Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits

Alvaro N Barbeira[1,†], Rodrigo Bonazzola[1,†], Eric R Gamazon[2,3,4,5,†], Yanyu Liang[1,†], YoSon Park[6,7,†], Sarah Kim-Hellmuth[8,9,10], Gao Wang[11], Zhuoxun Jiang[1], Dan Zhou[2], Farhad Hormozdiari[12, 13], Boxiang Liu[14], Abhiram Rao[14], Andrew R Hamel[12,15], Milton D Pividori[1], François Aguet[12], GTEx GWAS Working Group, Lisa Bastarache[16,17], Daniel M Jordan[18, 19, 20], Marie Verbanck[18, 19, 20, 21], Ron Do[18,19,20], GTEx Consortium, Matthew Stephens[11], Kristin Ardlie[12], Mark McCarthy[22], Stephen B Montgomery[23,24], Ayellet V Segrè[12, 15], Christopher D. Brown[6], Tuuli Lappalainen[9,10], Xiaoquan Wen[25], Hae Kyung Im[1,*]

1 Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA
2 Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
3 Data Science Institute, Vanderbilt University, Nashville, TN, USA
4 Clare Hall, University of Cambridge, Cambridge, UK
5 MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
6 Department of Genetics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
7 Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
8 Statistical Genetics, Max Planck Institute of Psychiatry, Munich, Germany
9 New York Genome Center, New York, NY, USA
10 Department of Systems Biology, Columbia University, New York, NY, USA
11 Department of Human Genetics, University of Chicago, Chicago, IL, USA
12 The Broad Institute of MIT and Harvard, Cambridge, MA, USA
13 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
14 Department of Biology, Stanford University, Stanford, California 94305, USA
15 Ocular Genomics Institute, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA
16 Department of Biomedical Informatics, Department of Medicine, Vanderbilt University, Nashville, TN, USA
17 Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, TN, USA
18 Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA
19 Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA
20 The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount sinai, New York, New York, USA
21 Université de Paris - EA 7537 BIOSTM, France
22 University of Oxford, United Kingdom
23 Department of Genetics, Stanford University, Stanford, CA, USA
24 Department of Pathology, Stanford University, Stanford, CA, USA
25 Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

†: These authors contributed equally to this work, alphabetic order;

∗: Correspondence to haky@uchicago.edu

Abstract

The resources generated by the GTEx consortium offer unprecedented opportunities to advance our understanding of the biology of human traits and diseases. Here, we present an in-depth examination of the phenotypic consequences of transcriptome regulation and a blueprint for the functional interpretation of genetic loci discovered by genome-wide association studies (GWAS). Across a broad set of complex traits and diseases, we find widespread dose-dependent effects of RNA expression and splicing, with higher impact on molecular phenotypes translating into higher impact downstream. Using colocalization and association approaches that take into account the observed allelic heterogeneity, we propose potential target genes for 47% (2,519 out of 5,385) of the GWAS loci examined. Our results demonstrate the translational relevance of the GTEx resources and highlight the need to increase their resolution and breadth to further our understanding of the genotype-phenotype link.

Introduction

In the last decade, the number of reproducible genetic associations with complex traits that have emerged from genome-wide association studies (GWAS) has substantially grown. Many of the identified associations lie in non-coding regions of the genome, suggesting that they influence disease pathophysiology and complex traits via gene regulatory changes. Integrative studies of molecular quantitative trait loci (QTL) (*1*) have pinpointed gene expression as a key intermediate molecular phenotype, and improved functional interpretation of GWAS findings, spanning immunological diseases (*2*), various cancers (*3, 4*), lipid traits (*5, 6*), and a broad array of other complex traits.

Large-scale international efforts such as the Genotype-Tissue Expression (GTEx) Consortium have provided an atlas of the regulatory landscape of gene expression and splicing variation in a broad collection of primary human tissues (*7–9*). Nearly all protein-coding genes in the genome now have at least one local variant identified to be associated with expression and a majority also have common variants affecting splicing (FDR $< 5\%$) (*9*).

In parallel, there has been an explosive growth in the number of genetic discoveries across a large number of phenotypes, prompting the development of integrative approaches to characterize the function of GWAS findings (*10–14*). Nevertheless, our understanding of underlying biological mechanisms for most complex traits substantially lags behind the improved efficiency of discovery of genetic associations, made possible by large-scale biobanks and GWAS meta-analyses.

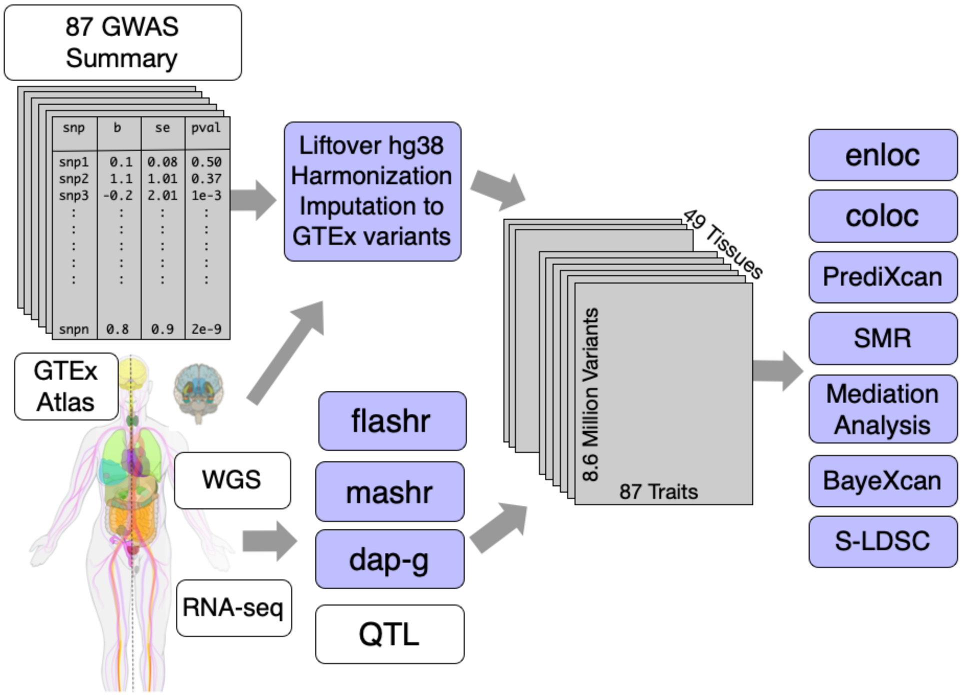
One of the primary tools for functional interpretation of GWAS associations has been integrative analysis of molecular QTLs. Colocalization approaches that seek to establish shared causal variants (e.g., eCaviar (*15*), *enloc* (*16*), and *coloc* (*17*)), enrichment analysis (S-LDSC (*18*) and QTLEnrich (*11*)) or mediation and association methods (SMR (*12*), TWAS (*13*) and PrediXcan (*19*)) have provided important insights, but they are often used in isolation, and there have been limited prior assessments of power and error rates associated with each (*20*). Their applications often fall short of providing a comprehensive, biologically interpretable view across multiple methods, traits, and tissues or offering guidelines that are generalizable to other contexts. Thus, a comprehensive assessment of expression and splicing QTLs for their contributions to disease susceptibility and other complex traits requires the development of novel methodologies with improved resolution and interpretability.

Here, we develop novel methods, approaches, and resources that elucidate how genetic variants associated with gene expression (cis-eQTLs) or splicing (cis-sQTLs) contribute to, or mediate, the functional mechanisms underlying a wide array of complex diseases and quantitative traits. Since splicing QTLs have largely been understudied, we perform a comprehensive integrative study of this class of QTLs, in a broad collection of tissues, and GWAS associations. We leverage full summary results from 87 GWAS for discovery analyses and use independent datasets for replication and validation. Notably, we find widespread dose-dependent effect of cis-QTLs on traits through multiple lines of evidence. We examine the importance of considering, or correcting for, false functional links attributed to GWAS loci due to neighboring but distinct causal variants. To identify predicted causal effects among the complex trait associated QTLs, we conduct systematic evaluation across different methods. Furthermore, we provide guidelines for employing complementary methods to map the regulatory mechanisms underlying genetic associations with complex traits.

Harmonized GWAS and QTL datasets

The final GTEx data release (v8) includes 54 primary human tissues, 49 of which included at least 65 samples and were used for cis-QTL mapping (Fig. 1) (*9*). This phase increases the number of available tissues relative to previous GTEx publications (v6p; 44 tissues) (*8*) and doubles the sample size from 7,051 RNA-Seq samples from 449 individuals to 15,253 samples from 838 individuals, now all with whole genome sequencing data as opposed to genotype imputation in v6p. Furthermore, the v8 core data resources now include splicing QTLs (*9*), allowing parallel analysis of both expression and splicing variation underlying complex traits. Using these resources, we investigated the contribution of expression and splicing QTLs in cis (eQTL and sQTL, respectively) to complex trait variance and etiology.

We retained 87 GWAS datasets representing 74 distinct complex traits for further analyses (table S1 and fig. S1) after stringent quality control (fig. S2; (*21*)) and data harmonization(fig. S3, fig. S4).



snp	b	se	pval
snp1	0.1	0.08	0.50
snp2	1.1	1.01	0.37
snp3	-0.2	2.01	1e-3
⋮	⋮	⋮	⋮
snpn	0.8	0.9	2e-9

Fig. 1. Overview of workflow for mapping complex trait associated QTLs. Full variant association summary statistics results from 114 GWAS were downloaded, standardized, and imputed to the GTEx v8 WGS variant calls (maf>0.01) for analyses. A total of 8.87 million imputed and genotyped variants were investigated to identify trait-associated QTLs. A total of 49 tissues, 87 traits, and 23,268 protein-coding genes and lncRNAs remained after stringent quality assurance protocols and selection criteria. A wide array of complex trait classes, including cardiometabolic, anthropometric, and psychiatric traits, were included.

Dose-dependent effects of expression and splicing regulation on complex traits

The robust enrichment of GWAS variants (fig. S5, fig. S6) and heritability in eQTLs and sQTLs has been established by multiple studies, including our analysis of GTEx v8 data (*9, 21*). This observation provides strong support for a causal role of expression and splicing regulation in complex traits. Importantly, transcriptome-based PrediXcan/TWAS methods implicitly assume that gene regulation affects complex traits in a dose-dependent manner. Nevertheless, there has been little formal support for this assumption. Here, we tested a dose-dependent effect on traits, i.e., whether e/sVariants with higher impact on gene expression or splicing lead to higher impact on a complex trait and a larger GWAS effect (Fig. 2A). We note these analyses were performed with fine-mapped variants (*21*). The dose-dependent effect was quantified by the genic mediating effect, β_g, which reflects how strongly the change in a given gene's dosage affects a trait, with a non-causal gene having a flat slope ($\beta_g = 0$).

To get a first-order approximation of the average mediating effect size (fig. S10), we calculated the correlation between the magnitude of the QTL effect size and that of the GWAS effect size for each tissue-trait pair, using fine-mapped QTLs with the largest posterior inclusion probability within each causal LD cluster (*21*). Importantly, this correlation reflects the mediated effect and corrects for LD contamination (see (*21*); fig. S9). As hypothesized, we found a significant positive correlation between the GWAS and QTL effects, consistently across all 87 by 49 trait-tissue pairs. The average correlations were 0.18 (s.e. $= 0.004$, $p < 1 \times 10^{-30}$) and 0.25 (s.e. $= 0.006$, $p < 1 \times 10^{-30}$) for expression and splicing, respectively (Fig. 2B and fig. S7). Averages were calculated taking into account correlation between tissues (*21*), and p-values were calculated against permuted null with matched local LD (*21*). These results provide the first line of evidence of the dose-response effect.

Correlating the eQTL and GWAS effect sizes across genes has additional noise arising from different genes having different levels of dosage sensitivity - i.e., a similar trait effect may arise from a small change in one gene's expression and a large change in another one. To account for this heterogeneity in mediating effect or, equivalently, dosage sensitivity, we modeled the slope (β_g) as a random variable following the normal distribution as $\beta_g \sim \mathcal{N}(0, \sigma^2_{\text{gene}})$, where the variance is a measure of the average mediated effect for each trait (*22*). These effects were significantly larger than the permuted null (expression: $p = 1.8 \times 10^{-9}$; splicing: $p = 2.5 \times 10^{-7}$; Fig. 2C). These results indicate that strong genetic effects on expression or splicing are more likely to have a strong association to complex traits, adding further support for a dose-dependent relationship between gene regulation and downstream traits (Fig. 2E).

Furthermore, the high degree of allelic heterogeneity in the GTEx data enables analysis of the GWAS contribution of multiple independent eQTL effects for the same gene (Fig. 2D, (*9, 21*)). Allelic heterogeneity allows a more precise analysis of dose-dependent effects through comparison of the dose sensitivity between primary and secondary eQTLs, estimated as the ratio of GWAS to eQTL effects, $\hat{\beta} = \hat{\delta}/\hat{\gamma}$ (fig. S11). This method is equivalent to Mendelian randomization approaches, estimating the likely causal effect of a gene on a trait. More graphically (Fig. 2A), we tested whether the points in a dose-response plot align along the corresponding slope line for each gene.

We found a significantly higher correlation in mediating effect between primary and secondary eQTLs for a given gene compared to a null distribution obtained by sampling GWAS effect sizes from a bivariate normal distribution to account for the small observed LD between primary and secondary eQTLs (Fig. 2D-E) while keeping the observed eQTL effect sizes ($p < 1 \times 10^{-30}$).

Interestingly, the correlation between primary and secondary eQTLs for non-colocalized genes (rcp $<$ 0.01), which were used as controls (*9, 21*), was significantly higher than this more accurate null, indicating that even eQTLs with very low colocalization probability include many genes that are likely causal. Given this concordance between multiple independent eQTLs, it is clear that with widespread allelic heterogeneity detected with currently available sample sizes, methods that assume single causal variants are highly limited. The approaches described here enable insights into how multiple regulatory effects converge to mediate the same trait association.

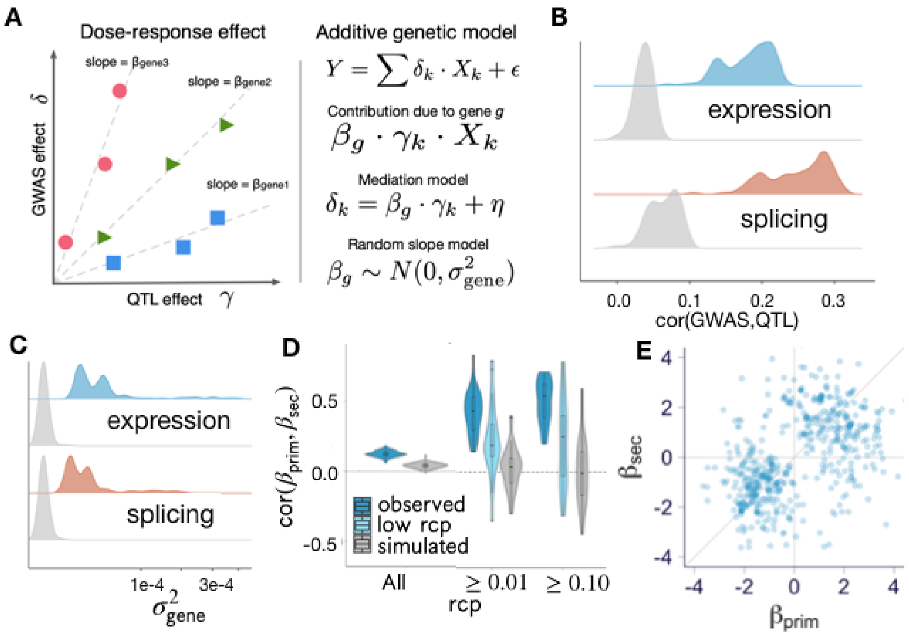


Fig. 2. Dose-dependent effects of QTLs on complex traits. Here all analyses were performed with fine-mapped variants. **(A)** Schematic representation of dose-response model. **(B)** Correlation between QTL and GWAS effects, Cor($|\hat{\delta}|, |\hat{\gamma}|$), median across 49 tissues is shown. Gray distribution represents permuted null with matched local LD. **(C)** Average mediated effects from mediation model (σ^2_{gene}, median across tissues). Gray distribution represents permuted null with matched local LD. **(D)** Correlation of mediated effects between primary (ordered by significance) and secondary eQTLs for different colocalization thresholds (rcp $\geq$ 0, 0.01, 0.10) in dark blue. Correlation for genes rcp $\leq$ 0.01 and matched LD is shown in light blue. Correlation for null calculated with simulated GWAS effects from bivariate normal with LD and observed QTL effects is shown in gray. **(E)** Mediated effects of secondary vs primary eQTLs of colocalized genes (rcp $>$ 0.10) in whole blood, genes for all 87 traits are shown.

Causal gene prediction and prioritization

In addition to genome-wide analyses that shed light on the molecular architecture of complex traits, QTL analysis of GWAS data can identify potential causal genes and molecular changes in individual GWAS loci. Towards this end, we analyzed colocalization and genetically predicted regulation association (Fig. 3A). After evaluating the performance of *coloc* and *enloc* (*16, 17*), we chose *enloc* as our primary approach, due to its use of hierarchical models to estimate colocalization priors (*16*) and its ability to account for multiple causal variants. The *coloc* assumption of a single causal variant drastically reduces performance especially in large QTL datasets such as GTEx with widespread allelic heterogeneity (fig. S27). We estimated the posterior regional colocalization probability (rcp), using *enloc*, for 12,072,964 (tissue, gene, GWAS locus, trait)-tuples and 67,943,800 (tissue, splicing event, GWAS locus, trait)-tuples. We used rcp>0.5 as a stringent evidence of colocalization.

In total, we identified 3,477 (15% of 23,963) unique genes colocalizing with GWAS hits (rcp > 0.5) across all traits and tissues analyzed (fig. S14A). 3,157 splicing events (1% out of 310,042) colocalized with GWAS hits, corresponding to 1,226 genes with at least one colocalized splicing event (5% of 23,963, fig. S14B).

To assess the performance of different colocalization approaches, we compared the fine-mapping results based on two large GWAS of height in European-ancestry individuals: GIANT (*23*) and UK Biobank. Colocalization of signals in two traits occurs when they share fine-mapped variants, i.e. variants with posterior causal probability greater than 0. We found that 85% of fine-mapped variants (posterior inclusion probability > 0.25) in GIANT had posterior probability of 0 in the UK Biobank, which implies that the colocalization probability contributed by these variants is 0. Notably, 48% of GIANT's fine-mapped loci had no overlap with the UK Biobank's loci, resulting in a colocalization probability of 0. Given the larger sample size in the UK Biobank, this low colocalization cannot be attributed to lack of power but is likely due in part to reference LD differences. Thus, colocalization is highly conservative and may miss many causal genes, and low colocalization probability should not be interpreted as evidence of lack of a causal link between the molecular phenotype and the GWAS trait.

A complementary approach to colocalization is to estimate GWAS association for genetically predicted gene expression or splicing (*19*). The GTEx v8 data provides an important expansion of these analyses, allowing generation of prediction models in 49 tissues with whole genome sequencing data to impute gene expression and splicing variation. We trained prediction models using a variety of approaches and selected the top performing one based on precision, recall, and other metrics (*21, 24*). Briefly, the optimal model uses fine-mapping probabilities for feature selection and exploits global patterns of tissue sharing of regulation (see (*21*); fig. S12) to improve the weights. Multi-SNP prediction models were generated for a total of 686,241 (gene, tissue) pairs and 1,816,703 (splicing event, tissue) pairs. With the increased sample size and improved models, we increased the number of expression models by 14% (median across tissues) relative to the GTEx v7 models Elastic Net models (fig. S13). Splicing models are available only for the v8 release.

Next, we computed the association between an imputed molecular phenotype (expression or splicing) and a trait to estimate the genic effect on the trait using S-PrediXcan (*25*). Given the widespread tissue-sharing of regulatory variation (*8*), we also computed S-MultiXcan scores (*10*) to integrate patterns of associations from multiple tissues and increase statistical power (*10*). Twenty eight percent of the genes tested with S-PrediXcan showed a significant association with at least one of the 87 traits at Bonferroni-corrected p-value threshold ($p < 0.05/686,241$; fig. S14). For splicing, about 15% (20,364 of 138,890) of tested splicing events showed a significant association ($p < 0.05/1,816,703$). Nearly all traits (94%; 82 out of 87) showed at least one S-PrediXcan significant gene-level association in at least one tissue (fig. S19 and S20). This resource of S-PrediXcan associations can be used to prioritize a list of putatively causal genes for follow-up studies.

To replicate the PrediXcan expression associations in an independent dataset, BioVU, which is a large-scale biobank tied to Electronic Health Records (*26, 27*), we selected seven traits with predicted high statistical power. Out of 947 gene-tissue-trait discoveries tested, 458 unique gene-tissue-trait triplets (48%) showed replication in this independent biobank (p < 0.05; (*21*)).

Altogether, these results provide abundant links between gene regulation and GWAS loci. To further quantify this, we considered approximately LD-independent regions (*28*) with a significant GWAS variant for each trait, and calculated the proportion of GWAS loci that contain an associated gene from S-PrediXcan ($p < 0.05$ / # genes, 2×10^{-6}) or a colocalized gene from *enloc* (rcp > 0.5). Across the traits, 72% (3,899/5,385) of GWAS loci had a S-PrediXcan expression association in the same LD region and 55% (2,125/3,899) had evidence of colocalization with an eQTL (table S3, table S4, fig. S17). For splicing, 62% (3,345/5,385) had a S-PrediXcan association and 34% (1,135/3,345) *enloc* colocalized with an sQTL (fig. S18). From the combined list of eGenes and sGenes, 47% of loci have a gene with both *enloc* and PrediXcan support. The distribution of the proportion of associated and colocalized GWAS loci across 87 traits is summarised in Fig. 3-E; for a typical complex trait, about 20% of GWAS loci contained a colocalized, significantly associated gene while 11% contained a colocalized, significantly associated splicing event. These results propose function for a large number of GWAS loci, but most loci remain without candidate genes, highlighting the need to expand the resolution of transcriptome studies.

Performance of causal gene prediction

Multiple studies have found an excess of deleterious rare variants associated with complex traits in genes that are in the vicinity of common variants associated with the same trait (*29–31*), suggesting that the dose-response curve at the regulatory range may be extrapolated to the rare, loss-of-function end (Fig. 3B). We thus leveraged this rare variant information to analyze the sensitivity and specificity of S-PrediXcan and *enloc* in finding causal genes in GWAS loci. Towards this end, we curated two "silver standard" sets of genes associated with specific traits, based on the OMIM (Online Mendelian Inheritance in Man) database (*32*) and rare variant association studies (*29, 33, 34*) (fig. S21, table S6). We analyzed the genes, within the silver standard sets, that have a GWAS association for a matched trait in the same LD block (*21, 28*), resulting in 1,592 OMIM gene-trait pairs and 101 rare variant based gene-trait pairs (table S11, table S12, fig. S22). Since only genes in the vicinity of an index gene can be discovered with cis-regulatory information, we selected 229 OMIM genes and 81 genes harboring rare variant associations that are located within the same LD block as the GWAS locus for a matched trait (fig. S23).

Both S-PrediXcan and *enloc* showed high sensitivity to identifying the silver standard genes (Fig. 3C, Fig. 3D). Compared to a random set of genes within the same LD block as the GWAS locus and OMIM gene, S-PrediXcan and *enloc* showed substantial enrichment of 2.5 and 4.6 folds for expression and 2.5 and 6.1 folds for splicing, respectively. For the rare variant silver standard, we found similar enrichment for PrediXcan (2.2 and 2.19 for expression and splicing, respectively) and *enloc* (14.7 and 21.7). We note comparison of this enrichment between the methods is not interpretable because the tresholds based on significance and colocalization probability are not comparable.

For applications such as target selection for drug development or follow-up experiments, another relevant metric is the precision or, equivalently, positive predictive value (PPV) – the probability that the gene-trait link is causal given that it is called significant or colocalized. Using the same threshold as for the sensitivity calculation, we found that 8.5% (73 out of 859) of PrediXcan significant genes and 11.7% (49 out of 419) of *enloc*-colocalized genes were also OMIM genes for matched traits.

These enrichment results were corroborated by ROC and precision-recall curves, which demonstrate that *enloc* and PrediXcan contribute to prediction of causal genes, and that combining *enloc* and PrediXcan improves the precision-recall trade-off (fig. S25). However, the overall prediction performance is modest, which is likely to be partially due to the fact that the OMIM gene list has an inherent bias. Our current understanding of gene function is biased towards protein-coding variants with very large effects, which is reflected in the list of OMIM genes. Genes associated to rare severe disease tend to be depleted of regulatory variation (*35, 36*), which will decrease the performance of a QTL-based method in a way that is unlikely to be generally applicable to GWAS genes that are more tolerant to regulatory variation (*36*).

To further investigate whether this predictive power could be improved by considering the proximity of the GWAS peaks to the OMIM genes, we performed a joint logistic regression of OMIM gene status on 1) the proximity of the top GWAS variant to the nearest gene, 2) posterior probability of colocalization, and 3) PrediXcan association significance between QTL and GWAS variants. To make the scale of the three features more comparable, we used their respective ranking within the locus with a threshold for genes with no evidence of colocalization or association. Among the 229 OMIM genes, 28.4% were the closest gene, 22.7% were the most colocalized, and 18.3% were the most significant fig. S24. All three features were significant predictors of OMIM gene status, with better ranked genes more likely to be OMIM genes (proximity $p = 2.0 \times 10^{-2}$, *enloc* $p = 6.1 \times 10^{-3}$, PrediXcan $p = 2.5 \times 10^{-4}$), indicating that each method provides an independent source of causal evidence. Similar results were obtained using splicing colocalization and association scores and the rare variant based silver standard, as shown in table S7. These results provide further empirical evidence that a combination of colocalization and association methods will perform better than individual ones. The significance of proximity is an indicator of the missing regulability, i.e. mechanisms that may be uncovered by a gene assignment that assays other tissue or cell type contexts, larger samples, and other molecular traits.

Predicted OMIM genes included well-known findings such as *PCSK9* for LDLR, with *PCSK9* significant and colocalized for relevant GWAS traits (LDL-C levels, coronary artery disease, and self-reported high cholesterol), and *Interleukins* and *HLA* subunits for asthma, both significant and colocalized for related immunological traits. Significantly associated and colocalized genes that predicted OMIM genes also included *FLG* (eczema), *TPO* (hypothyroidism), and *NOD2* (inflammatory bowel disease) (see table S11 for complete list). Prediction of genes in the rare variant based silver standard was similarly observed (see (*21*); fig. S26).

