The missing link between genetic association and regulatory function

Noah Connally¹,²,³, Sumaiya Nazeen¹,²,⁴, Daniel Lee¹,²,³, Huwenbo Shi³,⁵, John Stamatoyannopoulos⁶, Sung Chun⁷, Chris Cotsapas³,⁸,⁹*, Christopher A. Cassa²,³*, Shamil Sunyaev¹,²,³* ⁴

¹ Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA ⁵ Brigham and Women’s Hospital, Division of Genetics, Harvard Medical School, Boston, MA, USA ⁶ Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA ⁷ Brigham and Women’s Hospital, Department of Neurology, Harvard Medical School, Boston, MA, USA ⁸ Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA ⁹ Altius Institute, Seattle, WA, USA ¹⁰ Division of Pulmonary Medicine, Boston Children’s Hospital, Boston, MA, USA ¹¹ Department of Neurology, Yale Medical School, New Haven, CT, USA ¹² Department of Genetics, Yale Medical School, New Haven, CT, USA ¹³ * Co-corresponding authors

The genetic basis of most traits is highly polygenic and dominated by non-coding alleles. It is widely assumed that such alleles exert small regulatory effects on the expression of cis-linked genes. However, despite the availability of gene expression and epigenomic data sets, few variant-to-gene links have emerged. It is unclear whether these sparse results are due to limitations in available data and methods, or to deficiencies in the underlying assumed model. To better distinguish between these possibilities, we identified 220 gene-trait pairs in which protein-coding variants influence a complex trait or its Mendelian cognate. Despite the presence of expression quantitative trait loci near most GWAS associations, by applying a gene-based approach we found limited evidence that the baseline expression of trait-related genes explains GWAS associations, whether using colocalization methods (8% of genes implicated), transcription-wide association (2% of genes implicated), or a combination of regulatory annotations and distance (4% of genes implicated). These results contradict the hypothesis that most complex trait-associated variants coincide with homeostatic eQTLs, suggesting that better models are needed. The field must confront this deficit, and pursue this “missing regulation.”
Introduction

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,” in which no combination of common and rare variants can explain a large fraction of trait heritability. Further work has revealed unexpectedly high polygenicity for most human traits and very small effect sizes for individual variants. 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. Furthermore, genes in trait-associated loci are more likely to have genetic variants that affect 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. Combined, these observations have led to the inference that most trait-associated variants are eQTLs, and their effects arise from altering transcript abundance, rather than protein sequence. Equivalent sQTL (splice 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. The genetic variant causing expression changes may lie outside the locus and 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 trans-eQTLs are typically mediated by the variant influencing a gene in the region (a cis-eQTL). Thus, a model has emerged in which most trait-associated variants influence proximal gene regulation.

Here we argue that this unembellished model—in which GWAS peaks are mediated by the effects on the homeostatic expression assayed in tissue samples—is the exception rather than the rule. We highlight challenges of current strategies linking GWAS variants to genes and call for a reevaluation of the basic model in favor of more complex models possibly involving context-specificity with respect to cell types, developmental stages, cell states, or the constancy of expression effects.

Our argument begins with several observations that challenge the unembellished model. One challenge is the difference between spatial distributions of eQTLs, which are dramatically enriched in close proximity to genes, and GWAS peaks, which are usually farther away. Another is that expression levels mediate a minority of complex trait heritability. Finally, many studies have designed tools for colocalization analysis: a test of whether GWAS and eQTL associations are due to the same set of variants, not merely distinct variants in linkage disequilibrium. If the model is correct, most trait
associations should also be eQTLs, but across studies, only 5-40% of trait associations co-localize with eQTLs\(^{11-14}\).

Despite the doubts raised, the fact that most GWAS peaks do not colocalize with eQTLs cannot disprove the predominant, unembellished model. In a sense, negative colocalization results are confusing because their hypothesis is too broad. If we predict merely that GWAS peaks will colocalize with some genes' expression, it is not clear what is meant by a peak's failure to colocalize with any individual gene's expression.

Thus, a narrower, more testable hypothesis requires identifying genes we believe \textit{a priori} are biologically relevant to the GWAS trait. If these trait-linked genes have nearby GWAS peaks and eQTLs, failure to colocalize would be a meaningful negative result. Earlier studies tested all GWAS peaks; when a peak has no colocalization, the model is inconclusive. But trait-linked genes that fail to colocalize reveal that our method for detecting non-coding variation is, with current data, incompatible with our model for understanding it.

With this distinction in mind, we created a set of trait-associated genes capable of supporting or contradicting the model of non-coding GWAS associations acting as eQTLs. For this purpose the selection of genes becomes extremely important. Because the model attempts to explain the genetic relationship between traits and gene expression, true positives cannot be selected based on measurements of genetic association to traits (GWAS) or expression (eQTL mapping). With this restriction, 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 tied to a related Mendelian trait or the same complex trait. Using a model not based on expression, Mendelian genes are enriched in common-variant heritability for cognate complex traits\(^ {15}\). The genes and their coding variants may be detected in familial studies of cognate Mendelian disorders, or by aggregation in a burden test on the same complex phenotypes as GWAS\(^ {16}\).

