

# UCLA

## UCLA Previously Published Works

### Title

Opportunities and challenges for transcriptome-wide association studies

### Permalink

<https://escholarship.org/uc/item/70g2j0qd>

### Journal

Nature Genetics, 51(4)

### ISSN

1061-4036

### Authors

Wainberg, Michael  
Sinnott-Armstrong, Nasa  
Mancuso, Nicholas  
et al.

### Publication Date

2019-04-01

### DOI

10.1038/s41588-019-0385-z

Peer reviewed



Published in final edited form as:

*Nat Genet.* 2019 April ; 51(4): 592–599. doi:10.1038/s41588-019-0385-z.

## Opportunities and challenges for transcriptome-wide association studies

Michael Wainberg<sup>1</sup>, Nasa Sinnott-Armstrong<sup>2</sup>, Nicholas Mancuso<sup>3</sup>, Alvaro N. Barbeira<sup>4</sup>, David A. Knowles<sup>5,6</sup>, David Golan<sup>2</sup>, Raili Ermel<sup>7</sup>, Arno Ruusalepp<sup>7,8</sup>, Thomas Quertermous<sup>9</sup>, Ke Hao<sup>10</sup>, Johan L. M. Björkegren<sup>8,10,11,12,\*</sup>, Hae Kyung Im<sup>4,\*</sup>, Bogdan Pasaniuc<sup>3,13,14,\*</sup>, Manuel A. Rivas<sup>15,\*</sup>, Anshul Kundaje<sup>1,2,\*</sup>

<sup>1</sup>Department of Computer Science, Stanford University, Stanford, CA, USA.

<sup>2</sup>Department of Genetics, Stanford University, Stanford, CA, USA.

<sup>3</sup>Department of Pathology & Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA.

<sup>4</sup>Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA.

<sup>5</sup>New York Genome Center, New York, NY, USA.

<sup>6</sup>Department of Computer Science, Columbia University, New York, NY, USA.

<sup>7</sup>Department of Cardiac Surgery, Tartu University Hospital, Tartu, Estonia.

<sup>8</sup>Clinical Gene Networks AB, Stockholm, Sweden.

<sup>9</sup>Division of Cardiovascular Medicine, Stanford University, Stanford, CA, USA.

<sup>10</sup>Department of Genetics & Genomic Sciences, Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

<sup>11</sup>Department of Pathophysiology, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia.

<sup>12</sup>Integrated Cardio Metabolic Centre, Department of Medicine, Karolinska Institutet, Karolinska Universitetssjukhuset, Huddinge, Sweden.

---

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

\***Correspondence** should be addressed to J.L.M.B., [johan.bjorkegren@mssm.edu](mailto:johan.bjorkegren@mssm.edu); H.K.I., [haky@uchicago.edu](mailto:haky@uchicago.edu); B.P., [pasaniuc@ucla.edu](mailto:pasaniuc@ucla.edu); M.A.R., [mrivas@stanford.edu](mailto:mrivas@stanford.edu); A.K., [akundaje@stanford.edu](mailto:akundaje@stanford.edu).

**Author contributions**

M.W., M.A.R. and A.K. conceived the study. M.W., N.M. and A.N.B. performed analyses. N.S.-A., D.A.K. and D.G. provided intellectual input. R.E., A.R., T.Q., K.H. and J.L.M.B. provided assistance with analysis of the STARNET dataset. H.K.I., B.P., M.A.R. and A.K. supervised the study. M.W., H.K.I., B.P., M.A.R. and A.K. wrote the manuscript. All authors reviewed the manuscript.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Competing interests**

The authors declare no competing interests.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41588-019-0385-z>.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

<sup>13</sup>Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA.

<sup>14</sup>Department of Computational Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA.

<sup>15</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA, USA.

## Abstract

Transcriptome-wide association studies (TWAS) integrate genome-wide association studies (GWAS) and gene expression datasets to identify gene-trait associations. In this Perspective, we explore properties of TWAS as a potential approach to prioritize causal genes at GWAS loci, by using simulations and case studies of literature-curated candidate causal genes for schizophrenia, low-density-lipoprotein cholesterol and Crohn's disease. We explore risk loci where TWAS accurately prioritizes the likely causal gene as well as loci where TWAS prioritizes multiple genes, some likely to be non-causal, owing to sharing of expression quantitative trait loci (eQTL). TWAS is especially prone to spurious prioritization with expression data from non-trait-related tissues or cell types, owing to substantial cross-cell-type variation in expression levels and eQTL strengths. Nonetheless, TWAS prioritizes candidate causal genes more accurately than simple baselines. We suggest best practices for causal-gene prioritization with TWAS and discuss future opportunities for improvement. Our results showcase the strengths and limitations of using eQTL datasets to determine causal genes at GWAS loci.

---

GWAS have robustly associated thousands of genomic loci with complex traits. Despite this success, GWAS loci are often difficult to interpret: linkage disequilibrium (LD) often obscures the causal variants driving the association, and the causal genes mediating variant effects on the trait are rarely ascertainable from GWAS data alone<sup>1</sup>. This interpretation challenge has motivated the development of methods to prioritize causal genes at GWAS loci.

One such family of methods is TWAS, which leverage expression reference panels (eQTL cohorts with expression and genotype data) to discover gene-trait associations from GWAS datasets<sup>2-4</sup>. First, the expression panel is used to learn per-gene predictive models of expression variation by using allele counts of genetic variants in the gene's vicinity (typically within 500 kilobases or 1 megabase). These models are used to predict gene expression for each individual in the GWAS cohort. Finally, statistical associations are estimated between predicted gene expression and the trait (Fig. 1a). Expression prediction and association may be performed sequentially with individual-level GWAS data (PrediXcan<sup>2</sup>) or simultaneously with summary-level GWAS data (Fusion<sup>3</sup> and S-PrediXcan<sup>4</sup>). Closely related methods include SMR/HEIDI<sup>5-7</sup>, which performs Mendelian randomization (MR) from gene expression to trait, and GWAS-eQTL colocalization methods such as Sherlock<sup>8</sup>, coloc<sup>9,10</sup>, QTLMatch<sup>11</sup>, eCaviar<sup>12</sup>, enloc<sup>13</sup> and RTC<sup>14</sup>, which discover genes whose expression is regulated by the same causal variants that underlie a GWAS hit.

TWAS have garnered substantial interest within genetics and have been conducted for many traits and tissues<sup>15,16</sup>. Although TWAS methods are statistical tests associating genetically predicted expression and disease risk, with no guarantees of causality, a key reason for their appeal is the promise of prioritizing candidate causal genes (genes mediating the phenotypic effects of causal genetic variants) and tissues underlying GWAS loci. Unfortunately, there is a prevalent misconception that TWAS are causal-gene tests and that TWAS associations represent bona fide causal genes; in the following sections, we provide guidelines for interpreting TWAS results, highlighting scenarios in which TWAS accurately prioritize candidate causal genes and others for which TWAS-prioritized genes are likely to be non-causal.

