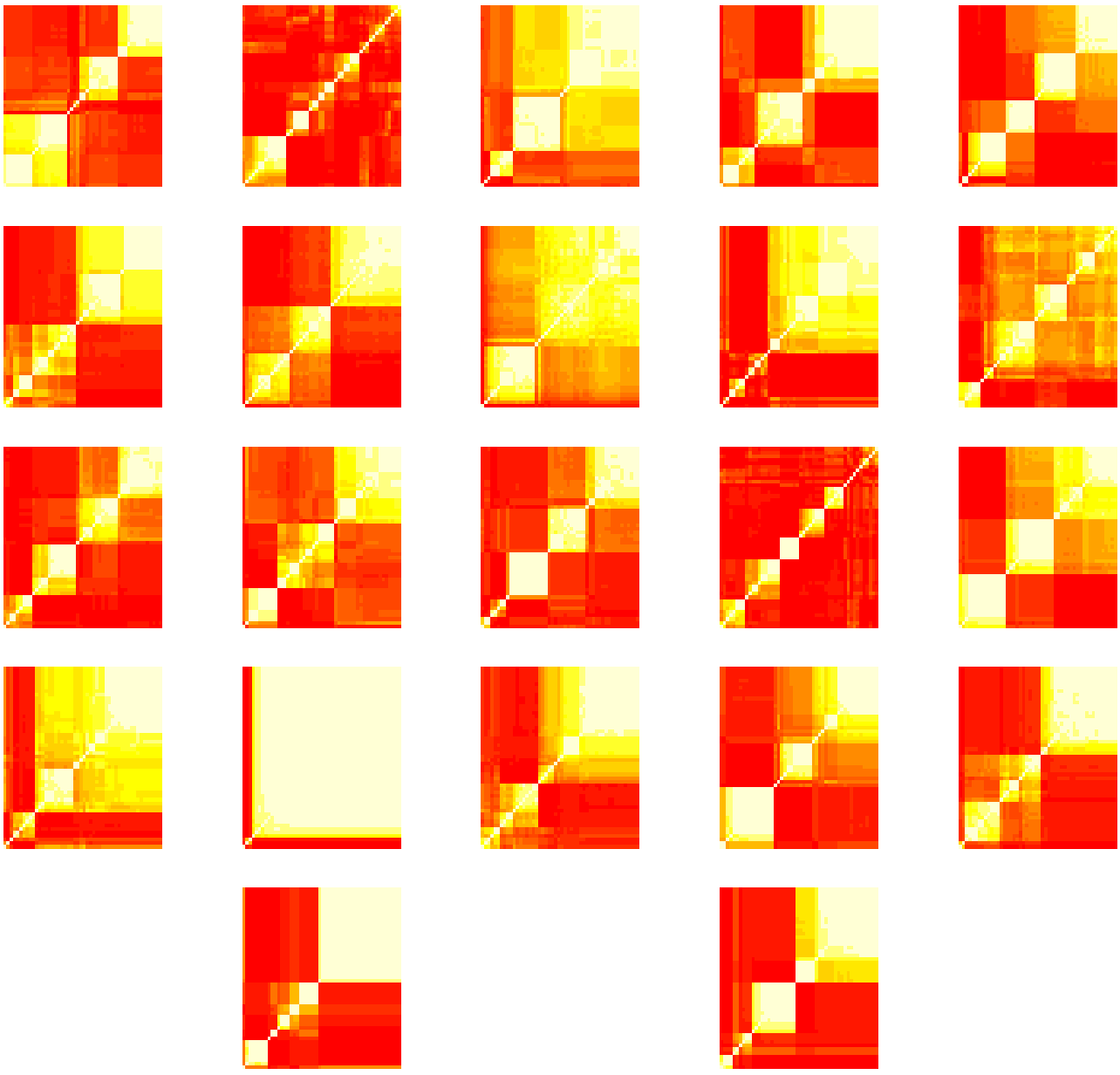

Figure S3: **LD structure of assembled genome used in the simulation study - breakdown by LD blocks** Each figure in the panel shows LD structure within an LD block, consisting of 50 SNPs, in the assembled genome.

