

Chromatin accessibility variation provides insights into missing regulation underlying immune-mediated diseases

Reviewed Preprint

v1 • June 19, 2024

Not revised

Raehoon Jeong, Martha L. Bulyk 

Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA • Bioinformatics and Integrative Genomics Graduate Program, Harvard University, Cambridge, MA 02138, USA • Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

 https://en.wikipedia.org/wiki/Open_access Copyright information

Abstract

Most genetic loci associated with complex traits and diseases through genome-wide association studies (GWAS) are noncoding, suggesting that the causal variants likely have gene regulatory effects. However, only a small number of loci have been linked to expression quantitative trait loci (eQTLs) detected currently. To better understand the potential reasons for many trait-associated loci lacking eQTL colocalization, we investigated whether chromatin accessibility QTLs (caQTLs) in lymphoblastoid cell lines (LCLs) explain immune-mediated disease associations that eQTLs in LCLs did not. The power to detect caQTLs was greater than that of eQTLs and was less affected by the distance from the transcription start site of the associated gene. Meta-analyzing LCL eQTL data to increase the sample size to over a thousand led to additional loci with eQTL colocalization, demonstrating that insufficient statistical power is still likely to be a factor. Moreover, further eQTL colocalization loci were uncovered by surveying eQTLs of other immune cell types. Altogether, insufficient power and context-specificity of eQTLs both contribute to the 'missing regulation.'

eLife assessment

This paper addresses a question regarding the low overlap between genetic variants linked to human complex diseases and variants linked to differences in gene expression. Some of the analyses supporting the main claims are **convincing**, and the key conclusions are **valuable** and of interest to readers in the fields of human genetics and functional genomics. However, chromatin accessibility QTL (caQTL) also carry the limitation of not identifying the genes that directly mediate the influence on disease phenotypes.

<https://doi.org/10.7554/eLife.98289.1.sa2>

Introduction

More than a decade of genome-wide association studies (GWAS) has revealed several properties of the genetic architecture of complex traits and diseases^{1,2}. Most (~93%) of the genetic associations are detected in the noncoding portion of the genome³, and disease heritability is concentrated in putative regulatory regions⁴. Expression quantitative trait loci (eQTLs), which are loci associated with gene expression levels, are enriched for trait associations^{5,6}. Complex traits are also characterized by their extreme polygenicity, where individual genetic association has only a small effect on the trait⁷. Altogether, these observations have led to a prevalent theory that causal genetic variants affect regulation of key genes across the genome, where each gene explains a modest proportion of trait variation⁸. There are experimental strategies aimed at nominating putative causal genes at noncoding GWAS loci^{9–11}. As an alternate approach, an eQTL signal colocalizing with the GWAS signal illustrates the effect of the causal variant on gene expression and suggests that the affected gene contributes to the trait. Detection of disease-associated eQTLs thus can identify putative disease genes, helping to elucidate disease mechanisms and develop therapeutics targeting them¹².

Although it has become expected that eQTLs will be discovered in most noncoding GWAS loci, only a minority of trait-associated loci have been explained by eQTLs^{13–16}. The Genotype-Tissue Expression (GTEx) eQTL study across 49 human tissues recognized that, for a typical complex trait, about 20% of GWAS loci contained a colocalized eQTL in the *cis* region (*i.e.*, 1 Mb) around the gene (*i.e.*, *cis*-eQTL)¹⁴. Even when focusing just on putatively causal genes, the rate of colocalization was very low (8%)¹⁵. Furthermore, the proportion of trait heritability mediated by *cis*-eQTLs ($h^2_{\text{med}}/h^2_{\text{SNP}}$) of assayed gene expression was estimated to be only about 11% on average¹⁶. We will call the missing link between genetic association to traits and regulatory function of the associated noncoding variants as ‘missing regulation’, as Connally and colleagues introduced¹⁵. To be able to detect eQTLs in the unexplained disease-associated loci, a better understanding of the possible reasons for why they have been missing is essential.

There are many possible explanations for why disease-associated loci are missing colocalized eQTLs^{15,17–19} and for why $h^2_{\text{med}}/h^2_{\text{SNP}}$ estimates for eQTLs are relatively low¹⁶. First, statistical power to detect disease-associated eQTLs may be insufficient¹⁹. For example, negative selection against gene expression variation may lead to challenges in detecting eQTLs for trait-relevant genes^{18,20}. Since disease genes are likely to be dosage sensitive, there would be selection against their having large eQTL effects²⁰. Consequently, only weak eQTL variants, often in regions distal to the gene’s promoter²¹, may remain and affect traits. In fact, eQTL-mediated heritability was enriched in genes showing mutational constraint and those with lower *cis*-heritability¹⁶. These weaker eQTL effects would require larger sample sizes to be detected with statistical significance and to show colocalization¹⁹. Second, causal eQTL effects may be specific to cell types that have not been assayed¹⁷. For example, immune cell eQTLs are highly cell-type specific, and eQTL effects specific to some immune cell types may mediate immune disease risk²². Specificity of eQTL effects can also be limited to specific cell states^{23,24}. Detecting cell-type or cell-state specific eQTL effects requires the necessary gene expression data sets from the relevant cell types and states¹⁷, which has been a limiting resource for such analyses.

To investigate why disease-associated eQTL signals have been missing, we focused on immune-mediate diseases (IMDs) as a model set of complex traits. We aimed to collect IMD-associated loci that are expected to show eQTL signals in some cell type. Since active regulatory elements coordinate target gene expression²⁵, we reasoned that variants that affect chromatin phenotypes at regulatory elements, such as transcription factor (TF) binding^{26–28} and chromatin accessibility^{29,30}, have the potential to impact gene expression³¹. These

chromatin phenotypes may show detectable genetic effects even when an eQTL effect in the same cell type was not identified in the locus³². For example, only about 20% of lymphoblastoid cell lines' (LCLs') PU.1 binding QTLs (bQTLs) that colocalized with blood cell traits' association showed an eQTL effect for a nearby gene in LCLs³³.

Here, we analyzed genetic and functional genomic (*i.e.*, ATAC-seq and RNA-seq) data in LCLs. LCLs are derived from B lymphocytes, and their *cis*-regulatory elements were enriched for variants associated with some IMDs^{34,35}. We evaluated whether chromatin accessibility QTLs (caQTLs) in LCLs potentially explain IMD associations using mediated heritability analysis¹⁶ and colocalization^{36,37}. Then, we searched for disease-associated loci that were significant caQTLs, but not eQTLs.

We examined whether the various potential reasons for missing eQTLs can account for IMD-associated loci that are explained by caQTLs but not eQTLs. First, we explored the extent to which eQTLs may have been missed because of limited statistical power. We compared *cis*-heritability of colocalized caQTLs and eQTLs stratified by distance between the accessible region and the transcription start site of the associated gene. We also investigated whether meta-analysis of published LCL eQTL summary statistics, in order to effectively increase the sample size, can uncover previously missed eQTLs. Second, we surveyed whether cell-type specificity of regulatory variant effect may account for the missing regulation. We surveyed various immune cell eQTL data to identify loci with which they colocalize even if LCL eQTLs did not colocalize with those loci.

Through this study, we present how regulatory QTLs beyond eQTLs, such as caQTLs, can be effective in detecting the potential molecular consequences of disease-associated variants. Moreover, results from inspecting disease-associated loci where genetic effects are detected on chromatin accessibility but not on expression suggest reasons why the effects on gene expression may have been missed. These results provide insights on which strategies may be effective in uncovering more genes that underlie diseases.

Results

Accessible chromatin in LCLs explains a significant proportion of immune-mediated disease heritability

We aimed to evaluate whether variants that alter chromatin accessibility in LCLs may explain genetic associations to IMD. First, we verified whether accessible regions in LCLs are enriched for IMD heritability. We reanalyzed 100 LCL ATAC-seq samples³⁰ to define accessible regions in this cell type. With stratified LD score regression (S-LDSC)³⁸, we estimated their heritability enrichment across 13 IMDs, including 11 autoimmune diseases — autoimmune thyroid disease (ATD)³⁹, celiac disease (CEL)⁴⁰, Crohn's disease (CD)⁴¹, inflammatory bowel disease (IBD)⁴¹, juvenile idiopathic arthritis (JIA)⁴², multiple sclerosis (MS)⁴³, primary biliary cholangitis (PBC)³⁹, rheumatoid arthritis (RA)⁴⁴, systemic lupus erythematosus (SLE)⁴⁵, ulcerative colitis (UC)⁴¹, and vitiligo (VIT)⁴⁶ — and 2 allergic diseases — allergy (ALL)⁴⁷ and asthma (AST)⁴⁷. We also analyzed GWAS data for 3 non-immune diseases — type 2 diabetes (T2D)⁴⁸, coronary artery disease (CAD)⁴⁹ and schizophrenia (SCZ)⁵⁰ — for comparison. Single nucleotide polymorphisms (SNPs) in accessible regions in LCLs were significantly enriched for IMD heritability ($p < 0.003125$ (Bonferroni-corrected), S-LDSC; **Figure 1A**) and there was no significant enrichment for non-immune diseases ($p > 0.05$, S-LDSC). These results indicate that accessible regions in LCLs harbor many variants specifically associated with IMDs, and therefore that LCLs share IMD-associated accessible regions with those of the causal cell type(s).