As a motivating example illustrating both the successes and interpretational challenges of TWAS, consider *C4A*, a causal gene for schizophrenia. Variants at the *C4A* locus contribute to schizophrenia risk by increasing the brain expression of *C4A*<sup>17</sup>. A TWAS has strongly associated *C4A* with schizophrenia on the basis of brain expression data from the Genotype-Tissue Expression (GTEx) project<sup>18</sup>. Notably, *C4A* is by far the most significantly associated gene within 100 kilobases in brain tissues. *C4A* is also the most significantly associated gene in any tissue (Supplementary Table 1), even compared with other closely related genes in the complement system (*C4B*, *CFB* and *C2*). However, 8 of the 12 other genes within 100 kilobases are at least marginally significant ( $P < 0.05$ ) in some brain tissue, and 11 of 12 are highly significant ( $P < 5 \times 10^{-5}$ ) in at least one tissue. *C4A* is also more significantly associated with schizophrenia in the pancreas than in any brain tissue.

## TWAS-significant loci contain multiple associated genes

GWAS are well known to rarely identify single variant-trait associations but instead to identify blocks of associated variants in LD (Fig. 1b). Analogously, TWAS frequently identify multiple hit genes per locus<sup>16</sup> (Fig. 1c).

To explore this phenomenon, we performed TWAS in two traits and two tissues with Fusion and S-PrediXcan, by using GWAS summary statistics for low-density lipoprotein (LDL) cholesterol<sup>19</sup> and Crohn's disease<sup>20</sup>, and the 522 liver and 447 whole-blood expression samples from the Stockholm-Tartu Atherosclerosis Reverse Networks Engineering Task (STARNET) cohort<sup>21</sup> (Supplementary Fig. 2 and Supplementary Note). We grouped hit genes within 2.5 megabases and found some loci with a single hit gene but others with as many as 11 hit genes (Supplementary Fig. 3).

## Correlated expression across individuals may cause false hits

We explored the extent to which co-regulation can lead to multi-hit loci. Co-regulation is conventionally measured by correlating the expression of a pair of genes across individuals. Do genes with correlated expression with a strong TWAS hit also tend to be TWAS hits? We analyzed the *SORT1* locus in LDL/liver (TWAS  $P = 1 \times 10^{-243}$ ; Fig. 2a), the strongest hit locus across all four Fusion TWAS.

Although *SORT1* has strong evidence of causality, its locus contains eight hit genes in addition to *SORT1*, and their TWAS  $P$  values are highly related to their expression

correlation with *SORT1* (Spearman correlation = 0.75; Fig. 2b). A similar pattern holds for S-PrediXcan (Supplementary Fig. 5). The two most correlated genes, *PSRC1* and *CELSR2*, were previously noted<sup>22</sup> to share an eQTL with *SORT1* in the liver (rs646776). Given *SORT1*'s strong evidence of causality and the other genes' lack of strong literature evidence, the most parsimonious (though certainly not the only) explanation is that most or all other genes are non-causal and are prioritized only because of correlation with *SORT1*.

## Correlated predicted expression may also cause false hits

However, expression correlation is not the whole story: TWAS tests for association with genetically predicted expression, not total expression. Total expression includes genetic, environmental and technical components, and the genetic component includes contributions from common cis eQTLs (the only component reliably detectable in current TWAS methods), rare cis eQTLs and trans eQTLs. Predicted expression represents only a small component of total expression: a large-scale twin study<sup>23</sup> has found that common cis eQTLs explain only approximately 10% of genetic variance in expression.

Predicted expression correlations between same-locus genes are generally slightly higher than total expression correlations, sometimes substantially so (Fig. 3a and Supplementary Figs. 4 and 5d). A gene pair can have correlated predicted expression if the same causal eQTL regulates both genes or if two causal eQTLs in LD each regulate one of the genes<sup>24</sup>. Although only the first case counts as mechanistic co-regulation, we consider both cases together, because they are not designed to be distinguishable by TWAS: the two genes' TWAS models can rely on distinct variants even in the first case or rely on the same variant even in the second case. For instance, given a causal eQTL in near-perfect LD with another variant, an L1-penalized linear expression model (for example, LASSO or ElasticNet) may place the most weight on only one of the two variants, but which variant is chosen could change depending on statistical fluctuations in the training set.

Predicted expression correlation may lead to non-causal genes being prioritized before causal genes, even if the total expression correlation is low. This type of confounding has also been observed in gene-set analysis<sup>25</sup>. For instance, *SARS* is the main outlier in Fig. 2b and is as significant as *SORT1* despite having a total expression correlation of only ~0.2, because of its high predicted expression correlation of ~0.9 (Fig. 3a). *SARS* is also an outlier in PrediXcan for the same reason (Supplementary Fig. 5d).

Another example is the *IRF2BP2* locus in LDL/liver (Fig. 3b). *IRF2BP2* encodes an inflammation-suppressing regulatory factor with causal evidence from mouse models. *RP4-781K5.7* is a largely uncharacterized long non-coding RNA that lacks evidence of function; most long non-coding RNAs are non-essential for cell fitness<sup>26</sup>, and current evidence is compatible with most non-coding RNAs being non-functional<sup>27</sup>. Despite a negligible total expression correlation between the two genes (-0.02), *IRF2BP2*'s Fusion expression model includes GWAS hit rs556107 with a negative weight, whereas *RP4-781K5.7*'s includes the same variant (plus two linked variants) with a positive weight (Fig. 3c), thus resulting in almost perfectly anti-correlated predicted expression (-0.94) and both genes being TWAS hits. *IRF2BP2* and *RP4-781K5.7* are also both hits with S-PrediXcan, and both S-PrediXcan

and Fusion place the largest weight on rs556107 but with opposite signs (Supplementary Fig. 6).

We simulated expression and trait data ( $n_{\text{trait}} = 50,000$  individuals;  $n_{\text{expression}} = 500$ ) for 1,000 random genomic loci by using the FOCUS simulation framework<sup>24</sup> and conducted TWAS by using L2-penalized linear regression (Supplementary Note). As expected, a larger predicted expression correlation increased the probability of having a larger TWAS  $z$  score than that of the causal gene (Supplementary Table 2). However, this probability remained modestly high even when the predicted expression correlation was low, thus implying that predicted expression, though better than true expression, still imperfectly captures co-regulation.

### Shared GWAS variants may cause false hits

More generally, pairs of gene models may share variants (or at least LD partners) even if the predicted expression correlation is low, because other variants distinct between the models may ‘dilute’ the correlation. For instance, at the *NOD2* locus for Crohn’s disease/ whole blood, *NOD2* is a known causal gene, but four other genes are also Fusion hits (Fig. 4a), none of which have strong causal evidence (though rare variants in *ADCY7* have been associated with ulcerative colitis<sup>28</sup>). The model for the strongest hit gene, *BRD7*, places the most weight on rs1872691, the strongest GWAS hit in *NOD2*’s model (Fig. 4b). However, *NOD2*’s model places the most weight on two weaker GWAS hits, rs7202124 and rs1981760. Thus, even though co-regulation with *NOD2* may explain why *BRD7* is a TWAS hit, this co-regulation is not captured by the metrics that we discussed: both the predicted expression ( $-0.03$ ) and total expression ( $0.05$ ) correlations are near 0. The same five genes are also S-PrediXcan hits, and *NOD2* and *BRD7*’s models share the same rs1872691 variant, as with Fusion (Supplementary Fig. 7).

