Title: **The missing link between genetic association and regulatory function**

Noah Connally[1,2,3], Sumaiya Nazeen[1,2,4], Daniel Lee[1,2,3], Huwenbo Shi[3,5], John Stamatoyannopoulos[6,7], Sung Chun[8], Chris Cotsapas[3,9,10]*, Christopher Cassa[1,3]*, Shamil Sunyaev[1,2,3]*

[1] Brigham and Women's Hospital, Division of Genetics, Harvard Medical School, Boston, MA, USA

[2] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[3] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[4] Brigham and Women's Hospital, Department of Neurology, Harvard Medical School, Boston, MA, USA

[5] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[6] Altius Institute for Biomedical Sciences, Seattle, WA, USA

[7] Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA

[8] Division of Pulmonary Medicine, Boston Children's Hospital, Boston, MA, USA

[9] Department of Neurology, Yale Medical School, New Haven, CT, USA

[10] Department of Genetics, Yale Medical School, New Haven, CT, USA

* Co-corresponding authors (C. Cotsapas: cotsapas@broadinstitute.org; C. Cassa: ccassa@bwh.harvard.edu; S. Sunyaev: ssunyaev@rics.bwh.harvard.edu)

The genetic basis of most complex traits is highly polygenic and dominated by non-coding alleles, and it is widely assumed that such alleles exert small regulatory effects on the expression of *cis*-linked genes. However, despite availability of expansive gene expression and epigenomic data sets, few variant-to-gene links have emerged. We identified 139 genes in which protein-coding variants cause severe or familial forms of nine human traits. We then computed the association between common complex forms of the same traits and non-coding variation, revealing that most such traits are also associated with non-coding variation in the vicinity of the same genes. However, we found colocalization evidence—the same variant influencing both the physiological trait and gene expression—for only 7% of genes, and transcriptome-wide association evidence with correct direction of effect for only 4% of genes, despite an abundance of eQTLs in most loci. Fine mapping variants to regulatory elements and assigning these to genes by linear distance similarly failed to implicate most genes in complex traits. These results contradict the hypothesis that most complex trait-associated variants coincide with currently ascertained expression quantitative trait loci. The field must confront this deficit, and pursue the "missing regulation."

Modern complex trait genetics has uncovered surprises at every turn, including the paucity of associations between traits and coding variants of large effect, and the "mystery of missing heritability," where no combination of common and rare variants can explain a large fraction of trait heritability[1]. Further work has revealed unexpectedly high

polygenicity for most human traits and very small effect sizes for individual variants. Bulk enrichment analyses have demonstrated that a large fraction of heritability resides in regions with gene regulatory potential, predominantly tissue-specific accessible chromatin and enhancer elements, suggesting that trait-associated variants influence gene regulation[2–4]. Furthermore, genes in trait-associated loci are more likely to have genetic effects on their expression levels (expression QTLs, or eQTLs), and the variants with the strongest trait associations are more likely also to be associated with transcript abundance of at least one proximal gene[5]. Combined, these observations have led to the inference that most trait-associated variants are eQTLs, exerting their effect on phenotype by altering transcript abundance, rather than protein sequence. The mechanism may involve a knock-on effect on gene regulation, with the variant altering transcript abundances for genes elsewhere in the genome (a *trans*-eQTL), but the consensus view is that this must be mediated by the variant influencing a gene in the region (a *cis*-eQTL)[6]. As most eQTL studies profile cell populations or tissues from healthy donors at homeostatic equilibrium, the further assumption has been tacitly made that these trait-associated variants affect genes in *cis* under resting conditions. Equivalent QTL analyses of exon usage data have revealed a more modest overlap with trait-associated alleles, suggesting that a fraction of trait-associated variants influence splicing, and hence the relative abundance of different transcript isoforms, rather than overall expression levels. Thus, a model has emerged where most trait-associated variants influence proximal gene regulation.

Several observations have challenged this basic model. One challenge comes from the difference between spatial distributions of eQTLs, which are dramatically enriched in close proximity of genes, and GWAS peaks, which are usually distal[7]. Another comes from colocalization analyses, attempting to map shared genetic associations between human traits and gene expression. If the model is correct, most trait associations should also be eQTLs; trait and expression phenotype should thus share an association in that locus (rather than two association peaks overlapping). However, only 5-40% of trait associations co-localize with eQTLs in relevant tissues or cell types[6,8–10], and only 15% of genes colocalize with any of 74 different complex traits[11]. Finally, expression levels mediated a minority of complex trait heritability[12]. This has led to the suggestion that most trait-associated alleles influence gene regulation in a context-specific manner[13]—either altering expression during development or in response to specific physiological stimuli—or that they act indirectly in *trans* to affect the regulation of a small number of genes involved in trait biology (the omnigenic model[14,15]). Without a set of true positive cases, in which the gene driving trait variation is known, it remains difficult to assess either the basic model or the proposed variations.