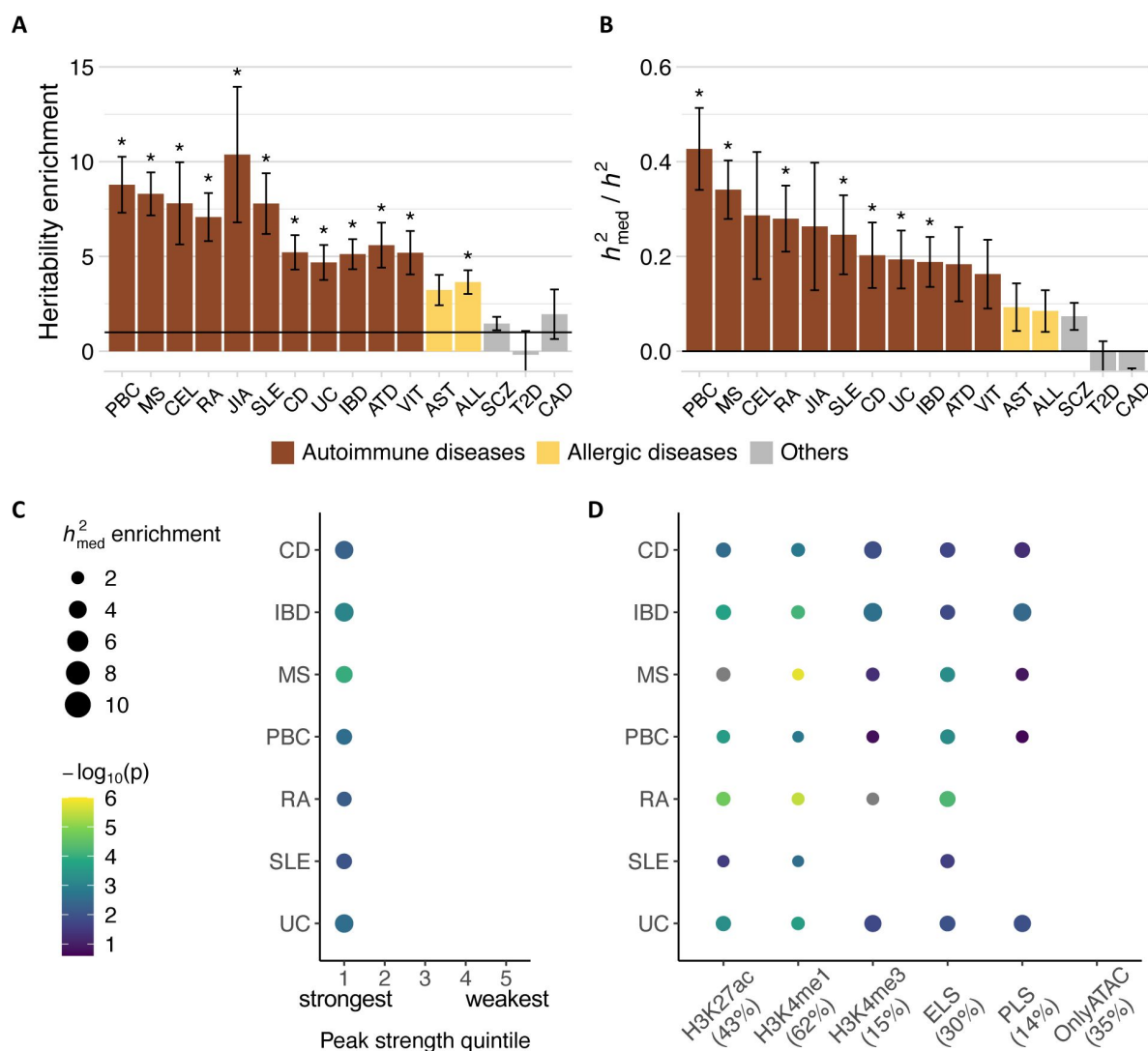


Figure 1.

Immune disease heritability mediated by chromatin accessibility in LCLs.

(A–B) Heritability enrichment in accessible regions in LCLs, based on (A) stratified LD score regression (S-LDSC) and (B) the proportion of heritability mediated by caQTLs in LCLs based on mediated expression score regression (MESR). For both, error bars represent jackknife standard errors of the mean. The color of the bars indicates disease type. *: $p < 0.003125$ (Bonferroni-corrected).

(C) Mediated heritability enrichment of accessible regions by peak strength quintile. The strongest peaks are in the 1st quintile, and the weakest peaks are in the 5th quintile. Only enrichment values with FDR < 5% (based on q-value) are shown. (D) Mediated heritability enrichment of accessible regions by histone mark annotation. The percentages in parentheses represent the proportion of accessible regions with the indicated histone mark. Only enrichment values with FDR < 5% (based on q-value) are shown. Color and size of the points are on the same scale as in (C). PBC, primary biliary cholangitis; MS, multiple sclerosis; CEL, celiac disease; RA, rheumatoid arthritis; JIA, juvenile idiopathic arthritis; SLE, systemic lupus erythematosus; CD, Crohn's disease; UC, ulcerative colitis; IBD, inflammatory bowel disease; ATD, autoimmune thyroid disease; VIT, vitiligo; AST, asthma; ALL, allergies; SCZ, schizophrenia; T2D, type 2 diabetes; CAD, coronary artery disease.

Next, we applied mediated expression score regression (MESC)¹⁶ to investigate the causal relationship between caQTLs in LCLs and IMD associations. Compared to S-LDSC analysis that tests for heritability enrichment of SNPs with some functional annotation (e.g., accessible regions), MESC analysis specifically estimates the heritability that is mediated (i.e., h^2_{med}) by the SNPs' cis-effects on a molecular phenotype (e.g., caQTLs). We estimated that caQTLs in LCLs mediate 16.3–42.7 % of autoimmune disease heritability and 8.5–9.4 % of allergic disease heritability (**Figure 1B**). For non-immune diseases, the estimates were lower and not significant ($p > 0.003125$ (Bonferroni-corrected), MESC). Interestingly, schizophrenia showed a nominally significant proportion of caQTL-mediated heritability in LCLs ($p < 0.05$, MESC), consistent with the hypothesis that B cells may play some role in schizophrenia pathogenesis^{50,51}. Our results indicate that LCLs are a valid cell type in which to search for caQTLs that mediate genetic risk for 7 IMDs — CD, IBD, MS, PBC, RA, SLE, and UC — but not for allergic diseases. In subsequent analyses, we focused on 7 IMDs — CD, IBD, MS, PBC, RA, SLE, and UC — that showed significant caQTL-mediated heritability ($p < 0.003125$ (Bonferroni-corrected), MESC).

Regions with higher levels of accessibility and active histone marks explain most of caQTL-mediated heritability

To understand which features characterize accessible regions that mediate IMD heritability, we estimated h^2_{med} enrichment (proportion of h^2_{med} / proportion of peaks)¹⁶ in specific sets of accessible regions. We found that peaks with a larger number of nonredundant sequencing reads (i.e., “stronger” peaks) in LCLs showed stronger h^2_{med} enrichment (**Figure 1C**) and thus likely affect IMD-relevant gene expression more than “weaker” peaks do. This observation is consistent with the “Activity-by-Contact” model⁹, in which peaks with greater chromatin accessibility and H3K27ac ChIP-seq signal are predicted to have proportional effects on target gene expression.

Next, we considered peaks with the active histone marks H3K27ac, H3K4me1, or H3K4me3^{28,52}. Consistent with prior observations that putative cell-type-specific regulatory elements marked with H3K27ac and H3K4me1 are enriched for relevant disease associations^{34,35}, we found that caQTLs with H3K27ac and H3K4me1 marks in LCLs were significantly enriched for mediated IMD heritability (q-value < 0.05 , MESC; **Figure 1D**). Strikingly, both peak sets explained almost all of caQTL-mediated IMD heritability (**Supplementary Figure 1**). Peaks with H3K4me3 marks, representative of promoters⁵³, also showed significant h^2_{med} enrichment for most IMDs (q-value < 0.05 , MESC). Peaks with promoter-like signatures (i.e., H3K27ac and H3K4me3)⁵⁴ and those with enhancer-like signatures (i.e., H3K27ac, but no H3K4me3)⁵⁴ were also enriched for all IMD heritability (q-value < 0.05 , MESC). Conversely, peaks without any of the three active histone marks were completely depleted of caQTL-mediated IMD heritability (**Supplementary Figure 1**). These ‘ATAC only’ peaks were shorter, weaker, and further away from the TSS compared to peaks with active histone marks (**Supplementary Figure 2**). Altogether, these results indicate that peaks characterized as putatively active regulatory elements explain nearly all of caQTL-mediated IMD heritability.

caQTLs share IMD heritability with eQTLs and explain more of IMD heritability than do eQTLs

The model that gene regulatory activity explains a significant fraction of noncoding genetic associations to IMDs is supported by our findings that caQTLs mediate a significant proportion of IMD heritability and that those with active histone marks show strong h^2_{med} enrichment. This is in contrast with relatively low average h^2_{med} / h^2_{SNP} estimates ($\sim 11\%$) previously having been observed for eQTLs across 48 human tissues in GTEx and various human traits⁵⁵. To directly compare the proportion of IMD heritability mediated by caQTLs and eQTLs in the same cell type (i.e., LCLs), we additionally applied MESC to gene expression data from LCLs (i.e., Geuvadis data)⁵⁶.

Across the 7 autoimmune diseases, the estimated proportion of heritability mediated by eQTLs ($h^2_{\text{med; eQTL}} / h^2_{\text{SNP}}$) ranged from 9% to 22% (**Figure 2A**). For all 7 diseases, we estimated that eQTLs mediated less heritability than did caQTLs, even though the caQTLs' smaller sample size would potentially bias the estimates toward zero.¹⁶ A possible explanation is that some IMD-associated regulatory variants may show detectable effects on chromatin accessibility, but not on gene expression, in LCLs at the current sample size ($n = 373$); such loci may account for the missing regulation.¹⁵

We anticipated that IMD-associated variants that affect gene expression in *cis* do so by modulating regulatory element's activity. Therefore, we investigated whether eQTL-mediated IMD heritability is shared by caQTL-mediated signals (**Figure 2B**). We performed MESC on both caQTLs and eQTLs together to estimate the amount of IMD heritability mediated by both collectively (**Methods**). For the 7 IMDs, the combined $h^2_{\text{med; caQTL} \cup \text{eQTL}} / h^2_{\text{SNP}}$ was only slightly higher (2.2–9.0 %) than the estimates for just caQTLs ($h^2_{\text{med; just caQTL}} / h^2_{\text{SNP}}$; **Figure 2A**), suggesting that approximately 56–82% of eQTL-mediated heritability is shared with caQTL-mediated heritability (i.e., $h^2_{\text{med; caQTL} \cap \text{eQTL}} / h^2_{\text{med; eQTL}}$; **Supplementary Figure 3**). These estimates are consistent with substantial sharing of caQTL- and eQTL-mediated IMD heritability. Nevertheless, 9–27% of IMD heritability is explained just by caQTLs, while only 2–9% of IMD heritability is explained just by eQTLs.

Many IMD-associated loci show colocalization with caQTLs but not with eQTLs

We applied colocalization analysis³⁷ to identify IMD-associated loci that share genetic signals with caQTLs or eQTLs in LCLs. We selected candidate loci of 200-kb windows for each IMD with the following conditions: (1) lead IMD association at $p < 10^{-6}$, (2) lead caQTL or eQTL association at $p < 10^{-4}$, and (3) at least one variant simultaneously showed caQTL or eQTL χ^2 statistics greater than $0.8 \times$ lead χ^2 statistics for the caQTL or eQTL, respectively, and IMD association χ^2 statistics greater than $0.8 \times \chi^2$ statistics for the IMD lead variant in the locus. We applied gwas-pw³⁷ and considered loci with posterior probability of colocalization (PPA3) > 0.98 to be colocalized⁵⁷. Some loci colocalized with only either a caQTL or an eQTL, while others colocalized with both (**Figure 2C** and **Supplementary Tables 1-2**).