For genes whose coding variants can cause detectable phenotypic change, the strong expectation is that a variant of small effect influences the gene identified by its rare coding variants. As an example, \textit{APOE} and \textit{LDLR} are both low-density lipoprotein receptor genes\(^ {17,18}\). Coding variants in \textit{APOE} and \textit{LDLR} can lead to the Mendelian disorder familial hypercholesterolemia\(^ {18,19}\). Even in the absence of a Mendelian coding variant, experiments in animal models have found that the overexpression of these genes reduces cholesterol levels\(^ {20-22}\). GWAS on human subjects have found significant associations near \textit{APOE} and \textit{LDLR}, so it seems reasonable to suspect that any noncoding effects in these loci may be mediated by these genes. This general
Results

To test the model that trait-associated variants influence baseline gene expression, we assembled a list of putatively causative genes for seven polygenic common traits with available large-scale GWAS data, each of which also has an extreme form in which coding variants of large effect alter one or more genes with well-characterized biology (Table 1). Our selection included four common diseases: type II diabetes, 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, which have Mendelian pediatric forms characterized by severity of presentation; and breast cancer, where germline coding mutations (e.g., BRCA1) or somatic tissue (e.g., PIK3CA) are sufficient for disease. We also chose three quantitative traits: low and high density lipoprotein levels (LDL and HDL); and height.

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, which could make the effect of a coding variant appear to be regulatory instead. 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 (height, LDL, HDL); for those studies without available genotype data, we computed conditional associations from summary statistics using COJO (Methods). With both methods, the resulting GWAS associations should reflect only non-coding variants.

After controlling for coding variation, we examined whether these genes are more likely than chance to be in close proximity to variants associated with the polygenic form of each trait. In agreement with existing literature, we observe a significant enrichment for all traits in our combined Mendelian and Backman et al. gene sets (Fig. 1 Supp. 1).

Of our 220 genes, 147 (67%) fell within 1 Mb of a GWAS locus for the cognate complex trait, over three times as many as the 43 predicted by a random null model (95% confidence interval: 31.5-54.5). Our window of 1 Mb represents roughly the upper bound for distances identified between enhancer-promoter pairs, but most pairs are closer, so we would expect enrichment to increase as the window around genes decreases; this proves to be the case. At a distance of 100 kb, we find 104 putatively causative genes (47%), though the null model predicts only 11 (95% CI 4.5-17.0), an
order-of-magnitude enrichment (Fig. 1 Supp. 1). Given their known causal roles in the severe forms of each phenotype, these results suggest that the 147 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 147 genes in relevant tissues. Controlling for the number of tests we conducted, 134 (91.1%) of these genes had an eQTL in at least one relevant tissue at a false discovery rate of Q < 0.05 (Methods). If these variants act through changes in gene expression, phenotypic associations should be driven by some of the same variants as eQTLs in relevant tissue types. We therefore looked for co-localization between our GWAS signals and eQTLs in relevant tissues (Table 2) drawn from the GTEx Project, using three well-documented methods: coloc11, JLIM12, and eCAVIAR14. We found support for the co-localization of trait and eQTL association for only 7 genes out of 147 (4.8%) for coloc; 10/147 (6.8%) for JLIM; and 8/147 (5.4%) for eCAVIAR. Accounting for overlap, this represents only 18/220 putatively causative genes (8.2%) or 18/147 (12.2%) putatively causative genes near GWAS peaks, even without full multiple-hypothesis testing correction (Methods), which is not obviously better than random chance. We note that prior estimates of the fraction of GWAS associations colocalizing with eQTLs (25%-40%11,12,14,31) do not directly evaluate the ability to find causative genes. By contrast, our estimate of the number of putatively causative genes that colocalize with eQTLs tests the consistency of our knowledge, models, and data.

A potential weakness of our approach is the restriction of our search to pre-defined tissues. We believe this is necessary in order to avoid the disadvantages of testing each gene-trait pair in each tissue—either a large number of false positives, or a severe multiple-testing correction that may lead to false negatives. However, restricting to the set of tissues with a known biological role and available expression data almost certainly leaves out tissues with relevance in certain contexts. Some of the tissues we do use have smaller sample sizes, limiting their power to detect eQTLs with smaller effects.

To address potential shortcomings from the available sample of tissue contexts, we incorporated the Multivariate Adaptive Shrinkage Method (MASH)32. MASH is a Bayesian method that takes genetic association summary statistics measured across a variety of conditions and, by determining patterns of similarity across conditions, updates the summary statistics of each individual condition. In our case, if an eQTL is...
difficult to find in a tissue of interest, incorporating information from other tissues may help us detect it. Unlike meta-analysis, this method generates summary statistics that still correspond to a specific tissue.

We ran MASH on every locus used in our earlier analysis, using data from all non-brain GTEx tissues (Methods). Rerunning coloc with these modified statistics increased the number of GWAS-eQTL colocalizations across all genes by 26% (from 389 to 489). However, the 100 new colocalizations identified only four additional putatively causative genes (Fig. 1 Supp. 2). These results indicate that tissue type selection was not the limiting factor in our analysis.

Transcriptome-wide association studies (TWAS) are another class of methods applied to identify causative genes under GWAS peaks using gene expression. TWAS measures genetic correlation between traits and is not designed to avoid correlations caused by LD, which gives it higher power in the case of allelic heterogeneity or poorly typed causative variants. However, while sensitive, TWAS analyses typically yield expansive result sets that include many false positives and are sensitive to the number of tissue types. Results from the FUSION implementation of TWAS across all tissues identified our putatively causative genes as likely tied to the GWAS peak in 66/220 loci (30%). However, only 4/220 (1.8%) genes were identified by FUSION when we restricted the analysis to relevant tissues.