One source of true positives is to identify genes that are both in loci associated with a complex trait and are also known to harbor coding mutations causing severe or early onset forms of related traits (e.g. related Mendelian disorders). The strong expectation is that a variant of small effect influences the gene identified in the severe form of the trait. This expectation is supported by several lines of evidence. Comorbidity between Mendelian and complex traits has been used to identify common variants associated

with the complex traits[16]. A handful of genes have been conclusively identified in both Mendelian and complex forms of the same trait, including *APOE*, which is involved in cholesterol metabolism[17,18], and *SNCA*, which contributes to Parkinson's disease risk. Early genome-wide association studies (GWAS) found associations near genes identified through familial studies of severe disease[19,20], and more recent analyses have found that GWAS associations are enriched in regions near causative genes for cognate Mendelian traits in both blood traits[8] and a diverse collection of 62 traits[21].

To test the model that trait-associated variants influence baseline gene expression, therefore, we assembled a list of such "putatively causative" genes. We selected nine polygenic common traits with available large-scale GWAS data, each of which also has an extreme form in which coding mutations of large effect size affect one or more genes with well-characterized biology (Table 1). Our selection included four common diseases: type II diabetes[22], where early onset familial forms are caused by rare coding mutations (insulin-independent MODY; neonatal diabetes; maternally inherited diabetes and deafness; familial partial lipodystrophy); ulcerative colitis and Crohn disease[23,24], which have Mendelian pediatric forms characterized by severity of presentation; and breast cancer[25], where coding mutations in the germline (e.g. *BRCA1*) or somatic tissue (e.g. *PIK3CA*) are sufficient for disease. We also chose five quantitative traits: low and high density lipoprotein levels (LDL and HDL); systolic and diastolic blood pressure; and height. We selected 139 genes harboring large-effect-size coding variants for one of the nine phenotypes (Table 1). These genes were identified in familial studies, and, for breast cancer, using the MutPanning method[26].

We first examined whether these genes are more likely than chance to be in close proximity harboring variants associated with the polygenic form of each trait. In agreement with existing literature[21], we observe a highly significant enrichment. However, in well-powered GWAS, even relatively rare large-effect coding alleles (mutations in *BRCA1* which cause breast cancer, for instance) may be detectable as an association to common variants. To account for this possibility, we computed association statistics in each GWAS locus conditional on coding variants. We applied a direct conditional test to datasets with available individual-level genotype data; for those studies without available genotype data, we computed conditional associations from summary statistics using COJO[27]. After controlling for coding variation, we still detected a highly significant enrichment of our genes under GWAS peaks. Of our 139 genes, 89 (64%) fell within 1 Mb of a GWAS locus for the cognate complex trait. After fine-mapping the GWAS associations in each locus using the SuSiE algorithm[28], we found that 23/139 (17%) putative causal genes are closer to the GWAS fine-mapped SNPs (posterior inclusion probability > 0.7) than any other gene in the locus, as measured from the transcription start site. Given their known causal roles in the severe forms of each phenotype, we thus suggest that the 89 genes near GWAS signals are likely to be the targets of trait-associated non-coding variants. For example, we see a significant GWAS association between breast cancer risk and variants in the estrogen receptor (*ESR1*) locus even after controlling for coding variation; the baseline expression model would thus predict that non-coding risk alleles alter *ESR1* expression to drive breast cancer risk.

We next looked for evidence that the trait-associated variants were also altering the expression of our 89 genes in relevant tissues. If these variants act through changes in gene expression, phenotypic associations should be driven by the same variants as eQTLs in relevant tissue types. We therefore looked for co-localization between our GWAS signals and eQTLs in relevant tissues (Supplementary table 1) drawn from the GTEx Project, using three well-documented methods: coloc[10], JLIM[9], and eCAVIAR[29]. We found support for the colocalization of trait and eQTL association for only four (coloc), six (JLIM), and three (eCAVIAR) of our 89 putatively causative genes, even before correcting for multiple-hypothesis testing, which is not obviously better than random chance. We note that our estimates of the number of putatively causative *genes* with colocalization of eQTL and GWAS signal is conceptually distinct from and not directly comparable to the existing estimates of the fraction of *GWAS associations* colocalizing with eQTLs. This distinction matters because it illuminates the role of eQTLs in known trait biology rather than examining the locus for the presence of a colocalizing eQTL which may or may not be relevant to the complex trait.

A different way to identify potential causative genes under GWAS peaks using gene expression is the transcriptome-wide association study design (TWAS)[30–32]. This approach measures local genetic correlation between a complex trait and gene expression. Though not designed to avoid correlation signals caused by LD[33], the approach has higher power than colocalization methods in cases of allelic heterogeneity or poorly typed causative variants[30]. We used the FUSION implementation of TWAS,

which accounts for the possibility of multiple cis-eQTLs linked to the trait-associated variant by jointly calling sets of genes predicted to include the causative gene, to interrogate our 89 loci[32].

FUSION included our putatively causative genes in the set of genes identified as likely relevant to the GWAS peak in 42/89 (47%) loci. Genes were often identified as hits in multiple tissues, but with an inconsistent direction of effect—that is, increased gene expression correlated with an increase in the quantitative trait or disease risk in some tissues, but a decrease in others. This may indicate that different tissues have relevant genes that are different, but still called within the same joint set. Because of this possibility, and the known biological role of many of our genes, we restricted our results to tissues with established relevance to our traits. Only 9/89 (10%) genes were identified by FUSION when we restricted the analysis to relevant tissues, and of these, only five had a direction of effect on the complex trait consistent with what is known from hypomorphic and amorphic Mendelian mutations. This fact, combined with the inconsistent direction of effect across tissues, may indicate that even when putatively causative genes fall within a set of genes jointly called by TWAS, their baseline expression may not be mediating the association.