Supplemental Methods

1 Single-variant TWAS Scan

This section provides details for the single-variant TWAS scan procedure used in the paper. It serves as a representative baseline for all TWAS scan approaches. The simplicity of the procedure is particularly appealing and helps elucidate some key common features of TWAS scan analysis.

We view the described procedure under the special constraint that only a single variant is allowed for gene expression prediction. Under this setting, the resulting TWAS scan p-value is simply the single-SNP GWAS association p-value of the most significant eQTL SNP identified from the single-SNP eQTL mapping.

We note that the best single-SNP gene expression prediction model is represented by the most significant eQTL SNP based on the training eQTL data. This is because the single-SNP association p-value is monotonic to the R^2 value (or PVE) of the corresponding simple regression model, which directly measures the model's predictive ability. We denote the resulting optimal single-SNP prediction model by

$$\hat{\boldsymbol{y}}_e = \hat{\beta}_e \, \boldsymbol{g}_e, \tag{1}$$

where $\hat{\boldsymbol{y}}_e$ and $\boldsymbol{g}_e$ denote the vectors of predicted gene expression levels and the genotypes in the eQTL samples, respectively (assuming all gene expression phenotypes are pre-centered).

In a separate GWAS sample, TWAS scan measures the correlation between phenotype vector, $\boldsymbol{y}_c$ and the vector of the predicted gene expression, $\hat{\boldsymbol{y}}_e = \hat{\beta}_e \boldsymbol{g}'_e$. Because the estimated eQTL effect, $\hat{\beta}_e$, is a constant with respect to the GWAS samples, it follows that

$$\mathrm{Corr}(\boldsymbol{y}_c, \hat{\boldsymbol{y}}_e) = \mathrm{Corr}(\boldsymbol{y}_c, \boldsymbol{g}'_e). \tag{2}$$

This quantity and the corresponding standard error can be equivalently measured by a simple regression model, regressing $\boldsymbol{y}_c$ on the selected eQTL SNP genotypes. Consequently, upon selecting the predictive eQTL SNP, its single-SNP association evidence for the GWAS trait, z_{gy} is a valid test statistic for TWAS scan analysis.

Alternatively, the procedure can be explained by an optimal single instrument Mendelian randomization (MR) testing procedure. The MR testing procedure examines the null hypothesis that the specified genetic instrument is uncorrelated with the complex trait of interest (i.e., the outcome) [1, 2]. It should be clear that the most significant eQTL SNP is the optimal (or the strongest) if candidate instruments must be single genetic variants (i.e., no composite instrument allowed). This reasoning also leads to the same procedure.

1.1 Comparison to SMR

Summary data-based Mendelian Randomization (SMR) also utilizes a single genetic variant for testing. Thus, it falls into the category of single-SNP TWAS procedure. It is also derived from the principle of MR but instead relies on the MR estimation procedure to derive the test statistic. For a target SNP, let z_{gx} and z_{gy} denote the z-scores derived from the single-SNP analyses of eQTL mapping and GWAS, respectively. The resulting test statistic derived by [3] is given by:

$$z_{gy}^2 \frac{z_{gx}^2}{z_{gy}^2 + z_{gx}^2} \sim \chi_1^2 \tag{3}$$

under the null.

It is clear that the test statistic from the optimal single-variant TWAS approach follows that

$$z_{gy}^2 \sim \chi_1^2, \tag{4}$$

under the MR null hypothesis (i.e., no causal relation from the gene expression, x, to the complex trait, y).

However, because

$$z_{gy}^2 \frac{z_{gx}^2}{z_{gy}^2 + z_{gx}^2} \leq z_{gy}^2 \qquad \forall z_{gx}, z_{gy}, \tag{5}$$

it implies that the described procedure is universally more powerful. This result should not be surprising, as our derivation has also indicated the optimality of the proposed procedure.