To confirm that the colocalized genes are relevant to IMD, we tested for their enrichment of Gene Ontology (GO) annotation terms for specific biological processes⁵⁸. Considering all genes within 500 kb of the IMD GWAS lead variants at colocalized loci as background, the genes that showed eQTL colocalization for any IMD were enriched for various immune responses and signaling processes, such as “positive regulation of immune system process” and “regulation of lymphocyte activation” (**Supplementary Figure 4**), indicating that the colocalized genes in LCLs are involved in immune function. For example, *IL6R* and *IL12A* encode direct or indirect targets of approved drugs – Tocilizumab and Ustekinumab – for autoimmune diseases like RA⁵⁹ and CD⁶⁰. These two genes showed colocalization with both caQTLs and eQTLs in CD and PBC GWAS, respectively (**Supplementary Figure 5A-B**). Increased *IL6R* expression was associated with higher risk for CD, and increased *IL12A* expression was associated with lower risk for PBC and SLE. The former observation is in line with Tocilizumab, a monoclonal antibody to IL-6 receptor, showing efficacy in CD patients⁵⁹, although it is not pursued for approval because of potential side effects⁶¹. Interestingly, *IL6R* and *IL12A* eQTLs did not colocalize with the association signals of RA and CD, respectively, which are the diseases for which these drugs are approved (**Supplementary Figure 5C-D**). Moreover, *ELMO1*, which previously had not been associated with autoimmune diseases, showed eQTL colocalization with RA, PBC, and MS association signals (**Figure 2D**). In all three, decreased *ELMO1* expression was associated with increased disease risk. In mice, *Elmo1* was required for polarization and migration of B and T lymphocytes⁶².

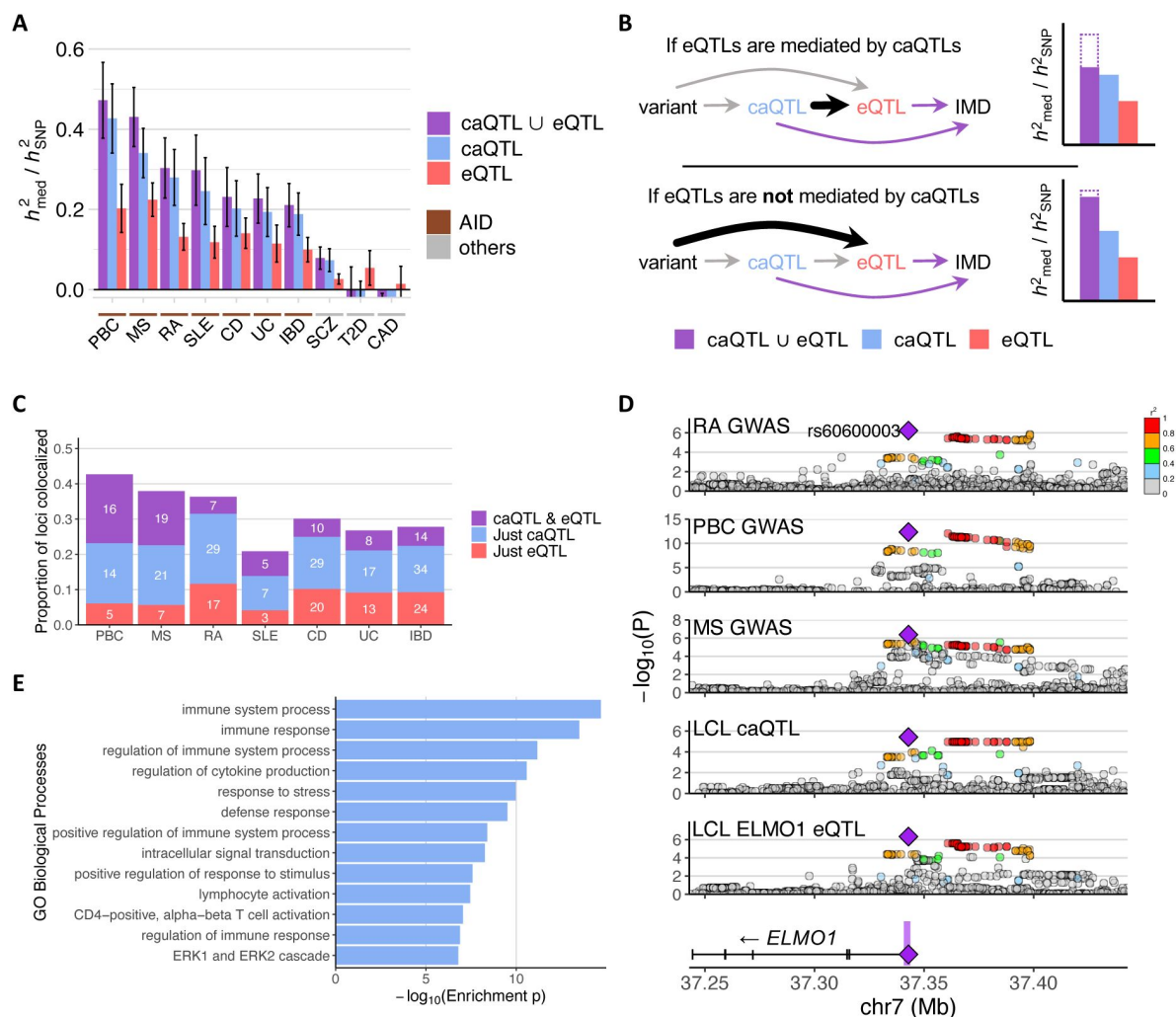


Figure 2.

IMD heritability mediated by caQTLs and eQTLs.

(A) h^2_{med} / h^2_{SNP} estimates of various IMDs for caQTLs, eQTLs, and their union. AID: autoimmune disease. Disease abbreviations along the x-axis are as in [Figure 1](#). (B) Schema of the potential causal relationships between genetic variants, caQTLs, eQTLs, and IMD risk. The two diagrams depict possible h^2_{med} / h^2_{SNP} trends depending on the causal relationship between caQTLs and eQTLs. (C) Number of IMD-associated loci colocalized with caQTLs or eQTLs. The proportion is out of the total number of IMD-associated ($p < 10^{-6}$) loci. (D) *ELMO1* locus plot showing association to RA, PBC, MS, chromatin accessibility, and *ELMO1* expression in LCLs. Purple shading in the gene plot at the bottom indicates the caQTL peak, and the purple diamond is the lead variant (rs60600003) that is within that peak. The other variants are colored by the degree of linkage disequilibrium (LD) with the annotated variant. (E) Enrichment of the Biological Process Gene Ontology (GO) terms of genes in proximity to IMD-colocalized caQTLs without eQTL colocalization.

Across the IMDs, there were many loci that colocalized with a caQTL but not with an eQTL (**Figure 2C**). These “caQTL only” loci showed enrichment for immune response genes in *cis* compared to all accessible regions in LCLs^{63,64}, even though the colocalized eQTLs were enriched for immune response genes as well (**Figure 2E** and **Supplementary Figure 4**), indicating that IMD-relevant genes without eQTL colocalization in Geuvadis LCL data⁵⁶ are likely found in these loci.

Distance to transcription start sites affect eQTLs but not caQTLs

Why might there be loci with caQTL colocalization only, despite the caQTL data having fewer samples than the eQTL data (100 vs. 373)? Limited statistical power can prevent some eQTLs from being detected and showing significant colocalization¹⁹. As hypothesized by Mostafavi and colleagues, disease-relevant eQTLs may be weaker and more distal¹⁸. To understand the extent to which this effect may result in many loci showing colocalization only with caQTLs, we compared the *cis*-heritability (h^2_{cis}) of caQTLs and eQTLs depending on the distance from the ATAC peak to the transcription start site (TSS) of the gene (*i.e.*, peak-to-TSS distance). We considered all caQTLs and eQTLs regardless of disease association.

We identified pairs of caQTLs and eQTLs that colocalized with each other³⁷, which implies that the regulatory variant modulating chromatin accessibility also affects gene expression. Then, the distance between the ATAC peak and the TSS of the eQTL gene (*i.e.*, eGene) is the distance between a regulatory element and its target gene's TSS. We stratified the pairs into peak-to-TSS distance quintiles and compared the eQTL h^2_{cis} distribution of the first quintile (*i.e.*, closest pairs) with that of the later quintiles. We observed that eQTL h^2_{cis} distribution decreased with increasing distance of the paired ATAC peaks from the TSS ($p = 1.0 \times 10^{-4}$, 2.1×10^{-10} , 1.6×10^{-15} , and 1.7×10^{-20} , respectively, one-sided Wilcoxon rank-sum test; **Figure 3A**), consistent with the negative relationship between promoter-enhancer genomic distance and impact on gene expression^{65,66}. This result also explains why discovered eQTLs are concentrated near the promoter, where the variants are more likely to show stronger effects⁶⁷. In contrast, caQTL h^2_{cis} distribution was similar across peak-to-TSS distances ($p > 0.05$, one-sided Wilcoxon rank-sum test; **Figure 3A**). For all distance quintiles, caQTL h^2_{cis} was significantly higher than that of the paired eQTLs ($p = 4.0 \times 10^{-23}$, 1.3×10^{-24} , 1.5×10^{-19} , 1.1×10^{-28} , and 2.4×10^{-49} , respectively, one-sided paired Wilcoxon rank-sum test; **Figure 3A**), and the contrast between them was greater at more distant quintiles, suggesting that the statistical power to detect and colocalize eQTLs is increasingly lower than that for caQTLs for regulatory effects far from the TSS.

Overall, caQTL h^2_{cis} had a significant positive effect ($p = 6.6 \times 10^{-11}$, linear regression) and peak-to-TSS distance had a negative effect ($p = 2.4 \times 10^{-23}$, linear regression) on eQTL h^2_{cis} (**Figure 3B**). Thus, for regulatory variants that showed both caQTL and eQTL signals, those with larger effects on chromatin accessibility tended to exhibit larger effects on gene expression, but their eQTL effects diminished with increasing distance from TSSs.

Next, we investigated how caQTL-mediated IMD heritability is distributed with respect to TSS. If caQTLs beyond the typical *cis*-eQTL window of 1 megabase (Mb) around the genes' TSS explain some proportion of IMD heritability, then *cis*-eQTL analyses might require a wider window to detect disease-associated eQTLs. Across the 7 diseases, caQTLs within 500 kb of the TSS of expressed genes explained almost all of the caQTL-mediated IMD heritability (92–100%; **Figure 3C**), indicating that regulatory variants are most likely within 500 kb of the target gene's TSS and supporting the use of a 1 Mb window for *cis*-eQTL analyses. Depending on the disease, 41–66% of the caQTL-mediated IMD heritability was detected in distal peaks further than 10 kb from the TSS of expressed genes, further supporting the analysis of regulatory variants beyond promoter regions.