Given the paucity of expression-mediated GWAS peaks, we asked whether GWAS variants in our study loci reside in likely regulatory sites. Taking the 128 genes in the Mendelian subset of putatively causative genes, we fine-mapped each nearby GWAS association using the SuSiE algorithm. For 37 of these genes, we identified at least one high-confidence fine-mapped SNP (PIP>0.7) within 100kb of the transcription start site. We tested whether these fine-mapped SNPs fall within (i) regulatory DNA marked by DNase I hypersensitive sites (DHSs); (ii) a narrowly mapped active histone modification feature (H3K27ac, H3K4me1, or H3K4me3); or (iii) sites marked as an “enhancer” by ChromHMM (Methods). As many as 32/37 (86%) genes identified this way have a fine-mapped SNP within a candidate regulatory region across all the tissue types examined, or 25/37 (68%) when restricting to phenotype-relevant tissues (Fig. 3; Supp. File 1, Supp. File 2). Despite this strong evidence that these GWAS associations arise from effects on regulatory DNA, only 5/25 loci (20%) demonstrably correspond to expression effects in our eQTL analysis.

In order to more analogously compare our regulatory feature analysis to our eQTL analysis, we computed “activity-by-distance” (ABD)—a simplification of the “activity-by-contact” method that provides a measure connecting a distal regulatory region to a
target gene$^{30,43}$ (Methods; Fig. 2). Taking each locus’s feature with the highest ABD score, we implicate 5/37 (14%) of our Mendelian subset of genes. As such, even when a GWAS association and trait-relevant gene are in the same locus, they are difficult to link, whether using eQTLs or current approaches to integrating chromatin data with target genes.

**Discussion**

Overall, our results are strongly consistent with the idea that complex traits are governed by non-coding genetic variants whose effects on phenotype are mediated by their contribution to the regulation of nearby genes. However, these same results are inconsistent with a model that such effects on gene regulation arise from the genetic influences on baseline gene expression that are captured by eQTLs.

The enrichment of putatively causative genes—selected based on existing biological knowledge—near GWAS peaks supports their role in complex traits. Additionally, the enrichment of fine-mapped GWAS variants in likely regulatory regions marked by DHSs and other chromatin features lends support to the general model of GWAS associations arising from effects on gene regulation and expression. However, the inability of varied statistical methods to connect GWAS associations and expression argues against the idea that the causative GWAS variants exert their effects via the homeostatic or blended effects on gene expression that are captured by bulk-tissue eQTLs of the sort discovered by broad expression-data collection projects.

Many explanations have been suggested for the limited success of expression-centric methods to explain the mechanisms of GWAS variants. Undirected, broad approaches—including most GWAS-eQTL linking studies—are designed to be largely independent of *a priori* biological knowledge and hypotheses. This unconstrained focus is ideal for discovery because it delivers the largest number of positive findings, but it is ill-suited for providing explanations for negative results—when you don’t know what you were looking for, it’s hard to explain why you didn’t find it. By testing only loci for which there is a strongly suspected contributing gene, we were better able to distinguish which factors might prevent us from identifying GWAS-to-gene links using expression alone.

As a result, we conclude that a number of explanations often considered when evaluating expression-based variant-to-gene methods are not applicable in the context we examined. These include (i) non-expression-mediated mechanisms, (ii) lack of statistical power for GWAS, (iii) the absence of eQTLs for relevant genes, (iv) and underpowered methods for linking expression to GWAS (Table 3).
Our results suggest that there remains to be discovered a major component connecting variants to gene expression, which we term “missing regulation.” We propose that this currently lacking component could be exposed by considering more nuanced experimental and analytical models of gene expression. One likely model involves context-dependent gene expression—expression whose relevant effects are confined to (i) particular refined cell types or anatomical partitions, or (ii) cell states such as responses to perturbations, whether exogenous (e.g., environmental change) or endogenous (e.g., cell differentiation). A complementary model may incorporate heterogeneity of gene expression or the variance of expression across relatively short time scales. These models and others may depend on or be augmented by thresholding or buffering of variant effects, which may produce changes in gene expression with non-linear effects on phenotype. A summary of proposed models can be found in Table 4.

We propose that finding the “missing regulation” will require not only identifying novel eQTLs explaining GWAS peaks, but also explaining the phenotypic irrelevance of “red herring eQTLs”—i.e., eQTLs for putatively causative genes that fall near GWAS peaks but do not colocalize with them. Above, we use the example genes APOE and LDLR. Both these genes harbor coding variants causing Mendelian hypercholesterolemia, and both have non-coding variants that GWAS have tied to LDL levels. Both have eQTLs in trait-relevant tissues. For APOE, these points cohere into an explanation: the LDL-association is an eQTL for the lipid-binding gene. But for LDLR—and for most genes—the association, the mechanism, and the gene cannot be tied together.

Importantly, our results do not diminish the importance or general utility of eQTLs. Rather, they suggest that current models are deficient in two respects: (i) they fail to unify trait-associated non-coding variants with known trait-associated genes, and (ii) they fail to explain the non-effects of identified “red herring” eQTLs. These deficiencies highlight a need for new approaches to the role of gene regulation in complex traits.

One long-standing goal of GWAS has been to discover genes contributing to complex traits, but low rates of positive findings for expression-based variant-to-gene methods have constrained this possibility. Among other challenges, this has limited the benefit of GWAS and expression data for disease-gene mapping and drug discovery. Another practical question raised is the value of current large-scale public datasets. Compared to genotypes, expression data are relatively difficult to collect, particularly from specialized cell contexts. If the most relevant models are shown to depend on effects not observable in homeostatic eQTL mapping, the field may need to consider prioritizing other biological contexts and forms of expression data.