It should be noted that the SMR offers additional inference features, i.e., a gene-to-trait effect estimate and its corresponding standard error. Nevertheless, because our focus is on the TWAS scan/testing procedure, the proposed single-variant TWAS approach offers much simpler computation and results in a more powerful test.

2 Simulation Details

2.1 Simulation to Illustrate LD Hitchhiking Effect in TWAS Scan

Our simulation scheme is informed by real data analysis. Firstly, we identify the SNP rs2871960 as one of the most significant genetic associations ($z = -34.2$) for standing height in the UK Biobank data. In the GTEx data, this SNP maps to the *cis*-region of a single gene, *ZBTB38* (Ensembl ID: ENSG00000177311), in the Muscle Skeletal tissue. Additionally, we identify 38 neighboring genes within 8 Mb of the SNP. The list of all 39 genes is provided in Supplemental Table S3.

In this simulation, we consider all 22,662 *cis*-SNPs of the 39 genes. The eQTL dataset is directly taken from the GTEx muscle skeletal tissue with 706 individuals. We note that rs2871960 is unlikely to be the true causal eQTL for ENSG00000177311 in the muscle skeletal tissue based on the fine-mapping result. Specifically, it does not fall into any signal cluster of the gene and the PIP $= 2.75 \times 10^{-3}$ (despite its single-SNP testing p-value reaching 10^{-11}). Additionally, the SNP-level colocalization probability for this SNP is also quite low (2.72×10^{-3}). We simulate a complex trait ($\boldsymbol{y}_h$) for the 706 GTEx individuals using their true genotypes ($\boldsymbol{g}$) for SNP rs2871960 and the following simple linear regression model,

$$\boldsymbol{y}_h = -1.5\,\boldsymbol{g} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \mathrm{N}(0, I). \tag{6}$$

That is, the only causal SNP for the complex trait is assumed to be rs2871960, with its genetic effect fixed at -1.5. Note that the genotypes for the gene expression and complex trait studies are perfectly matched. However, because the complex trait is independently simulated, the overall scheme fits the two-sample design for integrative analysis. The particular genetic effect value matches the observed complex trait z-score in the UK Biobank with a much-reduced sample size. To further illustrate that all

significant TWAS findings from the simulated dataset are due to the LD hitchhiking effect, we regress out the genotypes of the causal SNP and treat the residuals as a new complex trait phenotype. Finally, we analyze both datasets by PTWAS scan and report the corresponding p-values for each examined gene.

2.2 Simulation to evaluate locus-level colocalization analysis

In this simulation, we assemble a genomic region with a "known" LD structure from the genotype data of 838 GTEx samples. We classify the artificial genomic region into 22 LD blocks, with each block containing 50 consecutive common SNPs from a unique chromosome. We intend to capture natural LD patterns within each block and minimize LD between blocks (as the genotypes are taken from distinct chromosomes). The LD structure of the assembled genomic region is shown in Supplementary Figures S1 and S2.

For each simulated dataset, we independently generate phenotype data for a molecular ($\boldsymbol{y}_e$) and a complex trait ($\boldsymbol{y}_c$) using the following linear regression models

$$\begin{aligned} \boldsymbol{y}_e &= v_e \mathbf{1} + \sum_{i=1}^{1100} \alpha_i \, \boldsymbol{g}_i + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(0, I) \\ \boldsymbol{y}_c &= \mu_c \mathbf{1} + \sum_{i=1}^{1100} \beta_i \, \boldsymbol{g}_i + \boldsymbol{e}, \quad \boldsymbol{e} \sim \mathrm{N}(0, I) \end{aligned} \tag{7}$$