Our results so far are consistent with trait-associated variants altering the regulation of causative genes in ways that are not well-represented by steady-state gene expression measurements. We thus tried to find fine-mapped GWAS variants that appear in regulatory sites within +/- 1 Mb windows around the transcription start sites (TSS) of our

putatively causative genes. We found that 73 fine-mapped variants with a high posterior probability of association (PIP > 0.7) to a trait fall within a narrow peak of H3K27ac, H3K4me1, or H3K4me3 chromatin modification features. Despite our 1 Mb window, all identified features are located within a 100 kb window around the transcription starts sites of 27/89 (30%) putatively causative genes (two of these genes, *ATG16L1* and *CARD9*, are putatively causative for both CD and UC). Extending our search to include not only fine-mapped variants within chromatin modification features, but also those within 500 bp of features, identifies only two additional putatively causative genes. Restricting our analysis to chromatin features in relevant tissues, 46 fine-mapped variants fall within chromatin features, corresponding to 24 putatively causative genes.

Combining activity and proximity signals, we evaluated an “activity-by-distance” measure, a simplified version of the “activity-by-contact” method[34]. Activity-by-distance uses linear distance along the genome instead of the chromatin contact frequency between feature and TSS. Among the fine-mapped variants that fall inside chromatin modification features, 17 variants appear in the feature with the highest activity-by-distance score in the locus, corresponding to 11 genes.

Next, we relaxed the requirement of proximity to a specific feature and selected all enhancer regions annotated by the ChromHMM[35] method in any measured cell or tissue type. Overall, within +/- 1 Mb windows of our putatively causative genes 120/335 fine-mapped variants fall in an enhancer region (i.e. enhancer, bivalent enhancer, genetic enhancer) highlighted by ChromHMM’s core 15-state model. These enhancers

correspond to 43 putatively causative genes. Restricting our analysis to relevant tissues, 51/335 fine-mapped variants fall in enhancers, corresponding to 26 putatively causative genes.

In sum, we observe that fine-mapped variants appear near sites of regulatory activity—suggested by the presence of activating chromatin marks—for a sizable minority of our loci. However, 54/89 (61%) putatively causative genes, no fine-mapped variants are associated with regulatory regions according to either chromatin marks or ChromHMM. Furthermore, because we connect regulatory features to genes based solely on proximity, it is possible that our finding of 35 genes represents an over-estimate.

Overall, our results do not support the assertion that most common non-coding variants associated with human traits alter baseline gene expression in trait-relevant tissues. Several explanations may account for this: incorrect assumptions, lack of statistical power, biological context, and alternative regulatory mechanisms. We discuss each below.

Incorrect assumptions: it is possible that our putatively causative genes may simply not be causative in complex trait forms. This would invalidate our underlying premise that they should be targets of trait-associated variants in the common, complex forms of phenotypes. This implies that in the vast majority of cases, a common variant associated with the polygenic form of a trait near a gene known to cause a severe form

actually targets a different gene. For instance, the risk alleles driving the breast cancer GWAS signal near *BRCA2*, do not alter *BRCA2* expression in breast tissue, but instead influence another gene. This would also explain why 42 putatively causal genes do not fall near a GWAS peak. The implication is that the underlying biological causes of an extreme phenotypic presentation are different from the causes of the polygenic form across all nine of the traits we have studied. This, to our minds, stretches credulity given the highly significant enrichment of our genes near significant GWAS loci for cognate phenotypes. We suggest it is more likely that our putatively causative genes are relevant but influenced in some other way by polygenic risk alleles. More parsimonious explanations for the 42 genes are that currently available GWAS are incompletely powered, and thus have not detected association with alleles in those loci; or that strong purifying selection acting on noncoding regions of these genes is preventing noncoding variants from reaching population frequencies detectable by GWAS.

Lack of statistical power: it is possible that complex trait GWAS are insufficiently powered to allow accurate fine-mapping and hence accurate colocalization; that eQTL studies do not detect all eQTLs; that epigenetic studies do not identify all elements; or that colocalization and regulatory element mapping methods lack power to detect overlaps. However, we have ascertained GWAS associations at genome-wide significance, and fine-map the majority of these signals using a Bayesian approach; and the GTEx Consortium eQTL studies have reached saturation for eGene discovery[6].