**Figures**
Figure 1. Putatively causative genes identified by each method category. The leftmost column in each half of the plot displays the entire group of putatively causative genes for our Mendelian set of genes and our Backman et al. set of genes respectively, as well as noting how many are unique to each set or shared between the two sets. The second column in each half indicates how many genes from each set have a nearby GWAS peak, or have both a nearby GWAS peak and an eQTL. The remaining columns indicate how many genes were identified through colocalization, TWAS, or chromatin methods, while noting how many of these genes are unique vs. shared between the Mendelian and Backman sets.
**Figure 1 Supplement 1. Enrichment of Mendelian genes near GWAS peaks.**

A) As the window around GWAS peaks shrinks, the enrichment of Mendelian genes within the window becomes increasingly significant, while the enrichment of non-matching trait pairs used as controls (gray lines; **Methods**) is not consistently increased. Some controls achieve nominal significance (dotted horizontal line), but none reach significance once multiple-testing is corrected for (solid horizontal line). B) As above, but for genes from Backman et al. (2021)\textsuperscript{16}. C) The combined gene lists from parts A and B. Note that, accounting for multiple test correction (based on the total number of tests across all panels) height does not reach significance using the Mendelian gene list, while T2D is barely significant using the Backman list. However, combining the lists increases power and demonstrates significance for all traits.
Figure 1 Supplement 2. Change in coloc hits when adjusting eQTL statistics using MASH.

By using the Bayesian method MASH to update our measurements of eQTLs based on tissues with similar expression patterns, we increased the number of colocalizations found. However, even in tissues in which the number of genes identified increased substantially, we did not meaningfully increase the number of putatively causative genes identified.
Following the fine-mapping of GWAS variants, three parallel methods were used. The first identified fine-mapped variants falling within regions annotated as enhancers by ChromHMM. The second identified variants within histone modification features, and evaluated their relevance using an ABD score that combined the strength of the feature (i.e., the strength of the acetylation or methylation peak) with its genomic distance to the gene of interest (Methods). The third repeated both of these—checking for fine-mapped variants within a region and calculating the ABD score—for DNase I hypersensitivity sites.
Figure 3. Genes identified as associated with a complex trait by each method. Columns “Mend” and “Backman” indicate whether a gene is from the Mendelian set of putatively causative genes, the Backman et al. set, or both. Subsequent columns indicate whether a gene was identified as a hit using each of our methods: JLIM, coloc, eCaviar, TWAS, and chromatin analysis.
<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDL</td>
<td>APOB&lt;sup&gt;45,4&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>APOC2&lt;sup&gt;48&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>APOE&lt;sup&gt;49&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>LDLR&lt;sup&gt;50&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>LPL&lt;sup&gt;51,52&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PCSK9&lt;sup&gt;53&lt;/sup&gt;</td>
</tr>
<tr>
<td>HDL</td>
<td>ABCA1&lt;sup&gt;54–56&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>APOA1&lt;sup&gt;57&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>CETP&lt;sup&gt;58&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>LIPC&lt;sup&gt;59–61&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>LIPG&lt;sup&gt;62&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PLTP&lt;sup&gt;63&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>SCARB1&lt;sup&gt;64,65&lt;/sup&gt;</td>
</tr>
<tr>
<td>Height</td>
<td>ANTXR1&lt;sup&gt;66,67&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>ATR&lt;sup&gt;68,69&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>BLM&lt;sup&gt;70,71&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>CDC6&lt;sup&gt;72&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>CDT1&lt;sup&gt;72,73&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>CENPJ&lt;sup&gt;74&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>COL1A1&lt;sup&gt;75&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>COL1A2&lt;sup&gt;76,77&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>COMP&lt;sup&gt;78,79&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>CREBBP&lt;sup&gt;80–82&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>DNA2&lt;sup&gt;83&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>EP300&lt;sup&gt;84,85&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>EVC&lt;sup&gt;86,87&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>EVC2&lt;sup&gt;87,88&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>FBN1&lt;sup&gt;89–92&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>FGFR3&lt;sup&gt;93–95&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>FKBP10&lt;sup&gt;96–98&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>GHR&lt;sup&gt;99–102&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>KRAS&lt;sup&gt;103–105&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>NBN&lt;sup&gt;106,107&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>NIPBL&lt;sup&gt;108,109&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>ORC1&lt;sup&gt;72,73,110&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>ORC4&lt;sup&gt;72,73&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>ORC6L&lt;sup&gt;72,111&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PCNT&lt;sup&gt;112–114&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PLOD2&lt;sup&gt;115–117&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PTPN11&lt;sup&gt;118–120&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>RAD21&lt;sup&gt;121–123&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>RAF1&lt;sup&gt;124,125&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>RECQL4&lt;sup&gt;126–128&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>RIT1&lt;sup&gt;129–131&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>ROR2&lt;sup&gt;132–134&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>SLC26A2&lt;sup&gt;135–137&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>SMAD4&lt;sup&gt;138–140&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>SRCAP&lt;sup&gt;141,142&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>WRN&lt;sup&gt;143–145&lt;/sup&gt;</td>
</tr>
<tr>
<td>Crohn disease</td>
<td>ATG16L&lt;sup&gt;146&lt;/sup&gt;</td>
</tr>
<tr>
<td>Ulcerative colitis</td>
<td>ATG16L1&lt;sup&gt;163&lt;/sup&gt;</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td>CARD9&lt;sup&gt;147&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>IL23&lt;sup&gt;R&lt;/sup&gt;&lt;sup&gt;155,164&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>IRGM&lt;sup&gt;156&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PRDM1&lt;sup&gt;161&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PTPN22&lt;sup&gt;162&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>RNF186&lt;sup&gt;165,166&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type II diabetes</th>
<th>ABCC8&lt;sup&gt;16r&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLK&lt;sup&gt;168&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>CEL&lt;sup&gt;169,170&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>EIF2AK3&lt;sup&gt;171–173&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>GATA4&lt;sup&gt;174&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>GATA6&lt;sup&gt;175,176&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>GCK&lt;sup&gt;177&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>GLIS3&lt;sup&gt;178&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>HNF1A&lt;sup&gt;179,180&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>HNF1B&lt;sup&gt;181,182&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>HNF4A&lt;sup&gt;183,184&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>IER3IP1&lt;sup&gt;185–187&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>INS&lt;sup&gt;188&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>KCNJ11&lt;sup&gt;189,190&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>KLF11&lt;sup&gt;191&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>LMNA&lt;sup&gt;192&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>NEUROD1&lt;sup&gt;193&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>NEUROG3&lt;sup&gt;194–196&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PAX4&lt;sup&gt;197–199&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PDX1&lt;sup&gt;200–202&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PPARG&lt;sup&gt;203,204&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>PTF1A&lt;sup&gt;205&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>RFX6&lt;sup&gt;206,207&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>SLC19A2&lt;sup&gt;208–210&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>SLC2A2&lt;sup&gt;211,212&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>WFS1&lt;sup&gt;213–215&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>ZFP57&lt;sup&gt;216,217&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Breast cancer (selected using MutPanning&lt;sup&gt;218&lt;/sup&gt;)</th>
<th>AKT1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ARID1A</td>
</tr>
<tr>
<td></td>
<td>ATM</td>
</tr>
<tr>
<td></td>
<td>BRCA1</td>
</tr>
<tr>
<td></td>
<td>BRCA2</td>
</tr>
<tr>
<td></td>
<td>CBFB</td>
</tr>
<tr>
<td></td>
<td>CDH1</td>
</tr>
<tr>
<td></td>
<td>CDKN1B</td>
</tr>
</tbody>
</table>
Table 1. Putatively causative Mendelian genes.