In sum, the power to detect eQTLs diminishes with increasing distance of the variant from the TSS, but the power to detect caQTLs is largely invariant regardless of peak-to-TSS distance (**Figure 3D**). Since $h^2_{med; caQTL}$ are distributed mostly within 500 kb of genes' TSS, the IMD loci

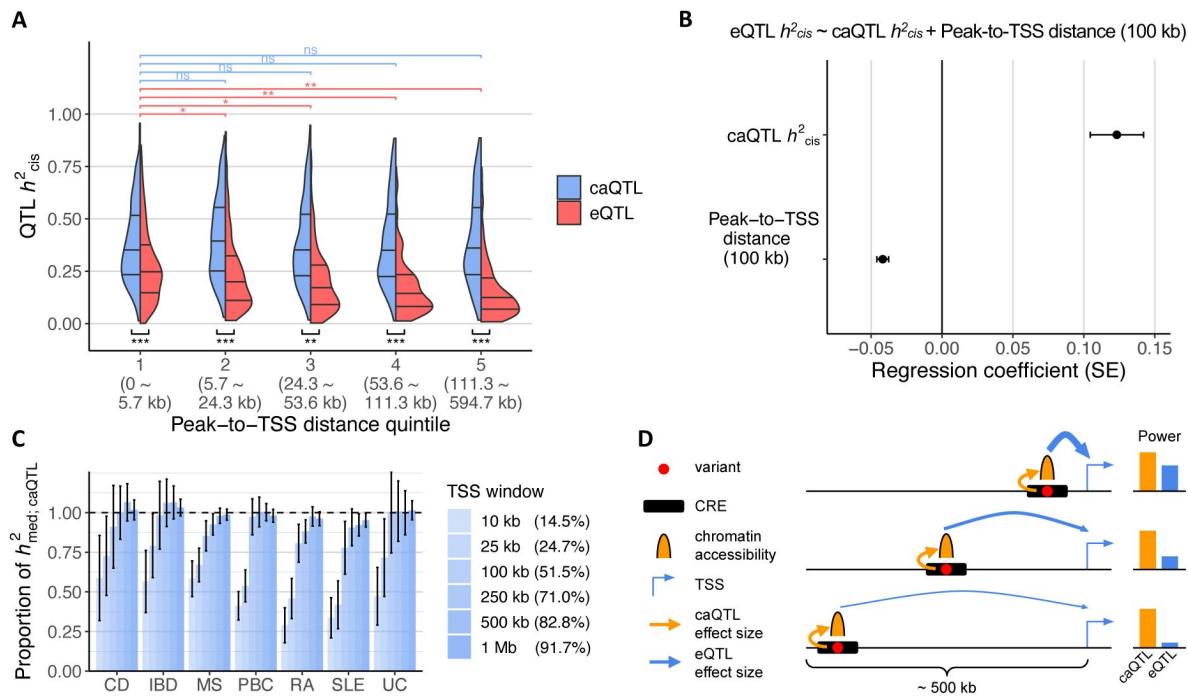


Figure 3.

Effect of peak-to-TSS distance on *cis*-heritability of caQTLs and eQTLs and IMD heritability.

(A) Distribution of *cis*-heritability (h^2_{cis}) of caQTLs and eQTLs by peak-to-TSS distance quintiles. The ranges of peak-to-TSS distance are shown in parentheses. The comparisons shown on the top (in respective colors) are between the nearest and each subsequent quintile of the respective QTL h^2_{cis} distribution (i.e., one-sided Wilcoxon rank-sum test). The comparisons shown on the bottom (in black) are between caQTL and eQTL h^2_{cis} distribution (one-sided paired Wilcoxon rank-sum test). *: $p < 10^{-4}$, **: $p < 10^{-10}$, ***: $p < 10^{-20}$, and ns: $p > 0.05$. (B) Regression estimates and their standard errors of the linear regression model testing the effects of caQTL h^2_{cis} and peak-to-TSS distance on eQTL h^2_{cis} . Peak-to-TSS distance was expressed in units of 100 kb to neatly visualize the effect size estimates. SE: standard error. (C) Proportion of caQTL-mediated IMD heritability explained by ATAC peaks within various TSS windows. Percentage for each TSS window denotes the proportion of ATAC peaks in that window. Disease abbreviations along the x-axis are as in Figure 1. (D) A model of the relationship between peak-to-TSS distance and power to detect a corresponding caQTL or eQTL. The thickness of the arrows indicates the variant effect size on chromatin accessibility (yellow) or gene expression (blue). TSS, transcription start site; CRE, *cis*-regulatory element.

colocalizing only with caQTLs could still be weak, undetected eQTLs. Under this model, we predicted that increasing the power to detect eQTLs in LCLs, such as increasing sample size, may lead to further eQTL colocalizations in loci in which we observed only caQTL colocalization.

Increasing the sample size reveals some eQTL colocalization

For genetic association studies, increasing the sample size is a way to increase statistical power. Therefore, we meta-analyzed four LCL eQTL summary statistics^{55,56,68,69}, leading to a total sample size of 1,128 individuals. We performed colocalization analysis using the meta-analyzed summary statistics to evaluate whether effectively increasing the sample sizes would uncover more disease-associated eQTLs in LCLs, especially in loci where a caQTL already showed IMD colocalization. Up to 6 additional loci showing eQTL colocalization were thus detected for each IMD (**Figure 4A** and **Supplementary Table 3**). For example, *CIITA* is the class II major histocompatibility complex transactivator, which causes severe immunodeficiency if dysfunctional⁷⁰. The *CIITA* locus is associated with IBD right below the genome-wide significance level ($rs10445003$, $p = 7.5 \times 10^{-8}$), and it colocalized with a caQTL signal, but initially not with any eQTL (**Supplementary Figure 6A**). However, the meta-analyzed statistics showed a stronger association to *CIITA* expression ($p = 6.7 \times 10^{-8}$) than without meta-analysis ($p = 5.4 \times 10^{-4}$) and exhibited a significant colocalization. Interestingly, two of the causal CD genes that previously lacked colocalized eQTLs¹⁵, *CARD9* and *ATG16L1*, showed significant colocalization in the meta-analyzed LCL eQTL data (**Supplementary Figure 6B-C**).

We hypothesized that increased sample size would improve the power to detect weaker and distal eQTL colocalization. Comparison of the accessibility peak's distance to the paired eQTL gene's (eGene) TSS showed that the newly detected eQTLs tended to be more distal ($p = 0.06$, one-sided Wilcoxon rank-sum test; **Figure 4B**). However, compared to the distribution of the peak-to-TSS distance for all caQTL-eQTL pairs showing colocalization, IMD-associated loci that showed caQTL and eQTL colocalization had greater peak-to-TSS distance on average ($p = 0.002$ for Geuvadis and $p = 5.9 \times 10^{-6}$ for the meta-analyzed data; **Figure 4B**). For example, a RA-associated locus near *POU3F1*, a neuronal transcription factor that is also induced by interferon⁷¹, colocalized with a distal eQTL located about 126 kb upstream of its promoter, after meta-analysis strengthened the eQTL association ($p < 10^{-10}$; **Figure 4C**). These results suggest that additional IMD-associated loci with distal, weaker eQTLs in LCLs might be found if eQTL data were generated for a larger number of individuals.

IMD loci that colocalized with caQTLs but not eQTLs showed lower levels of active histone marks

Despite uncovering more eQTL colocalizations through meta-analysis, more than 40% of the caQTL-colocalized loci nevertheless showed no colocalization with an eQTL in LCLs (**Figure 4B**). We investigated whether these “caQTL only” peaks might be inactive regulatory elements. We quantified the active histone mark levels for H3K27ac, H3K4me1, and H3K4me3 at colocalized caQTL peaks in LCLs and then compared their levels between the “caQTL and eQTL” and “caQTL only” loci. On average, H3K27ac marks were stronger at “caQTL and eQTL” peaks, supporting that the corresponding regulatory elements might be more active (**Supplementary Figure 7**). H3K4me3 marks were detected more often at “caQTL and eQTL” peaks, leading to stronger average signal. In contrast, H3K4me1 levels were highly similar between the two sets of peaks. Although “caQTL only” peaks generally showed lower levels of active histone marks, several individual “caQTL only” peaks showed comparable levels. These peaks could be inactive *cis*-regulatory elements in LCLs that affect gene expression in a different cellular context. Therefore, we next examined whether those caQTLs appear as eQTLs in other immune cell types.

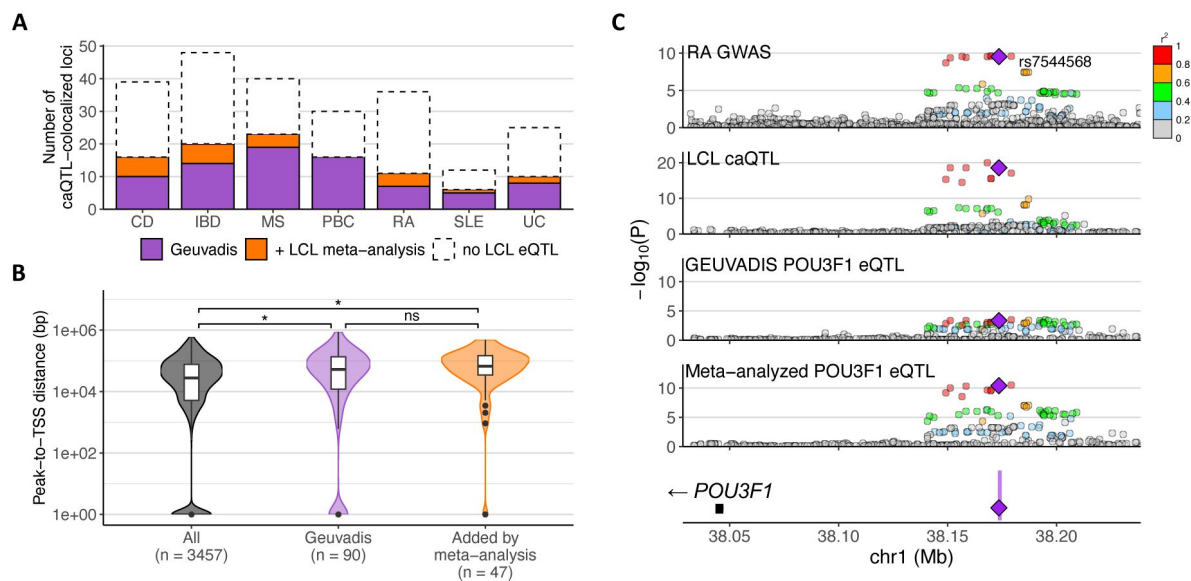


Figure 4.

Additional colocalization of caQTL-colocalized IMD loci with meta-analyzed LCL eQTL data.

(A) Number of caQTL-colocalized IMD loci that showed eQTL colocalization in LCLs. Disease abbreviations along the x-axis are as in [Figure 1](#). (B) Distribution of peak-to-TSS distance of all caQTL-eQTL pairs and of those colocalized with IMD association. The number of loci in each category is shown in parentheses. *: $p < 0.01$, ns: $p > 0.05$. (C) *POU3F1* eQTL that became significantly colocalized with RA association by meta-analyzing LCL eQTL data. Purple shading in the gene plot at the bottom indicates the caQTL peak, and the purple diamond is the lead variant (rs60600003) that is within that peak. The other variants are colored by the degree of linkage disequilibrium (LD) with the annotated variant.

Various immune cell types exhibit eQTL colocalization where LCLs did not

We downloaded eQTL summary statistics generated from 26 naïve and stimulated immune cell types^{22,23,72–74} to search for eQTLs that may correspond to the remaining, IMD-colocalized caQTLs (**Supplementary Table 4**). The profiled cell types range from B cells and monocytes to subtypes of T cells, as well as stimulated T cells and macrophages. We tested for colocalization of these eQTLs to IMD associations.