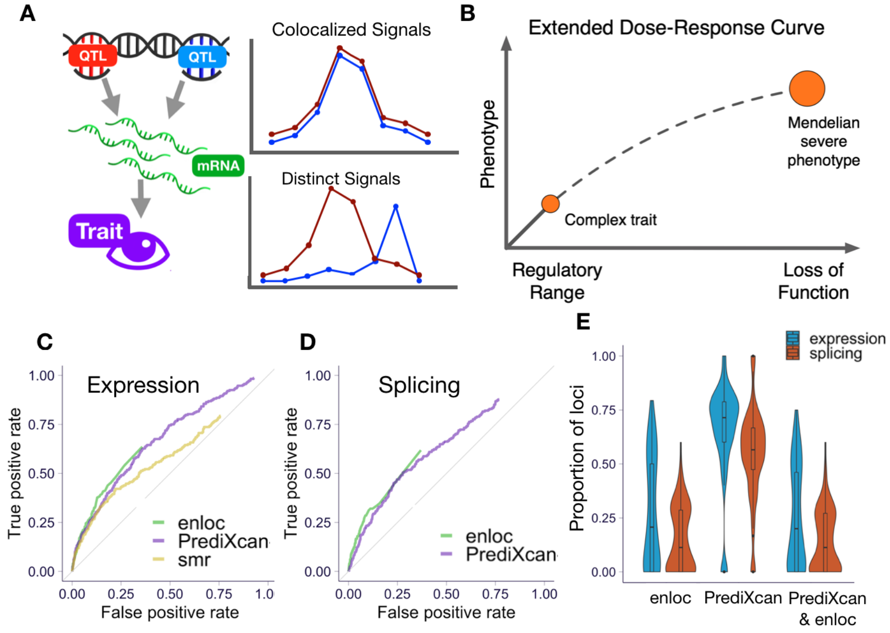


Fig. 3. Identifying and validating predicted causal genes. (A) Schematic representation of association and colocalization approaches. **(B)** Schematic representation of extrapolation of the dose-response curve to Mendelian end of phenotype-genotype spectrum (*37*). **(C-D)** Performance of *enloc*, PrediXcan, and SMR on expression (C) and splicing (D) data to predict causal genes using the OMIM silver standard. **(E)** Proportion of GWAS-associated loci per trait that contain colocalized and S-PrediXcan-associated signals for expression and splicing.

Tissue enrichment of GWAS signals

A systematic survey of regulatory variation across 49 human tissues promises to facilitate the identification of the tissues of action for complex traits. However, because of the broad sharing of regulatory variation across tissues and the reduced significance of tissue-specific eQTLs, causal tissue identification has been challenging. Here we used sparse factors from *mashR* representing patterns of tissue sharing of eQTLs (*21*), to classify each gene-trait association into one of 15 tissue classes (fig. S28). Using the pattern of tissue classes of non-colocalized genes (rcp = 0) as the expected null, we assessed whether significantly associated and colocalized genes (PrediXcan significant and rcp > 0.01) were over-represented in certain tissue classes (Fig. 4). Consistent with previous reports (*11, 38*), we identified several instances in which the most significant tissue is supported by current biological knowledge. For example, blood cell count traits were enriched in whole blood, neuroticism and fluid intelligence in brain/pituitary, hypothyrodism in thyroid, coronary artery disease in artery, and cholesterol-related traits in liver. Taken together, these results show the potential of leveraging regulatory variation to help identify tissues of relevance for complex traits.

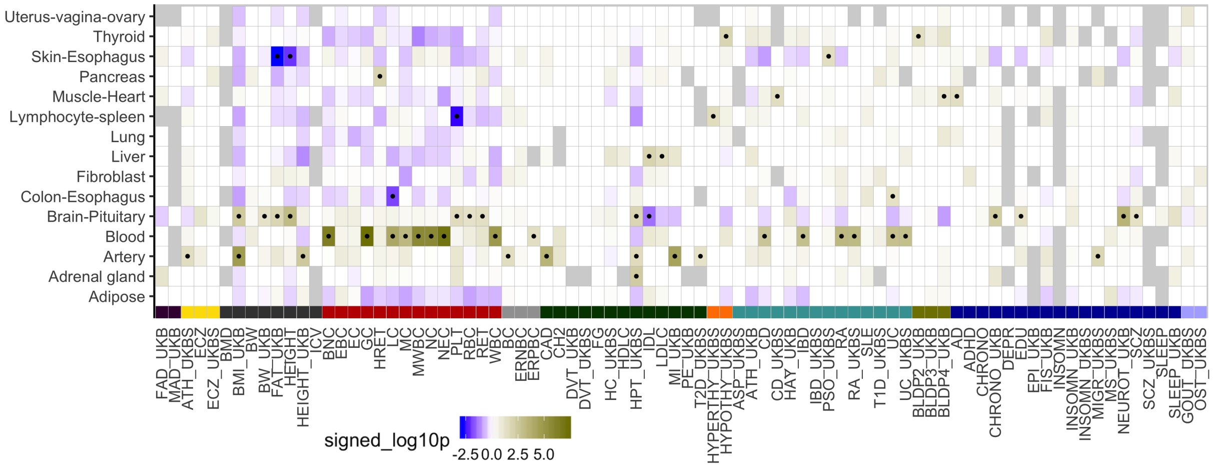


Fig. 4. Identifying trait-relevant tissues using tissue-specific enrichment. Enrichment of tisssue-specific association and colocalization compared to the pattern of tissue-specificity of non-colocalized genes. Over-representation of the tissue class for PrediXcan-significant and colocalized genes is indicated by dark yellow while depletion is indicated by blue. Black dots label the tissue class-trait pairs passing the nominal p-value significance threshold of 0.05. Abbreviation: S1. Trait category colors: S1.

Discussion

We examined in-depth the phenotypic consequences of transcriptome regulation and provide novel computational methodologies and best-practice guidelines for using the GTEx resources to interpret GWAS results. We provide a systematic empirical demonstration of the widespread dose-dependent effect of expression and splicing on complex traits, i.e., variants with larger impact at the molecular level had larger impact at the trait level. Furthermore, we found that target genes in GWAS loci identified by *enloc* and PrediXcan were predictive of OMIM genes for matched traits, implying that for a proportion of the genes, the dose-response curve can be extrapolated to the rare and more severe end of the genotype-trait spectrum. The observation that common regulatory variants target genes also implicated by rare coding variants underscores the extent to which these different types of genetic variants converge to mediate a spectrum of similar pathophysiological effects and may provide a powerful approach to drug target discovery.

We implemented association and colocalization methods that leverage the observed allelic heterogeneity of expression traits. After extensive comparison using two independent sets of silver standard gene-trait pairs, we conclude that combining *enloc*, PrediXcan, and proximity ranking outperforms the individual approaches. The significance of the proximity ranking is a sign of the "missing regulability" emphasizing the need to expand the resolution, sample size, and range of contexts of transcriptome studies as well as to examine other molecular mechanisms.

We caution that the increased power offered by this release of the GTEx resources also brings higher risk of false links due to LD contamination and that naive use of eQTL or sQTL association p-values to assign function to a GWAS locus can be misleading. Colocalization approaches can be effective in weeding out LD contamination but given the current state of the methods and the lack of LD references from source studies, they can also be overtly conservative. Importantly, fine-mapping and colocalization approaches can be highly sensitive to LD misspecification when only summary results are used (*39*). The GWAS community has made great progress in recognizing the need to share summary results, but to take full advantage of these data, improved sharing of LD information from the source study as well as from large sequencing reference datasets, is also required. We highlight the importance of considering more than one statistical evidence to determine the causal mechanisms underlying a complex trait.

Finally, we generated several resources that can open the door for addressing key questions in complex trait genomics. We present a catalog of gene-level associations, including potential target genes for nearly half of the GWAS loci investigated here that provides a rich basis for studies on the functional mechanisms of complex diseases and traits. We provide a database of optimal gene expression imputation models that were built on the fine-mapping probabilities for feature selection and that leverage the global patterns of tissue sharing of regulation to improve the weights. These imputation models of expression and splicing, which to date has been challenging to study, provide a foundation for transcriptome-wide association studies of the human phenome – the collection of all human diseases and traits – to further accelerate discovery of trait-associated genes. Collectively, these data thus represent a valuable resource, enabling novel biological insights and facilitating follow-up studies of causal mechanisms.

Authors

* alphabetic order

Lead Analysts* *Equal contribution* Alvaro N Barbeira, Rodrigo Bonazzola, Eric R Gamazon, Yanyu Liang, YoSon Park

Analysts* François Aguet, Lisa Bastarache, Ron Do, Gao Wang, Andrew R Hamel, Farhad Hormozdiari, Zhuoxun Jiang, Daniel Jordan, Sarah Kim-Hellmuth, Boxiang Liu, Milton D Pividori, Abhiram Rao, Marie Verbanck, Dan Zhou

GTEx GWAS Working Group* François Aguet, Kristin Ardlie, Alvaro N Barbeira, Rodrigo Bonazzola, Christopher D Brown, Lin Chen, Eric R Gamazon, Kevin Gleason, Andrew R Hamel, Farhad Hormozdiari, Hae Kyung Im, Sarah Kim-Hellmuth, Tuuli Lappalainen, Yanyu Liang, Boxiang Liu, Dan L Nicolae, Yoson Park, Milton D Pividori, Abhiram Rao, John M. Rouhana, Ayellet V Segrè, Xiaoquan Wen

Senior Leadership* Kristin Ardlie, Christopher D. Brown, Hae Kyung Im, Tuuli Lappalainen, Mark McCarthy, Stephen Montgomery, Ayellet V Segrè, Matthew Stephens, Xiaoquan Wen

Manuscript Writing Group* Eric R Gamazon, Hae Kyung Im, Tuuli Lappalainen, Yanyu Liang, YoSon Park

Corresponding Author* Hae Kyung Im

GTEx Consortium

Laboratory and Data Analysis Coordinating Center (LDACC): François Aguet[1], Shankara Anand[1], Kristin G Ardlie[1], Stacey Gabriel[1], Gad Getz[1,2], Aaron Graubert[1], Kane Hadley[1], Robert E Handsaker[3,4,5], Katherine H Huang[1], Seva Kashin[3,4,5], Xiao Li[1], Daniel G MacArthur[4,6], Samuel R Meier[1], Jared L Nedzel[1], Duyen Y Nguyen[1], Ayellet V Segrè[1,7], Ellen Todres[1]

Analysis Working Group (funded by GTEx project grants): François Aguet[1], Shankara Anand[1], Kristin G Ardlie[1], Brunilda Balliu[8], Alvaro N Barbeira[9], Alexis Battle[10,11], Rodrigo Bonazzola[9], Andrew Brown[12,13], Christopher D Brown[14], Stephane E Castel[15,16], Don Conrad[17,18], Daniel J Cotter[19], Nancy Cox[20], Sayantan Das[21], Olivia M de Goede[19], Emmanouil T Dermitzakis[12,22,23], Barbara E Engelhardt[24,25], Eleazar Eskin[26], Tiffany Y Eulalio[27], Nicole M Ferraro[27], Elise Flynn[15,16], Laure Fresard[28], Eric R Gamazon[29,30,31,20], Diego Garrido-Martín[32], Nicole R Gay[19], Gad Getz[1,2], Aaron Graubert[1], Roderic Guigó[32,33], Kane Hadley[1], Andrew R Hamel[7,1], Robert E Handsaker[3,4,5], Yuan He[10], Paul J Hoffman[15], Farhad Hormozdiari[34,1], Lei Hou[35,1], Katherine H Huang[1], Hae Kyung Im[9], Brian Jo[24,25], Silva Kasela[15,16], Seva Kashin[3,4,5], Manolis Kellis[35,1], Sarah Kim-Hellmuth[15,16,36], Alan Kwong[21], Tuuli Lappalainen[15,16], Xiao Li[1], Xin Li[28], Yanyu Liang[9], Daniel G MacArthur[4,6], Serghei Mangul[26,37], Samuel R Meier[1], Pejman Mohammadi[15,16,38,39], Stephen B Montgomery[28,19], Manuel Muñoz-Aguirre[32,40], Daniel C Nachun[28], Jared L Nedzel[1], Duyen Y Nguyen[1], Andrew B Nobel[41], Meritxell Oliva[9,42], YoSon Park[14,43], Yongjin Park[35,1], Princy Parsana[11], Ferran Reverter[44], John M Rouhana[7,1], Chiara Sabatti[45], Ashis Saha[11], Ayellet V Segrè[1,7], Andrew D Skol[9,46], Matthew Stephens[47], Barbara E Stranger[9,48], Benjamin J Strober[10], Nicole A Teran[28], Ellen Todres[1], Ana Viñuela[49,12,22,23], Gao Wang[47], Xiaoquan Wen[21], Fred Wright[50], Valentin Wucher[32], Yuxin Zou[51]

Analysis Working Group (not funded by GTEx project grants): Pedro G Ferreira[52,53,54], Gen Li[55], Marta Melé[56], Esti Yeger-Lotem[57,58]

Leidos Biomedical - Project Management: Mary E Barcus[59], Debra Bradbury[60], Tanya Krubit[60], Jeffrey A McLean[60], Liqun Qi[60], Karna Robinson[60], Nancy V Roche[60], Anna M Smith[60], Leslie Sobin[60], David E Tabor[60], Anita Undale[60]

Biospecimen collection source sites: Jason Bridge[61], Lori E Brigham[62], Barbara A Foster[63], Bryan M Gillard[63], Richard Hasz[64], Marcus Hunter[65], Christopher Johns[66], Mark Johnson[67], Ellen Karasik[63], Gene Kopen[68], William F Leinweber[68], Alisa McDonald[68], Michael T Moser[63], Kevin Myer[65], Kimberley D Ramsey[63], Brian Roe[65], Saboor Shad[68], Jeffrey A Thomas[68,67], Gary Walters[67], Michael Washington[67], Joseph Wheeler[66]

Biospecimen core resource: Scott D Jewell[69], Daniel C Rohrer[69], Dana R Valley[69]

Brain bank repository: David A Davis[70], Deborah C Mash[70]

Pathology: Mary E Barcus[59], Philip A Branton[71], Leslie Sobin[60]

ELSI study: Laura K Barker[72], Heather M Gardiner[72], Maghboeba Mosavel[73], Laura A Siminoff[72]

Genome Browser Data Integration & Visualization: Paul Flicek[74], Maximilian Haeussler[75], Thomas Juettemann[74], W James Kentv[75], Christopher M Lee[75], Conner C Powell[75], Kate R Rosenbloom[75], Magali Ruffier[74], Dan Sheppard[74], Kieron Taylor[74], Stephen J Trevanion[74], Daniel R Zerbino[74]

eGTEx groups: Nathan S Abell[19], Joshua Akey[76], Lin Chen[42], Kathryn Demanelis[42], Jennifer A Doherty[77], Andrew P Feinberg[78], Kasper D Hansen[79], Peter F Hickey[80], Lei Hou[35,1], Farzana Jasmine[42], Lihua Jiang[19], Rajinder Kaul[81,82], Manolis Kellis[35,1], Muhammad G Kibriya[42], Jin Billy Li[19], Qin Li[19], Shin Lin[83], Sandra E Linder[19], Stephen B Montgomery[28,19], Meritxell Oliva[9,42], Yongjin Park[35,1], Brandon L Pierce[42], Lindsay F Rizzardi[84], Andrew D Skol[9,46], Kevin S Smith[28], Michael Snyder[19], John Stamatoyannopoulos[81,85], Barbara E Stranger[9,48], Hua Tang[19], Meng Wang[19]

NIH program management: Philip A Branton[71], Latarsha J Carithers[71,86], Ping Guan[71], Susan E Koester[87], A. Roger Little[88], Helen M Moore[71], Concepcion R Nierras[89], Abhi K Rao[71], Jimmie B Vaught[71], Simona Volpi[90]

Affiliations 1. The Broad Institute of MIT and Harvard, Cambridge, MA, USA

2. Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA, USA

3. Department of Genetics, Harvard Medical School, Boston, MA, USA

4. Program in Medical and Population Genetics, The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA

5. Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA

6. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

7. Ocular Genomics Institute, Massachusetts Eye and Ear, Harvard Medical School,

Boston, MA, USA
8. Department of Biomathematics, University of California, Los Angeles, Los Angeles, CA, USA
9. Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA v 10. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
11. Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
12. Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland
13. Population Health and Genomics, University of Dundee, Dundee, Scotland, UK
14. Department of Genetics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
15. New York Genome Center, New York, NY, USA
16. Department of Systems Biology, Columbia University, New York, NY, USA
17. Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA
18. Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, Missouri, USA
19. Department of Genetics, Stanford University, Stanford, CA, USA
20. Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
21. Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
22. Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland
23. Swiss Institute of Bioinformatics, Geneva, Switzerland
24. Department of Computer Science, Princeton University, Princeton, NJ, USA
25. Center for Statistics and Machine Learning, Princeton University, Princeton, NJ, USA
26. Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA
27. Program in Biomedical Informatics, Stanford University School of Medicine, Stanford, CA, USA
28. Department of Pathology, Stanford University, Stanford, CA, USA

29. Data Science Institute, Vanderbilt University, Nashville, TN, USA
30. Clare Hall, University of Cambridge, Cambridge, UK
31. MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
32. Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Catalonia, Spain
33. Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
34. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
35. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
36. Statistical Genetics, Max Planck Institute of Psychiatry, Munich, Germany
37. Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA, USA
38. Scripps Research Translational Institute, La Jolla, CA, USA
39. Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA
40. Department of Statistics and Operations Research, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain
41. Department of Statistics and Operations Research and Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA
42. Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA
43. Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
44. Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona. Spain.
45. Departments of Biomedical Data Science and Statistics, Stanford University, Stanford, CA, USA
46. Department of Pathology and Laboratory Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, USA
47. Department of Human Genetics, University of Chicago, Chicago, IL, USA
48. Center for Genetic Medicine, Department of Pharmacology, Northwestern University, Feinberg School of Medicine, Chicago, IL, USA
49. Department of Twin Research and Genetic Epidemiology, King's College London,

London, UK

50. Bioinformatics Research Center and Departments of Statistics and Biological Sciences, North Carolina State University, Raleigh, NC, USA

51. Department of Statistics, University of Chicago, Chicago, IL, USA

52. Department of Computer Sciences, Faculty of Sciences, University of Porto, Porto, Portugal

53. Instituto de Investigação e Inovação em Sauúde, Universidade do Porto, Porto, Portugal

54. Institute of Molecular Pathology and Immunology, University of Porto, Porto, Portugal

55. Columbia University Mailman School of Public Health, New York, NY, USA

56. Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

57. Department of Clinical Biochemistry and Pharmacology, Ben-Gurion University of the Negev, Beer-Sheva, Israel

58 National Institute for Biotechnology in the Negev, Beer-Sheva, Israel

59. Leidos Biomedical, Frederick, MD, USA

60. Leidos Biomedical, Rockville, MD, USA

61. UNYTS, Buffalo, NY, USA

62. Washington Regional Transplant Community, Annandale, VA, USA

63. Therapeutics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA

64. Gift of Life Donor Program, Philadelphia, PA, USA

65. LifeGift, Houston, TX, USA

66. Center for Organ Recovery and Education, Pittsburgh, PA, USA

67. LifeNet Health, Virginia Beach, VA. USA v 68. National Disease Research Interchange, Philadelphia, PA, USA v 69. Van Andel Research Institute, Grand Rapids, MI, USA

70. Department of Neurology, University of Miami Miller School of Medicine, Miami, FL, USA

71. Biorepositories and Biospecimen Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD, USA

72. Temple University, Philadelphia, PA, USA

73. Virgina Commonwealth University, Richmond, VA, USA

74. European Molecular Biology Laboratory, European Bioinformatics Institute, Hinx-

ton, United Kingdom v 75. Genomics Institute, UC Santa Cruz, Santa Cruz, CA, USA
76. Carl Icahn Laboratory, Princeton University, Princeton, NJ, USA
77. Department of Population Health Sciences, The University of Utah, Salt Lake City, Utah, USA
78. Schools of Medicine, Engineering, and Public Health, Johns Hopkins University, Baltimore, MD, USA
79. Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
80. Department of Medical Biology, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia
81. Altius Institute for Biomedical Sciences, Seattle, WA, USA v 82. Division of Genetics, University of Washington, Seattle, WA, University of Washington, Seattle, WA, USA
83. Department of Cardiology, University of Washington, Seattle, WA, USA
84. HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA
85. Genome Sciences, University of Washington, Seattle, WA, USA
86. National Institute of Dental and Craniofacial Research, Bethesda, MD, USA
87. Division of Neuroscience and Basic Behavioral Science, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA
88. National Institute on Drug Abuse, Bethesda, MD, USA
89. Office of Strategic Coordination, Division of Program Coordination, Planning and Strategic Initiatives, Office of the Director, National Institutes of Health, Rockville, MD, USA
90. Division of Genomic Medicine, National Human Genome Research Institute, Bethesda, MD, USA

Acknowledgements

We thank the donors and their families for their generous gifts of organ donation for transplantation, and tissue donations for the GTEx research project; Mariya Khan and Christopher Stolte for the illustration in Figure 1; and Laura Vairus for the illustrations in Figures 2 and 3.

The consortium was funded by GTEx program grants: HHSN268201000029C (F.A., K.G.A., A.V.S., X.Li., E.T., S.G., A.G., S.A., K.H.H., D.Y.N., K.H., S.R.M., J.L.N.), 5U41HG009494 (F.A., K.G.A.), 10XS170 (Subcontract to Leidos Biomedical) (W.F.L., J.A.T., G.K., A.M., S.S., R.H., G.Wa., M.J., M.Wa., L.E.B., C.J., J.W., B.R., M.Hu., K.M., L.A.S., H.M.G., M.Mo., L.K.B.), 10XS171 (Subcontract to Leidos Biomedical) (B.A.F., M.T.M., E.K., B.M.G., K.D.R., J.B.), 10ST1035 (Subcontract to Leidos Biomedical) (S.D.J., D.C.R., D.R.V.), R01DA006227-17 (D.C.M., D.A.D.), Supplement to University of Miami grant DA006227. (D.C.M., D.A.D.), HHSN261200800001E (A.M.S., D.E.T., N.V.R., J.A.M., L.S., M.E.B., L.Q., T.K., D.B., K.R., A.U.), R01MH101814 (M.M-A., V.W., S.B.M., R.G., E.T.D., D.G-M., A.V.), U01HG007593 (S.B.M.), R01MH101822 (C.D.B.), U01HG007598 (M.O., B.E.S.), R01MH107666 (H.K.I.), P30DK020595 (H.K.I.). E.R.G. is supported by the National Human Genome Research Institute (NHGRI) under Award Number 1R35HG010718 and by the National Heart, Lung, and Blood Institute (NHLBI) under Award Number 1R01HL133559. E.R.G. has also significantly benefitted from a Fellowship at Clare Hall, University of Cambridge (UK) and is grateful to the President and Fellows of the college for a stimulating intellectual home. S.K.-H. is supported by the Marie-Sklodowska Curie fellowship H2020 Grant 706636. R.D.: R35GM124836, R01HL139865, 15CVGPSD27130014. D.M.J.: T32HL00782. Y.Pa. is supported by the NHGRI award R01HG10067. A.R.H. was supported by the Massachusetts Lions Eye Research Fund Grant. Computation was performed at the high performance cluster of the Center for Research Informatics at the University of Chicago, funded by the Biological Sciences Division and CTSA UL1TR000430.

Disclosure

F.A. is an inventor on a patent application related to TensorQTL; S.E.C. is a co-founder, chief technology officer and stock owner at Variant Bio; E.T.D. is chairman and member of the board of Hybridstat LTD.; B.E.E. is on the scientific advisory boards of Celsius Therapeutics and Freenome; G.G. receives research funds from IBM and Pharmacyclics, and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, POLYSOLVER and TensorQTL; S.B.M. is on the scientific advisory board of Prime Genomics Inc.; D.G.M. is a co-founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme; H.K.I. has received speaker honoraria from GSK and AbbVie.; T.L. is a scientific advisory board member of Variant Bio with equity and Goldfinch Bio. P.F. is member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomes, Ltd. P.G.F. is a partner of Bioinf2Bio. E.R.G. receives an honorarium from *Circulation Research*, the official journal of the American Heart Association, as a member of the Editorial Board, and has performed consulting for the City of Hope / Beckman Research Institute. R.D. has received research support from AstraZeneca and Goldfinch Bio, not related to this work.

Code and data availability

The code for methods applied in this paper can be downloaded from the github repository (https://github.com/hakyimlab/gtex-gwas-analysis).

Data availability statement

Genotype-Tissue Expression (GTEx) project raw whole transcriptome and genome sequencing data are available via dbGaP accession number phs000424.v8.p1. All processed GTEx data are available via GTEx portal. Imputed summary results, *enloc*, *coloc*, PrediXcan, MultiXcan, *dap-g*, and prediction models are available in https://github.com/hakyimlab/gtex-gwas-analysis and links therein.