The upper bound on the power of colocalization methods, under near-ideal circumstances, is 66% at $P < 0.01$ (Barbeira et al. 2020). Under more typical conditions, the portion of GWAS peaks which colocalize with an eQTL is 25% or higher[9,10,29]. As not all GWAS peaks will share a causative SNP with a *cis*-eQTL, these estimates represent a lower bound on power, with empirical power likely to be much higher. Given our assumption that putatively causative genes are mediating association signals, we would expect that 25% of these associations would colocalize, and that in each case, the gene they colocalize with is our putatively causative gene. We would thus expect *at least* 22/89 (25%) of putatively causative genes near a polygenic trait association signal to have a colocalizing eQTL in relevant tissue. Here, we report all associations without correcting for multiple testing, so we would expect substantially more colocalizations. We thus cannot attribute the absence of such events to lack of power. This conclusion is supported directly by our analyses: coloc explicitly tests the hypothesis that GWAS and eQTL signals are distinct, and finds strong statistical support for this hypothesis in three times as many loci as it finds evidence for colocalization. This suggests that, in many cases, genetically induced changes to baseline expression of putatively causative genes do not translate into downstream phenotypic effects. At the same time, most GWAS peaks over these genes are not eQTLs in available tissues.

The power of TWAS is comparable to colocalization methods in cases of a single typed causative SNP. Its relative power increases in cases of poorly-typed SNPs, allelic heterogeneity, or apparent heterogeneity (when multiple SNPs tag a single untyped

causative SNP)[30]. Thus, the paucity of TWAS signals in the correct tissue and with the correct direction of effect cannot be explained by low power.

Biological context: causative eQTLs may only manifest in certain developmental windows, under specific conditions, or in a crucial cell subpopulation. We used data from the GTEx project, which profiled bulk post-mortem adult tissue samples. If causative eQTLs are only present in early development, or under specific exposures or conditions not applicable to the GTEx donors, they would not be captured in these contexts, even though *cis*-eQTLs have been detected for essentially every gene in the genome in the GTEx data[6].

Single-cell RNA sequencing (scRNA-seq) studies have identified some eQTLs present in only a subset of the cell types captured in bulk-tissue analysis, but these appear to be limited—van der Wjist et al. found that 60% of cell type-specific eQTLs replicate in bulk-tissue analysis, and their use of scRNA-seq found only 13% more eQTLs than bulk-tissue analysis[36]. It has also been posited that cell type-specific eQTLs may be enriched in disease association[37]. Additionally, genes causal for disease tend to have more enhancers, which may lead to more complex spatiotemporal expression[38]. Nonetheless, using this tendency to explain the many putatively causative genes whose expression was not linked to GWAS requires us to believe most genes both have *cis*-eQTLs that do not show up in bulk-tissue analysis, and lack those *cis*-eQTLs which do show up in bulk-tissue analysis. Additionally, nearly all genes identified through

proximity to a fine-mapped variant chromatin mark peak were identified in relevant tissues, suggesting that our selection of tissue is correct.

A new cell-type TWAS method, which leverages large sample sizes for human bulk tissues and high-resolution mouse scRNA-seq data to infer cell-type-specific gene expression for each GTEx sample with respect to each Tabula Muris cell type under an empirical Bayes framework and produce gene expression prediction models at cell-type resolution, found no additional disease-associated gene in type II diabetes, and only one, targeting *FGFR2*, in breast cancer (albeit not in breast mammary tissue; Huwenbo Shi and Alkes Price, unpublished correspondence). This argues against context-specific eQTLs being the most prevalent effect of trait-associated variants.

It is possible for eQTLs to change or disappear over the course of development[39]. Because colocalization and TWAS methods rely on eQTL-mapping, such dynamic eQTLs present a potential blind spot. Chromatin marks provide an orthogonal source of information generally. Furthermore, because chromatin marks within a tissue—especially H3K4me3—can remain stable across developmental time[40], they provide specific value in addressing this blind spot.

Alternative regulatory mechanisms: finally, it is conceivable that most non-coding trait-associated variants act not on expression levels, but on other aspects of gene regulation. For example, splicing QTLs (sQTLs) are enriched in GWAS peaks to the same extent as eQTLs[41,42]. However, only 29% of our trait-associated variants that are

highly likely to be causal (fine-mapping posterior probability > 0.7) fall in introns, despite introns composing 45% of the genome[43]. Thus sQTLs do not immediately appear as a viable hypothesis to explain the majority of trait-associated variation.

We thus have to explain the observation that putatively causative genes are often near GWAS signals driven by non-coding variants, and that these genes are influenced by baseline eQTLs in relevant tissues, but that trait-associated variants are not driving those eQTLs. This result questions the basic assumption that trait variants act by perturbing baseline gene expression, so that eQTLs in GWAS peaks are necessarily relevant to the mapped trait. That these genes are more likely than chance to be near such non-coding trait-associated variants suggests that both the structure and regulation of these genes is relevant to complex traits. However, our results demonstrate that the mechanism by which our genes influence complex traits is generally not their baseline expression.

Regardless of the root cause, our results have consequences for efforts to uncover the biology underlying human traits by linking variants to molecular function through baseline expression measurements. These variant-to-function methods are currently the most common computational strategies for identifying the biological significance and therapeutic potential of non-coding genetic associations. Though they have successfully identified many genes of biological consequence and clinical promise, most causative genes likely go undiscovered. Given the difficulties many tissues present in obtaining expression data across diverse developmental and environmental contexts, the

limitations of examining baseline expression may present a difficult obstacle to overcome.