Each gene includes reference(s) to the known biological role of its coding variants, as established in familial studies, *in vitro* experiments, and/or animal models. Genes from Backman *et al.* are not included here, but can be found in Figure 3.

<table>
<thead>
<tr>
<th>Mendelian trait</th>
<th>GWAS trait</th>
<th>Tissues examined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>Breast cancer</td>
<td>Breast mammary tissue</td>
</tr>
<tr>
<td>Crohn disease</td>
<td>Crohn disease</td>
<td>Small intestine terminal ileum Colon Sigmoid Colon Transverse</td>
</tr>
<tr>
<td>Ulcerative colitis</td>
<td>Ulcerative colitis</td>
<td>Small intestine terminal ileum Colon Sigmoid Colon Transverse</td>
</tr>
<tr>
<td>Dyslipidemia Hyperlipidemia Tangier’s disease</td>
<td>HDL</td>
<td>Liver Adipose (subcutaneous) Whole blood</td>
</tr>
<tr>
<td>Dyslipidemia Hyperlipidemia</td>
<td>LDL</td>
<td>Liver Adipose (subcutaneous) Whole blood</td>
</tr>
</tbody>
</table>
Table 2. Tissue-trait pairs
Tissues were selected for each trait based on a priori knowledge of disease biology.

<table>
<thead>
<tr>
<th>Mendelian short stature</th>
<th>Height</th>
<th>Skeletal muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monogenic diabetes</td>
<td>Type II diabetes</td>
<td>Pancreas</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Skeletal muscle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adipose (subcutaneous)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Small intestine terminal ileum</td>
</tr>
</tbody>
</table>

### Violated assumptions

<table>
<thead>
<tr>
<th>Genes implicated via coding variants are irrelevant for non-coding associations</th>
<th>- Our genes are enriched for GWAS associations even after removing the effects of coding variants.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Loss-of-function variants, which underlie many Mendelian-trait genes, can be thought of as large-effect eQTLs.</td>
</tr>
<tr>
<td></td>
<td>- Genes identified from Backman et al. are not based on cognate phenotypes, but the same complex phenotypes as GWAS.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Regulatory mechanisms other than cis-eQTLs</th>
<th>- Splice QTLs are consistently found to explain less phenotypic variance than eQTLs, and they cannot explain the many GWAS associations that fall within intergenic regions.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Trans-eQTLs are believed to rely on their effects as cis-eQTLs for other genes; the few exceptions to this model (e.g., CTCF binding sites) are not broadly applicable.</td>
</tr>
</tbody>
</table>

### Insufficient power

<table>
<thead>
<tr>
<th>Lack of GWAS power</th>
<th>- GWAS have been shown to have sufficient power to identify small effects even in rare variants.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- 2/3 of the genes we used have nearby GWAS associations, reflecting a strong enrichment and indicating that GWAS discovery is not a limiting factor.</td>
</tr>
<tr>
<td></td>
<td>- Our analysis is conditioned on the presence of GWAS associations.</td>
</tr>
</tbody>
</table>

| Lack of eQTL mapping power | - GTEx is well powered for eQTL discovery in bulk tissue$^{6,33}$. |
|                           | - 93% of our genes have a mapped cis-eQTL in a relevant tissue. |
Lack of power for colocalization and TWAS methods

- Simulations show that colocalization and TWAS methods are well-powered\textsuperscript{12–14,34,219}.
- They are robust to levels of LD mismatch higher than what would be expected given our datasets\textsuperscript{12,37,219}.
- Some, though not all, of the methods are robust to allelic heterogeneity\textsuperscript{12,37}.

Table 3. Proposed explanations for negative results under the unembellished model.