URLs

1000 Genomes Project Reference for LDSC,
https://data.broadinstitute.org/alkesgroup/LDSCORE/1000G_Phase3_plinkfiles.tgz;
1000 Genomes Project Reference with regression weights for LDSC,
https://data.broadinstitute.org/alkesgroup/LDSCORE/1000G_Phase3_weights_hm3_no_MHC.tgz;
BioVU, https://victr.vanderbilt.edu/pub/biovu/?sid=194;
eCAVIAR, https://github.com/fhormoz/caviar;
QTLEnrich, https://github.com/segrelabgenomics/eQTLEnrich;
flashr, https://gaow.github.io/mnm-gtex-v8/analysis/mashr_flashr_workflow.html#flashr-prior-covariances;
Gencode, https://www.gencodegenes.org/releases/26.html;

GTEx GWAS subgroup repository, https://github.com/broadinstitute/gtex-v8;
GTEx portal, http://gtexportal.org;
Hail, https://github.com/hail-is/hail;
HapMap Reference for LDSC, https://data.broadinstitute.org/alkesgroup/LDSCORE/w_hm3.snplist.bz2;
LD score regression (LDSD regression), https://github.com/bulik/ldsc;
MetaXcan, https://github.com/hakyimlab/MetaXcan;
Mouse Phenotype Ontology, http://www.informatics.jax.org/vocab/mp_ontology;
NHGRI-EBI GWAS catalog, https://www.ebi.ac.uk/gwas/;
picard, http://picard.sourceforge.net/;
PLINK 1.90, https://www.cog-genomics.org/plink2;
PrediXcan, https://github.com/hakyim/PrediXcan;
pyliftover, https://pypi.org/project/pyliftover/;
Storeyś qvalue R package, https://github.com/StoreyLab/qvalue;
Summary GWAS imputation, https://github.com/hakyimlab/summary-gwas-imputation;
TORUS, https://github.com/xqwen/torus;
UK Biobank GWAS, http://www.nealelab.is/uk-biobank/;
UK Biobank, http://www.ukbiobank.ac.uk/;

References

1. D. L. Nicolae, *et al.*, *PLoS genetics* **6**, e1000888 (2010).
2. H. Guo, *et al.*, *Human Molecular Genetics* **24**, 3305 (2015).
3. L. Wu, *et al.*, *Nature genetics* **50**, 968 (2018).
4. J. Gong, *et al.*, *Nucleic acids research* **46**, D971 (2018).
5. E. E. Pashos, *et al.*, *Cell stem cell* **20**, 558 (2017).
6. M. Caliskan, *et al.*, *American journal of human genetics* **105**, 89 (2019).
7. L. J. Carithers, *et al.*, *Biopreservation and Biobanking* **13**, 311 (2015).
8. GTEx Consortium, *et al.*, *Nature* **550**, 204 (2017).
9. GTEx Consortium, *Journal* **550** (2019).
10. A. N. Barbeira, *et al.*, *PLOS Genetics* **15**, 1 (2019).
11. E. R. Gamazon, *et al.*, *Nature genetics* **50**, 956 (2018).
12. Z. Zhu, *et al.*, *Nature genetics* **48**, 481 (2016).
13. A. Gusev, *et al.*, *Nature Genetics* **48**, 245 (2016).
14. X. Wen, *Ann. Appl. Stat.* **10**, 1619 (2016).
15. F. Hormozdiari, *et al.*, *American journal of human genetics* **99**, 1245 (2016).
16. X. Wen, R. Pique-Regi, F. Luca, *PLoS Genetics* **13**, e1006646 (2017).
17. C. Giambartolomei, *et al.*, *PLOS Genetics* **10**, 1 (2014).
18. B. K. Bulik-Sullivan, *et al.*, *Nature Genetics* **47**, 291 (2015).
19. E. R. Gamazon, *et al.*, *Nature Genetics* **47**, 1091 (2015).
20. M. Wainberg, *et al.*, *Nature genetics* **51**, 592 (2019).
21. See supplementary materials .

22. (2019).

23. A. R. Wood, *et al.*, *Nature Genetics* (2014).

24. GTEx Consortium, *Journal* **550** (2020).

25. A. N. Barbeira, *et al.*, *Nature Communications* (2018).

26. D. M. Roden, *et al.*, *Clinical Pharmacology and Therapeutics* (2008).

27. J. C. Denny, *et al.*, *Nature Biotechnology* **31**, 1102 (2013).

28. T. Berisa, J. K. Pickrell, *Bioinformatics* **32**, 283 (2016).

29. E. Marouli, *et al.*, *Nature* **542**, 186 (2017).

30. C. Fuchsberger, *et al.*, *Nature* **536**, 41 (2016).

31. A. Keinan, A. G. Clark, *Science* **336**, 740 (2012).

32. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, *Nucleic Acids Research* (2005).

33. D. J. Liu, *et al.*, *Nature genetics* **49**, 1758 (2017).

34. A. E. Locke, *et al.*, *Nature* p. 1 (2019).

35. K. J. Karczewski, *et al.*, *bioRxiv* (2019).

36. P. Mohammadi, *et al.*, *Science* **366**, eaay0256 (2019).

37. R. M. Plenge, E. M. Scolnick, D. Altshuler, *Nature Publishing Group* **12**, 581 (2013).

38. H. Ongen, *et al.*, *Nature genetics* **49**, 1676 (2017).

39. C. Benner, *et al.*, *The American Journal of Human Genetics* **101**, 539 (2017).

40. A. Dobin, *et al.*, *Bioinformatics* **29**, 15 (2013).

41. B. Li, C. N. Dewey, *BMC Bioinformatics* **12**, 323 (2011).

42. D. S. DeLuca, *et al.*, *Bioinformatics* **28**, 1530 (2012).

43. M. D. Robinson, A. Oshlack, *Genome Biology* **11**, R25 (2010).

44. H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, O. Delaneau, *Bioinformatics* (2016).

45. Y. I. Li, *et al.*, *Nature Genetics* **50**, 151 (2018).

46. D. Lee, T. B. Bigdeli, B. P. Riley, A. H. Fanous, S. A. Bacanu, *Bioinformatics* **29**, 2925 (2013).

47. B. Pasaniuc, *et al.*, *Bioinformatics* **30**, 2906 (2014).

48. C. Bycroft, *et al.*, *Nature* **562**, 203 (2018).

49. A. Buniello, *et al.*, *Nucleic Acids Research* (2019).

50. J. Bowden, G. D. Smith, S. Burgess, *International Journal of Epidemiology* (2015).

51. W. Wang, M. Stephens, *arXiv preprint arXiv:1802.06931* (2018).

52. S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions, *Tech. rep.*, Nature Publishing Group (2018).

53. G. Wang, A. K. Sarkar, P. Carbonetto, M. Stephens, *bioRxiv* p. 501114 (2018).

54. J. Friedman, T. Hastie, R. Tibshirani, *Journal of Statistical Software* **33**, 1 (2010).

55. A. L. Price, *et al.*, *Nature Genetics* **38**, 904 (2006).

56. S. Köhler, *et al.*, *Nucleic acids research* **42**, D966 (2013).

57. L. Bastarache, *et al.*, *Science* **359**, 1233 (2018).

58. K. Musunuru, *et al.*, *Nature* **466**, 714 (2010).

59. M. Nikpay, *et al.*, *Nature genetics* **47**, 1121 (2015).

60. V. Anttila, *et al.*, *Nature genetics* **45**, 912 (2013).

61. S. Kidambi, S. B. Patel, *Xenobiotica; the fate of foreign compounds in biological systems* **38**, 1119 (2008).

62. L. Yu, *et al.*, *Proceedings of the National Academy of Sciences* **99**, 16237 (2002).

63. K. R. Wilund, L. Yu, F. Xu, H. H. Hobbs, J. C. Cohen, *Journal of lipid research* **45**, 1429 (2004).

64. G. M. P. Peloso, *et al.*, *Circulation: Genomic and Precision Medicine* (2019).

65. G. Walldius, I. Jungner, *Journal of internal medicine* **255**, 188 (2004).

66. J. H. Contois, *et al.*, *Clinical chemistry* **55**, 407 (2009).

67. M. Leslie, *Science (New York, N.Y.)* **358**, 1237 (2017).

68. J. Malone, *et al.*, *Bioinformatics* **26**, 1112 (2010).

1 Supplementary Materials

Alvaro N Barbeira[1,†], Rodrigo Bonazzola[1,†], Eric R Gamazon[2,3,4,5,†], Yanyu Liang[1,†], YoSon Park[6,7,†], Sarah Kim-Hellmuth[8,9,10], Gao Wang[11], Zhuoxun Jiang[1], Dan Zhou[2], Farhad Hormozdiari[12, 13], Boxiang Liu[14], Abhiram Rao[14], Andrew R Hamel[12,15], Milton D Pividori[1], François Aguet[12], GTEx GWAS Working Group, Lisa Bastarache[16,17], Daniel M Jordan[18, 19, 20], Marie Verbanck[18, 19, 20, 21], Ron Do[18,19,20], GTEx Consortium, Matthew Stephens[11], Kristin Ardlie[12], Mark McCarthy[22], Stephen B Montgomery[23,24], Ayellet V Segrè[12, 15], Christopher D. Brown[6], Tuuli Lappalainen[9,10], Xiaoquan Wen[25], Hae Kyung Im[1,*]

1 Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA
2 Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
3 Data Science Institute, Vanderbilt University, Nashville, TN, USA
4 Clare Hall, University of Cambridge, Cambridge, UK
5 MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
6 Department of Genetics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
7 Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
8 Statistical Genetics, Max Planck Institute of Psychiatry, Munich, Germany
9 New York Genome Center, New York, NY, USA
10 Department of Systems Biology, Columbia University, New York, NY, USA
11 Department of Human Genetics, University of Chicago, Chicago, IL, USA
12 The Broad Institute of MIT and Harvard, Cambridge, MA, USA
13 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
14 Department of Biology, Stanford University, Stanford, California 94305, USA
15 Ocular Genomics Institute, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA
16 Department of Biomedical Informatics, Department of Medicine, Vanderbilt University, Nashville, TN, USA
17 Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, TN, USA
18 Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA
19 Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA
20 The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount sinai, New York, New York, USA
21 Université de Paris - EA 7537 BIOSTM, France
22 University of Oxford, United Kingdom
23 Department of Genetics, Stanford University, Stanford, CA, USA
24 Department of Pathology, Stanford University, Stanford, CA, USA
25 Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

†: These authors contributed equally to this work, alphabetic order;
∗: Correspondence to haky@uchicago.edu

Contents

1.1 Genotype-Tissue Expression (GTEx) Project

All processed Genotype-Tissue Expression (GTEx) Project v8 data have been made available on dbGAP (accession ID: phs000424.v8). Primary and extended results generated by consortium members are available on the Google Cloud Platform storage accessible via the GTEx Portal (see URLs). The GTEx Project v8 data, based on 17,382 RNA-sequencing samples from 54 tissues of 948 post-mortem subjects, has established the most comprehensive map of regulatory variation to date. In addition to the larger sample size and greater tissue coverage compared to v6, v8 data also included whole-genome sequencing data, facilitating high resolution QTL map of 838 subjects for 49 tissues with at least 70 samples. We mapped complex trait associations for 23,268 cis-eGenes and 14,424 cis-sGenes (*9*). We did not include trans QTLs in our analyses due to limited power after correcting for confounders and potential pleiotropic effect in complex trait associations. Below, we briefly describe the whole-genome sequencing, RNA-sequencing and QTL data processing protocols. Detailed description of subject ascertainment, sample procurement, and sequencing data processing are available elsewhere (*9*).

Whole-genome sequence data processing and quality control

Out of 899 WGS samples sequenced at an average coverage of 30x on HiSeq200 (68 samples) and HiSeqX (all other samples), variant call files (VCF) for 866 GTEx donors were included in downstream analyses after excluding one each from 30 duplicate samples and three donors. Of these, 838 subjects with RNA-seq data were included for QTL mapping and subsequent complex trait association analyses in our study. All whole-genome sequencing data were mapped to GRCh38/hg38 reference.

RNA-Seq data processing and quality control

Whole transcriptome RNA-Seq data were aligned using STAR (v2.5.3.a; (*40*)). For STAR index, GENCODE v26 (GRCh38; see URLs) was used with the sjdbOverhang 75 for 76-bp paired-end sequencing protocol. Default parameters were used for RSEM (see URLs; (*41*)) index generation. GTEx utilized Picard (see URLs) to mark and remove potential PCR duplicates and RNA-SeQC (*42*) to process post-alignment quality control. RSEM was then used for per-sample transcript quantification. Subsequently, read counts were normalized between samples using TMM (*43*). For eQTL analyses, latent factor covariates were calculated using PEER as follows: 15 factors for N<150 per tissue; 30 factors for 150<=N<250; 45 factors for 250<=N<350; and 60 factors for N>=350. Finally, fastQTL (*44*) was used for cis-eQTL mapping in each tissue. Only protein-coding, lincRNA, and antisense biotypes as defined by Gencode v26 were considered for further analyses. To study alternative splicing, GTEx applied LeafCutter (version 0.2.8; (*45*)) using default parameters to quantify splicing QTLs in cis with intron excision ratios (*9*).

1.2 Genome-wide association studies (GWAS)

Harmonization

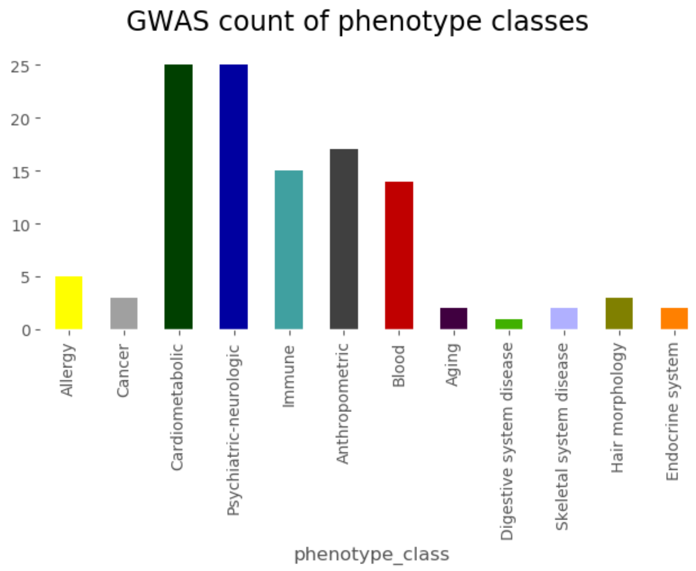
The process followed for the harmonization and imputation are depicted in fig. S2. For each standardized GWAS summary statistics, we mapped all variants to hg38 (GRCh38) references using *pyliftover* (see URLs). For missing chromosome or genomic position information in the original GWAS summary statistics file, we queried dbSNP build 125 (hg17), dbSNP build 130 (hg18/GRCh36), and dbSNP build 150 (hg19/GRCh37) using the provided variant rsID information and the original reference build of the GWAS summary statistics file. Variants with missing chromosome, genomic position, and rsID information were excluded from further analyses. Only autosomal variants were included in our analyses. Missing allele frequency information was filled using the allele frequencies estimated in the GTEx (v8) individuals of genotype-based European genetic ancestry (here onwards, GTEx-EUR) whenever possible. We excluded variants with discordant reference and alternate allele information between GTEx and the GWAS study. We included only the alleles with the highest MAF among multiple alternate alleles if the variant was reported as multiallelic in GTEx. When more than one GWAS variant mapped to a given GTEx variant (i.e., the same chromosomal location in hg38), only the one with the highest significance was retained. For binary traits, if the sample size was present but the number of cases was missing, we filled the missing count with the sample size and number of cases reported in the paper. For continuous traits, if the file contained the sample size for each variant, the reported number was used. If not, we filled this value using the number reported in the corresponding publication. If only some variants were missing, we filled the missing value with the median of all reported values.

Imputation of GWAS summary statistics

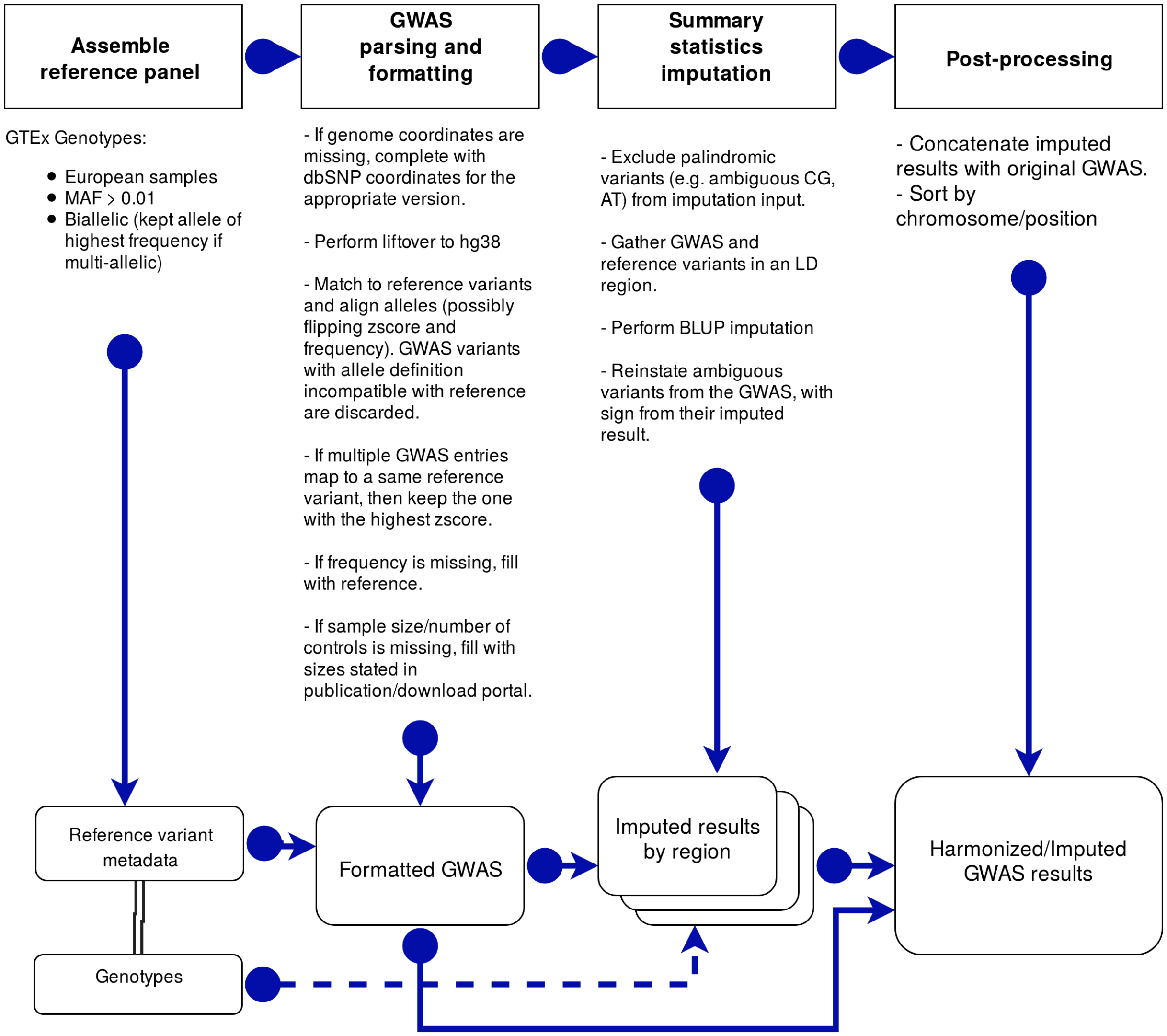
To standardize the number of variants across trait-tissue pairs, all processed GWAS results were imputed. We implemented the Best Linear Unbiased Prediction (BLUP) approach (*46,47*) in-house (https://github.com/hakyimlab/summary-gwas-imputation) to impute z-scores for those variants reported in GTEx without matching data in the GWAS summary statistics. This algorithm does not impute raw effect sizes (β coefficients). The imputation was performed in specific regions assumed to have sufficiently low correlations between them, defined by approximately independent linkage disequilibrium (LD) blocks (*28*) lifted over to hg38/GRCh38.

Only GTEx variants with MAF > 0.01 in GTEx-EUR subjects were used in downstream analyses. Covariance matrices (reference LD information) were estimated on these GTEx-EUR subjects. The corresponding (pseudo-)inverse matrices for covariances C were calculated via Singular Value Decomposition (SVD) using ridge-like regularization $C + 0.1I$. To avoid ambiguous strand issues homogeneously, palindromic variants (i.e. CG) were excluded from the imputation input. Thus, an imputed z-score was generated for palindromic variants available in the original GWAS; for them, we report the absolute value of the original entry with the sign from the imputed z-score. The sample size that we report for the imputed variants is the same as the sample size for the observed ones if it is reported as constant across variants, or their median if it changes across the observed variants, which occurs in the case of meta-analyses.

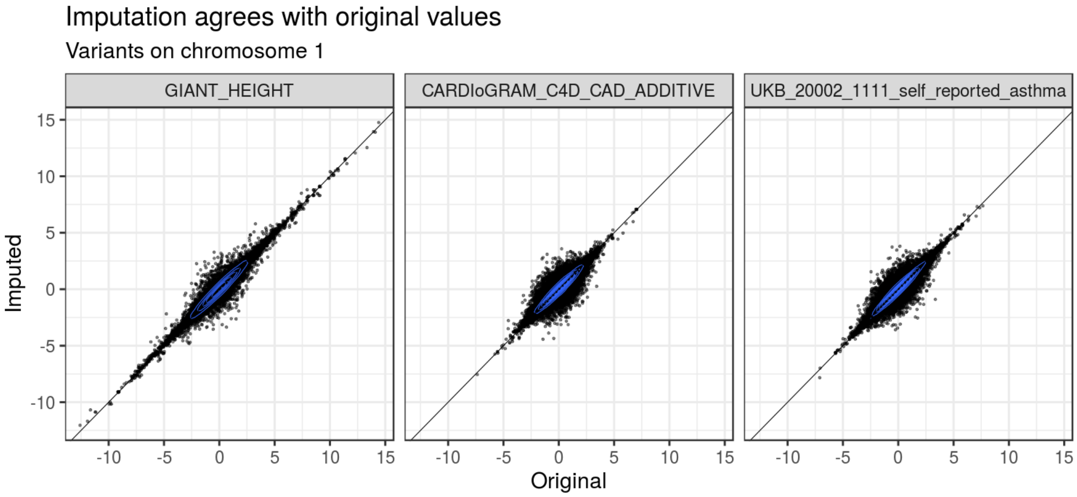
We initially considered publicly available GWAS summary statistics for 114 complex traits provided by large-scale consortia and the UK Biobank (*48*) (table S9). Of these, 27 studies with a relatively small intersection of variants with the GTEx panel (number of variants$< 2 \times 10^6$, compared to almost 9×10^6 variants available in GTEx) exhibited significant deflation of their association p-values(fig. S4). Thus, all analyses focused on 87 traits where missing variants could be properly imputed unless otherwise stated explicitly (table S1). We observed noteworthy association prediction performance across the selected 87 traits (e.g., with a median $r^2 = 0.90$ (IQR = 0.0268) between the original and imputed zscores on chromosome 1). The median slope was 0.94 (IQR = 0.0164), as the imputed zscore values tend to be more conservative than the original ones. Imputation quality was consistent across traits, depending strongly on the number of input available variants (fig. S3).



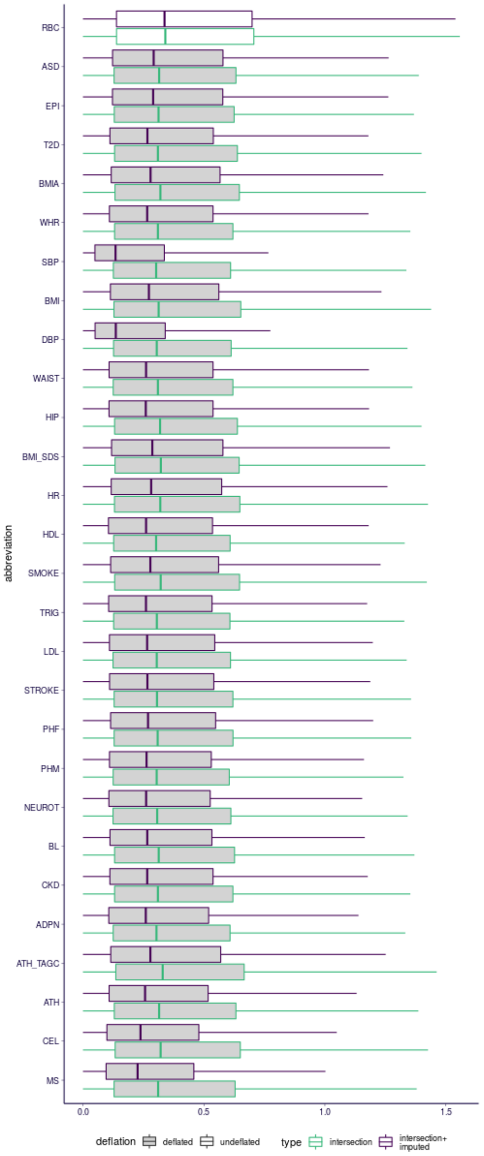
Supplementary Fig. S1. GWAS trait categories. Categories of the traits with full GWAS summary statistics used in the analysis. See list of traits in S1.


Assemble reference panel
GWAS parsing and formatting
Summary statistics imputation
Post-processing
GTEx Genotypes:
● European samples
● MAF > 0.01
● Biallelic (kept allele of highest frequency if multi-allelic)
- If genome coordinates are missing, complete with dbSNP coordinates for the appropriate version.
- Perform liftover to hg38
- Match to reference variants and align alleles (possibly flipping zscore and frequency). GWAS variants with allele definition incompatible with reference are discarded.
- If multiple GWAS entries map to a same reference variant, then keep the one with the highest zscore.
- If frequency is missing, fill with reference.
- If sample size/number of controls is missing, fill with sizes stated in publication/download portal.
- Exclude palindromic variants (e.g. ambiguous CG, AT) from imputation input.
- Gather GWAS and reference variants in an LD region.
- Perform BLUP imputation
- Reinstate ambiguous variants from the GWAS, with sign from their imputed result.
- Concatenate imputed results with original GWAS.
- Sort by chromosome/position
Reference variant metadata
Genotypes
Formatted GWAS
Imputed results by region
Harmonized/Imputed GWAS results

Supplementary Fig. S2. Workflow of GWAS results processing.



Supplementary Fig. S3. GWAS imputation quality Original versus imputed zscores for palindromic variants in chromosome 1 for 3 traits.



Supplementary Fig. S4. GWAS imputation deflation This figure compares the distribution of p-values for 28 GWAS traits before and after imputation. Vertical scale shows -log10(p-value) of variant association. The 27 traits that exhibited deflation are filled in gray. An undeflated trait (e.g., Red Blood Cell count) is included for comparison. See trait abbreviation list in Table S9.