There are limited mechanistic models to explain the function of non-coding variants besides their action as *cis*-eQTLs. Besides sQTLs, another possibility is *trans*-eQTLs that are not mediated by a *cis* effect on a gene, such as variants affecting CTCF binding sites[37], but this fails to explain the enrichment in GWAS signal near putatively causative genes. Though it is likely that power and context play a role in the lack of overlap we observe, for the reasons above it seems improbable that they explain it entirely. Cumulatively, our analysis shows that whilst gold standard genes are often the closest to a genetic association, more sophisticated analyses incorporating functional genomic data fail to identify them as relevant to the trait in meaningful numbers. There are currently no prominent models to fill this gap, but we must remember that complex trait genetics has overturned our assumptions time and time again.

Figures

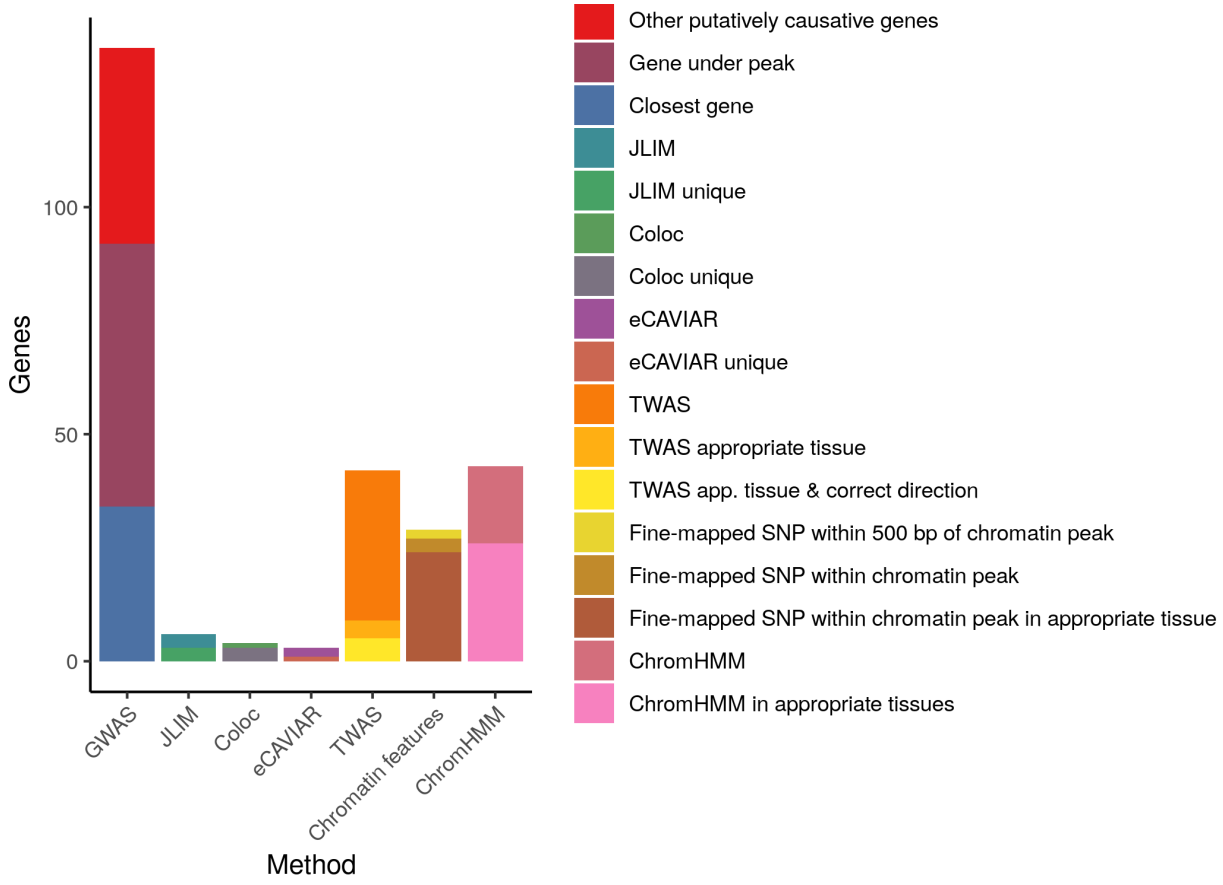
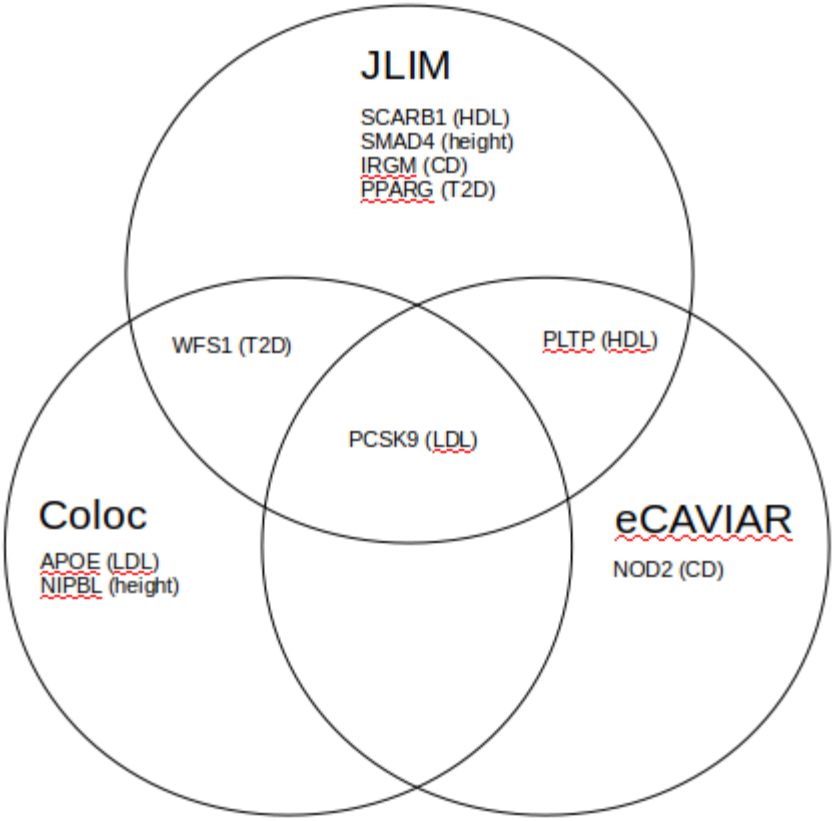