Only 2 SNPs located in different LD blocks have non-zero genetic effects for each trait. The exact effects for the two causal SNPs of each trait are independently sampled from the distribution $\mathrm{N}(0, 1)$. As a result, the percentage of variance explained (PVE) by genetics for each trait is approximately 0.35 on average. Coincidentally, we find that under this signal-to-noise ratio, the GLCP threshold at the 5% level FDR control is very close (only slightly higher) to the commonly used value of 0.50 in colocalization analysis. We simulate 5,000 datasets with the causal eQTLs and the GWAS hits located in distinct LD blocks (i.e., no colocalization). They serve as a baseline to examine potential false-positive findings. For the other 2,500 datasets, we explicitly select one causal eQTL SNP and one GWAS SNP colocalized at a single variant while the remaining casual eQTL SNP and GWAS SNP are placed in distinct LD

blocks. Let $\boldsymbol{g}_1$ denote the colocalized SNP, the generative models for the phenotypes become

$$\begin{aligned} \boldsymbol{y}_e &= v_e \mathbf{1} + \alpha_1 \boldsymbol{g}_1 + \alpha_2 \boldsymbol{g}_2 + \boldsymbol{\epsilon} \\ \boldsymbol{y}_c &= \mu_c \mathbf{1} + \beta_1 \boldsymbol{g}_1 + \beta_3 \boldsymbol{g}_3 + \boldsymbol{e} \end{aligned} \tag{8}$$

Note that the colocalization in (8) induces a structural (mean) equation between $\boldsymbol{y}_e$ and $\boldsymbol{y}_c$, i.e.,

$$\boldsymbol{y}_c = \mu' \mathbf{1} + \beta' \boldsymbol{y}_e + \boldsymbol{e}', \tag{9}$$

which is often used in TWAS simulations [4, 3].

3 TWAS-FOCUS analysis of LD-hitchhiking data

We analyze the simulated LD-hitchhiking data using the algorithm implemented in the software package FOCUS [5]. Among the 39 candidate genes, a single gene, *ZBTB38* (Ensembl ID: ENSG00000177311), receives posterior probability (PIP) ≈ 1.00, the remaining 38 genes' PIP values are all approximately 0. That is, FOCUS implicates *ZBTB38* as the sole "causal" gene for the simulated complex trait. In comparison, the proposed joint analysis approach implicates no gene from the simulated data. The maximum GLCP = 0.07 is attained for *ZBTB38*, indicating weak locus-level colocalization evidence between the causal eQTL and the sole causal GWAS hit.

Like the widely-used conditional analysis in GWAS, FOCUS effectively identifies LD-hitchhiking effects between genes. That is, once the strongest TWAS gene is controlled, the remaining genes no longer show TWAS association. However, FOCUS can behave differently from the proposed approach when assessing the TWAS gene showing the strongest association signal.

It is worth noting that the LD hitchhiking effect can occur within a single gene (i.e., when the causal eQTL and the GWAS hits are in weak to modest LD) and lead to strong TWAS signals. The statistical model employed by FOCUS is closely connected to the MR-Egger method [6] when dealing with a single gene. It is considered more robust than the traditional MR/IV analysis method by allowing direct effects (also known as "SNP pleiotropic effects") from the instruments (i.e., eQTLs) to the complex trait of

interest. However, in the presence of within-gene LD hitchhiking, Barfield *et al.* [7] shows that the InSIDE assumption (instrument strength independent of direct effect, [8]) is violated, and the FOCUS model can lead to inflated false positives for inferring causal genes. Our observation of the simulated data seems to confirm their finding.

In principle, the FOCUS model can be modified to incorporate the proposed strategy, e.g., by constraining the validity of the instruments. Future work is needed to implement such ideas fully.

Supplemental References

References

[1] Katan, M. Apoupoprotein e isoforms, serum cholesterol, and cancer. *The Lancet* **327**, 507–508 (1986).

[2] VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M. & Kraft, P. Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.)* **25**, 427 (2014).

[3] Zhu, Z. *et al.* Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics* **48**, 481 (2016).

[4] Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48**, 245 (2016).

[5] Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nature genetics* **51**, 675–682 (2019).

[6] Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology* **44**, 512–525 (2015).

[7] Barfield, R. *et al.* Transcriptome-wide association studies accounting for colocalization using egger regression. *Genetic epidemiology* **42**, 418–433 (2018).

[8] Burgess, S. & Thompson, S. G. Interpreting findings from mendelian randomization using the mr-egger method. *European journal of epidemiology* **32**, 377–389 (2017).