Table S1: List of 87 GWAS datasets

Category	Trait	Abbreviation	Sample_Size
Psychiatric-neurologic	Alzheimers Disease	AD	54162
Psychiatric-neurologic	Attention Deficit Hyperactivity Disorder	ADHD	53293
Psychiatric-neurologic	Chronotype	CHRONO	128266
Psychiatric-neurologic	Chronotype UKB	CHRONO_UKB	337119
Psychiatric-neurologic	Depressive Symptoms	DEPR	180866
Psychiatric-neurologic	Education Years	EDU	293723
Psychiatric-neurologic	Epilepsy UKB	EPI_UKB	337119
Psychiatric-neurologic	Fluid Intelligence Score UKB	FIS_UKB	337119
Psychiatric-neurologic	Insomnia In Both Sexes	INSOMN	113006
Psychiatric-neurologic	Insomnia UKB	INSOMN_UKB	337119
Psychiatric-neurologic	Insomnia UKBS	INSOMN_UKBS	337119
Psychiatric-neurologic	Migraine UKB	MIGR_UKB	337119
Psychiatric-neurologic	Migraine UKBS	MIGR_UKBS	337119
Psychiatric-neurologic	Multiple Sclerosis UKBS	MS_UKBS	337119
Psychiatric-neurologic	Neuroticism UKB	NEUROT_UKB	337119
Psychiatric-neurologic	Parkinsons Disease UKBS	PD_UKBS	337119
Psychiatric-neurologic	Psychological Problem UKBS	PSY_UKBS	337119
Psychiatric-neurologic	Schizophrenia	SCZ	150064
Psychiatric-neurologic	Schizophrenia UKBS	SCZ_UKBS	337119
Psychiatric-neurologic	Sleep Duration	SLEEP	128266
Psychiatric-neurologic	Sleep Duration UKB	SLEEP_UKB	337119
Anthropometric	BMI UKB	BMI_UKB	337119
Anthropometric	Birth Weight	BW	143677
Anthropometric	Birth Weight UKB	BW_UKB	337119
Anthropometric	Body Fat Percentage UKB	FAT_UKB	337119
Anthropometric	Bone Mineral Density	BMD	49988
Anthropometric	Height	HEIGHT	253288
Anthropometric	Intracraneal Volume	ICV	30717
Anthropometric	Standing Height UKB	HEIGHT_UKB	337119
Cardiometabolic	CH2DB NMR	CH2	24154
Cardiometabolic	Coronary Artery Disease	CAD	184305
Cardiometabolic	Deep Venous Thrombosis UKB	DVT_UKB	337119
Cardiometabolic	Deep Venous Thrombosis UKBS	DVT_UKBS	337119
Cardiometabolic	Fasting Glucose	FG	46186
Cardiometabolic	Fasting Insulin	INSUL	38238
Cardiometabolic	HDL Cholesterol NMR	HDLC	19270
Cardiometabolic	Heart Attack UKB	MI_UKB	337119
Cardiometabolic	High Cholesterol UKBS	HC_UKBS	337119
Cardiometabolic	Hypertension UKBS	HPT_UKBS	337119
Cardiometabolic	LDL Cholesterol NMR	LDLC	13527
Cardiometabolic	Pulmonary Embolism UKB	PE_UKB	337119
Cardiometabolic	Triglycerides NMR	IDL	21559
Cardiometabolic	Type 2 Diabetes UKBS	T2D_UKBS	337119
Blood	Eosinophil Count	EC	173480
Blood	Granulocyte Count	GC	173480
Blood	High Light Scatter Reticulocyte Count	HRET	173480
Blood	Lymphocyte Count	LC	173480
Blood	Monocyte Count	MC	173480
Blood	Myeloid White Cell Count	MWBC	173480
Blood	Neutrophil Count	NC	173480
Blood	Platelet Count	PLT	173480
Blood	Red Blood Cell Count	RBC	173480
Blood	Reticulocyte Count	RET	173480
Blood	Sum Basophil Neutrophil Count	BNC	173480
Blood	Sum Eosinophil Basophil Count	EBC	173480
Blood	Sum Neutrophil Eosinophil Count	NEC	173480
Blood	White Blood Cell Count	WBC	173480
Cancer	Breast Cancer	BC	120000
Cancer	ER-negative Breast Cancer	ERNBC	120000
Cancer	ER-positive Breast Cancer	ERPBC	120000
Allergy	Asthma UKBS	ATH_UKBS	337119
Allergy	Eczema	ECZ	116863
Allergy	Eczema UKBS	ECZ_UKBS	337119
Immune	Ankylosing Spondylitis UKBS	ASP_UKBS	337119
Immune	Asthma UKB	ATH_UKB	337119
Immune	Crohns Disease	CD	20833
Immune	Crohns Disease UKBS	CD_UKBS	337119
Immune	Hayfever UKB	HAY_UKB	337119
Immune	Inflammatory Bowel Disease	IBD	34652
Immune	Inflammatory Bowel Disease UKBS	IBD_UKBS	337119
Immune	Psoriasis UKBS	PSO_UKBS	337119
Immune	Rheumatoid Arthritis	RA	80799
Immune	Rheumatoid Arthritis UKBS	RA_UKBS	337119
Immune	Systemic Lupus Erythematosus	SLE	23210
Immune	Type 1 Diabetes UKBS	T1D_UKBS	337119
Immune	Ulcerative Colitis	UC	27432
Immune	Ulcerative Colitis UKBS	UC_UKBS	337119
Aging	Fathers Age At Death UKB	FAD_UKB	337119
Aging	Mothers Age At Death UKB	MAD_UKB	337119
Digestive system disease	Irritable Bowel Syndrome UKBS	IBS_UKBS	337119
Endocrine system disease	Hyperthyroidism UKBS	HYPERTHY_UKBS	337119
Endocrine system disease	Hypothyroidism UKBS	HYPOTHY_UKBS	337119
Skeletal system disease	Gout UKBS	GOUT_UKBS	337119
Skeletal system disease	Osteoporosis UKBS	OST_UKBS	337119
Morphology	Balding Pattern 2 UKB	BLDP2_UKB	337119
Morphology	Balding Pattern 3 UKB	BLDP3_UKB	337119
Morphology	Balding Pattern 4 UKB	BLDP4_UKB	337119

NHGRI-EBI GWAS catalog

To investigate the downstream effects of GWAS loci using the resources from the GTEx consortium, we obtained the list of trait-associated SNPs from the GWAS catalog (*49*) (downloaded on 9/7/2018), which, at download, contained 80,727 entries. To measure the enrichment of e/sQTL in GWAS Catalog, we computed the proportion of e/sQTL in GWAS catalog relative to the proportion of e/sQTL among all GTEx V8 variants. And we obtained the uncertainty measurement of proportion and enrichment fold using block jackknife. See (*9*) for details.

Validation of findings in BioVU Biobank

For replication analyses in BioVU (*27*), we selected 11 complex traits and 10 tissues with the largest sample size and estimated statistical power to detect true associations.

1.3 On summarizing across traits and tissues

Many of our analyses generate one statistic for each of the 4,263 (87 × 49) trait-tissue pairs. These can have a complex error structure with a wide range of standard errors and correlation between tissues. Thus, the usual "iid" (independent and identically distributed) assumption behind common statistical tests is not appropriate. For summarizing across traits for a given tissue, we assumed independence across traits but took into account the different standard errors. For summarizing across trait-tissue pairs, we allowed both correlation between tissues and correlation between traits, and corrected for different standard errors. More specifically, let S_{tp} be some statistic estimated in trait p and tissue t with standard error $\mathrm{se}(S_{tp})$.

Summarizing across traits for a given tissue. Here we describe the procedure to summarize results that have one statistic (along with its standard error) per trait-tissue pair. For each tissue t, we summarized $S_{t1}, \cdots, S_{tP}$ by fitting the following linear model:

$$S_{tp} = \mu_S^t + \epsilon_{tp} \tag{1}$$

$$\epsilon_{tp} \sim N(0, \mathrm{se}(S_{tp})^2 \times \sigma_t^2) \tag{2}$$

Hence, we obtain $\hat{\mu}_S^t$ and $\mathrm{se}(\hat{\mu}_S^t)$ as the summary of $S_{t1}, \cdots, S_{tP}$ estimates aggregated across traits, which is essentially a weighted average across traits.

Summarizing across trait and tissue pairs. Similarly, we summarized $S_{11}, \cdots, S_{tp}, \cdots, S_{TP}$ by fitting the following linear model.

$$S_{tp} = \mu_S + \mu_S^t + \mu_S^p + \epsilon_{tp} \tag{3}$$

$$\mu_S^t \sim N(0, \sigma_T^2) \tag{4}$$

$$\mu_S^p \sim N(0, \sigma_P^2) \tag{5}$$

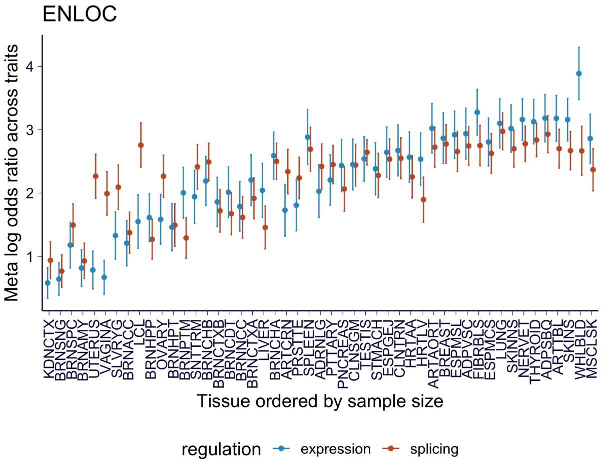
$$\epsilon_{tp} \sim N(0, \text{se}(S_{tp})^2 \times \sigma^2), \tag{6}$$

where μ_S^t is the tissue-specific random intercept (this accounts for tissue-specific features common across traits) and μ_S^p is the trait-specific random intercept (this accounts for trait-specific characteristics and thus accounts for the correlation between tissues for a given trait). The estimated $\hat{\mu}_S$ and $\text{se}(\hat{\mu}_S)$ is the average S_{tp} across all trait-tissue pairs accounting for the complex error structure.

Testing whether two statistics have different mean. For some analyses, we would like to test whether two quantities are different across all trait-tissue pairs (*e.g.* enrichment signal measured for sQTL as μ_{S_1} vs. the one measured for eQTL as μ_{S_2}, etc). For this purpose, we constructed the following paired test. First, we formed the test statistic $T^{tp} := S_{1,tp} - S_{2,tp}$ which, under the null $\mathcal{H}_0 : \mu_{S_1} = \mu_{S_2}$, has $T^{tp} \sim N(0, \text{se}(S_{1,tp})^2 + \text{se}(S_{2,tp})^2)$. Then, we summarized T^{tp} across all trait-tissue pairs by the procedure described in the previous paragraph where tissue- or trait-specific intercepts are introduced to account for the complicated correlation structure among T^{tp}'s. The resulting statistic T follows $T \sim N(0, \text{se}(T))$ under the null.

1.4 Enrichment across tissues

Detailed discussions regarding the enrichment analyses and methods to address LD contamination are described elsewhere (*9, 22*).



Supplementary Fig. S5. Enrichment of QTLs among complex trait associated variants. Enrichment estimates as *enloc* log odds ratio by tissue are summarized across traits (on y-axis) with error bar representing 95% confidence interval. Tissues (on horizontal axis) are ordered by sample size. Cis-expression results are shown in red and cis-splicing results are shown in green.

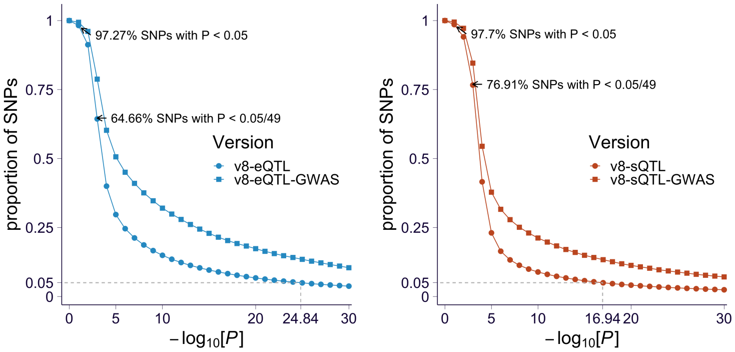
Proportion of QTLs by p-value cutoff

To estimate the proportion of SNPs considered as associated with expression (for at least one gene) at various p-value thresholds, we used the most significant p-value (tested using all GTEx individuals) for each SNP from all associations in all tissues (including all genes and variants tested). We plotted the proportion SNPs whose most significant p-value meets a p-value threshold for varying levels of this threshold (fig. S6). To test whether trait-associated SNPs are more likely to be e/sQTLs, we repeated the above procedure for the lead SNPs in the GWAS catalog.

Enrichment of GWAS catalog variants

To investigate the relevance of the QTLs in complex traits, we analyzed the database of trait-associated variants (defined as $p < 5 \times 10^{-8}$), as curated in the NHGRI-EBI GWAS catalog (see Methods; hereafter **GWAS catalog**), and tested enrichment of single tissue QTLs across complex traits represented in the GWAS catalog. Next, we examined the nominally associated loci that did not attain genome-wide significance in a subset of studies reported in the GWAS catalog.

Consistent with earlier reports (*1, 8*), we observed a 1.46 fold (s.e. = 0.02) enrichment of cis-eQTLs among the trait-associated variants (fig. S6), of which 63% are an eQTL in some tissue compared to 43% among all tested variants (MAF $>$ 0.01 in GTEx samples with European ancestry). Notably, splicing cis-QTLs showed a 1.87 fold (s.e. $=$ 0.06) enrichment with 37% of trait-associated variants as sQTLs compared to 20% of all tested variants.



Supplementary Fig. S6. Expression and splicing QTL enrichment among GWAS variants. The proportion of genetic variants associated with gene expression **(A)** and splicing **(B)** of at least one gene in at least one tissue for each p-value cutoff (on x-axis in $-\log_{10}(p)$ scale) is shown. The proportions for all tested variants are shown as squares and the proportions for the GWAS catalog variants are shown as circles.

We observed that the proportion of variants associated with expression and splicing at different significance threshold was much larger for trait-associated variants from the GWAS catalog than for the full set of tested common variants (fig. S6. The significance of this difference is reported elsewhere (*9*). Notably, as statistical power improved with increased sample sizes, spurious associations caused by trait-associated QTLs that are distinct from, but in linkage disequilibrium (LD) with, the trait causal variant(s) (LD contamination) also increased (*22*). At a nominal threshold, the proportion of common variants associated with the expression of a gene in some tissue increased from 92.7% in the V6 release (*8*) to 97.3% in V8. For splicing the proportion was 97.7%. We should caution that assigning function to a GWAS locus based on QTL association p-value alone, even with a more stringent threshold, could be misleading.

LD contamination

Here we illustrate how LD contamination affects functional interpretation of a GWAS locus. The lead BMI-associated variant (rs1558902, chr16_53769662_T_A_b38) is an eQTL associated with *FTO* expression in skeletal muscle (p=7.5×10^{-8}; FDR<0.05). However, fine-mapping of the region for BMI and *FTO* expression in muscle showed that the causal variant for each trait is likely to be distinct (with the credible set for BMI causal variants distinct from the credible set for for causal eQTLs in the locus). *FTO* expression fine-mapping assigned >99% probability of being causal to rs1861867, chr16_53814649_A_G_b38.

1.5 Fine-mapping QTL variants

We applied *dap-g* (*14*) to the 49 tissues to estimate the degree to which a variant might exert a causal effect on expression or splicing levels, using default parameter values. First, we selected genes annotated as protein-coding, lincRNA or pseudogenes. For each gene, we considered all variants within the cis-window (1Mbps) with $MAF > 0.01$, and used the same covariates as in the main eQTL analysis to correct for unwanted variation. This yielded a list of clusters (variants related by LD), and posterior inclusion probabilities (*pip*) that provide an estimate of the probability of a variant being causal. We repeated this process for splicing ratios from Leafcutter, using a cis-window ranging from 1Mbps upstream of the splicing event start location to 1Mbps downstream of the end location. We used individual-level data for GTEx-EUR subjects both for expression and splicing. Sample sizes ranged from 65 in kidney cortex to 602 in skeletal muscle tissues.

1.6 Mediation analysis

Modeling effect mediated by regulatory process

We compared the magnitude of GWAS and cis-QTL effect sizes, which is the basis of multi-SNP Mendelian randomization approaches (*50*).

To formalize the relationship between the GWAS effect size (δ) and the QTL effect size (γ), we assumed an additive genetic model for the GWAS trait. Specifically, for variant k,

$$Y = \sum_k \delta_k \cdot X_k + \epsilon, \tag{7}$$

where X_k is the allele count of variant k, Y is the trait, and ϵ is the un-explained variation. We decomposed GWAS effect size into its mediated and un-mediated components,

$$\delta_k = \sum_{g \in \mathcal{G}_k} \beta_g \gamma_{k,g} + \nu_k, \tag{8}$$

where $\mathcal{G}_k$ represents the set of genes regulated by variant k with corresponding QTL effect size as $\gamma_{k,g}$, and ν_k is the un-mediated effect of variant k on trait. And β_g is the downstream effect of gene g on the trait.

Selection of fine-mapped variants as instrumental variables

To investigate the relationship between GWAS and QTL effect sizes in the transcriptome, we generated a set of fine-mapped QTL signals derived from *dap-g* fine-mapping performed in the GTEx-EUR individuals (see Section 1.5) to serve as proxy for causal QTLs. For splicing, we utilized sQTLs at the splicing event/variant level rather than the gene/variant level. In particular, we selected the top variant within each 25% credible set of a gene or splicing event and filtered out the QTLs with *pip* less than 1%. For each of the selected QTLs, we used the QTL effect size estimated from the marginal test (using the GTEx-EUR individuals) and GWAS effect size obtained from the imputed z-score from the GWAS imputation by $\hat{\beta} \approx z/\sqrt{f(1-f)N}$, where f is the allele frequency and N is the GWAS sample size.

Correlation between GWAS and QTL effect sizes

For each trait-tissue pair, we calculated the Pearson correlation of the magnitude of observed GWAS effect size and of cis-eQTL effect size, $\widehat{\text{Cor}}(|\hat{\delta}_k|, |\hat{\gamma}_k|)$, for the list of selected fine-mapped QTLs, as described in Section 1.6. The observed Pearson correlation captures the mediated effect (see details in Section 1.6). To obtain a null distribution for the correlation, we computed the Pearson correlation under the shuffled data within each LD-score bin defined by quantiles (100 bins were used). The significance of the difference between observed and null distribution was calculated using the method described in 1.3.

Transcriptome-wide estimation of the downstream effect size

To estimate the transcriptome-wide contribution of the mediated effects on complex traits, we proposed a mixed-effects model on the basis of Eq. 8,

$$|\delta_k| = \beta_g \cdot (\text{sign}(\delta_k) \cdot \gamma_{k,g}) + b_0 + b_1 \cdot \sqrt{\text{LD-score}_k} + \epsilon \tag{9}$$

$$\beta_g \sim N(0, \sigma^2_{\text{gene}}) \tag{10}$$

$$\epsilon \sim N(0, \sigma^2), \tag{11}$$

where b_0, b_1 are the fixed effect capturing the un-mediated effect and β_g is the downstream effect (mediated effect) of the gene or splicing event g. In short, we assumed an infinitesimal model on the downstream effect and aimed at estimating σ^2_{gene} as the transcriptome-wide contribution of the mediated effect. For each tissue-trait pair, we fitted the model using selected fine-mapped QTLs, as described in Section 1.6, along with the corresponding $\hat{\delta}_k$, $\hat{\gamma}_{k,g}$. To obtain the distribution of σ^2_{gene} under the null, we performed the same calculation using shuffled GWAS effect sizes.

Concordance of mediated effects for allelic series of independent eQTLs

Under the mediation model in Eq. 8, we expect that for a given gene with multiple QTL signals, these signals should share the same downstream effect, β_g. Since the number of splicing events with multiple QTL signals was limited, we restricted this analysis to eQTLs only. We tested for concordance of downstream effect size obtained from the primary and secondary eQTL of a gene (ranked by QTL significance or QTL effect size estimate). Specifically, for a given trait and gene g, we defined the observed downstream effect for the kth variant as $\hat{\beta}_{k,g} = \hat{\delta}_k / \hat{\gamma}_{k,g}$. Thus, for each gene, we obtained $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$ as the observed downstream effect for the primary and secondary eQTLs if more than one eQTL signal was detected by *dap-g*. Ideally, for a mediating gene in a causal tissue (or a good proxy tissue), we would expect that $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$ should tend to have consistent value as compared to random. We measured the concordance in two ways: 1) correlation between $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$; 2) percent concordant, defined as the fraction of eQTL pairs having the same sign in $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$

Concordance as compared to non-colocalized genes with matched LD. To ensure that the concordance between $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$ was not driven by LD, we compared the concordance of primary and secondary eQTLs between colocalized and likely non-causal genes with similar LD pattern between primary and secondary eQTLs. Specifically, for each trait-tissue pair, we obtained primary and secondary eQTLs based on magnitude of effect size and measured the concordance as $\widehat{\text{Cor}}(\hat{\beta}_{prim}, \hat{\beta}_{sec})$. We computed the concordance for colocalized genes at various *enloc* rcp cutoffs (obtained from GTEx-EUR individuals). Furthermore, we randomly sampled the same number of genes with *enloc* rcp < 0.01 by matched LD and calculated the corresponding concordance as the null. To reduce the effect of outliers on concordance calculation, we removed genes with $\hat{\beta}_{prim}$ or

$\hat{\beta}_{sec}$ in the top and bottom 5%. We kept only trait-tissue pairs with more than 10 genes observed after removing outliers.

Concordance as compared to null with matched LD. The correlation between primary and secondary eQTLs (pairwise LD) could introduce correlation between primary and secondary $\hat{\delta}$'s and similarly to primary and secondary $\hat{\gamma}$'s, which would potentially contribute to concordance. To account for this confounding, for each gene, we simulated $\tilde{\delta}_{prim}$ and $\tilde{\delta}_{sec}$ preserving the correlation introduced by pairwise LD with

$$(\tilde{\delta}_{prim}, \tilde{\delta}_{sec}) \sim N(\Sigma \begin{bmatrix} \delta_{prim} \\ \delta_{sec} \end{bmatrix}, \Sigma) \tag{12}$$

$$\Sigma = \begin{bmatrix} 1 & \hat{R} \\ \hat{R} & 1 \end{bmatrix} \tag{13}$$

$$\delta_{prim}, \delta_{sec} \sim_{iid} N(0, 1) \tag{14}$$

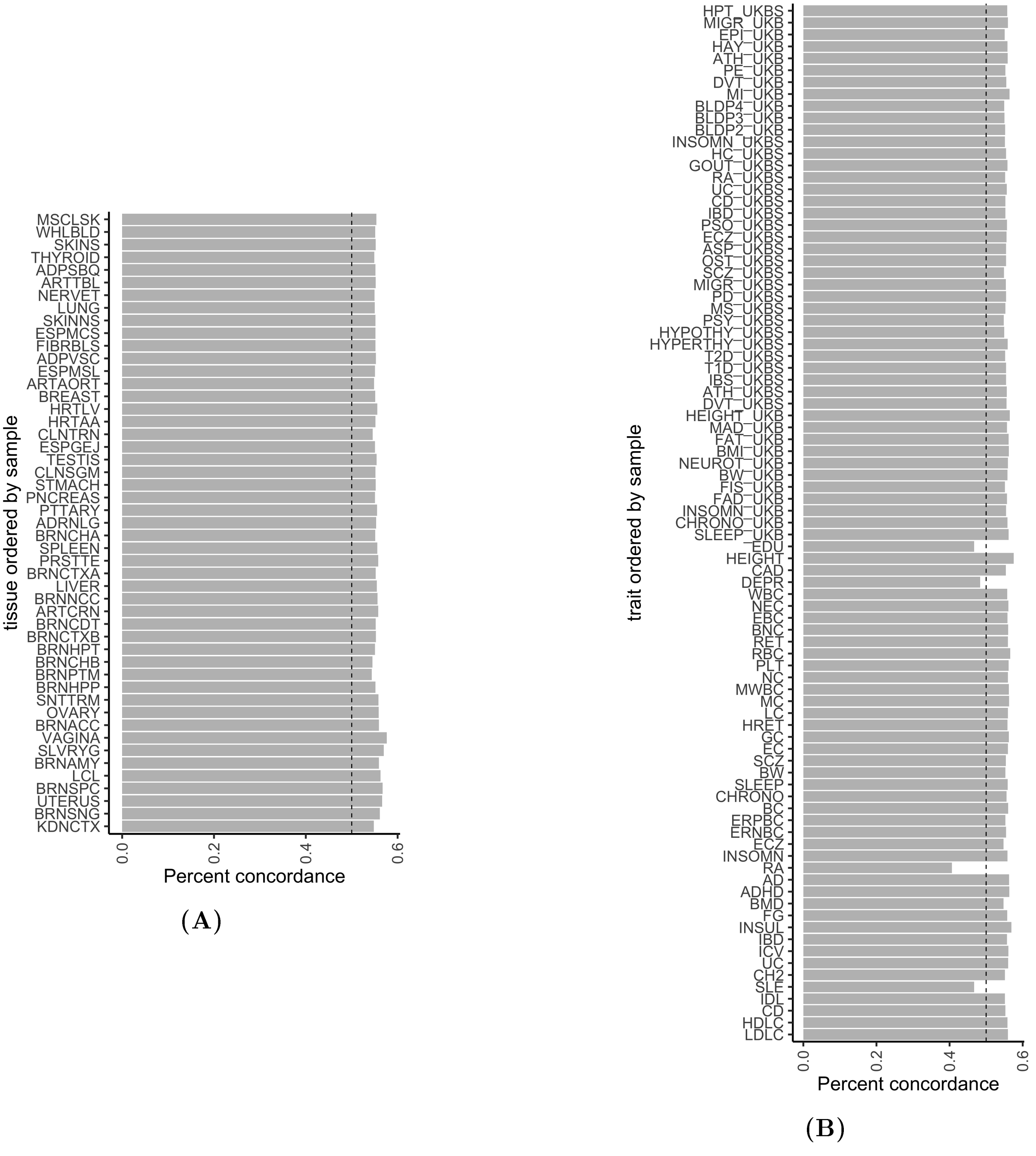
where $\hat{R}$ is the sample correlation (from GTEx-EUR individuals) of primary and secondary variants. This simulation scheme is equivalent to simulating phenotype as $\tilde{Y} \sim N(X\delta_{prim} + X\delta_{sec}, \sigma^2)$ and running GWAS on the GTEx-EUR genotypes.

Visualizing the concordance among *enloc* colocalized genes. To visualize the concordance of $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$, we first scaled $\hat{\delta}$ and $\hat{\gamma}$ by their standard deviation among all eQTLs selected in Section 1.6. Then, we extracted the set of genes with exactly two *dap-g* eQTLs (as described in 1.6) and labelled the two eQTLs as primary and secondary based on QTL significance or QTL effect size. We computed $\hat{\beta}_{prim}$ and $\hat{\beta}_{sec}$ and removed the genes with $\hat{\beta}_{prim}$ or $\hat{\beta}_{sec}$ in the top and bottom 5%. As a control, we also simulated random δ to compute simulated β_{sim} for downstream analysis. We further filtered the genes by selecting only those with *enloc* rcp > 0.1.