25–42% of the caQTL-colocalized IMD loci that were missing eQTL colocalizations in LCLs showed eQTL colocalizations in at least one immune cell type (**Figure 5A** and **Supplementary Table 5**). The overlap of LCL caQTLs with non-LCL immune cell eQTLs was greater than expected by chance for 5 of the 7 IMDs ($p < 0.00714$ (Bonferroni-corrected), Fisher’s exact test, for CD, IBD, PBC, RA, and UC; **Supplementary Table 6**), suggesting that the caQTLs found in LCLs may also show regulatory function in those immune cell types. Comparing across the datasets, we found that the number of loci with eQTL colocalization varied depending on the cell type, but that the effect of the sample size was more profound ($r^2 = 0.60–0.79$; **Figure 5B** and **Supplementary Table 7**). We meta-analyzed eQTL data of 3 immune cell types with multiple sources – naïve CD4⁺ T cell, monocyte, and memory regulatory T cell (Treg) – and this also increased the number of loci with significant eQTL colocalization (**Supplementary Table 8**). Altogether, these results suggest that although generating eQTL data in more cell types and cell states uncovers context-specific eQTLs, increasing the sample size should also be a priority to ensure sufficient statistical power.

We investigated the potential reasons why IMD loci that colocalized with caQTLs in LCLs showed eQTLs not in LCLs but in other immune cells. First, LCL caQTLs may correspond to gene regulatory elements that exert their effects on gene expression in a different cellular context. For instance, monocyte H3K27ac levels in the “caQTL only” loci where monocyte eQTLs colocalized were higher than those with no monocyte eQTLs (**Supplementary Figure 8**). Second, some examples, such as for *TNFSF15*, were due to cell-type specific gene expression (**Figure 6**): despite a significant colocalization of IBD with a caQTL in LCLs in the *TNFSF15* locus, disease-associated eQTL signal was detected only in monocytes (**Figure 6A**). Tumor necrosis factor-like cytokine 1A (TL1A), the protein encoded by *TNFSF15*, is secreted by monocytes and many other cells to activate helper T cells, regulatory T cells, and B cells⁷⁵. *TNFSF15* expression was low in LCLs (mean transcript per million [TPM] = 0.30), but higher in monocytes (mean TPM = 2.23). Of the genes that showed exclusively monocyte eQTL colocalization, those with low expression (mean TPM < 1) in LCLs generally showed higher expression in monocytes (**Figure 6B–C**).

However, low expression in LCLs was likely not the explanation for most cases of “monocyte only” eQTLs (blue points in **Figure 6B**) because most “monocyte only” eQTL genes were expressed at level higher than 1 TPM in LCLs.

Overall, expanding the eQTL search to various immune cell types increased the number of eQTL-colocalized loci among those that previously colocalized only with caQTLs (**Figure 5A**). On average, approximately 75% of the caQTL-colocalized loci ultimately showed eQTL colocalization. These results highlight the utility of eQTL data across a range of immune cell types for discovery of IMD-associated eQTLs.

Discussion

A lack of the link between many noncoding GWAS loci to the associated variants’ gene regulatory effects have posed challenges in understanding their genetic mechanism. A better evaluation of the potential reasons for the missing regulation will guide future data generation projects to

Figure 5.

Added utility of various immune cell eQTL data.

(A) Number of loci that additionally colocated with eQTLs by LCL meta-analysis (orange) and immune cell data (cyan) compared to the original analysis with Geuvadis LCL eQTL data (purple). The height of the bar is the proportion of loci with eQTL colocalization out of the total IMD loci with caQTL colocalization in the earlier analysis. Disease abbreviations along the x-axis are as in Figure 1. (B) Relationship between the number of MS GWAS loci with eQTL colocalization and sample size for each eQTL dataset. Meta-analyzed eQTL data are labeled with their cell types. NK cell, natural killer cell; Treg, regulatory T cell.

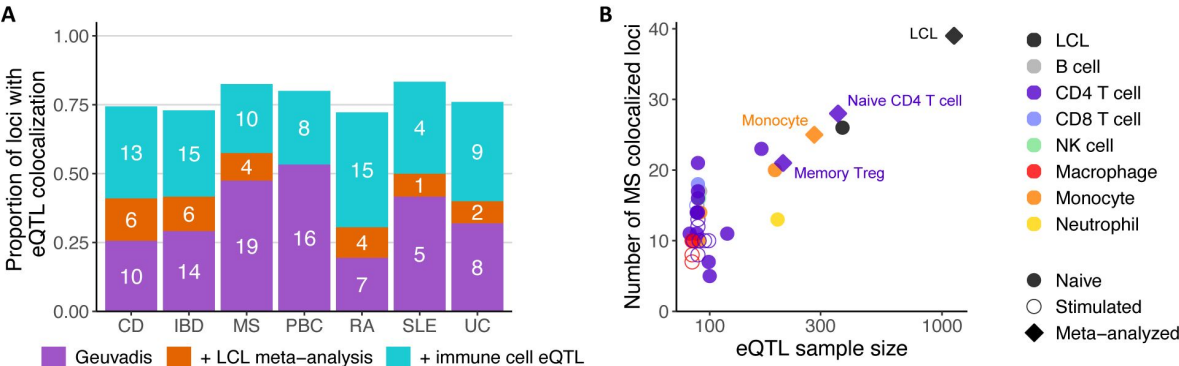
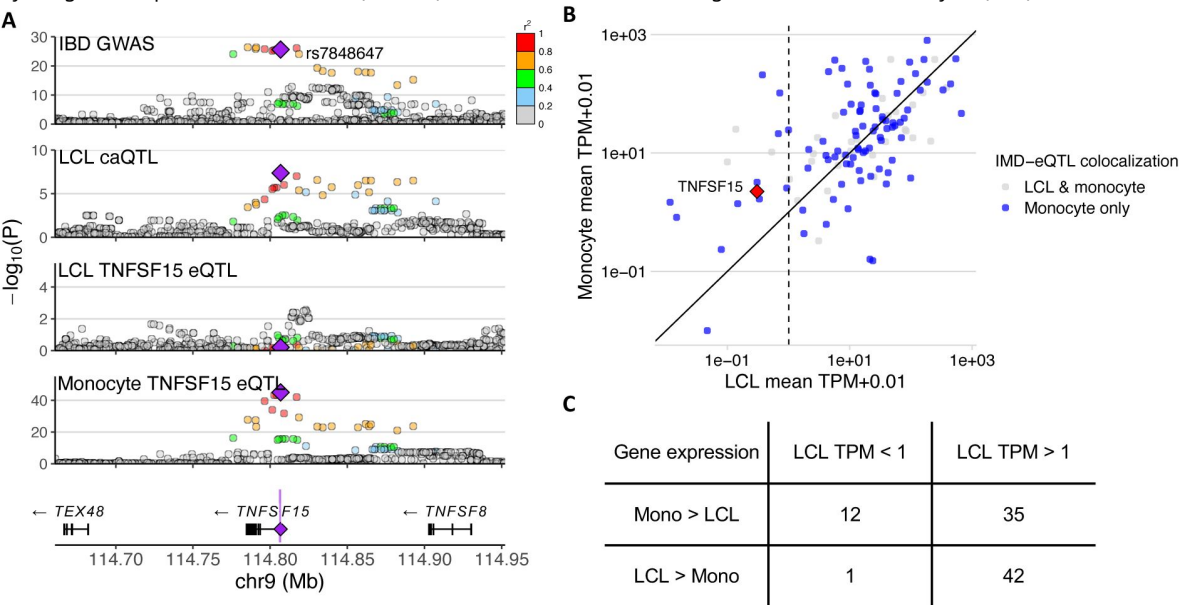


Figure 6.

IMD loci that colocated with an eQTL in monocytes, but not in LCLs, even though they colocated with caQTLs in LCLs.

(A) *TNFSF15* locus plot showing genetic association to IBD, chromatin accessibility in LCLs, and *TNFSF15* expression in LCLs and monocytes. Purple shading in the gene plot at the bottom indicates the caQTL peak, and the purple diamond shows a strongly associated variant (rs7848647) that is within that peak. The other variants are colored by the degree of LD with the annotated variant. (B) Expression levels of genes in LCLs and monocytes colored according to their eQTL colocalization outcome. TPM: transcripts per million. (C) Number of genes with eQTL colocalization in monocytes, but not in LCLs, separated by the gene's expression level in LCLs (column) and whether it is lower or higher than that in monocytes (row).



elucidate disease-associated loci. To determine why some loci might lack colocated eQTLs, we focused on chromatin accessibility, which is a molecular phenotype affected by regulatory variants more directly than are eQTLs. We approached this question with mediated heritability analysis¹⁶ and colocalization analysis^{36,37}.

We found that caQTLs in LCLs mediate a significant proportion of heritability for many autoimmune diseases. In contrast, LCLs did not appear to be an effective cell type to model gene regulatory effects in allergic diseases. The h^2_{med} / h^2_{SNP} estimates for caQTLs were higher than those of eQTLs in most autoimmune diseases, even though the smaller sample size of caQTL data (i.e., 100 vs. 373) could bias caQTLs' estimates toward zero¹⁶. We also showed that disease-associated chromatin accessibility effects often share the genetic signal with gene expression effects but that there are also many loci without an eQTL detected in LCLs.

By focusing on disease-associated caQTL that lacked significant eQTLs, we explored how additional colocated eQTL effects could be uncovered. First, increasing the sample sizes for the eQTL statistics via meta-analysis demonstrated that more eQTL colocalizations can be detected with increased statistical power and robustness¹⁹. These results are consistent with the hypothesis that disease-associated eQTLs are typically weaker and distal due to negative selection against large expression changes for causal genes¹⁸. Second, many caQTLs in LCLs without eQTLs in LCLs showed eQTL colocalization in other immune cell types. Context specificity of eQTLs has been widely considered to be the primary explanation for the difficulty of pinpointing disease-associated eQTLs^{17,22,23,73}. Our observation that many IMD-colocalized caQTLs in LCLs show eQTLs in other immune cell types suggests that caQTL effects may be shared across cell types whereas eQTL effects are more context-specific²³. eQTL data for the specific cellular contexts need to be generated in future studies to uncover more such examples. Our results also showed that having a large sample size in eQTL data is likely a key factor for discoveries of more disease-associated eQTLs.

Limitations of the study

The h^2_{med} / h^2_{SNP} estimates can be biased because of insufficient sample size of the QTL data. Thus, the proportion of mediated IMD heritability could change based on the specific caQTL and eQTL data. Colocalization analysis tests whether the QTL and GWAS data likely share genetic signal, and such shared signal could arise from either causal mediation or pleiotropy. Therefore, further experiments are needed to establish causality of colocated eQTL genes. Lastly, we hypothesized that caQTL effects are often shared across cell types, but we did not have access to relevant data to test this hypothesis.