Most generally, models need not even share the same GWAS variants (or LD partners) to have spurious hits. For instance, rs4643314, the strongest GWAS hit in *BRD7*’s Fusion model, is neither shared nor in strong LD with any variants in *NOD2*’s model, although it is in weak LD with rs1872691 (Fig. 4b). Although the most parsimonious explanation is that *BRD7* is also causal, and rs4643314 acts through *BRD7*, *BRD7* lacks evidence of causality. An alternate explanation is that only *NOD2* is causal, rs4643314 acts through *NOD2* (but also happens to co-regulate *BRD7*), and *NOD2*’s model erroneously fails to include it (a false negative). One trivial reason for false negatives is variants outside the 500 kilobase/1 megabase window included in the model, which can be solved by increasing the window. More problematic causes include bias in the expression panel (‘Discussion’) and, for methods using GWAS summary statistics, LD mismatch between the expression panel and GWAS. This scenario might occur even without any false negatives, for example, if a variant in LD with rs4643314 deleteriously affects *NOD2*’s coding sequence as well as regulating *BRD7*, because TWAS is not designed to detect coding effects. Figure 5 illustrates the various types of co-regulation that may lead to non-causal TWAS hits.

## Bias with expression panels from non-trait-related tissues

Tissues with large expression panels (whole blood or lymphoblastoid) are commonly used to maximize power, even when they are mechanistically less related to the trait. To date, our case studies have used expression from mechanistically related tissues: liver for LDL and whole blood for Crohn's disease. What if we swap these tissues and use tissues without a clear mechanistic relationship? The architecture of eQTLs differs substantially across tissues: even among strong eQTLs in GTEx ( $P \approx 1 \times 10^{-10}$ ), one-quarter show a switch in the most significantly associated gene across tissues<sup>18</sup>.

We curated candidate causal genes from the literature (Supplementary Table 3) at nine LDL/liver and four Crohn's disease/whole-blood Fusion TWAS loci and examined how the hit strengths changed when we swapped tissues (Fig. 6). Notably, almost every candidate causal gene (9 of 11 for LDL and 5 of 6 for Crohn's disease) was no longer a hit in the 'opposite' tissue, because of either insufficient expression ( $n = 4$ : *PPARG*, *LPA*, *LPIN3* and *SLC22A4*) or insufficiently heritable cis expression according to Fusion's likelihood-ratio test ( $n = 10$ : *SORT1*, *IRF2BP2*, *TNKS*, *FADS3*, *ALDH2*, *KPNB1*, *SLC22A5*, *IRF1*, *CARD9*, *STAT3*). This trend held globally, albeit less strongly: genome-wide, 3,085 of 5,858 LDL/liver genes (53%) dropped out after switching to whole blood, and 1,202 of 2,118 Crohn's disease/whole-blood genes (57%) dropped out after switching to liver. Just because a gene does not drop out, and is present in both tissues as a result of shared cross-tissue regulatory architecture, causality is not necessarily implied.

More problematically, 15 other genes at the same loci were still hits (eight in LDL/whole blood and seven in Crohn's disease/liver), five with  $P < 1 \times 10^{-20}$ . This result suggests that the strategy of conducting TWAS in a sub-optimal tissue with a large expression panel is especially problematic because even if there are hits at a locus, the causal gene may not be among them.

Combining the whole-blood and liver reference panels by averaging each individual's expression in the two tissues (equivalent, for L1- and/or L2-penalized regression, to concatenating the two panels) performed more poorly than using the mechanistically related tissue alone but better than using the less related tissue alone (Supplementary Fig. 8).

## TWAS improves causal-gene prioritization

We investigated TWAS's performance at ranking (prioritizing) causal genes at loci from the previous section. We compared Fusion to two simple gene-ranking baselines (Supplementary Table 4): transcription-start-site proximity to the most significant GWAS variant within 2.5 megabases of any gene at the locus ('proximity') and median expression across GTEx individuals in the liver (for LDL genes) or whole blood (for Crohn's disease genes) ('expression'). Genes with more significant TWAS  $P$  values, smaller distances to the lead GWAS variant or higher expression had higher rankings. The mean rank of the 17 candidate causal genes was 3.9 by random per-locus ranking, 2.0 by TWAS, 2.2 by proximity ( $P = 0.5$ , two-tailed Wilcoxon signed-rank test) and 2.9 by expression ( $P = 0.006$ ). Hence, Fusion outperforms both baselines but does not significantly outperform proximity.

## Suggested best practices and future opportunities

We highlighted two vulnerabilities—co-regulation and tissue bias—that affect TWAS’s performance in causal-gene prioritization. In this section, we discuss current best practices and future opportunities for their mitigation.

One emerging approach to address co-regulation repurposes GWAS fine-mapping to TWAS, on the basis of the analogy between LD in GWAS and co-regulation in TWAS. Fine-mapping of causal gene sets (FOCUS)<sup>24</sup> directly models predicted expression correlations and uses them to assign genes posterior probabilities of causality. At the *SORT1* locus, FOCUS includes *SORT1*, *SARS* and *CELSR2* in the 90%-credible set; at the *IRF2BP2* locus, FOCUS includes both *IRF2BP2* and *RP4-781K5.7* (Fig. 2d). We recommend using fine-mapping methods such as FOCUS or, at a minimum, considering relative association strengths (*P* values and effect sizes) at a locus when interpreting TWAS results. If individual-level data are available, inferring effects jointly through penalized regression (for example, LASSO or Ridge) offers a flexible alternative (Supplementary Tables 5 and 6). Nonetheless, TWAS fine-mapping is more challenging than GWAS fine-mapping: predicted expression only imperfectly captures cis expression, owing to both variance and bias in the expression modeling (Box 1).

To address tissue bias, we recommend using an expression panel from only the most mechanistically related tissue available, even when it is smaller than other tissues’. However, using a slightly less related tissue (for example, a different region of the brain) would be advisable if the sample size would be substantially increased; the trade-off between tissue bias and sample size should be evaluated on a case-by-case basis. When a trait’s most related tissue is not known a priori, a recent approach based on LD Score regression<sup>29</sup> can be used to select among multiple reference panels. Methods to handle cross-tissue pleiotropy and cell-type heterogeneity, discussed above in the context of fine-mapping, can also mitigate tissue bias. If no sufficiently large reference panels from closely related tissues are available, we recommend aggregating information across all available tissues in a tissue-agnostic manner<sup>4,30</sup>.