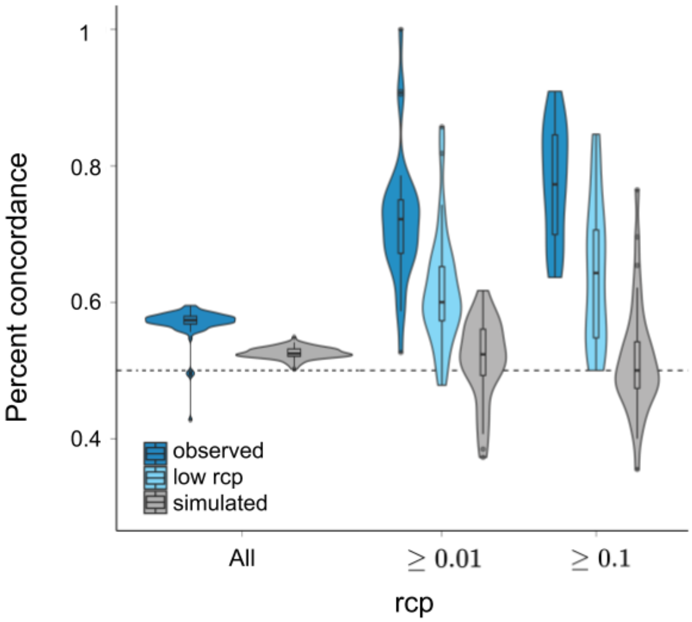
Widespread dose-dependent effects of expression and splicing regulation on complex traits (cont. from main text)

With multiple fine-mapped QTLs being detected in expression data, we proceeded to look into the effect of primary and secondary eQTLs. A third line of support for the dose-dependent effect was provided by the fact that primary eQTLs (ranked by effect size) showed, on average, larger GWAS effect sizes than secondary eQTLs (Fig. 2F).

Allelic series of independent eQTLs extensively replicate dose-response slopes (cont. from main text)



Supplementary Fig. S7. Proportion of genes with concordant sign in $\hat{\beta}$ between primary and secondary eQTLs (percent concordance) across tissues and traits. We computed the fraction of genes with primary and secondary eQTLs having concordant sign in $\hat{\beta} := \hat{\delta}/\hat{\gamma}$ (as percent concordance on y-axis). **(A)** The percent concordance for 49 tissues aggregated across 87 traits. **(B)** The percent concordance for 87 traits aggregated across 49 tissues.

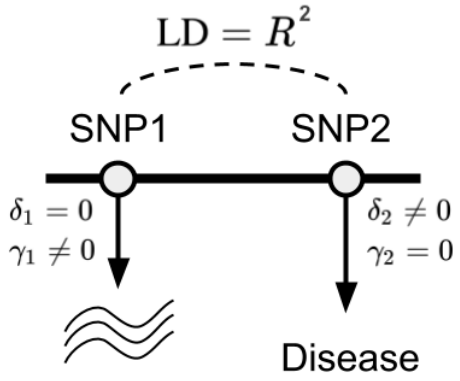


Supplementary Fig. S8. Percent concordance for genes showing colocalization evidence versus LD-matched non-colocalized genes and LD-matched simulated $\hat{\delta}_{prim}$ and $\hat{\delta}_{sec}$. The percent concordance (on y-axis) for genes with various *enloc* colcalization signal (on x-axis) is shown for 87 traits in whole blood in blue. The percent concordance obtained by sampling from non-colocalized genes (rcp < 0.01) with matched LD between primary and secondary QTLs is shown in light blue and that obtained by simulating $\hat{\delta}_{prim}$ and $\hat{\delta}_{sec}$ with observed pairwise LD is shown in gray.

Providing further support for the dose-dependent effect, the concordance of the mediated effects was consistently observed across traits and tissues and retained concordant directionality (Fig. 2D-E, figs. S8,S7), especially among colocalized genes (rcp > 0.1). Primary and secondary eQTLs ranked by eQTL effect size instead of p-value yielded similar patterns (fig. S11).

Correlation between GWAS and QTL effect sizes and mixed-effects model account for LD contamination

We illustrate the intuition behind the LD-contamination correction when the average mediated effects are estimated using the approximate method (correlation of absolute values) or the mixed-effects approach.



Supplementary Fig. S9. Diagram representation of LD contamination.

Consider the LD-contamination scenario where SNP 1 and SNP 2 are in LD with correlation R^2 (suppose LD is fixed) and have a non-zero effect on gene expression and trait, respectively (as shown in fig. S9). The marginal effect estimates of SNP 1, *i.e.* $\hat{\delta}_1$ and $\hat{\gamma}_1$, are given by

$$\hat{\delta}_1 = R\delta_2 + \epsilon_{\text{GWAS}} \tag{15}$$

$$\hat{\gamma}_1 = \gamma_1 + \epsilon_{\text{QTL}}, \tag{16}$$

where Eq. 15 holds because the marginal effect size depends on LD. To determine the covariance of the magnitude of the GWAS and QTL estimates for SNP 1, we consider $\text{E}(|\hat{\delta}_1||\hat{\gamma}_1|)$.

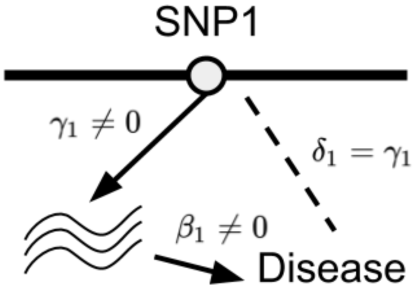
$$\begin{aligned} \text{E}(\hat{\delta}_1\hat{\gamma}_1 \; |R) &= \text{E}((R\delta_2 + \epsilon_{\text{GWAS}}) \cdot (\gamma_1 + \epsilon_{\text{QTL}}) \; |R) & (17) \\ &= \text{E}(R\delta_2\gamma_1 \; |R) + \text{E}(\epsilon_{\text{GWAS}}\gamma_1) + \text{E}(R\delta_2\epsilon_{\text{QTL}} \; |R) + \text{E}(\epsilon_{\text{GWAS}}\epsilon_{\text{QTL}}) & (18) \\ &= R \cdot \text{E}(\delta_2\gamma_1), & (19) \end{aligned}$$

where Eq. 19 holds since the last three terms in the previous line are zeros, due to the independence among ϵ_{GWAS}, ϵ_{QTL}, and true effect sizes, δ and γ.

Hence, the covariance of the GWAS and QTL effect sizes under the LD contamination scenario is

$$\begin{aligned}
\mathrm{Cov}(\hat{\delta}_1, \hat{\gamma}_1 \mid R) &= \mathrm{E}(\hat{\delta}_1 \hat{\gamma}_1 \mid R) - \mathrm{E}(\hat{\delta}_1 \mid R) \cdot \mathrm{E}(\hat{\gamma}_1 \mid R) && (20)\\
&= R \cdot \mathrm{E}(\delta_2 \gamma_1) - \mathrm{E}(\hat{\delta}_1 \mid R) \cdot \mathrm{E}(\hat{\gamma}_1 \mid R) && (21)\\
&= R \cdot \mathrm{E}(\delta_2 \gamma_1) - \mathrm{E}(R\delta_2 + \epsilon_{\mathrm{GWAS}}) \cdot \mathrm{E}(\gamma_1 + \epsilon_{\mathrm{QTL}}) && (22)\\
&= R \cdot \mathrm{E}(\delta_2 \gamma_1) - R \cdot \mathrm{E}(\delta_2) \cdot \mathrm{E}(\gamma_1) && (23)\\
&= R \cdot \mathrm{Cov}(\delta_2, \gamma_1), && (24)
\end{aligned}$$

which implies that conditioning on LD, the observed correlation between $\hat{\delta}$ and $\hat{\gamma}$ should be very small.



Supplementary Fig. S10. Diagram representation of mediation model.

Similarly, we can derive the correlation between GWAS and QTL effect size estimates under the simple mediation model shown in Supplementary Figure S10, where we have

$$\begin{aligned}
\hat{\delta}_1 &= \beta_1 \gamma_1 + \epsilon_{\mathrm{GWAS}} && (25)\\
\hat{\gamma}_1 &= \gamma_1 + \epsilon_{\mathrm{QTL}}, && (26)
\end{aligned}$$

where Eq. 25 follows by definition of the mediation model considering no direct effect. So,

$$\begin{aligned}
\mathrm{Cov}(\hat{\delta}_1, \hat{\gamma}_1 \mid \beta_1) &= \beta_1 \mathrm{E}(\gamma_1^2) - \beta_1 \mathrm{E}(\gamma_1)^2 && (27)\\
&= \beta_1 \mathrm{Var}(\gamma_1) && (28)
\end{aligned}$$

So, if we consider a gene locus, which naturally conditions on local LD and gene-level effect β, we can conclude that

$$\mathrm{Cov}(\hat{\delta}_1, \hat{\gamma}_1 \ |\text{gene locus}) = \mathrm{Cov}(\hat{\delta}_1, \hat{\gamma}_1 \ |\beta_1, R) \tag{29}$$

$$= \begin{cases} 0 & \text{LD contamination} \\ \mathrm{Var}(\gamma_1) & \text{Mediation model} \end{cases} \tag{30}$$

In practice, we do not have enough observations (*i.e.* independent QTLs) for each gene so that we cannot compute the above conditional correlation. Instead, motivated by this intuition, we developed two work-around approaches to capture the mediated effect across the transcriptome. First, we considered the correlation between GWAS and QTL effect sizes across the transcriptome. Essentially, when we take the correlation across all genes, we marginalize out the effect of β and R. Since the direction of β is arbitrary (with $\mathrm{E}(\beta) = 0$), we will not see the correlation between GWAS and QTL effect size even under the mediation model. To account for this fact, we proposed to examine the correlation between the magnitude of GWAS and QTL effect sizes, *i.e.* $\mathrm{Cor}(|\hat{\delta}|, |\hat{\gamma}|)$, which still captures the distinction between LD contamination and mediation model, since $\mathrm{Cor}(|\delta_1|, |\gamma_2|) = 0, \ \forall \delta_1 \perp \gamma_2$.

However, if the LD contamination goes into both GWAS and QTL effect sizes, $\mathrm{Cor}(|\delta|, |\gamma|)$ will be positive which is driven completely by LD. For instance, a region with high LD results in big GWAS and QTL effect sizes in magnitude and a region with low LD results in small GWAS and QTL effect sizes in magnitude. If we plot the magnitude of the QTL effect size against the one for GWAS across all regions with varying LD values, we will see the correlation as well. To account for this fact and measure the contribution of LD in the observed $\mathrm{Cor}(|\hat{\delta}|, |\hat{\gamma}|)$, we constructed permuted null by shuffling effect sizes within each LD-score bin.

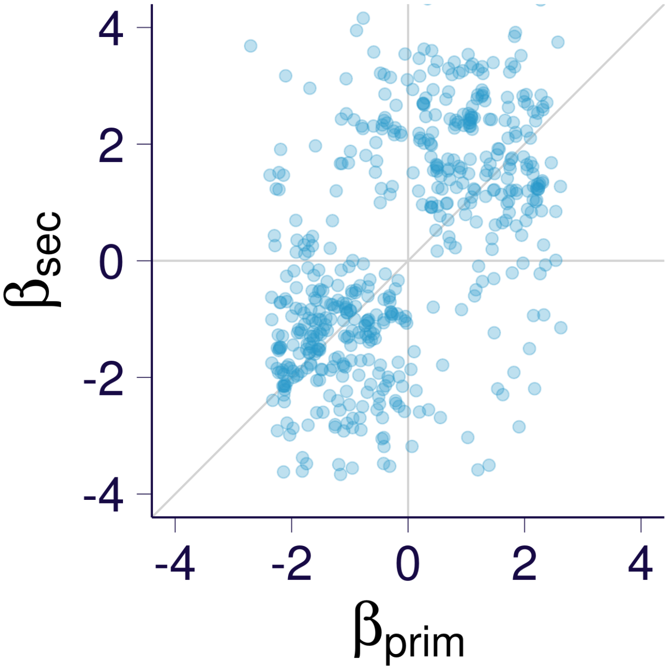
We also developed a mixed-effects approach. We model β as a random effect, $\beta \sim N(0, \sigma^2_{\text{gene}})$, and instead of averaging β's across the whole transcriptome (this is what $\mathrm{Cor}(\hat{\delta}, \hat{\gamma})$ does), we quantify the mediated effect by estimating σ^2_{gene}. Specifically, we fit

$$|\hat{\delta}_i| = \beta_g(\mathrm{sign}(\hat{\delta}_i) \cdot \hat{\gamma}_{gi}) + b_0 + b_1\sqrt{\text{LD-score}_i} + \epsilon_i \tag{31}$$

$$\beta_g \sim_{iid} N(0, \sigma^2_{\text{gene}}) \tag{32}$$

$$\epsilon_i \sim_{iid} N(0, \sigma^2), \tag{33}$$

where g indicates gene index and i indicates variant index. And b_0, b_1 are fixed effects accounting for the contribution of LD to the magnitude of GWAS effect size. To obtain the null for testing whether $\sigma^2_{\text{gene}} = 0$, we permute $\hat{\delta}_i$ and corresponding covariate (LD-score) altogether keeping the structure of grouping variants by gene the same. Essentially, the mixed-effects model in Eq. 31 is designed to capture the distinction between LD contamination and mediation shown in Eq. 30 in a transcriptome-wide manner.



Supplementary Fig. S11. Downstream effects of primary and secondary eQTLs highly correlated. Downstream effects of primary and secondary cis-eQTL signals ($\hat{\delta}/\hat{\gamma}$) which are ordered by $|\hat{\gamma}|$ are shown for all 87 traits in Whole Blood.

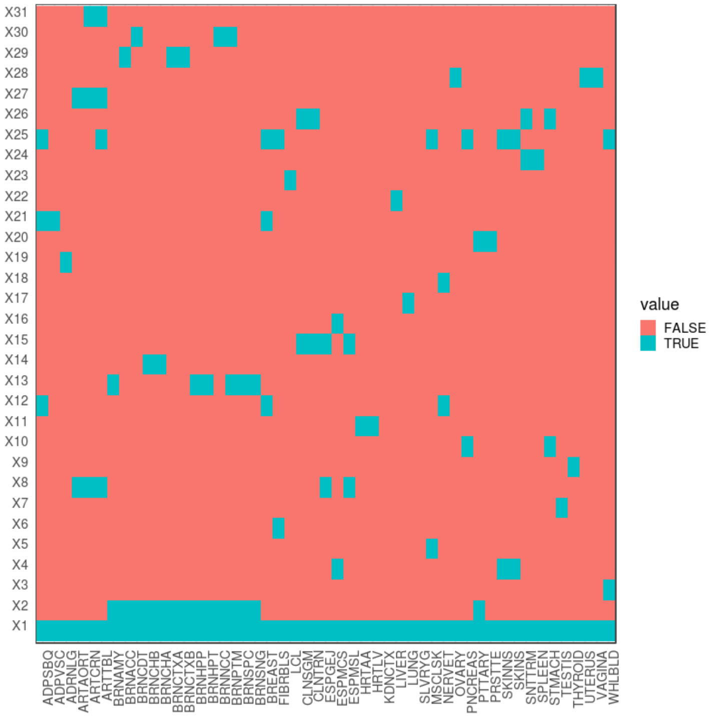
1.7 Patterns of regulation of expression across tissues

Identifying tissue-specificity of eQTL

We used FLASH Sparse Factor Analysis (*51*) to identify latent factors specific to different tissue clusters. Briefly, `flashr` was run on a set of top eQTLs (obtained from all GTEx individuals) per gene which had been tested in all 49 tissues (around 16,000 eQTLs in total were selected) which showed strong evidence of being active in at least one tissue. Specifically, for each selected variant-gene pair, the marginal effect size estimates were extracted for all 49 tissues regardless of whether it was significant in that tissue or not. The resulting estimated effect-size matrix (of dimension $\sim 16,000 \times 49$) was the input to `flashr` (with normal prior on loading and uniform with positive support as prior on factor) to obtain the sparse factors (see URLs). The `flashr` run yielded 31 FLASH factors (S12), which were used to assign the tissue-specificity of an eQTL. We defined the eQTL tissue-specificity by projecting the estimated effect-size vector across 49 tissues onto the FLASH factors and computed the quality of the projection, PVE, as $\text{PVE}_k = \frac{\|\vec{\beta}_k\|_2^2}{\|\vec{\beta}\|_2^2}$. PVE_k represented the quality score for using FLASH factor k to explain the cross-tissue pattern of eQTL. The eQTL was assigned to a FLASH factor k if PVE_k was maximal among all FLASH factors and $\text{PVE}_k > 0.2$ and for those with $\text{PVE}_k \leq 0.2$ in all FLASH factors, `NA` (short for not assigned) was assigned instead. These "not assigned" eQTLs had more complex tissue-sharing pattern than the factors captured in the FLASH analysis. To obtain an interpretable tissue-specificity category, we labeled *Factor1* as the shared factor, *Factor2*, *Factor13*, *Factor14*, *Factor29*, and *Factor30* as brain-specific factors, and the rest of the factor assignment as *other factors*.

Smoothing effect size estimates by leveraging global patterns of tissue sharing

We applied the multivariate adaptive shrinkage implemented in *mashr* (*52*) to smooth cis-eQTL effect size estimates (obtained from all GTEx individuals) by taking advantage of correlation between tissues. To fit the *mashr* model, we used the set of $\sim 16,000$ cis-eQTLs as stated in Section 1.7 to learn the *mashr* prior, and then fit the *mashr* model using $\sim 40,000$ randomly selected variant-gene pairs for the same set of eGenes. We learned data-driven *mashr* priors in three ways: 1) FLASH factors as described in Section 1.7; 2) PCA with number of PC = 3; 3) empirical covariance of observed z-scores. The data-driven covariances were further denoised by calling `cov_ed` in `mashr`. Furthermore, we included the set of canonical covariances as described in (*52*) as an additional *mashr* prior. We fit the *mashr* model using the set of randomly selected variant-gene pairs with the error correlation estimated by applying `estimate_null_correlation` function in `mashr` and the priors obtained above. The resulting *mashr* model was used to compute the posterior mean, standard deviation, and local false sign rate (LFSR) for a given variant-trait pair.



Supplementary Fig. S12. Tissue-specific factor estimation using flashr. We performed empirical Bayes matrix factorization (by `flashr`) on a set of the top cis-eQTLs (per gene), and we restricted factors to have non-negative values. We binarized the resulting factors by thresholding the tissue contribution to TRUE if it is at least 20% of the maximum. The pattern after thresholding is shown.

1.8 Causal gene prioritization

Two classes of methods can be used to identify the target genes of GWAS loci. One class is based on the colocalization of GWAS and QTL loci, which seeks to determine whether the causal variant for the trait is the same as the causal variant for the molecular phenotype. The other class is based on the association between the genetically regulated component of gene expression (or splicing) with the trait.

Colocalization

For a given variant associated with multiple traits such as gene expression (eQTL) and complex disease (trait-associated variant), extensive LD makes it challenging to identify the underlying true causal mechanisms. Thus, we conducted colocalization analysis using two independent approaches: *coloc* (*17*) and *enloc* (*16*)), to estimate whether a gene's expression or a splicing event shares a causal variant with a trait.

enloc

We computed Bayesian regional colocalization probability (rcp) using *enloc*, to estimate the probability of a GWAS region and a gene's cis window sharing causal variants. We leveraged the same *dap-g* results from 1.5, and split the GWAS summary statistics into approximately LD-independent regions (*28*), each region defining a GWAS locus. For each trait-tissue combination, we computed the rcp of every overlapping GWAS locus to a gene's or splicing event's cis window with *enloc* default parameters. The enrichment estimates obtained by *enloc* are shown in fig. S5.

For each trait, we counted the number of GWAS loci that contain a GWAS significant hit, and among these, the number of loci that additionally contain a gene with *enloc* colocalization $rcp > 0.5$. As shown in fig. S17C, across traits, a median 29% of loci with a GWAS signal contain an *enloc* colocalized signal. Given *enloc*'s conservative nature, we caution that $rcp < 0.5$ does not mean that there is no causal relationship between the molecular phenotype and the complex trait; rather, it should be interpreted as lack of sufficient evidence with current data. We summarize the findings in fig. S18. We observed a smaller proportion of GWAS loci containing a colocalized splicing event (median 11% across traits).

coloc

We computed *coloc* on all cis-windows with at least one eVariant (cis-eQTL per-tissue q-value< 0.05) or sVariant. For binary traits, case proportion and 'cc' trait type parameters were used. For continuous traits, sample size and 'quant' trait type parameters were used. In both cases, imputed or calculated z-scores were used as effect coefficients in Bayes factor calculations.

Coloc is very sensitive to the choice of priors. We used *enloc*'s enrichment estimates to define data-based priors in a consistent manner. First, we defined likely LD-independent blocks of variants using definitions provided previously (*28*). The probability of eQTL signal, $\Pr(d_i = 1)$, was estimated using *dap-g* (*14*). Subsequently, we calculated priors p_1,

p_2, and p_{12} for colocalization analyses as follows:

$$\begin{aligned}
p_1 &:= \Pr(\gamma_i = 1, d_i = 0) = \frac{\exp(\alpha_0)}{1+\exp(\alpha_0)} \times (1 - \Pr(d_i = 1)), \\
p_2 &:= \Pr(\gamma_i = 0, d_i = 1) = \frac{1}{1+\exp(\alpha_0+\alpha_1)} \times \Pr(d_i = 1), \text{ and} \\
p_{12} &:= \Pr(\gamma_i = 1, d_i = 1) = \frac{\exp(\alpha_0+\alpha_1)}{1+\exp(\alpha_0+\alpha_1)} \times \Pr(d_i = 1),
\end{aligned}$$

where α_0 and α_1 indicate intercept effect estimate and log odds ratio estimate for the enrichment using *enloc*, respectively.

We ran *coloc* using variants in the cis-window for each gene and the intersection with each GWAS trait, obtaining five probabilities for each (gene, tissue, trait) tuple: **P0** for the probability of neither expression nor GWAS having a causal variant; **P1** for the probability of only expression having a causal variant; **P2** for only the GWAS having a causal variant; **P3** for the GWAS and expression traits to have distinct causal variants; **P4** for the GWAS and expression traits to have a shared causal variant. We repeated this process using sQTL results.

1.9 Fine-mapping of GWAS using summary statistics

To investigate the robustness of fine-mapping, we fine-mapped "height" from the GIANT GWAS meta-analysis and "standing height" from the UK Biobank using `susieR` (*53*). We performed fine-mapping using `susie_bhat` within each LD block (*28*). We used GWAS effect sizes $\tilde{\beta}$ imputed from z-scores by $\tilde{\beta} = z/\sqrt{Nf(1-f)}$ and $\mathrm{se}(\tilde{\beta}) = \tilde{\beta}/z$, where f is allele frequency and N is GWAS sample size. The GTEx-EUR individuals were used as the reference LD panel. We applied the same approach to fine-map the BMI-associated FTO locus using the UK BioBank BMI data.

1.10 Association to predicted expression or splicing

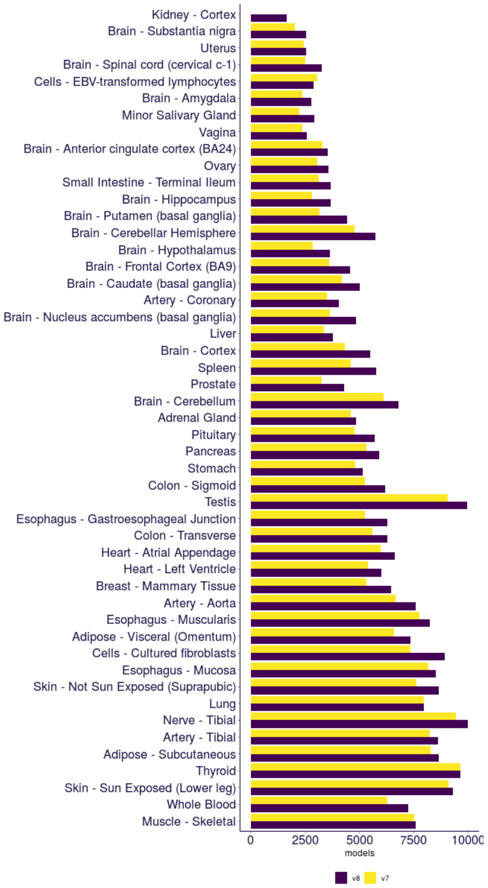
Prediction models

To predict expression, we constructed linear prediction models (*24*), using only individuals of European ancestry, and variants with MAF > 0.01, for genes annotated as protein-coding, pseudo-gene, or lncRNA. For each gene-tissue pair, we selected the variants with highest *pip* in their cluster, and kept those achieving pip > 0.01 in *dap-g* (*14*). We used *mashr* (*52*) effect sizes (as computed in 1.7) for each selected variant. For each model, we computed the covariance matrix between variants using only individuals of European ancestries, with sample sizes ranging from 65 (kidney - cortex) to 602 (skeletal muscle). This allowed us to build LD panels for every tissue, noting that GWAS studies are conducted on populations of predominantly European ancestries. For every gene, we also computed the covariance of all the variants present across the different tissue models, compiling a cross-tissue LD panel to compute the correlation between predicted expression levels across tissues. We refer to these models as *mashr* models. We compared the number of *mashr* models to the number of Elastic Net models from GTEx version 7 (fig. S13). We generated analogous prediction models for splicing ratios, as computed by Leafcutter (*45*), applying the same model-building methodology to the data from the sQTL analysis.

We generated a secondary set of prediction models based on fine-mapping information. For every gene-tissue pair, we selected the variants with the highest *pip* in each cluster achieving pip > 0.01 as explanatory variables. We performed an Elastic Net (*54*) regression of expression on these variables, for genes annotated as protein coding, pseudogene, or lncRNA. We employed a cross-validated strategy, and kept only models that achieved cross-validated correlation $\rho > 0.1$ and cross-validated prediction performance p-value $p < 0.05$. Each variant's effect size was penalized by a factor $1 - \text{pip}$, so that variants with higher probabilities were more likely to impact the model. Expression phenotypes were adjusted for unwanted variation using covariates such as gender, sequencing plaform, age, the top 3 principal components from genotype data, and PEER factors. The number of PEER factors was determined from the sample size: 15 for $n < 150$, 30 for $150 \leq n < 250$, 45 for $250 \leq n < 350$, 60 for $350 \leq n$. We obtained 686,241 models for different (gene, tissue) pairs. For each model, we computed tissue-level and cross-tissue covariances as in the *mashr* models.

We also generated analogous prediction models for splicing ratios, with the same model-building methodology applied to the data from the sQTL analysis, obtaining 1,816,703 (splicing event, tissue) pairs.

We constructed additional sets of prediction models comprised of a single snp, using the top eQTL per gene or the primary, secondary or tertiary eQTL arising from conditional or marginal analysis, in order to assess effect difference on the complex traits.



Supplementary Fig. S13. Number of models available in v8 MASHR family of models, compared to v7 Elastic Net family. Tissues are ordered by sample size.