Many explanations have been proposed for GWAS associations that are not explained by \textit{cis}-eQTLs. This table details explanations inconsistent with our results, which are explained in the left column and addressed in the right. Explanations involving more detailed models of gene regulation can be found in Table 4. Two of the explanations addressed here involve violations of the assumptions of our and other expression-based complex trait studies. If coding and non-coding variants affect fundamentally different biological pathways, or if trait associations rarely depend on \textit{cis}-eQTLs, our methods of mapping regulation to traits would have nothing to uncover. Even in the presence of eQTL-driven trait associations, insufficient power to detect trait associations, to detect eQTL associations, or to link the two would result in predominantly negative results.

<table>
<thead>
<tr>
<th>Extended models of gene regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Context dependency:</strong> A context-specific eQTL, invisible in bulk tissues analyzed to date, replaces or supplements the bulk tissue homeostatic eQTL</td>
</tr>
<tr>
<td><strong>Cell type</strong>\textsuperscript{220–234}</td>
</tr>
<tr>
<td>- Only a subset of cell types in the tissue contribute to the GWAS phenotype.</td>
</tr>
<tr>
<td>- An eQTL specific to such a cell type is causative for the phenotype.</td>
</tr>
<tr>
<td>- The eQTL either cannot be detected in bulk tissue because of the cell type’s low prevalence</td>
</tr>
<tr>
<td>- The appropriate cellular or anatomical context has not yet been analyzed.</td>
</tr>
<tr>
<td><strong>Developmental timing</strong>\textsuperscript{220,225–240}</td>
</tr>
<tr>
<td>- The GWAS phenotype depends on a specific point in cell/tissue development or differentiation.</td>
</tr>
<tr>
<td>- eQTLs present at the correct interval contribute to phenotype, but eQTLs observed at other points do not.</td>
</tr>
<tr>
<td><strong>Cell state or environment</strong>\textsuperscript{222,227,231,233,241–248}</td>
</tr>
<tr>
<td>- The causative eQTL has effects that are undetectable in steady-state expression under normal conditions.</td>
</tr>
<tr>
<td>- It may activate only in response to a specific environmental condition, such as immune activation or a metabolic shift.</td>
</tr>
</tbody>
</table>
Non-linear or non-homeostatic:  
The relationship between eQTL and genotype is indirect  

- There may be buffering that prevents a change in expression from producing a change in protein levels.  
- Expression below a certain level may not influence phenotype, rendering small eQTLs irrelevant.  
- Steady-state expression may be a poor model  
  - Phenotype may depend on the kinetics of expression, which could be cyclical or follow some other pattern.  
  - Expression may be stochastic, such that only a random subset of cells display the relevant expression pattern at any one time.

Table 4. Explaining negative results with more nuanced models of gene regulation.

To reconcile an expression-based model with our observations requires us to both explain the absence of trait-linked eQTLs as well as explaining away the inconsequence of eQTLs for trait-linked genes. The left-hand side lists additions or changes to the unembellished model, while the right contains explanations of the models and current relevant research.

Appendix 1. Evidence for the relationship between Mendelian and complex traits.

More generally, 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. Early GWAS found associations near genes identified through familial studies of severe disorders, and later implicated some of the same genes in complex and Mendelian forms of cardiovascular and neuropsychiatric traits. More recent analyses have found that GWAS associations are enriched in regions near causative genes for cognate Mendelian traits in blood traits, lipid traits and diabetes, as well as a diverse collection of 62 traits. Another recent method used transcriptomic, proteomic and epigenomic data to prioritize genes and found that, in a selection of nine phenotypes, selected genes were enriched for Mendelian genes causing similar traits. Together, these suggest that genes causing Mendelian traits also influence cognate complex traits, but not through the same coding mechanism.

Genes can also harbor coding variants tied to less severe forms of a trait. These coding variants are more difficult to identify individually, as their effect sizes are much smaller. However, the greater number of variants (in aggregate) and freedom from searching for severe segregating traits, allows the use of large population datasets. Backman et al. used burden testing on UK Biobank data to identify genes whose coding variation affects complex traits, finding many genes not identified through familial studies.

19
Supplementary File 1. Roadmap epigenomics aliases of tissue types used for functional genomic analysis.

Tissue types from the Roadmap Epigenomics Consortium do not perfectly match those from GTEx. However, there is overlap, and as with GTEx, we analyzed trait-relevant tissues.

Supplementary File 2. Tissue types and bio-samples from the DNase I hypersensitive sites index used for functional genomic analysis

Meuleman et al. assess DNase I hypersensitive sites across 438 cell and tissue types; we selected the above based on their relevance to our complex traits.

Supplementary methods

Gene selection

By manual literature search, we selected 128 genes harboring large-effect-size coding variants for one of the seven phenotypes (Table 1; specifically, we selected 128 gene-trait pairs, representing 121 unique genes). These genes were identified using familial studies, rare disease exome sequencing analyses, and, for breast cancer, using the MutPanning method (citations for each gene are included in Table 1). Review papers, as well as the OMIM database, were generally used as starting points, but an examination of the original literature was needed to confirm genes’ suitability. For example, though SMC3 is known to cause Cornelia de Lange syndrome, which is characterized in part by short stature, SMC3 mutations lead to a milder form of the syndrome, usually without a marked reduction in stature. Several of these phenotypes—height, HDL, cholesterol, breast cancer, and type II diabetes—were also analyzed in Backman et al., which, through burden testing, identified a total of 110 genes; after accounting for overlaps, this increased our set of putatively causative genes to 220.