When reference panels have highly dissimilar sizes across tissues, the tissue with the most significant TWAS *P* value cannot necessarily be assumed to be causal, because reference-panel size affects the *P* value. For this reason, we recommend considering TWAS effect size in addition to *P* value when investigating causal tissues for TWAS-associated genes. Even when all reference panels are similarly sized, the exact combination of tissue, cell type and context (for example, developmental stage and cellular stress) mediating the causal gene’s effect may not be captured by any panel, and this may be the case even if TWAS finds the correct causal gene (for example, *C4A* is correctly chosen on the basis of RNA-seq on adult samples even though its causal effect on schizophrenia probably occurs in adolescence). Furthermore, bias may alter the pattern of TWAS *P* values and effect sizes across tissues in unexpected ways. We caution against over-interpretation.

Several emerging topics in TWAS deserve further mention. Multi-tissue TWAS methods such as UTMOST<sup>30</sup> increase power by jointly training expression models across multiple



tissues. MultiXcan<sup>31</sup> fits a multivariate regression with phenotype as the outcome and a gene's expression across multiple tissues or contexts as the inputs to increase power. The adaptive sum of a powered score test<sup>32</sup> increases power by adaptively adjusting how much to exponentiate the weighted genotypes (genotypes times expression model weights) in the final expression-trait test, from  $\gamma = 1$  (for example, Fusion or PrediXcan) or  $\gamma = 2$  (for example, SKAT<sup>33</sup>) to  $\gamma = \infty$  (in which all weight is placed on the most significant GWAS variant, a method more appropriate than standard TWAS when there are few associated variants). Mogil et al.<sup>34</sup> have shown that between-population allele-frequency differences worsen cross-ancestry expression predictions, thus underscoring the importance of gathering diverse expression cohorts. Finally, the emerging ability to generate very large expression panels offers the promise of using trans eQTL signals to overcome the co-regulation problem<sup>35–36</sup>: although all genes at a locus may show GWAS signal at their cis eQTLs, owing to co-regulation, only the true causal genes are expected to show significant GWAS signal at their trans eQTLs as well.

## Discussion

In our case studies, we assumed that the single gene with substantial causal evidence was the sole causal gene at the locus, with some exceptions (*FADS1/2/3* and *SLC22A4/5-IRF1*). Nonetheless, other loci may contain multiple causal genes. Indeed, under an omnigenic model<sup>37</sup>, every gene may be causal to some degree, although TWAS identification of marginally causal genes as strong hits due to coregulation (effect size inflation) remains problematic. Furthermore, the expression of a 'non-causal' gene may causally influence expression of the causal gene merely by being transcribed, even if the gene is non-coding or its protein product is non-causal<sup>38</sup>.

Co-regulation and tissue bias affect other methods integrating GWAS and expression data. Testing of gene-trait associations based on MR<sup>5–7</sup> is vulnerable, because co-regulation, as a form of pleiotropy, violates one of the core assumptions of MR<sup>39</sup>. Although the HEIDI test<sup>5</sup> corrects for the case in which two genes have distinct but linked causal variants, it does not correct for the case in which they share the same causal variant. GWAS-eQTL colocalization methods such as Sherlock<sup>8</sup>, coloc<sup>9–10</sup>, QTLMatch<sup>11</sup>, eCaviar<sup>12</sup>, enloc<sup>13</sup> and RTC<sup>14</sup> are also vulnerable. The more tightly a pair of genes is co-regulated in cis, the more difficult it becomes to distinguish causality on the basis of GWAS and expression data alone. Our results underscore the need for computational and experimental methods that move beyond expression variation across individuals to complement TWAS in identifying causal genes at GWAS loci.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We gratefully acknowledge J. Pritchard, H. Tang and members of the laboratory of N. Zaitlen for helpful discussions. This work was funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) (grant PGSD3-476082-2015 to M.W.); a Stanford Bio-X Bowes fellowship (to M.W.); a Stanford Graduate Fellowship (to N.S.-A.); a National Defense Science & Engineering Grant (to N.S.-A.); NIH grants

1DP2OD022870 and U01HG009431 (to A.K.), 1U24HG008956 and 5U01HG009080 (to M.A.R.), R01HG009120 and R01MH115676 (to B.P.), R01MH107666, R01MH101820 and P30DK20595 (to H.K.I.), and R01HL125863 and R21TR001739 (to J.L.M.B.); NHGRI grant R01HG010140 (to M.A.R.); Leducq Foundation grant 12CVD02 (to J.L.M.B.); and American Heart Association grant A14SFRN20840000 (to J.L.M.B.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