S-PrediXcan

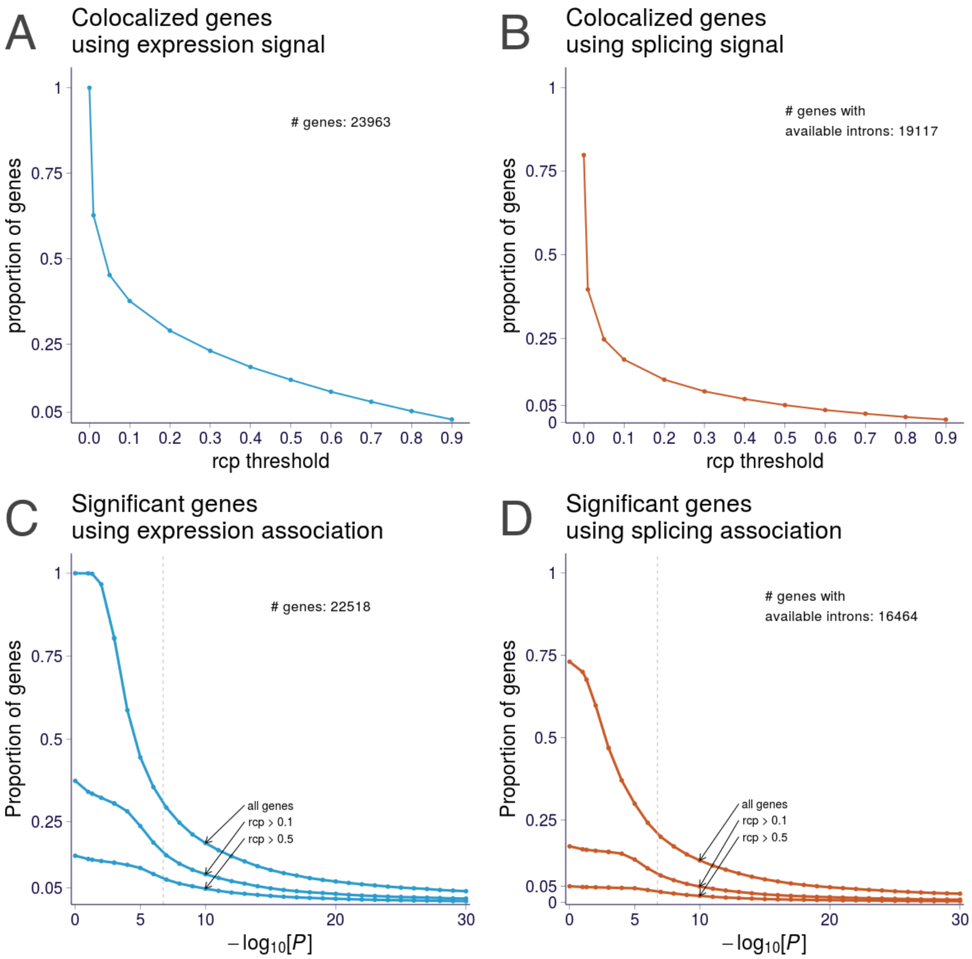
We performed S-PrediXcan analysis (*25*) on the 87 complex traits, using the GWAS summary statistics described in 1.2, to identify trait-associated genes (typically $p < 2.5 \times 10^{-7}$). We used the 49 models and LD panels described in 1.10, separately on each trait, to obtain 59,485,548 (gene, tissue, trait) tuples. Repeating this process to generate splicing event ratio models, we obtained 154,891,730 (splicing event, tissue, trait) tuples; for each trait, the Bonferroni-significance threshold was $p < 9.5 \times 10^{-8}$.

Colocalized and significantly associated genes

We assessed how many genes present evidence of trait association and colocalization, using both expression and splicing event. First, we counted the proportion of genes that showed a colocalized expression signal with any trait in any tissue, and observed 15% such genes at rcp>0.5. Then, for each gene, we considered the splicing event with highest colocalization value in any trait or tissue, and found evidence for 5% at rcp>0.5.

Then we repeated this process for S-PrediXcan associations at different signifcance thresholds. About 30% of genes showed a significant S-PrediXcan association to any trait, and only 8% when filtered for associations with rcp>0.5. When using the highest splicing association and colocalization value for a gene, these proportions were 20% and 3%, respectively.

These proportions gauge our power to predict causal genes affecting complex traits on the GTEx resource, with expression yielding more findings than splicing.



Supplementary Fig. S14. Proportion of genes with a colocalized or associated signal using expression or splicing event.
A shows the proportion of genes with colocalization evidence in expression data, for different rcp thresholds. 3,477 genes show evidence at rcp>0.5 (15% out of 23,963 genes with *enloc* results).
B shows the proportion of genes with colocalization evidence in splicing data; 1,277 genes (5% of all 23,963) show evidence at rcp>0.5.
C shows the proportion of genes with association evidence in expression data, additionally filtered by colocalization on different thresholds. About 30% of genes show associations at the bonferroni threshold ($p < 0.05/686,241$), while 8% also show colocalization evidence.
D shows the proportion with association and colocalization evidence in splicing data; about 20% show association evidence ($p < 0.05/1,816,703$) and 3% are also colocalized.

S-MultiXcan

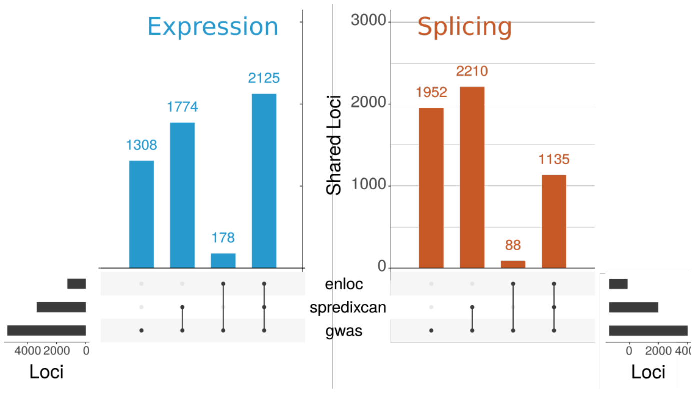
There is substantial sharing of eQTLs across tissues (*8*). Therefore, we applied S-MultiXcan (*10*), an approach to exploit the tissue sharing of regulatory variation, to improve our ability to identify trait-associated genes. The method extends the single-tissue S-PrediXcan approach, leveraging GWAS summary statistics and taking into account the correlation between tissues. We obtained association statistics for 1,958,220 (gene, trait) pairs and 11,986,329 (splicing event, trait) pairs.

PrediXcan replication in BioVU

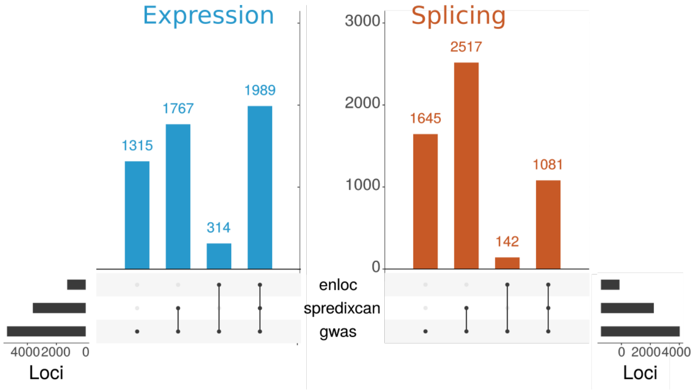
We replicated the significant gene-level associations for a prioritized list of traits (table S16) using BioVU (*27*), Vanderbilt University's DNA Biobank tied to a large-scale Electronic Health Records (EHR) database. We sought BioVU replication in the exact discovery tissues for the significant gene-trait associations. We restricted our analysis to subjects of European ancestries, using principal component analysis as implemented in EIGENSOFT (version 7.1.2; (*55*)). First, we estimated the genetically determined component of gene expression in the BioVU individuals using the PrediXcan imputation models. We then conducted association analysis for the prioritized traits using logistic regression, with sex and age as covariates.

Summary-data-based Mendelian Randomization (SMR) and HEIDI

We performed top-eQTL based Summary-data-based Mendelian Randomization (SMR) (*12*) analysis of the 4,263 tissue-trait pairs. SMR, which integrates summary statistics from GWAS and eQTL data, has been used to prioritize genes underlying GWAS associations.



Supplementary Fig. S15. Causal gene prioritization using S-PrediXcan and *enloc*. Summary of GWAS loci that also contain an associated S-PrediXcan or *enloc* signal, for expression (left) and splicing (right), using MASHR models.



Supplementary Fig. S16. Causal gene prioritization using S-PrediXcan and *enloc*. Summary of GWAS loci that also contain an associated S-PrediXcan or *enloc* signal, for expression (left) and splicing (right), using Elastic Net models.

Table S2: Expression and splicing prediction models using mashr-based models.

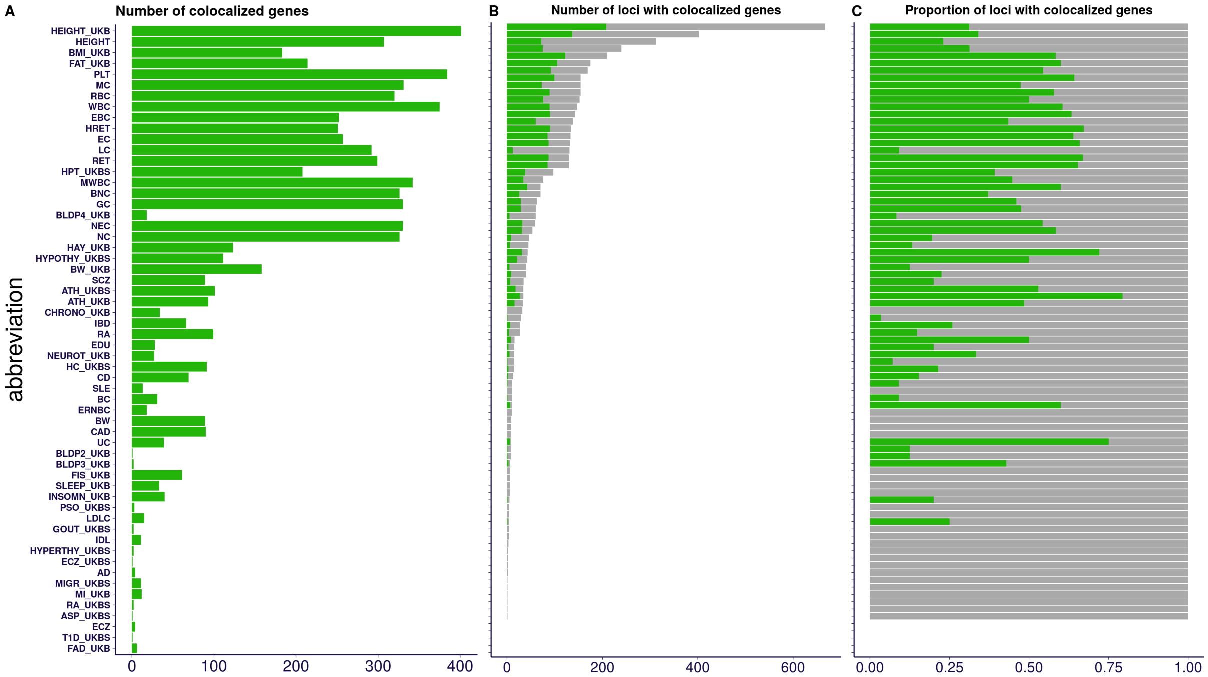
name	abbreviation	european samples	expression models	splicing models
Adipose - Subcutaneous	491	ADPSBQ	14732	42912
Adipose - Visceral (Omentum)	401	ADPVSC	14640	41720
Adrenal Gland	200	ADRNLG	13622	36754
Artery - Aorta	338	ARTAORT	14396	40474
Artery - Coronary	180	ARTCRN	13878	40579
Artery - Tibial	489	ARTTBL	14493	40690
Brain - Amygdala	119	BRNAMY	12814	24236
Brain - Anterior cingulate cortex (BA24)	135	BRNACC	13528	28806
Brain - Caudate (basal ganglia)	172	BRNCDT	14118	32127
Brain - Cerebellar Hemisphere	157	BRNCHB	13771	39862
Brain - Cerebellum	188	BRNCHA	13992	40747
Brain - Cortex	184	BRNCTXA	14284	35086
Brain - Frontal Cortex (BA9)	158	BRNCTXB	14091	32031
Brain - Hippocampus	150	BRNHPP	13526	27437
Brain - Hypothalamus	157	BRNHPT	13741	30326
Brain - Nucleus accumbens (basal ganglia)	181	BRNNCC	14062	32670
Brain - Putamen (basal ganglia)	153	BRNPTM	13694	28461
Brain - Spinal cord (cervical c-1)	115	BRNSPC	13096	28883
Brain - Substantia nigra	101	BRNSNG	12637	23677
Breast - Mammary Tissue	337	BREAST	14654	44613
Cells - Cultured fibroblasts	417	FIBRBLS	13976	36809
Cells - EBV-transformed lymphocytes	116	LCL	12398	37627
Colon - Sigmoid	274	CLNSGM	14363	41581
Colon - Transverse	306	CLNTRN	14582	41215
Esophagus - Gastroesophageal Junction	281	ESPGEJ	14285	41004
Esophagus - Mucosa	423	ESPMCS	14589	37186
Esophagus - Muscularis	399	ESPMSL	14603	40376
Heart - Atrial Appendage	322	HRTAA	14035	36322
Heart - Left Ventricle	334	HRTLV	13200	29470
Kidney - Cortex	65	KDNCTX	11164	24571
Liver	183	LIVER	12714	27011
Lung	444	LUNG	15058	44346
Minor Salivary Gland	119	SLVRYG	13884	38380
Muscle - Skeletal	602	MSCLSK	13381	31855
Nerve - Tibial	449	NERVET	15373	45478
Ovary	140	OVARY	13738	40857
Pancreas	253	PNCREAS	13695	31203
Pituitary	219	PTTARY	14647	42343
Prostate	186	PRSTTE	14450	41991
Skin - Not Sun Exposed (Suprapubic)	440	SKINNS	14932	42005
Skin - Sun Exposed (Lower leg)	517	SKINS	15204	42219
Small Intestine - Terminal Ileum	144	SNTTRM	14065	39864
Spleen	186	SPLEEN	14073	40290
Stomach	269	STMACH	14102	36624
Testis	277	TESTIS	17867	67784
Thyroid	494	THYROID	15303	45217
Uterus	108	UTERUS	13199	39485
Vagina	122	VAGINA	12969	36931
Whole Blood	573	WHLBLD	12623	24568
total			686241	1816703

Table S3: GWAS loci count for different prioritization method(s). Numbers of loci with associated/colocalized genes/splicing event detected by each method.

GWAS-significant (loci, trait) associations		5385
GWAS-significant unique loci		1167
enloc (loci, trait) colocalizations	expression	2303
enloc (loci, trait) colocalizations	splicing	1223
S-PrediXcan (loci, trait) associations	expression	3756
S-PrediXcan (loci, trait) associations	splicing	3598
S-PrediXcan & enloc (loci, trait) detections	expression	1989
S-PrediXcan & enloc (loci, trait) detections	splicing	1081

Table S4: Proportion of GWAS loci for different prioritizing method(s). proportion of GWAS-significant loci with colocalized/associated genes/splicing events.

method	molecular phenotype	# of loci with significant/colocalized gene/splicing-trait pairs	% of loci with significant/colocalized gene/splicing-trait pairs
enloc	expression	663	57%
enloc	splicing	435	37%
S-PrediXcan	expression	919	79%
S-PrediXcan	splicing	866	74%
S-PrediXcan & enloc	expression	594	51%
S-PrediXcan & enloc	splicing	386	33%

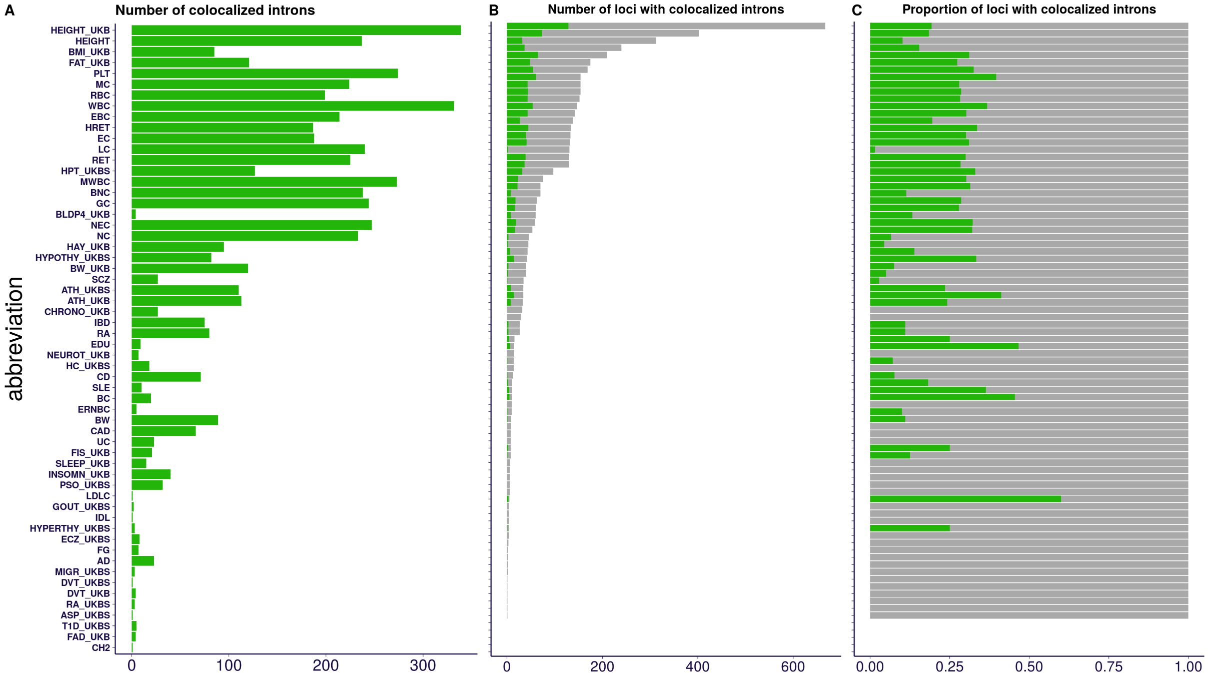


Supplementary Fig. S17. Colocalization of expression QTLs Colocalization for each of the 87 GWAS traits aggregated across the 49 tissues. GWAS loci are shown in gray, colocalized results are shown in dark green. The traits are ordered by number of GWAS-significant variants.

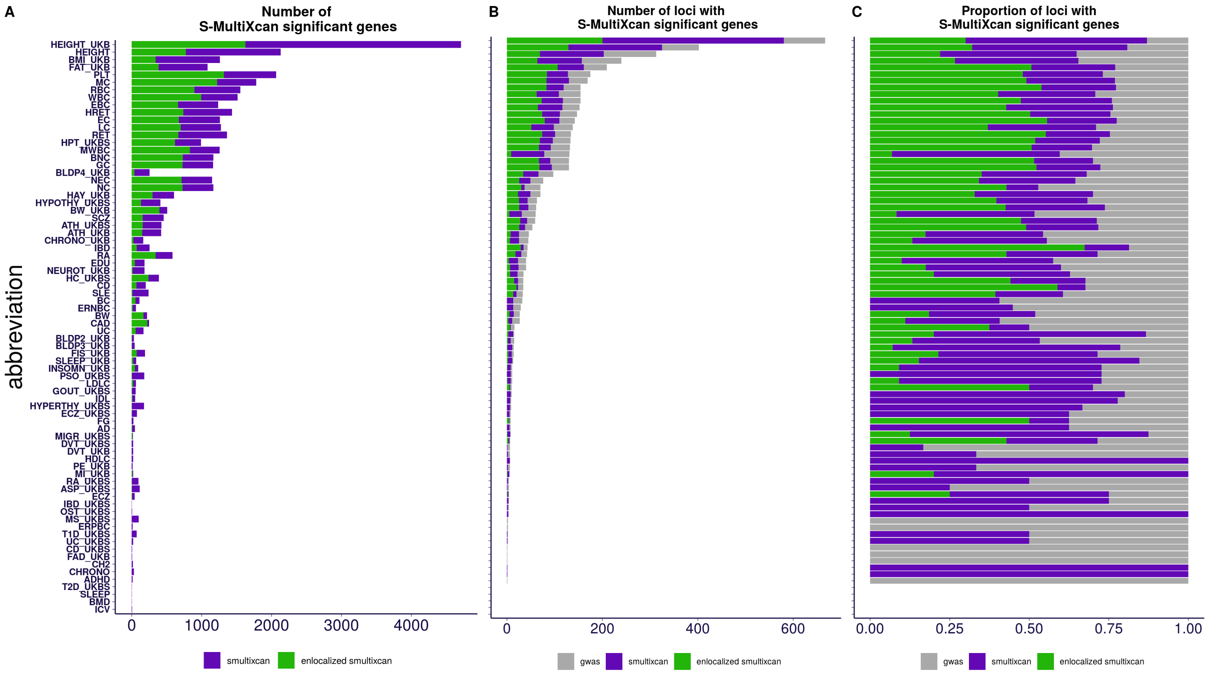
Panel A shows the number of colocalized genes, achieving *enloc rcp* > 0.5 in at least one tissue, for each GWAS trait. The number of colocalized results tends to increase with the number of GWAS-significant variants.

Panel B shows the number of loci (approximately independent LD regions from (*28*)) with at least one GWAS-significant variant (dark gray), and among them those with at least one gene reaching *rcp* > 0.5 (dark green).

Panel C shows the proportion of loci with at least one GWAS-significant hit that contain at least one colocalized gene. Across traits, a median of 21% of the GWAS loci contain colocalized results. See trait abbreviation list in Table S1. This figure is also presented in (*9*).



Supplementary Fig. S18. Colocalization of splicing QTLs for each of the 87 GWAS traits aggregated across the 49 tissues. The traits are ordered by number of GWAS-significant variants. GWAS loci are shown in gray, colocalized results are shown in dark green.
Panel A shows the number of colocalized splicing event, achieving *enloc rcp* > 0.5 in at least one tissue, for each GWAS trait. As with gene expression results, the number of colocalized results tends to increase with the number of GWAS-significant variants.
Panel B shows the number of loci (approximately independent LD regions from (*28*)) with at least one GWAS-significant variant (dark gray), and among them those with one splicing event achieving $rcp > 0.5$ (dark green).
Panel C shows the proportion of loci with at least one GWAS-significant hit loci with at least one colocalized splicing event. Across traits, a median of 11% of the GWAS loci contain a colocalized result, lower than the gene expression counterpart (29%), indicating a decreased power in the sQTL study. See trait abbreviation list in Table S1.



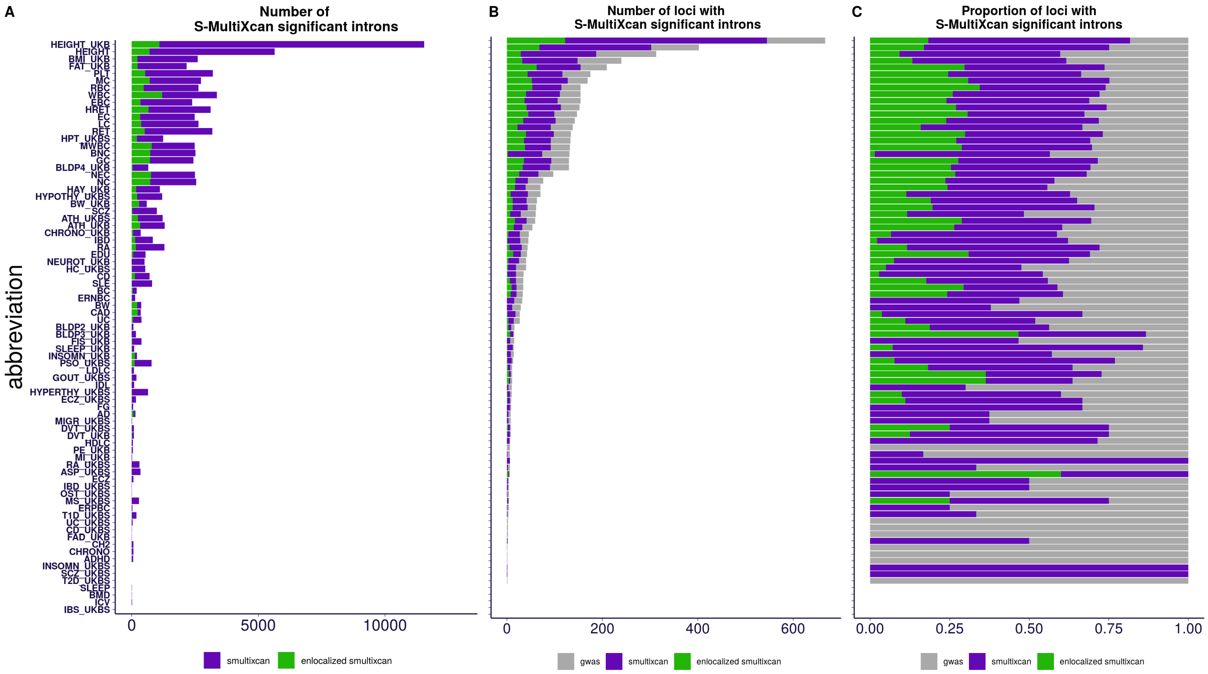
Supplementary Fig. S19. PrediXcan expression associations aggregated across tissues. This figure summarizes S-MultiXcan associations for each of the 87 traits using the gene expression models. The traits are ordered by number of GWAS-significant variants.

Panel A) shows in purple the number of S-MultiXcan significant genes, and in dark green the subset also achieving *enloc rcp* > 0.5 in any tissue. S-MultiXcan has a high power for detecting associations, but 12% (median across traits) of these genes show evidence of colocalization.

Panel B) shows the number of loci (approximately independent LD regions (*28*)) with a significant GWAS association (gray), a significant S-MultiXcan association (purple), and a significant S-MultiXcan association that is colocalized (dark green). Anthropometric and Blood traits tend to present the largest number of associated loci, with Height from two independent studies leading the number of associations.

Panel C) shows the proportion of loci with significant GWAS associations (gray) that contain S-Multixcan (purple) and colocalized S-MultiXcan associations (dark green). Across traits, a median of 70% of GWAS-associated loci show a S-MultiXcan detection, while 19% show a colocalized S-MultiXcan detection.

See trait abbreviation list in Table S1.



Supplementary Fig. S20. PrediXcan splicing associations aggregated across tissues. This figure summarizes S-MultiXcan associations for each of the 87 traits using splicing models. The traits are ordered by number of GWAS-significant variants.
Panel A) shows in purple the number of S-MultiXcan significant splicing events, and in dark green the subset also achieving *enloc rcp* > 0.5 in any tissue. The proportion of colocalized, significantly associated splicing events is typically 2%, much lower than the proportion from gene expression (12%).
Panel B) shows the number of loci (approximately independent LD regions (*28*)) with a significant GWAS association (gray), a significant S-MultiXcan association (purple), and a significant S-MultiXcan association that is colocalized (dark green). As in the case of expression models, Anthropometric and Blood traits tend to present the largest number of associated loci.
Panel C) shows the proportion of loci with significant GWAS associations (gray) that contain S-Multixcan (purple) and colocalized S-MultiXcan associations (dark green). Across traits, a median of 63% of GWAS-associated loci show an S-MultiXcan association, while 11% show a colocalized S-MultiXcan association. These proportions are lower than the corresponding ones for expression (70% and 19% respectively).
See trait abbreviation list in Table S1.