The inclusion of genes from Backman et al. ensures that our results are not dependent on an undetected bias in our selection. The set of genes chosen from familial studies offered the advantage that it was selected based on independent methods and data distinct from the large-scale genotyping studies that have characterized the GWAS era. The tradeoff to this was the impossibility of selecting genes through a fully systematic and non-arbitrary process. Because this work was performed in the UKBB, there is some overlap between their data and ours. However, our work did not use exomes, and most of the variants driving their findings are too rare to influence GWAS results. When this is not the case, our decision to condition on coding variants should make the effects used in our work independent from their findings.
Identifying coding variants

Because GWAS sample sizes are large enough to detect the low-frequency coding variants used to select some of our genes, it is possible that a coding SNP would distort the association signal of nearby eQTLs. To minimize this concern, we removed the effects of coding variants on GWAS. Many variants can fall within coding sequences in rare splice variants, so it is important to remove only those variants that appear commonly as coding. These coding SNPs were selected based on the pext (proportion of expression across transcripts) data. Two filters were used. First, we removed genes whose expression in a trait-relevant tissue was below 50% of their maximum expression across tissues. Second, we removed variants that fell within the coding sequence of less than 25% of splice isoforms in that tissue. The remaining variants were used to correct GWAS signal, as explained below.

GWAS

For height, LDL cholesterol, and HDL cholesterol, GWAS were performed using genotypic and phenotypic data from the UKBB. In order to avoid confounding, we restricted our sample to the 337K unrelated individuals with genetically determined British ancestry identified by Bycroft et al. The GWAS were run using Plink 2.0, with the covariates age, sex, BMI (for LDL and HDL only), 10 principal components, and coding SNPs.

Conditional analysis

Because UKBB has limited power for breast cancer, Crohn disease, ulcerative colitis, and type II diabetes, we used publicly available summary statistics. The Conditional and Joint Analysis (COJO) program can condition summary statistics on selected variants—in our case, coding variants—by using an LD reference panel. For this reference, we used TOPMed subjects of European ancestry. The ancestry of these subjects was confirmed with FastPCA and the relevant data were extracted using bcftools. Our conditional GWAS data are available at doi:10.5061/dryad.612jm644q

Enrichment analysis

At each distance, the number of Mendelian and non-Mendelian genes within that window around GWAS peaks are counted. P-values are calculated using Fisher’s exact test (Supp. Fig. 1). Because Mendelian genes may be unusually important beyond our chosen traits, we conduct a set of controls by measuring the enrichment of non-matching Mendelian and complex traits (CD genes & BC GWAS; BC genes & LDL GWAS; LDL genes & UC GWAS; UC genes & height GWAS; height genes & T2D GWAS; T2D genes & HDL GWAS; HDL genes & CD GWAS).

eQTL detection
eQTL summary statistics were taken from GTEx v7. Some methods detect colocalization with variants that are individually significant, but would not pass a genome-wide threshold. Because we tested only a subset of genes and, we used the Bejamini-Hochberg method to calculate the FDR based on the number of tests we conducted multiplied by a correction factor to account for variants that are tested in combination with a gene but are not reported (a factor of 20 closely matched the genome-wide FDR results for GTEx). With this method 204/220 (93%) of our genes displayed an eQTL, including 134/147 genes with a nearby GWAS peak (91%). Even using the FDR statistics of the GTEx project—which are based on the assumption of testing every gene in every tissue—107/220 (49%) of our genes and 76/147 (52%) of genes near GWAS peaks had an eQTL at Q < 0.05.

Colocalization

JLIM was run using GWAS summary statistics and GTEx v7 genotypes and phenotypes, for the tissues listed in Table 2. Coloc was run using GWAS and GTEx v7 summary statistics for the same tissues. eCAVIAR was run using GWAS and GTEx v7 summary statistics for these tissues, and a reference dataset of LD from UKBB. MASH was run incorporating data from all non-brain tissues, and coloc was re-run using the adjusted values for the same tissues as before.

MASH

Multivariate adaptive shrinkage (MASH) was applied to all GTEx tissues using the mashr R-package. We restricted this model to non-brain tissues—which include all of our trait-selected tissues—due to the known tendency of brain and non-brain tissues to cluster separately in expression analysis.

Fusion (TWAS)

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 220 loci. FUSION included our putatively causative genes in the set identified as likely relevant to the GWAS peak in 66/220 loci (30%). However, interpretation of this TWAS result is difficult. For many complex traits, TWAS returns a large number of findings (e.g., over 150 for LDL cholesterol and over 4,800 for height). This is in part due to the multiple genes jointly returned at a locus, and can also be a result of the large number of tissues and cell types included in the implementation of FUSION. Most hits are found in tissues without any clear relevance to the trait, and absent in relevant tissues—LDL, for example, has more TWAS associations between expression and eQTL in prostate adenocarcinoma (24 genes associated), brain pre-frontal cortex (23 genes associated), and transformed fibroblasts (21 genes associated) than it does in adipose (16 genes associated), blood
(11 genes associated), or liver (5 genes associated). Individual 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, which suggests that the gene in question may not be the one whose expression contributes to the complex trait. 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.

Fine-mapping GWAS hits

We fine-mapped the GWAS variants located within +/- 100 kb of our putatively causative genes by applying the SuSiE algorithm\(^\text{38}\) on the unconditional summary statistics from the GWAS of breast cancer, Crohn disease, ulcerative colitis, type II diabetes, height, LDL cholesterol, and HDL cholesterol. An LD reference panel from UKBB subjects of European ancestry was used for this analysis. Fine-mapped variants were annotated using snpEff (v4.3t). Only non-coding variants were kept for further analysis.