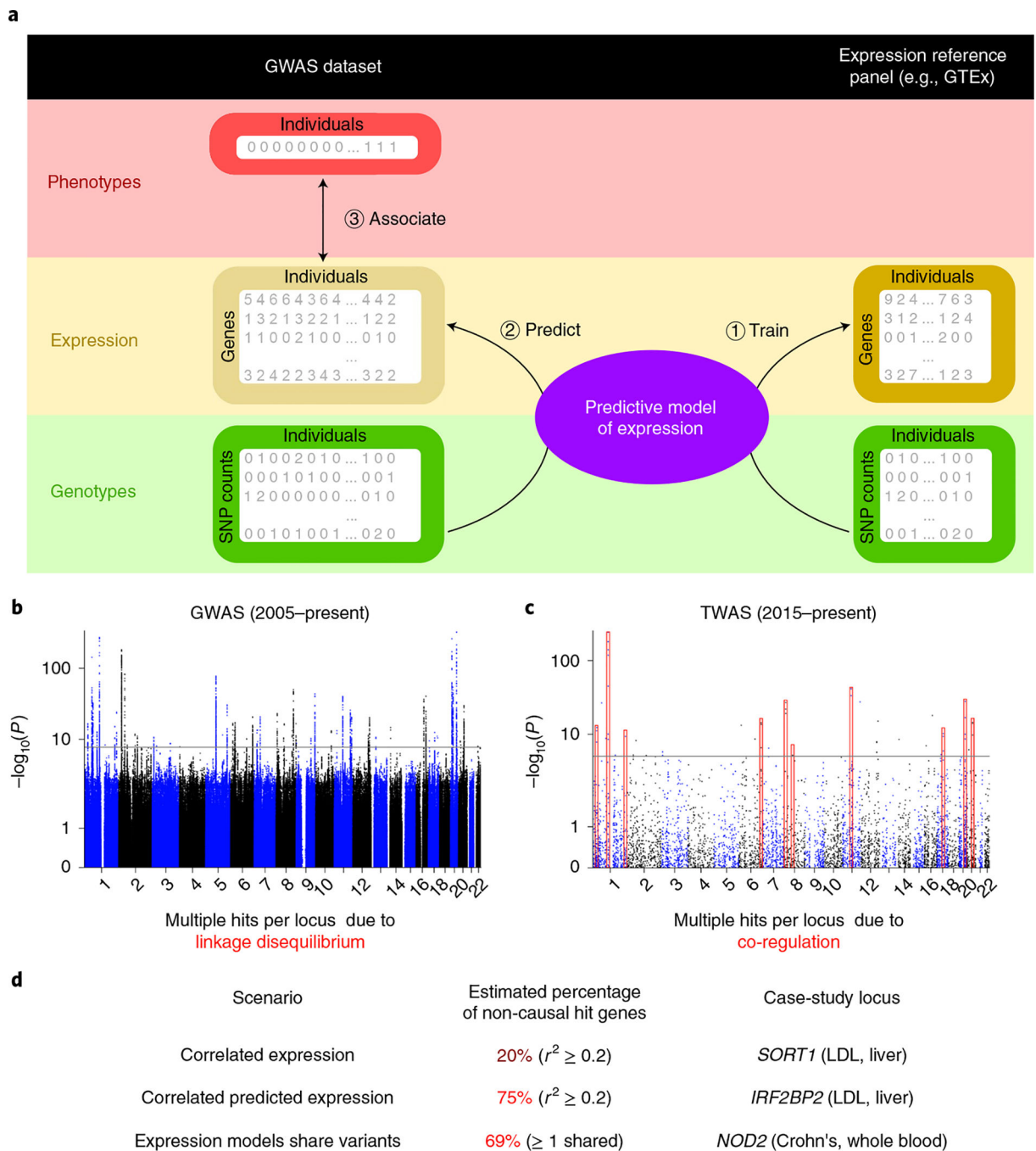
## References

1. Gallagher MD & Chen-Plotkin AS The post-GWAS era: from association to function. *Am. J. Hum. Genet* 102, 717–730 (2018). [PubMed: 29727686]
2. Gamazon ER et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet* 47, 1091–1098 (2015). [PubMed: 26258848]
3. Gusev A et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet* 48, 245–252 (2016). [PubMed: 26854917]
4. Barbeira AN et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun* 9, 1825 (2018). [PubMed: 29739930]
5. Zhu Z et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet* 48, 481–487 (2016). [PubMed: 27019110]
6. Hauberg ME et al. Large-scale identification of common trait and disease variants affecting gene expression. *Am. J. Hum. Genet* 100, 885–894 (2017). [PubMed: 28552197]
7. Pavlides JMW et al. Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Med* 8, 84 (2016). [PubMed: 27506385]
8. He X et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet* 92, 667–680 (2013). [PubMed: 23643380]
9. Wallace C et al. Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Hum. Mol. Genet* 21, 2815–2824 (2012). [PubMed: 22403184]
10. Giambartolomei C et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 10, e1004383 (2014). [PubMed: 24830394]
11. Plagnol V, Smyth DJ, Todd JA & Clayton DG Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics* 10, 327–334 (2009). [PubMed: 19039033]
12. Hormozdiari F et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet* 99, 1245–1260 (2016). [PubMed: 27866706]
13. Wen X, Pique-Regi R & Luca F Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet* 13, e1006646 (2017). [PubMed: 28278150]
14. Nica AC et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* 6, e1000895 (2010). [PubMed: 20369022]
15. 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). [PubMed: 28238358]
16. Gusev A et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet* 50, 538–548 (2018). [PubMed: 29632383]
17. Sekar A et al. Schizophrenia risk from complex variation of complement component 4. *Nature* 530, 177–183 (2016). [PubMed: 26814963]
18. G TEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* 550, 204–213 (2017). [PubMed: 29022597]
19. Willer CJ et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet* 45, 1274–1283 (2013). [PubMed: 24097068]
20. Liu JZ et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet* 47, 979–986 (2015). [PubMed: 26192919]

21. Franzen O et al. Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science* 353, 827–830 (2016). [PubMed: 27540175]
22. Musunuru K et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466, 714–719 (2010). [PubMed: 20686566]
23. Grundberg E et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet* 44, 1084–1089 (2012). [PubMed: 22941192]
24. Mancuso N et al. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet* 10.1038/s41588-019-0367-1 (2019).
25. de Leeuw CA, Neale BM, Heskes T & Posthuma D The statistical properties of gene-set analysis. *Nat. Rev. Genet* 17, 353–364 (2016). [PubMed: 27070863]
26. Liu SJ et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* 355, aah7111 (2017). [PubMed: 27980086]
27. Palazzo AF & Lee ES Non-coding RNA: what is functional and what is junk? *Front. Genet* 6, 2 (2015). [PubMed: 25674102]
28. Luo Y et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat. Genet* 49, 186–192 (2017). [PubMed: 28067910]
29. Finucane HK et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet* 50, 621–629 (2018). [PubMed: 29632380]
30. Barbeira AN et al. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet* 15, e1007889 (2019). [PubMed: 30668570]
31. Hu Y et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet* 51, 568–576 (2019). [PubMed: 30804563]
32. Xu Z, Wu C, Wei P & Pan W A powerful framework for integrating eQTL and GWAS summary data. *Genetics* 207, 893–902 (2017). [PubMed: 28893853]
33. Wu MC et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet* 89, 82–93 (2011). [PubMed: 21737059]
34. Mogil LS et al. Genetic architecture of gene expression traits across diverse populations. *PLoS Genet* 14, e1007586 (2018). [PubMed: 30096133]
35. Vosa U et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis Preprint at <https://www.biorxiv.org/content/10.1101/447367v1> (2018).
36. Wheeler HE et al. Imputed gene associations identify replicable transacting genes enriched in transcription pathways and complex traits Preprint at <https://www.biorxiv.org/content/10.1101/471748v1> (2018).
37. Boyle EA, Li YI & Pritchard JK An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186 (2017). [PubMed: 28622505]
38. Engreitz JM et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* 539, 452–455 (2016). [PubMed: 27783602]
39. Solovieff N, Cotsapas C, Lee PH, Purcell SM. & Smoller JW Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet* 14, 483–495 (2013). [PubMed: 23752797]
40. Bhutani K, Sarkar A, Park Y, Kellis M & Schork NJ Modeling prediction error improves power of transcriptome-wide association studies Preprint at <https://www.biorxiv.org/content/10.1101/108316v1> (2017).
41. Ongen H et al. Estimating the causal tissues for complex traits and diseases. *Nat. Genet* 49, 1676–1683 (2017). [PubMed: 29058715]
42. Claussnitzer M et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med* 373, 895–907 (2015).
43. Regev A et al. The Human Cell Atlas. *eLife* 6, e27041 (2017). [PubMed: 29206104]
44. Stegle O, Parts L, Piipari M, Winn J & Durbin R Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc* 7, 500–507 (2012). [PubMed: 22343431]