1.11 Regulatory mechanism extends to rare, Mendelian traits

OMIM-based curation

Table S5: 114 GWAS traits used for OMIM-based curation. Keywords of all 114 GWAS traits used for OMIM-based curation and analyses are listed.

Abbreviation	Keyword	Abbreviation	Keyword
Sleep_Duration_UKB	sleep duration	Sum_Eosinophil_Basophil_Count	
Chronotype_UKB	chronotype	Sum_Neutrophil_Eosinophil_Count	
Insomnia_UKB	insomnia	White_Blood_Cell_Count	white blood cell count
Fathers_Age_At_Death_UKB	aging	Coronary_Artery_Disease	coronary heart disease
Deep_Venous_Thrombosis_UKBS	venous thromboembolism	Chronic_Kidney_Disease	chronic kidney
Asthma_UKBS	asthma	Insomnia_In_Both_Sexes	insomnia
Irritable_Bowel_Syndrome_UKBS	irritable bowel	Type_2_Diabetes	type 2 diabetes
Type_1_Diabetes_UKBS	type 1 diabetes	Eczema	atopic dermatitis
Type_2_Diabetes_UKBS	type 2 diabetes	Birth_Length	
Hyperthyroidism_UKBS	hyperthyroidism	BMI_Childhood	bmi;body mass index
Hypothyroidism_UKBS	hypothyroidism	Birth_Weight	
Psychological_Problem_UKBS	psychiatric;psychological	Pubertal_Height_Female	
Multiple_Sclerosis_UKBS	multiple sclerosis	Pubertal_Height_Male	
Parkinsons_Disease_UKBS	Parkinson's	Intracraneal_Volume	intracranial volumn
Migraine_UKBS	migraine	Asthma	asthma
Schizophrenia_UKBS	Schizophrenia	Bone_Mineral_Density	bone mineral density
Osteoporosis_UKBS	osteoporosis	BMI_Active_Inds	bmi;body mass index
Ankylosing_Spondylitis_UKBS	ankylosing spondylitis	BMI_EUR	bmi;body mass index
Eczema_UKBS	eczema;dermatitis	Height	height
Psoriasis_UKBS	psoriasis	Hip_Circumference_EUR	hip circumference
Inflammatory_Bowel_Disease_UKBS	inflammatory bowel disease	Waist_Circumference_EUR	waist circumference
Crohns_Disease_UKBS	crohn's disease	Waist-to-Hip_Ratio_EUR	waist-to-hip
Ulcerative_Colitis_UKBS	ulcerative colitis	HDL_Cholesterol	hdl cholesterol
Rheumatoid_Arthritis_UKBS	rheumatoid arthritis	LDL_Cholesterol	ldl cholesterol
Gout_UKBS	gout	Triglycerides	triglycerides
High_Cholesterol_UKBS	total cholesterol	Neuroticism	neuroticism
Insomnia_UKBS	insomnia	Heart_Rate	heart rate
Fluid_Intelligence_Score_UKB	intelligence	Crohns_Disease	crohn's disease
Birth_Weight_UKB	birth weight	Inflammatory_Bowel_Disease	inflammatory bowel disease
Neuroticism_UKB	neuroticism	Ulcerative_Colitis	ulcerative colitis
BMI_UKB	bmi;body mass index	Alzheimers_Disease	alzheimer
Body_Fat_Percentage_UKB	body fat	Epilepsy	epilepsy
Balding_Pattern_2_UKB		Celiac_Disease	celiac disease
Balding_Pattern_3_UKB		Multiple_Sclerosis	multiple sclerosis
Balding_Pattern_4_UKB		Systemic_Lupus_Erythematosus	systemic lupus erythematosus
Mothers_Age_At_Death_UKB	aging	Stroke	stroke
Standing_Height_UKB	height	Chronotype	chronotype
Heart_Attack_UKB		Sleep_Duration	sleep duration
Deep_Venous_Thrombosis_UKB	venous thromboembolism	Fasting_Glucose	fasting glucose; fasting plasma glucose
Pulmonary_Embolism_UKB		Fasting_Insulin	fasting insulin
Asthma_UKB	asthma	CH2DB_NMR	
Hayfever_UKB		HDL_Cholesterol_NMR	hdl cholesterol
Epilepsy_UKB	epilepsy	Triglycerides_NMR	triglycerides
Migraine_UKB	migraine	LDL_Cholesterol_NMR	ldl cholesterol
Hypertension_UKBS	hypertension	Attention_Deficit_Hyperactivity_Disorder	attention deficit hyperactivity disorder
Adiponectin	adiponectin	Autism_Spectrum_Disorder	autism
Eosinophil_Count	eosinophil count	Schizophrenia	schizophrenia
Granulocyte_Count		Rheumatoid_Arthritis	rheumatoid arthritis
High_Light_Scatter_Reticulocyte_Count		Depressive_Symptoms	depression
Lymphocyte_Count	lymphocyte	Education_Years	education
Monocyte_Count	monocyte count;monocytes	Asthma_TAGC_EUR	asthma
Myeloid_White_Cell_Count		Systolic_Blood_Pressure	systolic blood pressure
Neutrophil_Count	neutrophil count;neutrophils	Diastolic_Blood_Pressure	diastolic blood pressure
Platelet_Count	platelet counts	ER-negative_Breast_Cancer	breast cancer
Red_Blood_Cell_Count	red blood cell count	ER-positive_Breast_Cancer	breast cancer
Reticulocyte_Count		Breast_Cancer	breast cancer
Sum_Basophil_Neutrophil_Count		Smoker	smoking behavior

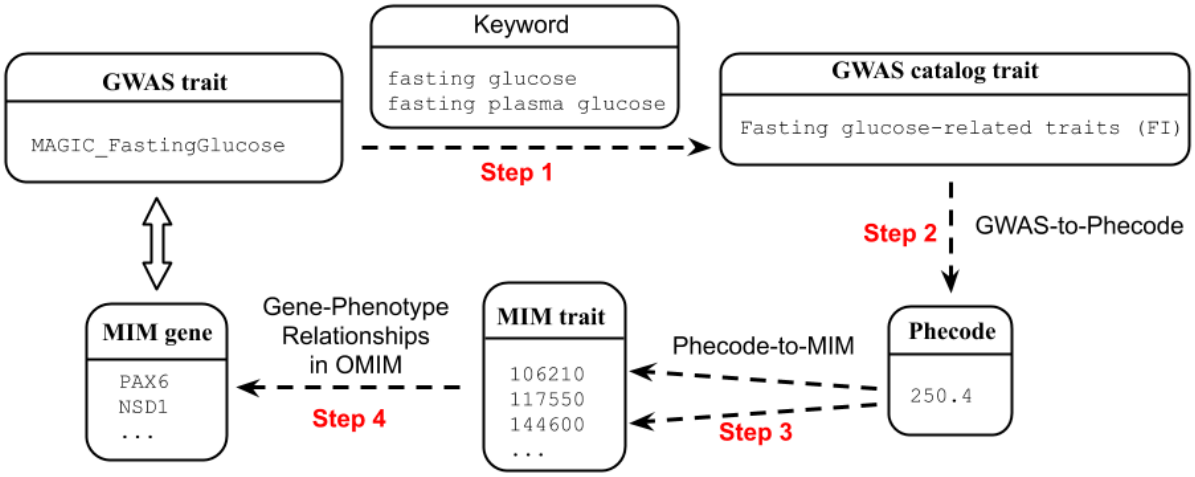
1.12 Curation of causal gene-trait pairs (silver standards)

We curated a list of 1,592 gene-trait pairs with evidence of causal associations in the OMIM database (hereafter, **OMIM genes**).

After matching traits, we retained 29 unique traits and 631 unique genes that are within the same LD block (*28*) as the GWAS hit (table S10). As an additional independent evaluation, we also curated a list of 'silver standard genes', using 101 gene-trait pairs from rare coding variant association studies (*29, 33, 34*) (table S13). The resulting 228 OMIM gene-trait pairs and 80 rare variant harboring gene-trait pairs are used in subsequent validation analyses.

Throughout this section, we limited our scope to only the protein-coding genes. We first describe the approach to define the OMIM-based silver standard. And then we describe the construction of a rare variant set based silver standard.

OMIM-based causal genes



Supplementary Fig. S21. Workflow of OMIM-based curation. The workflow of OMIM-based causal gene curation is shown where each box represents the trait description/identifier in different databases. The steps to obtain OMIM genes for `MAGIC_FastingGlucose`, one of our GWAS traits, is shown as a concrete realization of the workflow.

To obtain a curated set of trait-gene pairs from the OMIM database (*32*), we constructed a map between our GWAS traits and the OMIM traits and then mapped the OMIM traits to genes using the 'Gene-Phenotype Relationships' available in the OMIM database. Specifically, we constructed a keyword for each of the 114 GWAS traits (see TableThen, we matched the keyword to trait description in the GWAS catalog if the keyword occurred in the description sentence (as shown in fig. S21 step 1). The GWAS catalog-to-phecode map (as shown in S21 step 2) was created, using electronic health records (EHR) data (*27*), for a replication study of GWAS findings. The implementation of the map is described in detail in (*27*). Briefly, traits from the catalog (represented as free text) were mapped to the closest corresponding phecode. The phecode/catalog trait relationships were classified as "exact", "narrower" (if the phecode was more specific than the catalog trait), "broader" (if the phecode was more general than the catalog trait), and "proxy" (if the catalog trait was for a continuous measurement). All phecode/catalog trait relationships were included in the analysis.

The clinical descriptions from OMIM have been annotated with Human Phenotype Ontology (HPO) (*56*) terms. We created a map between phecodes and HPO terms used to describe OMIM diseases, as previously described (*57*), which gave rise to the mapping betweem phecodes and OMIM traits (as shown in S21 step 3). By combining these maps, we were able to relate GWAS catalog traits to OMIM disease descriptions, utilizing phecodes and HPO terms as intermediate steps.

For a subset of datasets with discovery (public) and replication (UKB) results in our collection, we kept the dataset with higher number of GWAS loci to avoid double counting. The number of GWAS loci was determined based on counting the lead variants, using the PLINK V1.9 command `-clump-r2 0.2 -clump-p1 5e-8` at genome-wide significance 5×10^{-8}) for each trait. Furthermore, for this analysis, we excluded GWAS traits with fewer than 50 GWAS loci. The full list of OMIM based trait-gene pairs is listed in S10.

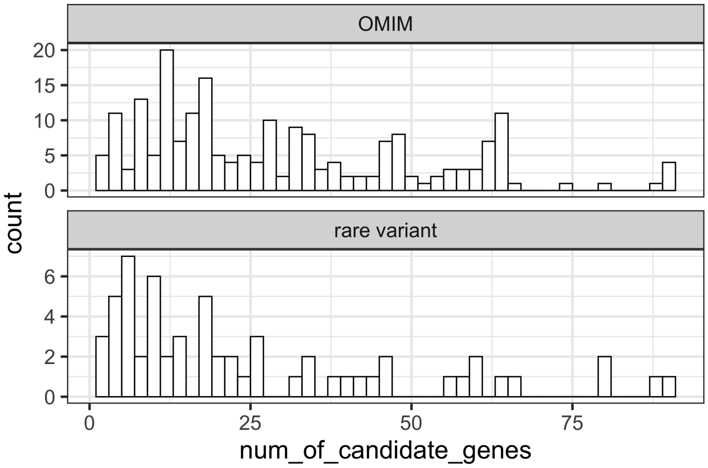
Rare variant association based causal genes

In addition to the OMIM-based curation, we collected a set of genes in which rare protein-coding variants were reported to be significantly associated with our list of complex traits. Here, we focused on rare variant association evidence reported on height and lipid traits (low-density lipid cholesterol, high-density lipid cholesterol, triglycerides, and total cholesterol levels) (*29, 33, 34*). In particular, we collected significant coding/splicing variants reported previously (*29*) and kept variants with effect allele frequency < 0.01 (Table S6: ExomeChip variants with Pdiscovery <2e-07 in the European-ancestry meta-analysis (N=381,625)). Similarly, we collected significant variants reported by (*33*) (Table S12: Association Results for 444 independently associated variants with lipid traits) and filtered out variants with minor allele frequency < 0.01. For the whole-exome sequencing study conducted in Finnish isolates (*34*), we extracted significant genes identified by a gene-based test using protein truncating variants (Table S9: Gene-based associations from aggregate testing with EMMAX SKAT-O with P<3.88E-6) and significant variants (Table S7: A review of all variants that pass unconditional threshold of P<5E-07 for at least one trait) with gnomAD MAF < 0.01. The full list of trait-gene pairs constructed from the process is available in Supplementary Table S13.

Constructing GWAS loci and candidate gene set for silver standard

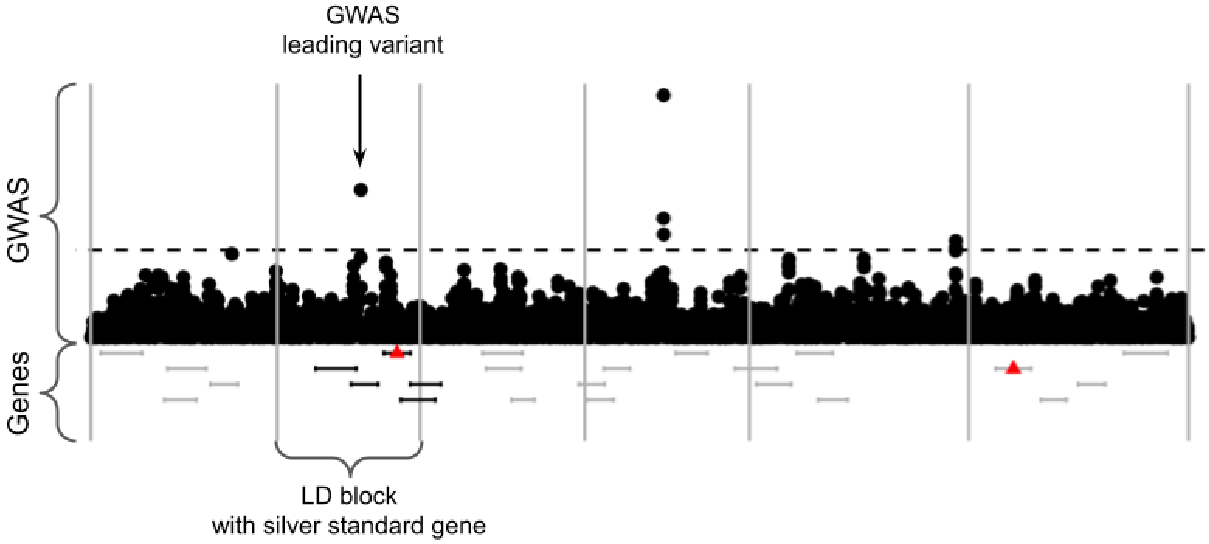
silver standard	trait	nloci	ngene
rare variant	Standing_Height_UKB	29	35
rare variant	LDL_Cholesterol	7	10
rare variant	High_Cholesterol_UKBS	6	8
rare variant	HDL_Cholesterol	12	18
rare variant	Triglycerides	6	9
OMIM	Deep_Venous_Thrombosis	2	2
OMIM	Asthma_UKBS	10	12
OMIM	Type_1_Diabetes_UKBS	1	2
OMIM	Hypothyroidism_UKBS	14	14
OMIM	Eczema_UKBS	4	5
OMIM	Psoriasis_UKBS	2	2
OMIM	Gout_UKBS	1	1
OMIM	High_Cholesterol_UKBS	6	8
OMIM	BMI_UKB	35	35
OMIM	Hypertension_UKBS	19	24
OMIM	Eosinophil_Count	7	7
OMIM	Lymphocyte_Count	2	2
OMIM	Monocyte_Count	1	1
OMIM	Neutrophil_Count	14	17
OMIM	White_Blood_Cell_Count	16	17
OMIM	Coronary_Artery_Disease	12	13
OMIM	Type_2_Diabetes	11	12
OMIM	Waist_Circumference_EUR	6	6
OMIM	LDL_Cholesterol	7	9
OMIM	Triglycerides	11	11
OMIM	Inflammatory_Bowel_Disease	7	8
OMIM	Ulcerative_Colitis	4	4
OMIM	Alzheimers_Disease	2	2
OMIM	Systemic_Lupus_Erythematosus	3	5
OMIM	Schizophrenia	1	1
OMIM	Rheumatoid_Arthritis	3	3
OMIM	Systolic_Blood_Pressure	2	2
OMIM	Diastolic_Blood_Pressure	3	3

Table S6: Count of GWAS loci with predicted causal effects overlapping likely functional genes. The number of GWAS loci and the number of silver standard genes included for analysis after taking the intersection between GWAS loci and silver standard genes are shown.



Supplementary Fig. S22. Distribution of the number of candidate genes per GWAS locus overlapping OMIM- and rare variant-based silver standard. The distributions of the number of candidate genes per GWAS locus are shown for OMIM-based curation (on the right) and rare variant association-based curation (on the left).

1.13 Performance of association and colocalization based methods using silver standards



Supplementary Fig. S23. Workflow on constructing the list of candidate genes for silver standard analysis. We started with GWAS summary statistics and extract LD blocks (boundaries of LD block are shown as gray vertical lines) that contain a variant with GWAS association $p < 5 \times 10^{-8}$. We further filtered out LD blocks which do not overlap with silver gene (labelled with red triangle). The candidate genes (shown as black horizontal bars) are those overlapping with the leftover LD blocks.

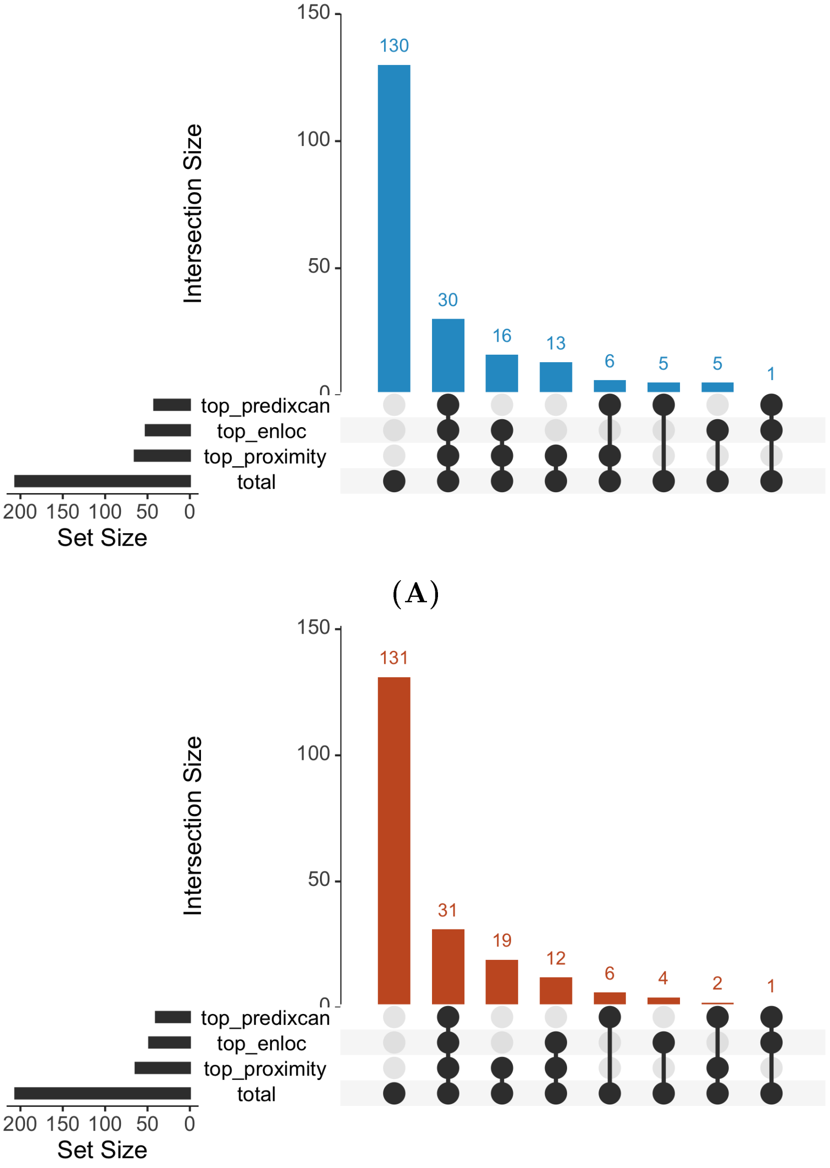
Construction of candidate genes and definition of the scores for various methods

We constructed a set of candidate genes for silver standard analysis following the workflow shown in S23. The rationale of the analysis was to focus on a common use case in practice, namely to prioritize genes within GWAS loci. First, for each trait we defined all LD blocks with genome-wide significant GWAS signals ($p < 5 \times 10^{-8}$) as GWAS loci. Then, we selected the variant with the highest significance as the GWAS lead variant for the locus (and randomly picked one in case of ties). Here we limited our scope to those GWAS loci containing silver standard genes since the current silver standard did not have enough information to test the rest, which were driven by genes without indication in OMIM database and/or rare variant associations, and potentially had smaller effect size as compared to the former. Thus, we kept only the GWAS loci overlapping with silver standard genes. The full list of silver standard genes, after the filtering procedure, can be found in S14 and S15. The number of GWAS loci and silver standard genes that remained after the above filtering steps can be found in Supplementary Table S6. From these GWAS loci containing silver standard genes, we extracted all genes overlapping with the loci as the set of candidate genes (the number of candidate genes per locus is shown in fig. S22).

For each of these candidate genes, we obtained the gene-level statistics for the corresponding traits from the application of various methods, *i.e.* *enloc*, *coloc*, SMR, and PrediXcan-*mashr* where we collapsed statistics across tissues by taking the 'best' scores (highest regional colocalization probability (rcp) in *enloc*; highest posterior probability under hypothesis 4 in *coloc*; smallest p-value in SMR and PrediXcan-*mashr*).

Regarding the results on splicing (with statistics reported at the intron excision event level), we obtained gene-level statistics by taking the 'best' score among all splicing events of the gene. We also summarized the cis-sQTL at the gene level using the same strategy.

Per locus prioritization



(A)

(B)

Supplementary Fig. S24. The number of OMIM genes ranked top within a GWAS locus by proximity, *enloc*, and PrediXcan. Results from expression are in **(A)** and those from splicing are in **(B)**.

Regression-based test on the per-locus rank To investigate the usefulness of the colocalization and association statistics reported by *enloc* and PrediXcan respectively, we performed logistic regression, as described in , to fit log odds of being a 'causal' gene against the ranking of: 1) proximity to GWAS lead variant (from close to distal), 2) rcp from *enloc* (from high to low), and 3) gene-level association p-value from PrediXcan-*mashr* or SMR (from significant to non-significant).

$$\text{logit}(\Pr(\text{causal}_i)) = \beta_0 + \beta_1 \cdot \text{rank}(\text{proximity}_i) + \beta_2 \cdot \text{rank}(\text{rcp}_i) + \beta_3 \cdot \text{rank}(\text{P-value}_i), \tag{34}$$

in which non-zero β_k meant that the kth variable contributed independently on predicting whether a gene was causal. Moreover, negative β_k indicated that the direction of contribution of the variable was as expected.

Regression-based test to investigate the independent contribution of proximity, colocalization, and association based methods

regulation	silver_standard	variable	coefficient	coefficient_se	pvalue
expression	OMIM	rank_proximity	-0.018	0.0081	0.03
expression	OMIM	predixcan_mashr_eur	-0.038	0.008	2.2×10^{-6}
expression	OMIM	enloc	-0.02	0.0093	0.031
splicing	OMIM	rank_proximity	-0.026	0.0073	0.00031
splicing	OMIM	predixcan_mashr_eur	-0.037	0.008	3.5×10^{-6}
splicing	OMIM	enloc	-0.012	0.0086	0.17
expression	rare variant	rank_proximity	-0.013	0.018	0.46
expression	rare variant	predixcan_mashr_eur	-0.043	0.016	0.0084
expression	rare variant	enloc	-0.043	0.02	0.032
splicing	rare variant	rank_proximity	-0.048	0.015	0.0015
splicing	rare variant	predixcan_mashr_eur	-0.018	0.013	0.15
splicing	rare variant	enloc	-0.02	0.015	0.2

Table S7: Predictive value of different per-locus prioritization methods. Results on regression-based test (logistic regression) in per-locus analysis are shown. The estimated log odds ratio of the rank of proximity (distance between GWAS leading variant and gene body), PrediXcan significance, and *enloc* rcp are shown in rows **rank_proximity**, **predixcan_mashr_eur**, and **enloc**.

Precision-recall and receiver operating characteristic curve

In addition to the per-locus analysis, we combined the gene-trait pairs from the per-locus analysis. We labelled the ones with silver standard genes for the corresponding trait as 1 and the rest as 0 so that we could plot PR and ROC curve for each association/colocalization based method by varying a universal threshold (i.e., either p-value or colocalization probability) across all analyzed traits and loci. These curves provide a measure of the predictive power of each method, but use of a universal cutoff across all traits and GWAS loci has limitations (see discussion in Section 1.13).