Methods

ATAC-seq data processing and peak calling

We downloaded LCL ATAC-seq data of British (GBR) samples (n = 100) from European Nucleotide Archive (ENA) under accession ERP110508³⁰. The available files were cram alignment files mapped to the b37 reference genome, so we extracted unique read pairs using SAMtools⁷⁶ and bamtofastq command from bedtools⁷⁷. The reads were paired-end and each 75 base pairs (bp) long. The data contained reads with Nextera transposase adapters, so we removed the adapter sequences and bases of poor quality at the 3' end using cutadapt. Trimmed reads with both pairs shorter than 20 bp were discarded. The command was “cutadapt -a file:\${forward} -A file:\${reverse} -e 0.25 -j 2 -q 15 --pair-filter=both -m 20”. (\${forward} and \${reverse} files contain the forward and reverse Nextera transposase adapters). We mapped the reads to the GRCh38 reference genome using Bowtie 2⁷⁸ with the “GRCh38_noalt_decoy_as” index provided on the tool's website. The command was “bowtie2 --very-sensitive --no-mixed --no-discordant -I 20 -X

2000". We kept only read alignments with mapping quality greater than 1. We also removed reads aligning to the mitochondrial genome, those overlapping ENCODE exclusion regions (file ID: ENCFF356LFX)⁷⁹ and potential PCR duplicates using scripts from WASP⁸⁰.

To represent peaks across the samples, we subsampled 3 million read pairs from each and pooled them. Then, we used MACS2⁸¹ with the BAMPE option for peak calling. The command was "macs2 callpeak -f BAMPE -g hs -q 0.05". We further used the 'bdgcmp' and 'bdgpeakcall' subcommand to find peaks that are at least 100 bp long (-l) and merge those that are less than 100 bp apart (-g). We also merged peaks similarly derived from individual samples using the 'merge' command in bedtools.

Furthermore, for the sake of comprehensiveness, we repeated these steps with LCL ATAC-seq data of Yoruban (YRI) samples⁸² and merged the peaks with the earlier peak set derived from GBR samples. In total, there were 443,403 peaks genome-wide.

RNA expression data preparation

We downloaded RNA expression level data of the LCL samples⁵⁶ from the Expression Atlas⁸³. The source data is available at European Nucleotide Archive under accession ERP001942. We retained TPM values of protein-coding and long noncoding RNAs in European samples for downstream analyses, including QTL analysis, mediated heritability analysis, and colocalization.

LCL samples' genotype preparation

To utilize the genotype calls of the highest quality, we downloaded the high-coverage 1000 Genomes (1kG) Project data⁸⁴. There was a total of 493 samples that had either ATAC-seq or RNA-seq data in this study. Of those, 479 samples had genotypes derived from the high-coverage whole-genome sequencing data. 14 samples had genotypes derived from microarrays⁸⁵. The samples that needed imputation are listed in **Supplementary Table 9**. We lifted over the microarray data based on the hg19 reference genome to GRCh38 and filtered for variants present in the high-coverage 1kG data. We first imputed microarray genotype data using the TOPMed imputation server⁸⁶ and extracted SNPs with imputation $R^2 \geq 0.5$ and imputed the rest of the variants, including short indels, using high-coverage data with Beagle5.2^{87,88}. After keeping only variants with DR2 ≥ 0.7 , we merged the imputed genotypes with the high-coverage 1kG genotypes.

Curation of IMD GWAS data

We downloaded GWAS summary statistics for 13 IMDs, including 11 autoimmune diseases (autoimmune thyroid disease (ATD)³⁹, celiac disease (CEL)⁴⁰, Crohn's disease (CD)⁴¹, inflammatory bowel disease (IBD)⁴¹, juvenile idiopathic arthritis (JIA)⁴², multiple sclerosis (MS)⁴³, primary biliary cholangitis (PBC)³⁹, rheumatoid arthritis (RA)⁴⁴, systemic lupus erythematosus (SLE)⁴⁵, ulcerative colitis (UC)⁴¹, and vitiligo (VIT)⁴⁶) and 2 allergic diseases (allergy (ALL)⁴⁷ and asthma (AST)⁴⁷). For each disease, we searched for more recent studies with larger sample sizes and prioritized those with genome-wide statistics, rather than those with only Immunochip variants⁸⁹. To compare with non-immune diseases, we also downloaded summary statistics for 3 non-immune diseases (type 2 diabetes (T2D)⁴⁸, coronary artery disease (CAD)⁴⁹ and schizophrenia (SCZ)⁵⁰). In this study, we analyzed only those GWAS summary statistics derived from cohorts of individuals with European ancestries.

Since the LCL samples' genotypes are based on the GRCh38 reference genome, we lifted over any GWAS data based on b37 reference genome to the GRCh38 genomic coordinates. Briefly, we formatted the summary statistics as bed files and used the liftOver tool⁹⁰ to convert them to GRCh38 genomic coordinates. Then, to ensure that the reference alleles match the sequences of the GRCh38 reference genome, we used gwas2vcf tool⁹¹.

Chromatin accessibility QTLs in lymphoblastoid cell lines

First, we quantified the chromatin accessibility levels at ATAC-seq peaks identified earlier. We counted the number of read fragments overlapping each peak using `featureCounts`⁹². For each sample, the read counts were normalized for library size using trimmed mean of M-values⁹³ so that the values are comparable across the samples. Then, the phenotype values were further normalized to follow a standard normal distribution across the samples, using quantile normalization. Peaks with counts per million (CPM) < 0.8 or counts < 10 for more than 20% of the samples were discarded.

Next, we performed a principal component analysis (PCA) on the phenotype matrix to derive potential latent covariates. We selected the number of principal components (PC) to incorporate in the regression model based on the Buja and Eyuboglu algorithm⁹⁴ that is implemented in `PCAFForQTL`⁹⁵. Ultimately, we accounted for sex, library size, 3 genotype PCs and 13 phenotype PCs in the QTL analysis. We performed genetic association tests on variants within 200 kilobases (kb) of the peak using `tensorQTL`⁹⁶. We discarded variants with minor allele frequency less than 5%.

IMD heritability enrichment in accessible regions of LCLs

We evaluated the relevance of accessible regions in LCLs to IMD heritability using stratified LD score regression (S-LDSC)³⁸. We used the baselineLD v2.2 annotation in hg38 and the European LD reference from the 1000 Genomes Project (downloaded from the S-LDSC website, <https://alkesgroup.broadinstitute.org/LDSCORE/GRCh38/>). We used the set of filtered ATAC-seq peaks that we tested for QTL associations. We accounted for 16 diseases (13 IMDs and 3 non-IMDs) for the Bonferroni-corrected *p*-value threshold.

Mediated heritability estimation of QTLs

We estimated the heritability mediated by QTLs (h^2_{med}) using mediated expression score regression (MESR)¹⁶. We denote heritability of tested SNPs as h^2_{SNP} .

caQTL-mediated heritability

We estimated the ‘expression scores’ for chromatin accessibility in LCLs using individual genotypes and phenotypes. We analyzed the same set of peaks and accounted for the same covariates as we did for the QTL analysis above. For mediated heritability estimation, we accounted for baselineLD v2.2 annotation in hg38 without the QTL annotations, as they could be redundant. The estimand of interest is the proportion of heritability mediated by caQTLs ($h^2_{\text{med}} / h^2_{\text{SNP}}$).

To evaluate whether certain peak sets are enriched for mediated heritability, we utilized the gene set analysis functionality. For peak strength, we considered the 95% percentile CPM value of each peak and stratified the peaks into quintiles. For histone marks, we first generated histone ChIP-seq peaks using data from Delaneau *et al.*⁵². Similar to calling ATAC-seq peaks, we sampled 3 million reads per sample and merged them before applying `MACS2`⁸¹. We downloaded 3 control ChIP-seq data from ENCODE to use as input (File IDs: ENCFF066RCS, ENCFF159XTB, and ENCFF850RIE)⁷⁹. Then, we curated sets of ATAC-seq peaks that overlapped H3K27ac, H3K4me1, and H3K4me3 ChIP-seq peaks by at least 1 bp. ATAC-seq peaks that overlapped H3K27ac but not H3K4me3 regions were labeled as ‘enhancer-like signature’, while those that overlapped H3K27ac and H3K4me3 regions were labeled as ‘promoter-like signature’. The estimand of interest is the proportion of mediated heritability explained by the peak set (peak set h^2_{med} / total h^2_{med}).

Comparison with eQTL-mediated heritability

First, we estimated $h^2_{\text{med}} / h^2_{\text{SNP}}$ of eQTLs (i.e., $h^2_{\text{med}; \text{eQTL}} / h^2_{\text{SNP}}$) the same way as we did for that of caQTLs. Then, we also estimated $h^2_{\text{med}} / h^2_{\text{SNP}}$ of caQTLs and eQTLs together (i.e., $h^2_{\text{med}; \text{caQTL} \cup \text{eQTL}}$) with MESC meta-analysis.¹⁶ caQTLs and eQTLs were also stratified as separate sets to account for potential differences in the relationship of QTL *cis*-heritability and h^2_{med} . This meta-analyzed estimate is effectively the amount of heritability mediated by either caQTLs or eQTLs in LCLs. This estimate reveals the overall relationship between heritability mediated by caQTLs and eQTLs. For instance, the estimate of the heritability mediated exclusively by caQTLs would be $h^2_{\text{med}; \text{just caQTL}} = h^2_{\text{med}; \text{caQTL} \cup \text{eQTL}} - h^2_{\text{med}; \text{eQTL}}$. The estimate of mediated heritability shared by caQTLs and eQTLs is $h^2_{\text{med}; \text{caQTL} \cap \text{eQTL}} = (h^2_{\text{med}; \text{caQTL}} + h^2_{\text{med}; \text{eQTL}}) - h^2_{\text{med}; \text{caQTL} \cup \text{eQTL}}$.

Colocalization analyses

caQTL and eQTL colocalization with IMD GWAS

First, we selected candidate colocalization loci by filtering for overlapping ‘significant’ variants. The candidate loci met the following conditions: (1) lead IMD association at $p < 10^{-6}$, (2) lead caQTL or eQTL association at $p < 10^{-4}$, and (3) at least one variant simultaneously showed caQTL or eQTL χ^2 statistics greater than $0.8 \times$ lead χ^2 statistics for the caQTL or eQTL and IMD association χ^2 statistics greater than $0.8 \times \chi^2$ statistics for the IMD lead variant in the locus. Then, we applied gwas-pw³⁷ on the variants within 100 kb of the lead variant. We considered loci with posterior probability of colocalization (PP3) > 0.98 to be colocalized⁵⁷.

Colocalization of caQTL with eQTL

We performed a colocalization analysis of caQTLs and eQTLs to curate a set of loci where the same genetic signal likely explains both associations. The pairs of caQTLs and eQTLs reveal the distance between the regulatory element and the target gene’s TSS. We selected candidate colocalization loci with: (1) IMD association at $p < 10^{-6}$, (2) QTL association at $p < 10^{-4}$, and (3) the two lead variants showed LD $r^2 > 0.8$. We applied gwas-pw³⁷ on the variants within 200 kb of the tested caQTL peak. We considered loci with posterior probability of colocalization (PP3) > 0.98 to be colocalized⁵⁷.