**Box 1 |****Sources of variance and bias in expression modeling**

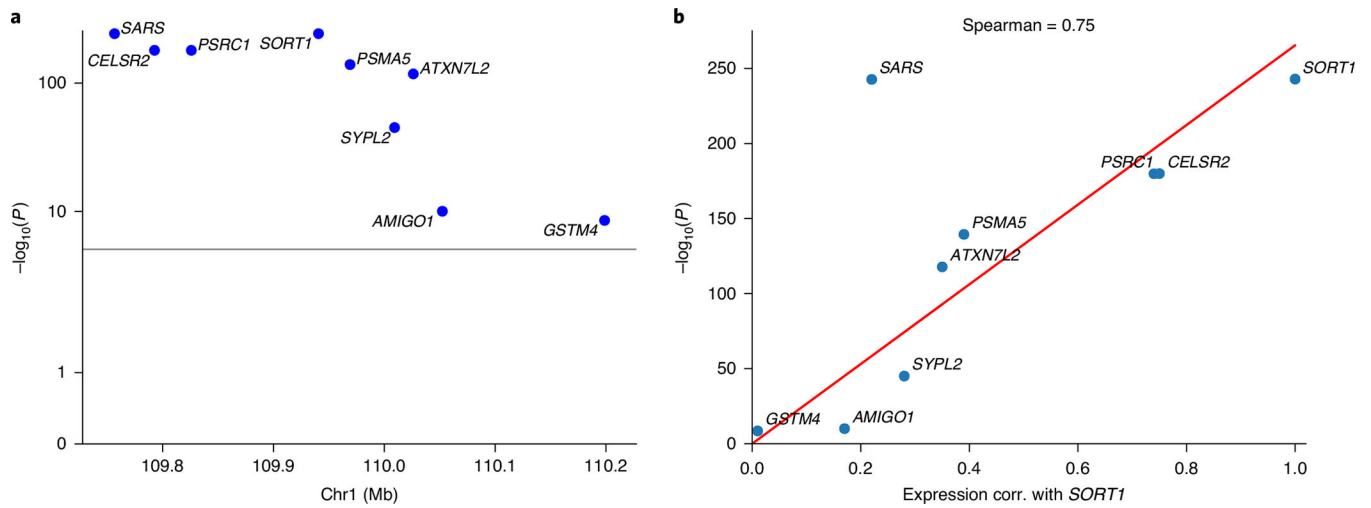
- **Finite-sized reference panel.** The main source of variance is the finite reference-panel size. This variance can be mitigated with Bayesian methods that explicitly model expression-prediction error<sup>40</sup>. This variance will decrease in the future as reference-panel sizes increase.
- **Pleiotropy across tissues.** Traits rarely act through a single tissue: some genes may be causal in tissues different from the reference panel's, thus introducing bias. Estimating causal tissues on a per-locus basis is an active area of research<sup>41</sup> and could be integrated into future TWAS fine-mapping.
- **Cell-type heterogeneity.** Most existing reference panels are heterogeneous, comprising multiple distinct cell types and states. Genes may be causal for only a single cell type/ state: a study identifying *IRX3* and *IRX5* as causal genes at the *FTO* locus has found genotype-expression associations in primary preadipocytes, representing a minority of adipose cells, but not whole adipose tissue<sup>42</sup>. Cell-type heterogeneity within/between samples (from blood and immune-cell infiltration, or genetically driven differences in cell-type proportions within a tissue) can worsen bias. Single-cell RNA sequencing enables reference panels for individual cell types/ states, most prominently through the Human Cell Atlas<sup>43</sup>.
- **Bias in expression quantification.** Time of day, physiological state (time since eating or exercise, or disease status) or cause of death may bias expression measurements, even after covariate correction through methods such as probabilistic estimation of expression residuals<sup>44</sup> (PEER). Covariates may be captured by a gene's expression model if they correlate with variants near the gene.



**Fig. 1 | TWAS, like GWAS, frequently has multiple significant associations per locus.**

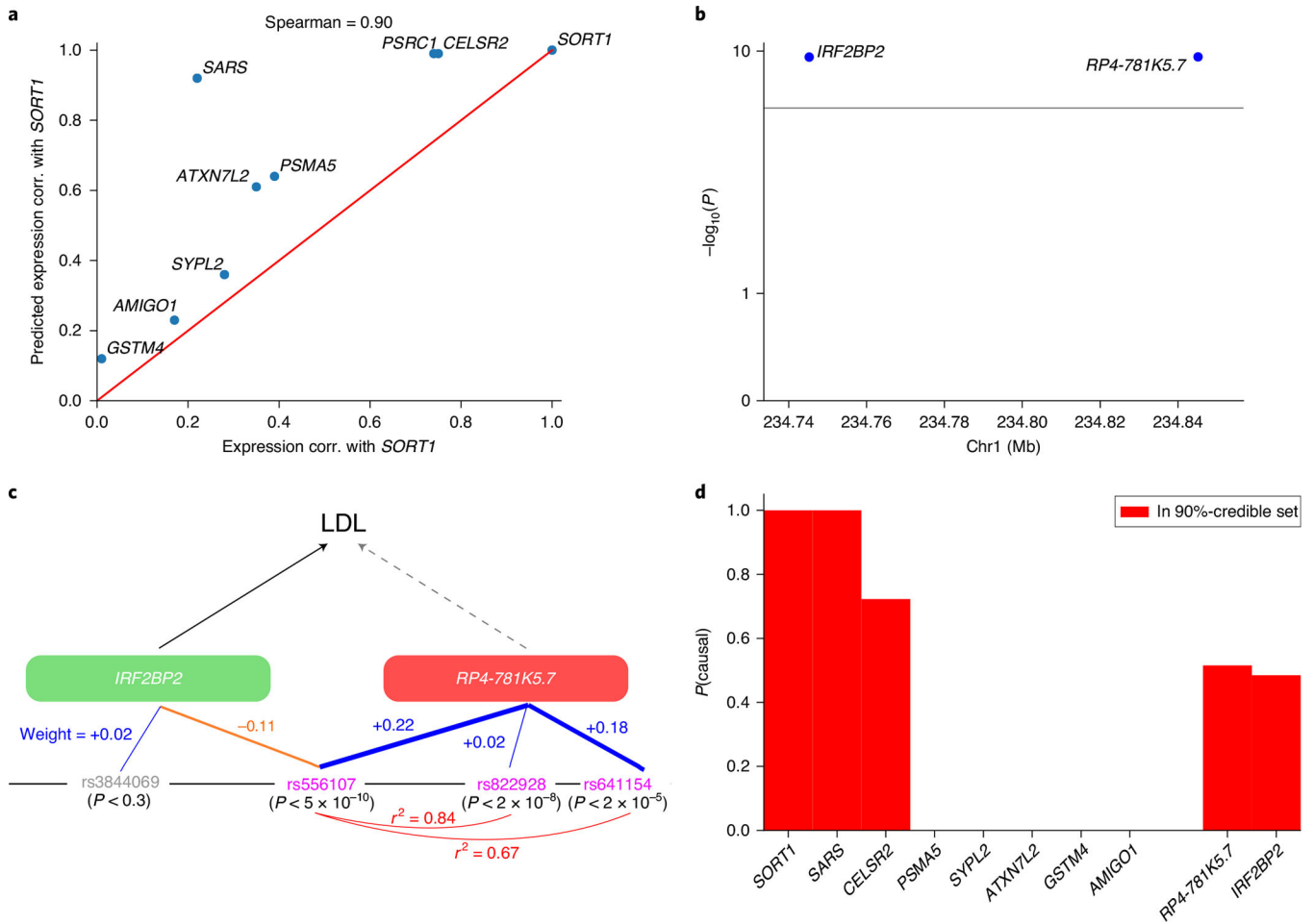
**a.** An overview of TWAS. Briefly, TWAS involves: (i) training a predictive model of expression from genotype on a reference panel such as GTEx; (ii) using this model to predict expression for individuals in the GWAS cohort; and (iii) associating this predicted expression with the trait. **b,c.** Manhattan plots of GWAS (**b**) and Fusion TWAS (**c**) for LDL cholesterol, using GWAS summary statistics from the Global Lipids Genetics Consortium and liver expression from the STARNET cohort (Supplementary Note). GWAS has multiple hits per locus, owing to LD, and TWAS has multiple hits per locus, owing to co-regulation