Functional genomic annotation of fine-mapped hits

We projected fine-mapped GWAS variants onto active regions of the genome, identified using three alternative approaches: (i) histone modification features, (ii) DNase I hypersensitive sites, and (iii) ChromHMM enhancers.

First, we looked at three histone modification marks, namely, acetylation of histone H3 lysine 27 residues (H3K27ac), mono-methylation of histone H3 lysine 4 residues (H3K4me1), and tri-methylation of lysine 4 residues (H3K4me3) from the Roadmap Epigenomics Project\(^\text{37}\) to identify functional enhancers which are key contributors of tissue-specific gene regulation. We downloaded imputed narrowPeak sets for H3K27ac, H3K4me1, and H3K4me3 from the Roadmap Epigenomics Project\(^\text{37}\) ftp site (https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidatedImputed/narrowPeak/) for 14 different tissue types (Supp. File 1). For each tissue type, we extracted the narrow peaks that are within +/- 5 Mb of our putatively causative genes. Then following the approach described in Fulco et al.\(^\text{36}\), we extended the 150 bp narrow peaks by 175 bp on both sides to arrive at candidate features of 500 bp in length. All features mapping to blacklisted regions (https://sites.google.com/site/anshulkundaje/projects/blacklists) were removed. Remaining features were re-centered around the peak and overlapping features were merged to give the final set of features per histone modification track. Mean activity/strength of a feature (\(A_F\)) was calculated by taking the geometric mean of the corresponding peak strengths from H3K27ac, H3K4me1, and H3K4me3 marks. We then combined these activity measurements with the linear distances between the features and the transcription start sites of causative
genes to compute “activity-by-distance” scores (a simplified version of ABC scores\textsuperscript{36}) for gene-feature pairs using the following formula.

\[
ABDscore_{F,G} = \frac{A_F \times 1/d_{F,G}}{\sum_{i} A_{G_i} \times 1/d_{G_i}}
\]

The ABD score can be thought of as a measure of the contribution of a feature, F to the combined regulatory network on gene, G. A high ABD score may serve as a proxy for an increased specificity between a chromatin feature and the gene of interest. We projected the fine-mapped variants onto the chromatin features in different tissue types to assess whether there is an enrichment of likely causal GWAS hits in regulatory features near our putatively causative genes. Both proximity (genomic distance) and specificity (ABD scores) were considered to determine the regulatory contribution of the fine-mapped hits.

Next, we looked at the DNase I hypersensitive sites (DHSs) which are considered to be generic markers of the regulatory DNA and can contain genetic variations associated with traits and diseases\textsuperscript{39}. We downloaded the index of human DHS along with biosample metadata from \url{https://www.meuleman.org/research/dhsindex/}. The index was in hg38 coordinates which were converted to hg19 coordinates using the online version of the hgLiftOver package (\url{https://genome.ucsc.edu/cgi-bin/hgLiftOver}). We created a DHS index for each tissue type relevant to the traits and diseases we analyzed by including all DHS that are present in at least one bio-sample from a certain tissue type (Supp. File 2). We then selected DHS that lie within +/- 100 kb of the TSS of our putatively causative genes. Since DHS are of variable widths, we recentered the summits in a 350 bp window and merged overlapping sites in the same way as we did for other chromatin marks. We calculated the mean activity ($A_F$) by averaging the strengths of all the merged sites. Next, we calculated the activity by distance score for each DHS and gene pair using the same formula described above. Finally, for each fine-mapped SNP, we identified all DHS that fall within +/- 100 kb of the SNP.

Finally, we used in-silico chromatin state predictions (chromHMM core 15-state model\textsuperscript{37}) for relevant tissue types (Supp. File 1) to identify active enhancer regions in the genome. Tissue-specific chromHMM annotations were downloaded from the Roadmap Epigenomics Project\textsuperscript{37} ftp site (\url{https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/}). We considered a fine-mapped variant to fall in an enhancer region if it was housed within a chromHMM segment described as either enhancer, or bivalent enhancer, or genic enhancer. Since chromHMM annotations are not accompanied by activity measurements, the ABD approach could not be applied here.


30


http://dx.doi.org/10.1056/NEJM198603133141102


95. Makrythanasis, P. et al. A Novel Homozygous Mutation in FGFR3 Causes Tall Stature, Severe Lateral Tibial Deviation, Scoliosis, Hearing Impairment, Camptodactyly, and


106. Varon, R. *et al.* Nibrin, a Novel DNA Double-Strand Break Repair Protein, Is Mutated in


139. Caputo, V. et al. A Restricted Spectrum of Mutations in the SMAD4 Tumor-Suppressor


Goff, C. L. *et al.* Not All Floating-Harbor Syndrome Cases are Due to Mutations in Exon 34 of SRCAP. *Hum. Mutat.* **34**, 88–92 (2013).


163. Fowler, E. V. et al. ATG16L1 T300A Shows Strong Associations With Disease


173. Durocher, F. *et al.* A novel mutation in the EIF2AK3 gene with variable expressivity in


177. Froguel, P. *et al.* Familial Hyperglycemia Due to Mutations in Glucokinase -- Definition of a Subtype of Diabetes Mellitus. http://dx.doi.org/10.1056/NEJM199303113281005


193. Malecki, M. T. *et al.* Mutations in NEUROD1 are associated with the development of


226. Neavin, D. *et al.* Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome*


Bonder, M. J. et al. Identification of rare and common regulatory variants in pluripotent


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 R35GM127131, R01HG010372, R01MH101244, and U01HG012009. 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.