Enrichment of biological processes

To test whether colocalized genes are likely relevant to autoimmune diseases, we surveyed which Biological Process Gene Ontology (GO) terms were overrepresented compared to all the genes within 500 kb of each IMD association signal. Enrichment of biological processes was evaluated using Protein Analysis Through Evolutionary Relationships (PANTHER)⁹⁷. The foreground list comprised all of the genes whose eQTL signal colocalized with one of the 7 autoimmune diseases (CD, IBD, MS, PBC, RA, SLE, and UC). The background list of genes was all of the genes within 500 kb of each IMD lead variant for which we observed colocalization. Moreover, we tested whether colocalized caQTLs without eQTLs are closer to genes related to immune processes than expected by chance. For this, we used Genomic Regions Enrichment of Annotations Tool (GREAT)⁶³. The foreground list comprised ATAC-seq peaks at IMD loci showing colocalization only with caQTLs and not with eQTLs. The background list is all ATAC-seq peaks identified in LCLs that we tested for caQTL association.

Relationship between peak-to-TSS distance and *cis*-heritability of caQTL and eQTL

The distance between the colocalized caQTL peak and the eGene's TSS (*i.e.*, peak-to-TSS distance) was determined to be the shortest distance from one end of the caQTL peak to the TSS. The pairs of caQTLs and eQTLs were split into quintiles based on their peak-to-TSS distance from closest to farthest.

MESC analysis uses REML implemented in GCTA⁹⁸ to estimate QTL *cis*-heritability. We referred to its output and compared the *cis*-heritability of caQTLs and eQTLs based on peak-to-TSS distance. To visualize the distribution of *cis*-heritability estimates, we grouped pairs of colocalized caQTLs and eQTLs based on peak-to-TSS distance quintiles.

LCL eQTL meta-analysis

We downloaded LCL eQTL data from 3 studies through eQTL Catalogue release 6⁹⁹. The sample sizes were 190, 147, and 418 for GENCORD⁶⁸, GTEX⁵⁵, and TwinsUK⁶⁹, respectively. We meta-analyzed the summary statistics using the inverse variance weighted fixed effects model. If a variant was missing in a subset of the studies, then only the available statistics were meta-analyzed. We used these meta-analyzed statistics to perform colocalization the same way as earlier colocalization analyses.

Colocalization analysis with immune cell eQTL data

We downloaded eQTL data for various immune cell types from the eQTL Catalogue release 6⁹⁹. The source studies from the eQTL Catalogue include BLUEPRINT⁷², DICE²², Alasoo *et al.*²³ and Bossini-Castillo *et al.*⁷⁴. The represented immune cell types include T cell subtypes, B cells, neutrophils, and monocytes. We also separately downloaded data for CD4⁺ T cells with and without stimulation⁷³. The selection of candidate loci and colocalization analysis on them followed the same procedure as that for other QTLs.

We also evaluated whether meta-analysis of eQTL data can increase the number of loci with eQTL colocalization. We meta-analyzed eQTL data for 3 immune cell types with multiple sources – naïve CD4⁺ T cell, monocyte, and memory regulatory T cell (Treg) – using the inverse variance weighted fixed effects model. Specifically, we meta-analyzed naïve CD4⁺ T cell eQTL summary statistics from Soskic *et al.*⁷³, BLUEPRINT⁷², and DICE²². We meta-analyzed monocyte eQTL summary statistics from BLUEPRINT⁷² and DICE²². We meta-analyzed memory Treg eQTL summary statistics from Bossini-Castillo *et al.*⁷⁴ and DICE²². We used these meta-analyzed statistics to perform colocalization the same way as earlier colocalization analyses.

Data availability

Processed data and code for generating the figures presented in the manuscript are available at <https://github.com/BulykLab/IMD-colocalization-manuscript-figures>. Genotype data are downloaded from the 1000 Genomes Project (<https://www.internationalgenome.org/data-portal/data-collection/30x-grch38>). LCL ATAC-seq data are downloaded from ENA (<https://www.ebi.ac.uk/ena/browser/view/PRJEB28318>). LCL RNA-seq count matrix is downloaded from Expression Atlas⁸³ (<https://www.ebi.ac.uk/gxa/experiments/E-GEUV-1/Results>). GWAS data for the studies listed in the Methods section are downloaded from the GWAS catalog¹⁰⁰ (<https://www.ebi.ac.uk/gwas/>). eQTL summary statistics are downloaded either from the eQTL Catalogue⁹⁹ (<https://www.ebi.ac.uk/eql/>) or the source publication.

Acknowledgements

We thank Vijay Sankaran, Shamil Sunyaev, Alexander Gusev, and members of the Raychaudhuri lab for helpful discussion. This work was funded by NIH grant R01 HG010501.

Author contributions

R.J. and M.L.B. conceived and designed the research project. R.J. performed all analyses and prepared the figures. M.L.B. supervised the research. R.J. and M.L.B. wrote the manuscript.

Declaration of interests

The authors declare no competing interests.

References

1. Visscher P. M., et al. (2017) **10 Years of GWAS Discovery: Biology, Function, and Translation** *Am. J. Hum. Genet* **101**:5–22
2. Claussnitzer M., et al. (2020) **A brief history of human disease genetics** *Nature* **577**:179–189
3. Maurano M. T., et al. (2012) **Systematic localization of common disease-associated variation in regulatory DNA** *Science* **337**:1190–1195
4. Gusev A., et al. (2014) **Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases** *Am. J. Hum. Genet* **95**:535–552
5. Nicolae D. L., et al. (2010) **Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS** *PLoS Genet* **6**
6. Hormozdiari F., et al. (2018) **Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits** *Nat. Genet* **50**:1041–1047
7. O'Connor L. J., et al. (2019) **Extreme Polygenicity of Complex Traits Is Explained by Negative Selection** *Am. J. Hum. Genet* **105**:456–476
8. Boyle E. A., Li Y. I., Pritchard J. K (2017) **An Expanded View of Complex Traits: From Polygenic to Omnigenic** *Cell* **169**:1177–1186
9. Nasser J., et al. (2021) **Genome-wide enhancer maps link risk variants to disease genes** *Nature* **593**:238–243
10. Weeks E. M., et al. (2023) **Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases** *Nat. Genet* **55**:1267–1276

11. Morris J. A., et al. (2023) **Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens** *Science* **380**
12. Plenge R. M., Scolnick E. M., Altshuler D (2013) **Validating therapeutic targets through human genetics** *Nat. Rev. Drug Discov* **12**:581–594
13. Chun S., et al. (2017) **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
14. Barbeira A. N., et al. (2021) **Exploiting the GTEx resources to decipher the mechanisms at GWAS loci** *Genome Biol* **22**
15. Connally N. J., et al. (2022) **The missing link between genetic association and regulatory function** *Elife* **11**
16. Yao D. W., O'Connor L. J., Price A. L., Gusev A (2020) **Quantifying genetic effects on disease mediated by assayed gene expression levels** *Nat. Genet* :1–8 <https://doi.org/10.1038/s41588-020-0625-2>
17. Umans B. D., Battle A., Gilad Y (2021) **Where Are the Disease-Associated eQTLs?** *Trends Genet* **37**:109–124
18. Mostafavi H., Spence J. P., Naqvi S., Pritchard J. K (2023) **Systematic differences in discovery of genetic effects on gene expression and complex traits** *Nat. Genet* **55**:1866–1875
19. Hukku A., et al. (2021) **Probabilistic Colocalization of Genetic Variants from Complex and Molecular Traits: Promise and Limitations** *Am. J. Hum. Genet* **108**:25–35
20. Glassberg E. C., Gao Z., Harpak A., Lan X., Pritchard J. K (2019) **Evidence for weak selective constraint on human gene expression** *Genetics* **211**:757–772
21. Dimas A. S., et al. (2009) **Common regulatory variation impacts gene expression in a cell type-dependent manner** *Science* **325**:1246–1250
22. Schmiedel B. J., et al. (2018) **Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression** *Cell* **175**:1701–1715
23. Alasoo K., et al. (2018) **Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response** *Nat. Genet* **50**:424–431
24. Nathan A., et al. (2022) **Single-cell eQTL models reveal dynamic T cell state dependence of disease loci** *Nature* **606**:120–128
25. Field A., Adelman K. (2020) **Evaluating Enhancer Function and Transcription** *Annu. Rev. Biochem* **89**:213–234
26. Kasowski M., et al. (2010) **Variation in transcription factor binding among humans** *Science* **328**:232–235
27. Kilpinen H., et al. (2013) **Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription** *Science* **342**:744–747

28. Waszak S. M., et al. (2015) **Population Variation and Genetic Control of Modular Chromatin Architecture in Humans** *Cell* **162**:1039–1050
29. Degner J. F., et al. (2012) **DNase-I sensitivity QTLs are a major determinant of human expression variation** *Nature* **482**:390–394
30. Kumasaka N., Knights A. J., Gaffney D. J (2019) **High-resolution genetic mapping of putative causal interactions between regions of open chromatin** *Nat. Genet* **51**:128–137
31. Albert F. W., Kruglyak L (2015) **The role of regulatory variation in complex traits and disease** *Nat. Rev. Genet* **16**:197–212
32. Wu Y., et al. (2023) **Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large fraction of GWAS signals shared with molecular phenotypes** *Cell Genomics* **100344** <https://doi.org/10.1016/j.xgen.2023.100344>
33. Jeong R., Bulyk M. L. (2023) **Blood cell traits' GWAS loci colocalization with variation in PU.1 genomic occupancy prioritizes causal noncoding regulatory variants** *Cell genomics* **3**
34. Roadmap Epigenomics Consortium, et al. (2015) **Integrative analysis of 111 reference human epigenomes** *Nature* **518**:317–329
35. Farh K. K. H., et al. (2015) **Genetic and epigenetic fine mapping of causal autoimmune disease variants** *Nature* **518**:337–343
36. Giambartolomei C., et al. (2014) **Bayesian test for colocalisation between pairs of genetic association studies using summary statistics** *PLoS Genet* **10**
37. Pickrell J. K., et al. (2016) **Detection and interpretation of shared genetic influences on 42 human traits** *Nat. Genet* **48**:709–717
38. Finucane H. K., et al. (2015) **Partitioning heritability by functional annotation using genome-wide association summary statistics** *Nat. Genet* **47**:1228–1235
39. Cordell H. J., et al. (2021) **An international genome-wide meta-analysis of primary biliary cholangitis: Novel risk loci and candidate drugs** *J. Hepatol* **75**:572–581
40. Dubois P. C. A., et al. (2010) **Multiple common variants for celiac disease influencing immune gene expression** *Nat. Genet* **42**:295–302
41. de Lange K. M., et al. (2017) **Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease** *Nat. Genet* **49**:256–261
42. López-Isac E., et al. (2021) **Combined genetic analysis of juvenile idiopathic arthritis clinical subtypes identifies novel risk loci, target genes and key regulatory mechanisms** *Ann. Rheum. Dis* **80**:321–328
43. Patsopoulos N. A., et al. (2019) **Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility** *Science* **365**
44. Ishigaki K., et al. (2022) **Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis** *Nat. Genet* **54**:1640–1651