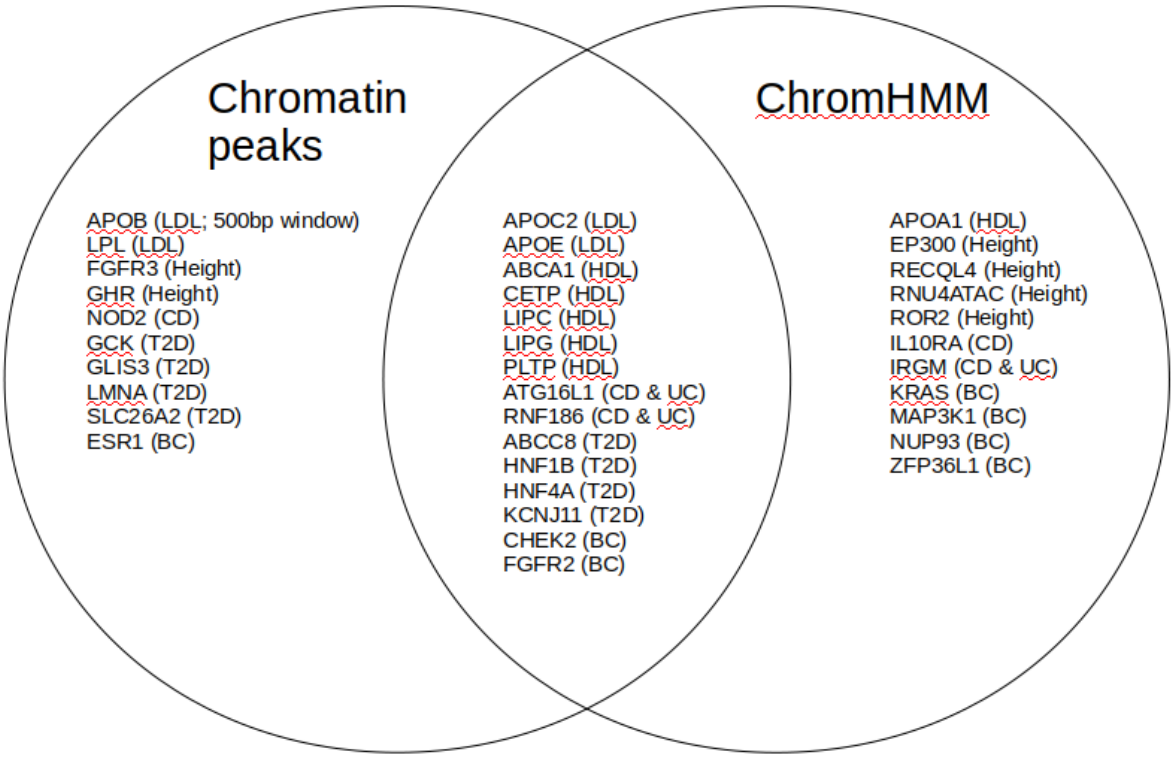
Figure 1. Putatively causative genes identified by each method.

The leftmost column displays the entire set of putatively causative genes, along with the subset near a linkage peak, and its subset of genes closest to the peak. For JLIM, Coloc, and eCAVIAR, the portion of genes that were the only gene to colocalize in their locus is noted. The numbers for these methods represent nominal significance thresholds. For TWAS results, the subsets of genes which are in an appropriate tissue and in an appropriate tissue in the right direction are indicated.