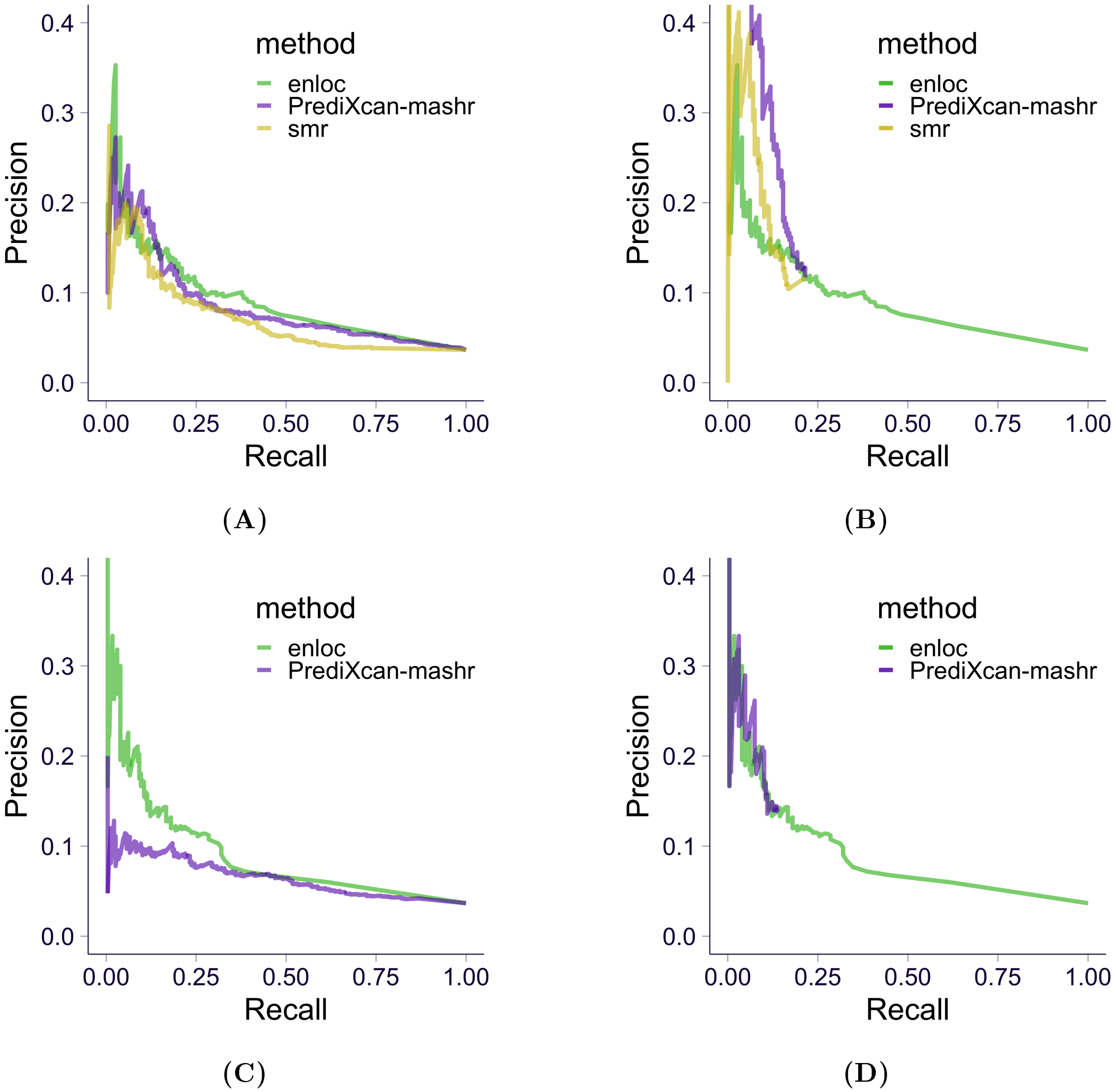
Enrichment and ROC curves

Regulation	Dataset	Method	ROC AUC	Enrichment
expression	OMIM	*coloc*	0.553	
expression	OMIM	*enloc*	0.669	4.56
expression	OMIM	PrediXcan	0.672	2.50
expression	OMIM	SMR	0.591	
expression	Rare variant	*coloc*	0.661	
expression	Rare variant	*enloc*	0.755	14.72
expression	Rare variant	PrediXcan	0.743	2.21
expression	Rare variant	SMR	0.629	
splicing	OMIM	*enloc*	0.650	6.10
splicing	OMIM	PrediXcan	0.632	2.54
splicing	Rare variant	*enloc*	0.714	21.76
splicing	Rare variant	PrediXcan	0.686	2.19

Table S8: Enrichment and AUC fo *coloc*, *enloc*, SMR, and PrediXcan

For expression, the areas under the curve (AUC) of were, in increasing performance, 0.553, 0.591, 0.669, and 0.672 for *coloc*, SMR, *enloc*, and PrediXcan using the OMIM silver standard 3C. AUC were higher when using the rare variant silver standard with SMR at the bottom of the ranking followed by *coloc*, PrediXcan, and enloc at the top S8. For splicing *enloc* had higher 0.650 vs. 0.632 for PrediXcan using OMIM silver standard and 0.714 and 0.686 using the rare variant silver standard.

Precision-recall curves of PrediXcan and *enloc* on silver standard gene sets



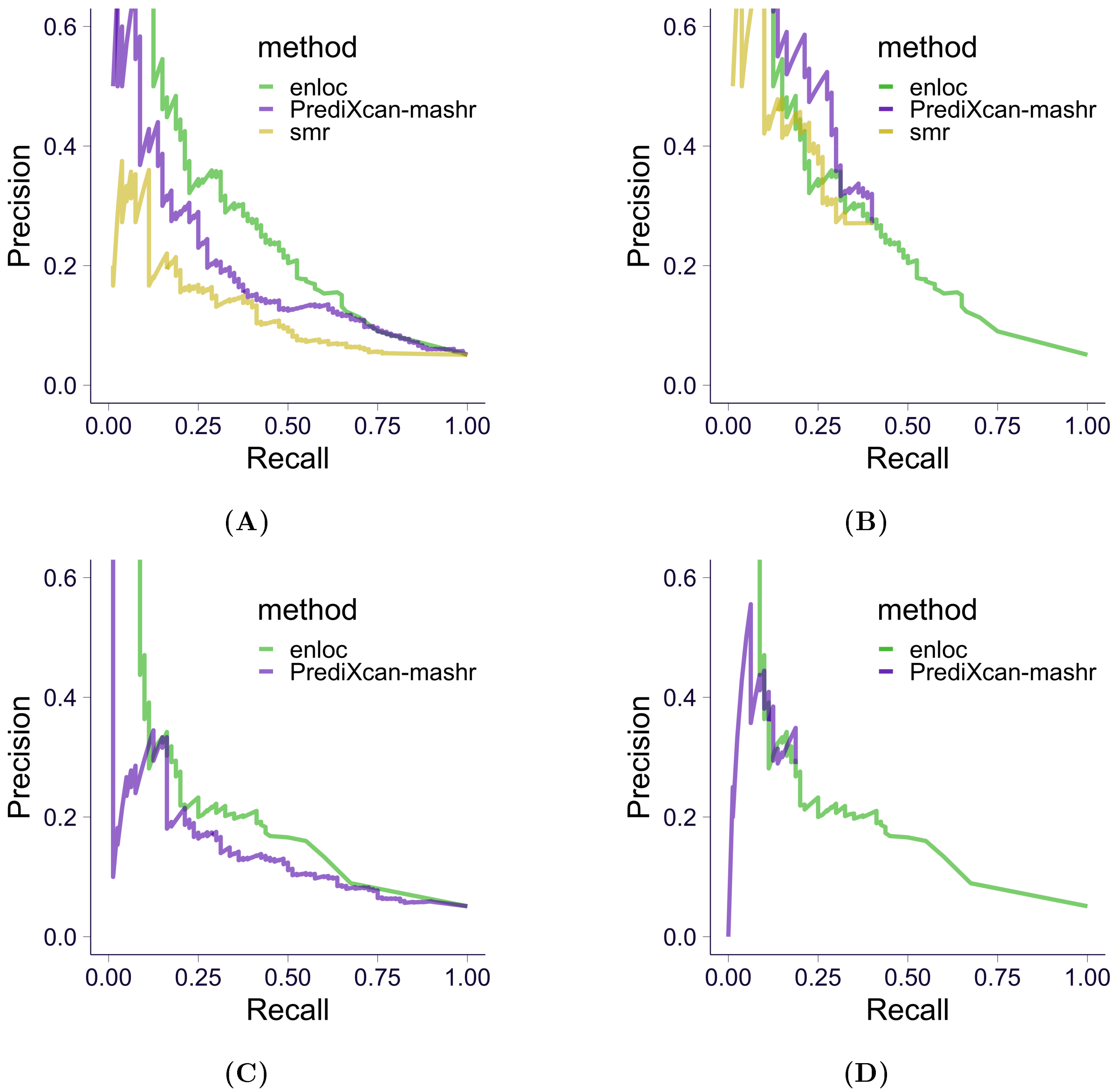
(A)

(B)

(C)

(D)

Supplementary Fig. S25. Precision-recall curves of colocalization/association based methods on OMIM silver standard. The results on expression data are shown in top row and the ones on splicing data are shown in bottom row. **(A,C)** Precision-recall curve of colocalization/association based methods. **(B,D)** Precision-recall curve of association based methods when pre-filtering with *enloc* rcp > 0.1.



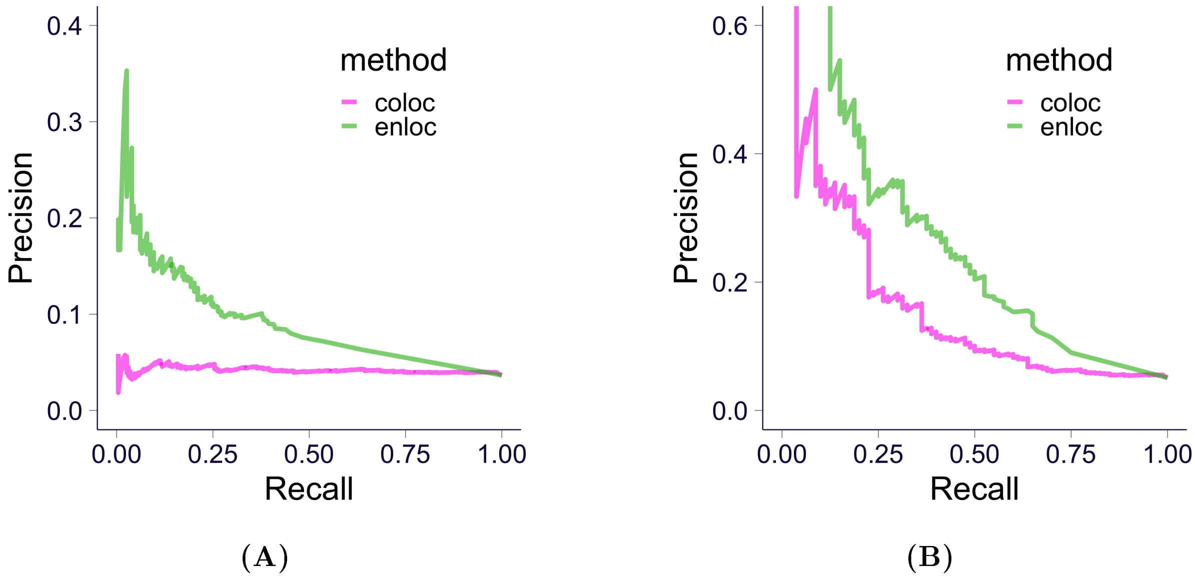
(A)

(B)

(C)

(D)

Supplementary Fig. S26. Precision-recall curves of colocalization/association based methods on rare variant-based silver standard. The results on expression data are shown in top row and the ones on splicing data are shown in bottom row. **(A,C)** Precision-recall curve of colocalization/association based methods. **(B,D)** Precision-recall curve of association based methods when pre-filtering with *enloc* rcp > 0.1.



Supplementary Fig. S27. Precision-recal curves of colocalization methods. Precision-recall curves of two colocalization methods (using expression data): ENLOC (blue) and COLOC (green) using OMIM silver standard (in **(A)**) and rare variant-based silver standard (in **(B)**).

Discussion on the limitation of applying universal cutoff in precision-recall/ROC curve

To apply a universal threshold to all loci and traits was a limitation of this approach. On the one hand, different GWAS traits may have different sample sizes so that the colocalization probabilities or association p-values are not comparable across traits. On the other hand, for the same trait, the magnitude of the mediated effect size (gene/splicing-level effect) at different loci may vary, which makes the colocalization probability or association p-value less comparable across loci. With these limitations, the comparison between methods was still informative but might favor the one that suffers less from the lack of comparability and being more stringent (*i.e. enloc*). Furthermore, the curves were not directly comparable to per-locus approaches since a per-locus approach intrinsically utilized the information that there was only one signal per locus. However, these curves provide insights into how these methods would perform (in terms of precision/power trade-off) if we applied a universal cutoff across loci, which is a typical use case in practice.

Unique causal gene assumption with silver standard In many practical applications, the investigator will be interested in identifying the causal gene that drives the signal at a given GWAS locus. Here we assume existence and uniqueness, i.e., that the causal gene is on the list of neighboring genes and that there is only one. These assumptions may fail, for example, if regulatory effects are unrelated to the causal mechanism or the causal molecular phenotypes have not been assayed or discovered. Furthermore, the assumption that there is only one causal gene is parsimonious and arguably reasonable, but in general we do not have hard evidence to rule out multiple genes contributing to the trait effect.

1.14 Predicted associations replicated in BioVU (cont.)

Among replicated loci are *SORT1* (liver, coronary artery disease rcp = 0.952; dicovery p = 2.041×10^{-19} BioVU p = 3.475×10^{-4}), which has a well-established associations to lipid metabolism and cardiovascular traits (*58*). Chromosome 6p24 region, which contains *PHACTR1*, has been previously associated with a constellation of vascular diseases, including coronary artery disease (*59*) and migraine headache (*60*). Notably, *PHACTR1* was significant in three different arteries (aorta artery, coronary artery and tibial artery) in two traits (coronary artery disease and migraine) in the replication analysis. In all six tissue-trait pairs, *PHACTR1* showed very high posterior probabilities in discovery analyses (rcp = 0.992 to 1.00). In our replication analysis, *PHACTR1* remained significant only for coronary artery disease associations (table S16, aorta artery, discovery p = 2.246×10^{-39}, BioVU p = 7.484×10^{-8}; coronary artery, discovery p = 1.952×10^{-37}, BioVU p = 2.047×10^{-7}; tibial artery, discovery p = 1.559×10^{-33}, BioVU p = 9.880×10^{-7}).

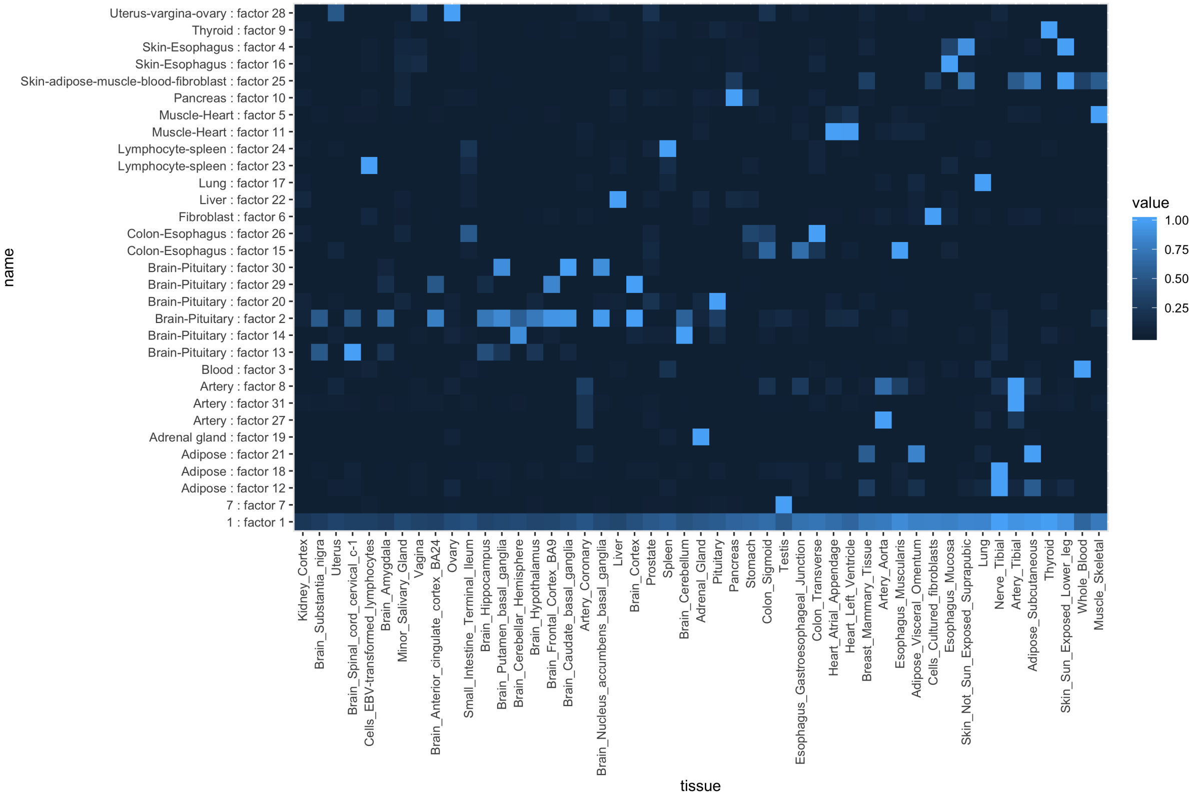
1.15 Validation of likely causal genes (cont.)

Of note, two members of the sterolin family, *ABCG5* and *ABCG8*, showed highly significant predicted causal associations using both PrediXcan and *enloc* for LDL-C levels and self-reported high cholesterol levels. *ABCG8* showed more significant associations in both datasets (chr2: 43838964 - 43878466; UKB self-reported high cholesterol: -log10($p_{\text{PrediXcan}}$) = 38.43, rcp = 0.985; GLGC LDL-C: -log10($p_{\text{PrediXcan}}$) = 71.40, rcp = 0.789), compared to *ABCG5* (chr2: 43812472 - 43838865; -log10($p_{\text{PrediXcan}}$) = 36.85, rcp = 0.941; -log10($p_{\text{PrediXcan}}$) = 80.80, rcp = 0.705). Mutations in either of the two ATP-binding cassette (ABC) half-transporters, *ABCG5* and *ABCG8*, lead to reduced secretion of sterols into bile, and ultimately, obstruct cholesterol and other sterols exiting the body (*61*). In mice with disrupted *Abcg5* and *Abcg8* (G5G8-/-), a 2- to 3-fold increase in the fractional absoprtion of dietary plan steols and extrememly low biliary cholesterol levels was observed, indicating that disrupting these genes contribute greatly to plasma cholesterol levels (*62*). The overexpression of human *ABCG5* and *ABCG8* in transgenic Ldlr-/- mice resulted in 30% reduction in hepatic cholesterol levels and 70% reduced atherosclerotic legion in the aortic root and arch (*63*) after 6-months on a Western diet.

Several other lipid-associated loci were also consistently predicted as causal across OMIM, the rare variant derived set, PrediXcan and *enloc*. Rare protein-truncating variants in *APOB* have been previously associated with reduced LDL-C and triglyceride levels and reduced coronary heart disease risk (*64*). Interestingly, *APOB* has been predicted as a causal gene in four related traits, coronary artery disease, LDL-C levels, triglyceride levels, and self-reported high cholesterol levels. Among the four traits, PrediXcan showed the highest association to LDL-C levels (-log10($p_{\text{PrediXcan}}$) = 130.89; rcp = 0.485) while self-reported high cholesterol showed the strongest evidence using *enloc* at nearly maximum posterior probability (-log10($p_{\text{PrediXcan}}$) = 93.66; rcp = 0.969). Although *APOB* has been suggested as a better molecular indicator of predicted cardiac events in place of LDL-C levels (*65, 66*), its translation has been surprisingly slow in clinical practice (*67*). Here, we provide an additional support for the crucial role *APOB* may play in predicting lipid traits.

1.16 Causal tissue analysis

We investigated the cross-tissue pattern of PrediXcan results across 49 tissues. For each trait-gene pair, the PrediXcan z-score can be represented as a 49×1 vector with each entry being the gene-level z-score in the corresponding tissue (if the prediction model of the gene is not available in that tissue, we filled in zero). To explore the tissue-specificity of the PrediXcan z-score vector, we proceeded by assigning the z-score vector to a tissue-pattern category and tested whether certain tissue-pattern categories were over-represented among colocalized PrediXcan genes as compared to non-colocalized genes. In particular, we used the FLASH factors identified from matrix factorization applied to the cis-eQTL effect size matrix, as described in Section 1.7 (as PrediXcan and cis-eQTL shared similar tissue-sharing pattern, data not shown). To obtain a set of detailed and biologically interpretable tissue-pattern categories from the 31 FLASH factors, we manually merged them into 18 categories as shown in fig. S28. For each trait, we projected the z-score vector of each gene to one of the 31 FLASH factors (as described in Section 1.7) so that the gene was assigned to the corresponding tissue-pattern category. We defined a 'positive' set of genes as the ones that met Bonferroni significance at $\alpha = 0.05$ in at least one tissue and *enloc* rcp $>$ 0.01 in at least one tissue, which could be thought as a set of candidate genes affecting the trait through expression level. We also constructed a 'negative' set of genes with *enloc* rcp $= 0$, which could be thought as a set of genes whose expressions were unlikely to affect the trait. We proceeded to test whether certain tissue-pattern categories were enriched in 'positive' set as compared to 'negative' set. Since the main focus of this analysis was tissue-specific patterns, we excluded *Factor1* (the cross-tissue factor) and *Factor25* (likely to be a tissue-shared factor capturing tissues with large sample size). Additionally, we excluded *Factor7* (testis), as it was unlikely to be the mediating tissue but might introduce false positives. We tested the enrichment of each tissue-pattern category by Fisher's exact test ('positive'/'negative' sets and in/not in tissue-patter category). Among 87 traits, 82 traits had *enloc* signal and the enrichment of these was calculated accordingly.



Supplementary Fig. S28. Factor analysis using flashr to identify causal tissues. Tissue-pattern categories generated from from FLASH applied to the cis-eQTLs are shown. These tissue categories (on y-axis) were the same ones used in the analysis of causal tissue identification. Tissues are ordered by sample size.

Table S9: GWAS Metadata contains relevant information concerning each GWAS study used. Full table available in Supplementary Material. Analyses used the 87 traits with deflation=0 unless explictly said otherwise. Columns are: **Tag**: Internal name to identify the study, **Deflation**: Deflation status after imputation (0 for no deflation, 1 for moderate deflation, 2 for extreme deflation), **PUBMED_Paper_Link**: PUBMED entry, **Pheno_File**: name of downloaded file, **Source_File**: actual name of GWAS summary statistics (i.e. downloaded files might contain several traits), **Portal**: URL to GWAS study portal, **Consortium**: Name of Consortium if any, **Link**: download link for the file, **Notes**: any special comment on the GWAS trait, **Header**: GWAS summary statistics header in case the file is malformed, **EFO**: Experimental Factor Ontology (*68*) entry if applicable, **HPO**: Human Phenotype Ontology (*56*) entry if applicable, **Description**: optional description of the study, **Trait**: trait name, **Sample_Size**: number of individuals included in the study, **Population**: types of populations present (EUR for European, AFR for African, EAS for East Asian, etc), **Date**: Date the file was downloaded, **Declared_Effect_Allele**: column specifying effect allele, **Genome_Reference**: Human Genome release used as reference (i.e. hg19, hg38), **Binary**: wether the trait is dichotomous, **Cases**: number of cases if binary trait, **abbreviation**: short string for figure and table display, **new_abbreviation**: additional abbreviation, **new_Trait**: additional trait name, **Category**: type of trait, **Color**: Hexadecimal color code for display

Supplementary tables in spreadsheet

Table S10: Presumed causal genes included in the OMIM database. Columns are: **trait**: Tag used for the trait, **pheno_mim**: MIM ID of the phenotype mapped to GWAS trait, **mim**: MIM ID of the corresponding gene, **entry_type**: Entry type in the OMIM database, **entrez_gene_id**: Gene ID based on Entrez database, **gene_name**: Official gene symbol, **ensembl_gene_id**: Gene ID based on Ensembl database, **gene_type**: Gene type based on Gencode, **gene**: Trimmed Gene ID based on Ensembl database.

Table S11: PrediXcan and enloc results for predicted causal genes selected based on OMIM. Columns are: **lead_var**: the most significant variant within the LD block, **trait**: trait name, **gene**: Ensembl ID for the gene, **is_omim**: Is included in the OMIM database. TRUE if included, FALSE if not, **proximity**: 0 if variant is in the gene, otherwise BPS from the gene boundary, **rank_proximity**: ranking by proximity within LD block (rank starts from 0 and the closer the lower rank), **percentage_proximity**: rank_proximity / number of genes in the locus, **predixcan_mashr_eur_score**: -log10 p-value (most significant across tissues is used) of PrediXcan-MASH trained on European data, **enloc_score**: rcp (max across tissues), **predixcan_mashr_eur_rank**: PrediXcan significance ranking within LD block (rank starts from 0 and the higher significance the lower rank), **enloc_rank**: enloc rcp ranking within LD block (rank starts from 0 and the higher rcp the lower rank), **predixcan_mashr_eur_percentage**: predixcan_mashr_eur_rank / number of genes in the locus, **enloc_percentage**: enloc_rank / number of genes in the locus, **gene_name**: Official gene symbol, **gene_type**: Gencode annotsted gene type, **chromosome**: Chromosome for the gene, **start**: Gencode annotated gene start position. All isoforms are combined, **end:** Gencode annotated gene end position. All isoforms are combined, **strand**: Gencode annotated gene strand.

Table S12: PrediXcan and enloc results for presumed causal genes in the rare variant based silver standard. Columns are: **lead_var**: the most significant variant within the LD block, **trait**: trait name, **gene**: Ensembl ID for the gene, **is_ewas**: Is included in the EWAS . TRUE if included, FALSE if not, **proximity**: 0 if variant is in the gene, otherwise BPS from the gene boundary, **rank_proximity**: ranking by proximity within LD block (rank starts from 0 and the closer the lower rank), **percentage_proximity**: rank_proximity / number of genes in the locus, **predixcan_mashr_score**: -log10 p-value (most significant across tissues is used) of PrediXcan-MASH trained on European data, **enloc_score**: rcp (max across tissues), **predixcan_mashr_rank**: PrediXcan significance ranking within LD block (rank starts from 0 and the higher significance the lower rank), **enloc_rank**: enloc rcp ranking within LD block (rank starts from 0 and the higher rcp the lower rank), **predixcan_mashr_percentage**: predixcan_mashr_eur_rank / number of genes in the locus, **enloc_percentage**: enloc_rank / number of genes in the locus, **gene_name**: Official gene symbol, **gene_type**: Gencode annotsted gene type, **chromosome**: Chromosome for the gene, **start**: Gencode annotated gene start position. All isoforms are combined, **end**: Gencode annotated gene end position. All isoforms are combined, **strand**: Gencode annotated gene strand.

Table S13: Genes suggested as causal by rare variant association studies. Columns are: **gene**: Trimmed gene ID based on Ensembl database, **nobs**: Number of times gene has been observed in the trait, **trait**: Tag for the trait name.

Table S14: OMIM genes included in the analysis. Columns are: **gene**, **trait**.

Table S15: Rare variant silver standard genes included in the analysis. Columns are: **gene**, **trait**.

Table S16: BioVU. Columns are: **gene**, **tissue**, **trait_map**: mapped trait, **pheno**: trait, **gene_name**, **p_discovery**, **rcp_discovery**, **beta_biovu**, **p_biovu**, **z_biovu**.

List of Figures

List of Tables