(which can also be driven in part by LD; described below), as explored in the main text. Clusters of multiple adjacent TWAS hit genes are highlighted in red. **d**, Three scenarios in which co-regulation can lead to multiple hits per locus, and the estimated percentage of non-causal hit genes subject to each scenario; each scenario is presented in a case study later in the text. To estimate the percentages, we grouped hits into 2.5-megabase clumps and made the approximation that genes that were not the top hit in multi-hit clumps were non-causal; we then calculated the percentage of these genes with total or predicted expression  $r^2 \geq 0.2$  or  $\geq 1$  shared variant with the top hit in their block, aggregating genes across the LDL/liver and Crohn's disease/whole-blood TWAS. The full distributions of the total and predicted expression correlations and number of shared variants are shown in Supplementary Fig. 1, separated by study.



**Fig. 2 | Co-regulation strongly predicts TWAS hit strength at the *SORT1* locus.**

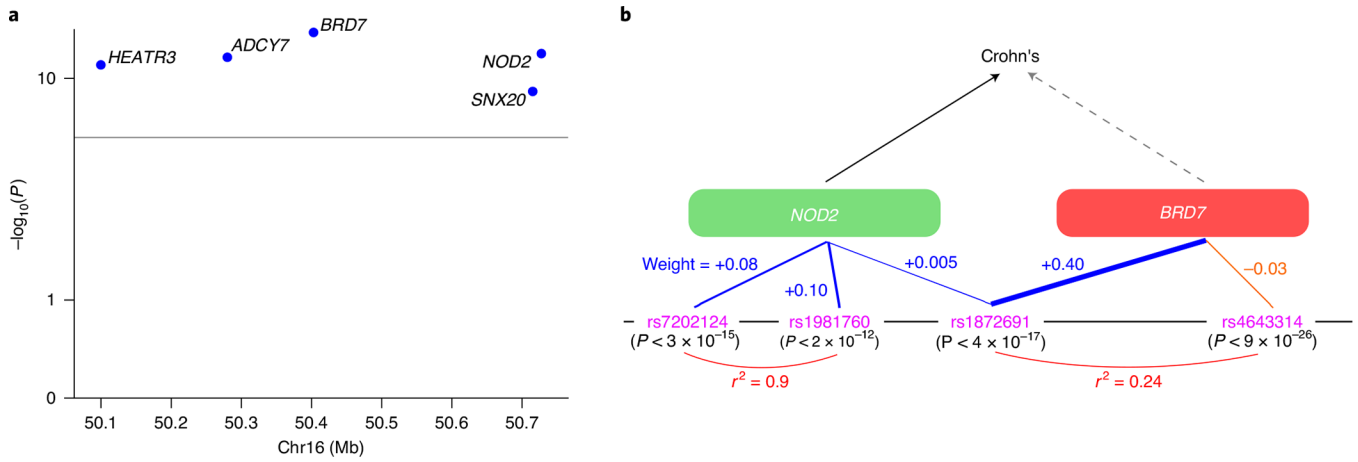
**a**, Fusion Manhattan plot of the *SORT1* locus. **b**, Expression correlation (corr.) with *SORT1* versus TWAS *P* value, for each gene in the *SORT1* locus. Chr, chromosome.



**Fig. 3 | Correlated predicted expression can cause non-causal hits even in the absence of correlated total expression.**

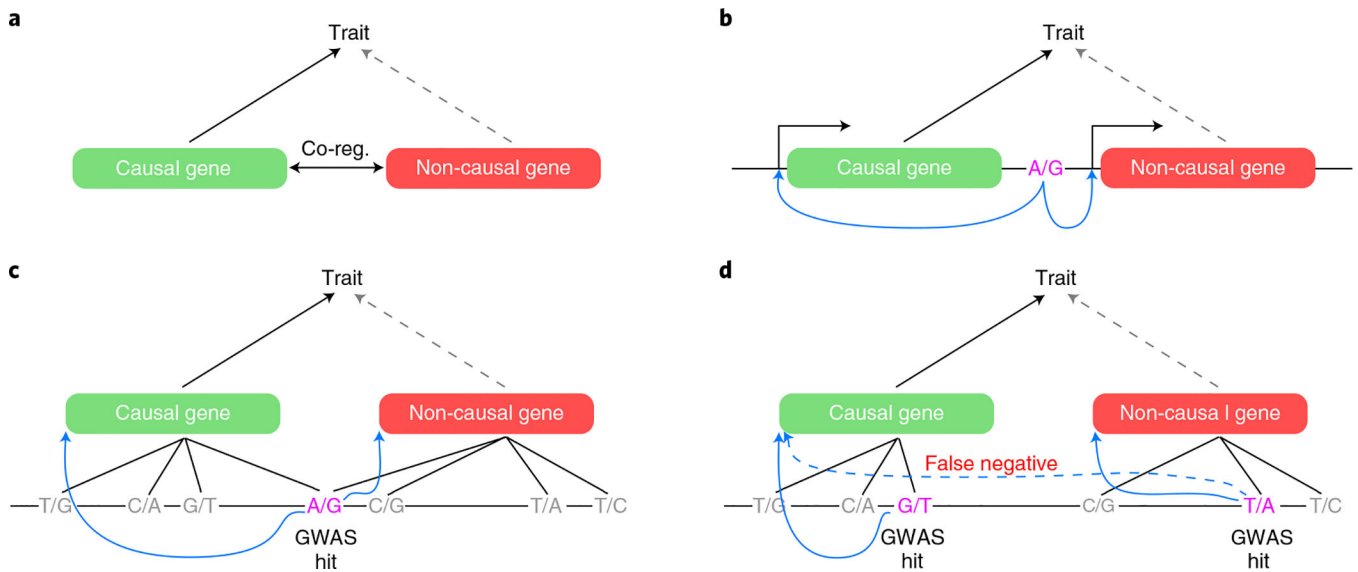
**a**, For nearby genes, Fusion-predicted expression correlations tend to be higher than total expression correlations, for example, at the *SORT1* locus. **b**, Fusion Manhattan plot of the *IRF2BP2* locus, where *RP4-781K5.7* is a likely non-causal hit due to predicted expression correlation with *IRF2BP2*. **c**, Details of the two genes' Fusion expression models: a line between a variant's rs number and a gene indicates that the variant is included in the gene's expression model with either a positive weight (blue) or a negative weight (orange); the thickness of the line increases with the magnitude of the weight; red arcs indicate LD. Pink rs numbers are GWAS hits (genome-wide significant or sub-significant), whereas gray rs numbers are not. For clarity, four variants with weights less than 0.05 in magnitude for *IRF2BP2* (rs2175594,  $P = 0.02$ , weight +0.03; rs2439500,  $P = 0.2$ , weight = +0.01; rs11588636,  $P = 0.3$ , weight = -0.03; and rs780256,  $P = 0.9$ , weight = -0.03) and five variants for *RP4-781K5.7* (rs478425,  $P = 0.01$ , weight = +0.02; rs633269,  $P = 0.02$ , weight = +0.01; rs881070,  $P = 0.06$ , weight = -0.02; rs673283,  $P = 0.1$ , weight = +0.004; and rs9659229,  $P = 0.1$ , weight = -0.04) are not shown. **d**, Estimated causal probability for each significant gene from Fusion at the *SORT1* and *IRF2BP2* loci, according to TWAS gene-based fine-mapping with the FOCUS method.