JLIM
SCARB1 (HDL)
SMAD4 (height)
IRGM (CD)
PPARG (T2D)
WFS1 (T2D)
PLTP (HDL)
PCSK9 (LDL)
Coloc
APOE (LDL)
NIPBL (height)
eCAVIAR
NOD2 (CD)

A)

B)


Chromatin peaks
ChromHMM
APOB (LDL; 500bp window)
LPL (LDL)
FGFR3 (Height)
GHR (Height)
NOD2 (CD)
GCK (T2D)
GLIS3 (T2D)
LMNA (T2D)
SLC26A2 (T2D)
ESR1 (BC)
APOC2 (LDL)
APOE (LDL)
ABCA1 (HDL)
CETP (HDL)
LIPC (HDL)
LIPG (HDL)
PLTP (HDL)
ATG16L1 (CD & UC)
RNF186 (CD & UC)
ABCC8 (T2D)
HNF1B (T2D)
HNF4A (T2D)
KCNJ11 (T2D)
CHEK2 (BC)
FGFR2 (BC)
APOA1 (HDL)
EP300 (Height)
RECQL4 (Height)
RNU4ATAC (Height)
ROR2 (Height)
IL10RA (CD)
IRGM (CD & UC)
KRAS (BC)
MAP3K1 (BC)
NUP93 (BC)
ZFP36L1 (BC)

C)

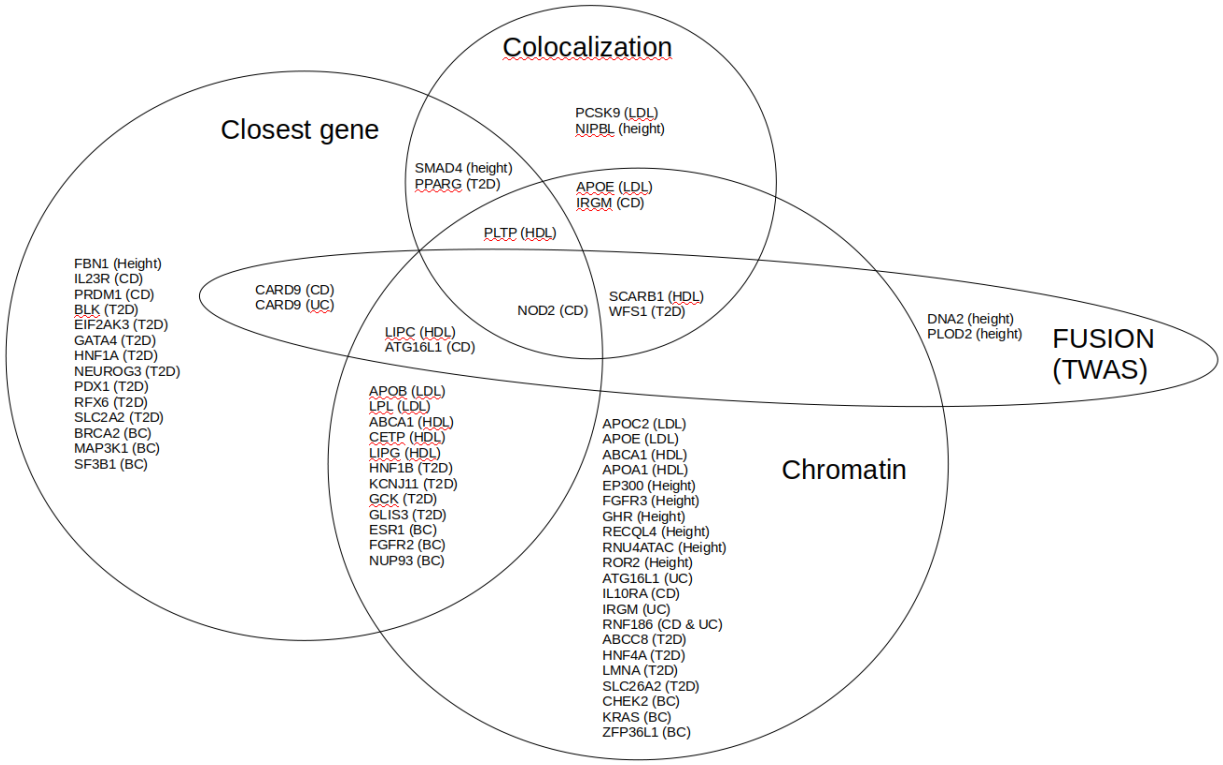


Figure 2. Genes identified as associated with a complex trait by each method.

A) Positive results for each of the three colocalization methods. B) Positive results for each of the two chromatin methods. C) Positive results for all methods, collapsing A) to "colocalization" and B) to "chromatin."

Phenotype	Genes
LDL	APOB APOC2 APOE

	LDLR LPL PCSK9
HDL	ABCA1 APOA1 CETP LIPC LIPG PLTP SCARB1
Height	ANTXR1 ATR BLM CDC6 CDT1 CENPJ COL1A1 COL1A2 COMP CREBBP DNA2 DTDST

	EP300 EVC EVC2 FAM157B FBN1 FGFR3 FKBP10 GHR KRAS NBN NIPBL ORC1 ORC4 ORC6L PCNT PLOD2 PTPN11 RAD21 RAF1 RECQL4 RIT1 RNU4ATAC ROR2

	SLC26A2 SMAD4 SMC3 SOS1 SRCAP WRN
Blood pressure (systolic and diastolic)	KCNJ1 SLC12A1 SLC12A3 WNK1 WNK4
Crohn disease	ATG16L1 CARD9 IL10 IL10RA IL10RB IL23R IRGM NOD2 PRDM1 PTPN22 RNF186