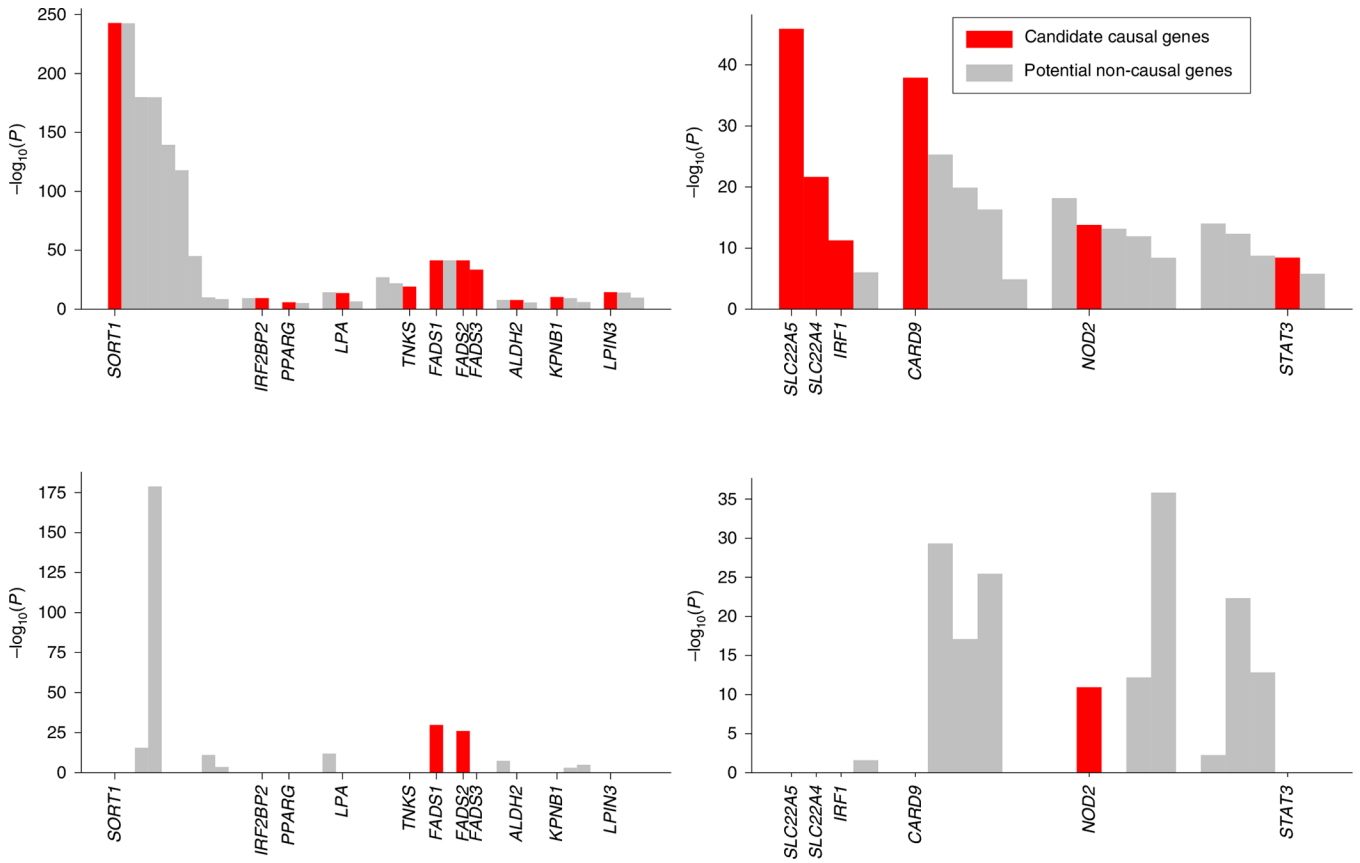
**Fig. 4 | Sharing of GWAS variants between expression models can contribute to non-causal hits even without correlated predicted expression.**

**a**, Fusion Manhattan plot of the *NOD2* locus. **b**, Details of the expression models of *NOD2* and *BRD7*; as in Fig. 2, a line between a variant's rs number and a gene indicates that the variant is included in the gene's expression model with either a positive weight (blue) or a negative weight (orange), with the thickness of the line increasing with the magnitude of the weight. Red arcs indicate LD.



**Fig. 5 |. co-regulation scenarios in TWAS that may lead to non-causal hits, from least to most general.**

**a.** Correlated expression across individuals: the causal gene has correlated total expression with another gene, which may become a non-causal TWAS hit. Co-reg., co-regulation. **b.** Correlated predicted expression across individuals: even if total expression correlation is low, predicted expression correlation may be high if the same variants (or variants in LD) regulate both genes and are included in both models. **c.** Sharing of GWAS hits: even if the two genes' models include largely distinct variants, and predicted expression correlation is low, only a single shared GWAS hit variant (or variant in LD) is necessary for both genes to be TWAS hits. **d.** Both models include distinct GWAS hits: in the most general case, the GWAS hits driving the signal at the two genes may not be in LD with each other, for instance if the non-causal gene's GWAS hit happens to regulate the causal gene as well, but this connection is missed by the expression modeling (a false negative), or if the causal gene's GWAS hit acts via a coding mechanism (not shown).



**Fig. 6 | Most candidate causal genes drop out after switching to a tissue with a less clear mechanistic relationship to the trait, owing to a lack of sufficient expression or sufficiently heritable expression.** Fusion TWAS  $P$  values at nine LDL/liver and four Crohn’s disease/whole-blood multi-hit loci, using expression from tissues with a clear (top row) or less clear or absent (bottom row) mechanistic relationship to the trait. Candidate causal genes are labeled and colored red.