Ulcerative colitis	ATG16L1 CARD9 IL23R IRGM PRDM1 PTPN22 RNF186
Type II diabetes	ABCC8 BLK CEL EIF2AK3 GATA4 GATA6 GCK GLIS3 HNF1A HNF1B HNF4A IER3IP1 INS KCNJ11 KLF11

	LMNA NEUROD1 NEUROG3 PAX4 PDX1 PPARG PTFA1 RFX6 SLC19A2 SLC2A2 WFS1 ZFP57
Breast cancer	AKT1 ARID1A ATM BRCA1 BRCA2 CBFB CDH1 CDKN1B CHEK2 CTCF

	ERBB2 ESR1 FGFR2 FOXA1 GATA3 GPS2 HS6ST1 KMT2C KRAS LRRC37A3 MAP2K4 MAP3K1 NCOR1 NF1 NUP93 PALB2 PIK3CA PTEN RB1 RUNX1 SF3B1 STK11 TBX3

	TP53 ZFP36L1

Table 1. Putatively causative genes

Supplementary methods

Identifying coding variants

Because many variants can fall within coding sequences in rare splice variants, coding SNPs were selected based on the pext (proportion of expression across transcripts) data[44]. Two filters were used. First, genes were considered only if their expression in a trait relevant tissue was at least 50% of their maximum expression across tissues. Second, variants were considered only if they fell within the coding sequence of at least 25% of splice isoforms in that tissue.

GWAS

For height, LDL cholesterol, and HDL cholesterol, GWAS were performed using unrelated individuals of European ancestry from UKBB. The GWAS was run in Plink 2.0[45], using age, sex, BMI (for LDL and HDL only), 10 principal components, and coding SNPs as covariates.

Conditional analysis

Analysis of breast cancer, Crohn disease, ulcerative colitis, and type II diabetes used publically available summary statistics. The summary statistics were corrected for

coding SNPs using an LD reference panel of TOPMed subjects of European ancestry[46]. These subjects were identified with FastPCA[47,48] and extracted using bcftools[49].

Colocalization

JLIM[9] was running using GWAS summary statistics and GTEx v7 genotypes and phenotypes. Coloc[10] was run using GWAS and GTEx v7 summary statistics. eCAVIAR[29] was run using GWAS and GTEx v7 summary statistics, and a reference dataset of LD from UKBB[50] (Weissbrod et al. 2021).

1. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
2. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).
3. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
4. Gusev, A. *et al.* Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
5. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLOS Genet.* **6**, e1000888 (2010).
6. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
7. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).

8. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214-1231.e11 (2020).

9. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).

10. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).

11. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *bioRxiv* 814350 (2020) doi:10.1101/814350.

12. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).

13. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).

14. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

15. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).

16. Blair, D. R. *et al.* A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell* **155**, 70–80 (2013).

17. Schneider, W. J. *et al.* Familial dysbetalipoproteinemia. Abnormal binding of mutant apoprotein E to low density lipoprotein receptors of human fibroblasts and membranes from liver and adrenal of rats, rabbits, and cows. *J. Clin. Invest.* **68**, 1075–1085 (1981).

18. Boerwinkle, E. & Utermann, G. Simultaneous effects of the apolipoprotein E polymorphism on apolipoprotein E, apolipoprotein B, and cholesterol metabolism. *Am. J. Hum. Genet.* **42**, 104–112 (1988).

19. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale

association analysis. *Nat. Genet.* **42**, 579–589 (2010).

20. Chan, Y. *et al.* Genome-wide Analysis of Body Proportion Classifies Height-Associated Variants by Mechanism of Action and Implicates Genes Important for Skeletal Development. *Am. J. Hum. Genet.* **96**, 695–708 (2015).

21. Freund, M. K. *et al.* Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits. *Am. J. Hum. Genet.* **103**, 535–552 (2018).

22. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

23. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

24. Goyette, P. *et al.* High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat. Genet.* **47**, 172–179 (2015).

25. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**, 572–581 (2020).

26. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).

27. Genetic Investigation of ANthropometric Traits (GIANT) Consortium *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).

28. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

29. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

30. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

31. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

32. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).

33. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).

34. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).

35. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

36. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).

37. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* (2020) doi:10.1016/j.tig.2020.08.009.

38. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).

39. Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).

40. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).

41. Walker, R. L. *et al.* Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. *Cell* **179**, 750-771.e22 (2019).

42. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).

43. Francis, W. R. & Wörheide, G. Similar Ratios of Introns to Intergenic Sequence across Animal Genomes. *Genome Biol. Evol.* **9**, 1582–1598 (2017).

44. Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).

45. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

46. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

47. Galinsky, K. J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).

48. Galinsky, K. J., Loh, P.-R., Mallick, S., Patterson, N. J. & Price, A. L. Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure. *Am. J. Hum. Genet.* **99**, 1130–1139 (2016).

49. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).

50. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).

Acknowledgements

We thank Alkes Price, Alex Bloemendal, Benjamin Neale, Bogdan Pasanuic, Sasha (Alexander Gusev), and Matt Warman for their helpful discussions. This research was supported by NIH grants HG010372, R35GM127131, R01HG010372, and R01MH101244. N.J.C was supported by NIH training grant T32GM74897. UK Biobank

was accessed under projects 14048 and 10438. TOPMed data were used under dbGaP project 28674.