45. Bentham J., et al. (2015) **Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus** *Nat. Genet* **47**:1457–1464
46. Jin Y., et al. (2016) **Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants** *Nat. Genet* **48**:1418–1424
47. Loh P. R., Kichaev G., Gazal S., Schoech A. P., Price A. L (2018) **Mixed-model association for biobank-scale datasets** *Nat. Genet* **50**:906–908
48. Morris A. P., et al. (2012) **Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes** *Nat. Genet* **44**:981–990
49. Schunkert H., et al. (2011) **Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease** *Nat. Genet* **43**:333–8
50. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) **Biological insights from 108 schizophrenia-associated genetic loci** *Nature* **511**:421–7
51. van Mierlo H. C., Broen J. C. A., Kahn R. S., de Witte L. D (2019) **B-cells and schizophrenia: A promising link or a finding lost in translation?** *Brain. Behav. Immun* **81**:52–62
52. Delaneau O., et al. (2019) **Chromatin three-dimensional interactions mediate genetic effects on gene expression** *Science* **364**
53. Heintzman N. D., et al. (2007) **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome** *Nat. Genet* **39**:311–318
54. ENCODE Project Consortium, et al. (2020) **Expanded encyclopaedias of DNA elements in the human and mouse genomes** *Nature* **583**:699–710
55. GTEx Consortium (2020) **The GTEx Consortium atlas of genetic regulatory effects across human tissues** *Science* **369**:1318–1330
56. Lappalainen T., et al. (2013) **Transcriptome and genome sequencing uncovers functional variation in humans** *Nature* **501**:506–511
57. Kundu K., et al. (2022) **Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for twelve immune-mediated diseases** *Nat. Genet* **44**:0–23
58. Thomas P. D., et al. (2022) **PANTHER: Making genome-scale phylogenetics accessible to all** *Protein Sci* **31**:8–22
59. Sanmartí R., Ruiz-Esquide V., Bastida C., Soy D (2018) **Tocilizumab in the treatment of adult rheumatoid arthritis** *Immunotherapy* **10**:447–464
60. Khanna R., Feagan B. G (2013) **Ustekinumab for the treatment of Crohn's disease** *Immunotherapy* **5**:803–15
61. Monemi S., et al. (2016) **Incidence of Gastrointestinal Perforations in Patients with Rheumatoid Arthritis Treated with Tocilizumab from Clinical Trial, Postmarketing, and Real-World Data Sources** *Rheumatol. Ther* **3**:337–352

62. Stevenson C., et al. (2014) **Essential role of Elmo1 in Dock2-dependent lymphocyte migration** *J. Immunol* **192**:6062–70
63. McLean C. Y., et al. (2010) **GREAT improves functional interpretation of cis-regulatory regions** *Nat. Biotechnol* **28**:495–501
64. Tanigawa Y., Dyer E. S., Bejerano G (2022) **WhichTF is functionally important in your open chromatin data?** *PLoS Comput. Biol* **18**
65. Fulco C. P., et al. (2019) **Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations** *Nat. Genet* **51**
66. Zuin J., et al. (2022) **Nonlinear control of transcription through enhancer-promoter interactions** *Nature* **604**:571–577
67. Vösa U., et al. (2021) **Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression** *Nat. Genet* **53**
68. Gutierrez-Arcelus M., et al. (2013) **Passive and active DNA methylation and the interplay with genetic variation in gene regulation** *Elife* **2013**:1–18
69. Buil A., et al. (2015) **Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins** *Nat. Genet* **47**:88–91
70. Dziembowska M., et al. (2002) **Three novel mutations of the CIITA gene in MHC class II-deficient patients with a severe immunodeficiency** *Immunogenetics* **53**:821–9
71. Hofmann E., et al. (2010) **Octamer-binding factor 6 (Oct-6/Pou3f1) is induced by interferon and contributes to dsRNA-mediated transcriptional responses** *BMC Cell Biol* **11**
72. Chen L., et al. (2016) **Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells** *Cell* **167**:1398–1414
73. Soskic B., et al. (2022) **Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation** *Nat. Genet* **54**:817–826
74. Bossini-Castillo L., et al. (2022) **Immune disease variants modulate gene expression in regulatory CD4+ T cells** *Cell Genomics* **2**
75. Xu W.-D., Li R., Huang A.-F (2022) **Role of TL1A in Inflammatory Autoimmune Diseases: A Comprehensive Review** *Front. Immunol* **13**
76. Danecek P., et al. (2021) **Twelve years of SAMtools and BCFtools** *Gigascience* **10**
77. Quinlan A. R., Hall I. M (2010) **BEDTools: a flexible suite of utilities for comparing genomic features** *Bioinformatics* **26**:841–2
78. Langmead B., Salzberg S. L (2012) **Fast gapped-read alignment with Bowtie 2** *Nat. Methods* **9**:357–359
79. Dunham I., et al. (2012) **An integrated encyclopedia of DNA elements in the human genome** *Nature* **489**:57–74

80. Van De Geijn B., Mcvicker G., Gilad Y., Pritchard J. K (2015) **WASP: Allele-specific software for robust molecular quantitative trait locus discovery** *Nat. Methods* **12**:1061–1063
81. Zhang Y., et al. (2008) **Model-based analysis of ChIP-Seq (MACS)** *Genome Biol* **9**
82. Banovich N. E., et al. (2018) **Impact of regulatory variation across human iPSCs and differentiated cells** *Genome Res* **28**:122–131
83. Papatheodorou I., et al. (2020) **Expression Atlas update: from tissues to single cells** *Nucleic Acids Res* **48**:D77–D83
84. Byrska-Bishop M., et al. (2022) **High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios** *Cell* **185**:3426–3440
85. Auton A., et al. (2015) **A global reference for human genetic variation** *Nature* **526**:68–74
86. Das S., et al. (2016) **Next-generation genotype imputation service and methods** *Nat. Genet* **48**:1284–1287
87. Browning B. L., Zhou Y., Browning S. R (2018) **A One-Penny Imputed Genome from Next-Generation Reference Panels** *Am. J. Hum. Genet* **103**:338–348
88. Browning B. L., Tian X., Zhou Y., Browning S. R (2021) **Fast two-stage phasing of large-scale sequence data** *Am. J. Hum. Genet* **108**:1880–1890
89. Cortes A., Brown M. A (2011) **Promise and pitfalls of the Immunochip** *Arthritis Res. Ther* **13**
90. Kent W. J., et al. (2002) **The human genome browser at UCSC** *Genome Res* **12**:996–1006
91. Lyon M. S., et al. (2021) **The variant call format provides efficient and robust storage of GWAS summary statistics** *Genome Biol* **22**
92. Liao Y., Smyth G. K., Shi W (2014) **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features** *Bioinformatics* **30**:923–30
93. Robinson M. D., Oshlack A (2010) **A scaling normalization method for differential expression analysis of RNA-seq data** *Genome Biol* **11**
94. Buja A., Eyuboglu N (1992) **Remarks on Parallel Analysis** *Multivariate Behav. Res* **27**:509–40
95. Zhou H. J., Li L., Li Y., Li W., Li J. J (2022) **PCA outperforms popular hidden variable inference methods for molecular QTL mapping** *Genome Biol* **23**
96. Taylor-Weiner A., et al. (2019) **Scaling computational genomics to millions of individuals with GPUs** *Genome Biol* **20**
97. Mi H., et al. (2019) **Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0)** *Nat. Protoc* **14**:703–721
98. Yang J., Lee S. H., Goddard M. E., Visscher P. M (2011) **GCTA: a tool for genome-wide complex trait analysis** *Am. J. Hum. Genet* **88**:76–82
99. Kerimov N., et al. (2021) **A compendium of uniformly processed human gene expression and splicing quantitative trait loci** *Nat. Genet* **53**:1290–1299

100. Sollis E., et al. (2023) **The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource** *Nucleic Acids Res* **51**:D977–D985

Editors

Reviewing Editor

Murim Choi

Seoul National University, Seoul, Korea, the Republic of

Senior Editor

Murim Choi

Seoul National University, Seoul, Korea, the Republic of

Reviewer #1 (Public Review):

Most human traits and common diseases are polygenic, influenced by numerous genetic variants across the genome. These variants are typically non-coding and likely function through gene regulatory mechanisms. To identify their target genes, one strategy is to examine if these variants are also found among genetic variants with detectable effects on gene expression levels, known as eQTLs. Surprisingly, this strategy has had limited success, and most disease variants are not identified as eQTLs, a puzzling observation recently referred to as "missing regulation".

In this work, Jeong and Bulyk aimed to better understand the reasons behind the gap between disease-associated variants and eQTLs. They focused on immune-related diseases and used lymphoblastoid cell lines (LCLs) as a surrogate for the cell types mediating the genetic effects. Their main hypothesis is that some variants without eQTL evidence might be identifiable by studying other molecular intermediates along the path from genotype to phenotype. They specifically focused on variants that affect chromatin accessibility, known as caQTLs, as a potential marker of regulatory activity.

The authors present data analyses supporting this hypothesis: several disease-associated variants are explained by caQTLs but not eQTLs. They further show that although caQTLs and eQTLs likely have largely overlapping underlying genetic variants, some variants are discovered only through one of these mapping strategies. Notably, they demonstrate that eQTL mapping is underpowered for gene-distal variants with small effects on gene expression, whereas caQTL mapping is not dependent on the distance to genes. Additionally, for some disease variants with caQTLs but no corresponding eQTLs in LCLs, they identify eQTLs in other cell types.

Altogether, Jeong and Bulyk convincingly demonstrate that for immune-related diseases, discovering the missing disease-eQTLs requires both larger eQTL studies and a broader range of cell types in expression assays. It remains to be seen what fractions of the missing disease-eQTLs will be discovered with either strategy and whether these results can be extended to other diseases or traits.

It should be noted that the problem of "missing regulation" has been investigated and discussed in several recent papers, notably Umans et al., *Trends in Genetics* 2021; Connally et al., *eLife* 2022; Mostafavi et al., *Nat. Genet.* 2023. The results reported by Jeong and Bulyk are not unexpected in light of this previous work (all of which they cite), but they add valuable empirical evidence that mostly aligns with the model and discussions presented in Mostafavi et al.

<https://doi.org/10.7554/eLife.98289.1.sa1>

Reviewer #2 (Public Review):

Summary:

eQTLs have emerged as a method for interpreting GWAS signals. However, some GWAS signals are difficult to explain with eQTLs. In this paper, the authors demonstrated that caQTLs can explain these signals. This suggests that for GWAS signals to actually lead to disease phenotypes, they must be accessible in the chromatin. This implies that for GWAS signals to translate into disease phenotypes, they need to be accessible within the chromatin.

However, fundamentally, caQTLs, like GWAS, have the limitation of not being able to determine which genes mediate the influence on disease phenotypes. This limitation is consistent with the constraints observed in this study.

(1) For reproducibility, details are necessary in the method section.

- Details about adding YRI samples in ATAC-seq: For example, how many samples are there, and what is used among public data? There is LCL-derived iPSC and differentiated iPSC (cardiomyocytes) data, not LCL itself. How does this differ from LCL, and what is the rationale for including this data despite the differences?

- caQTL is described as having better power than eQTL despite having fewer samples. How does the number of ATAC peaks used in caQTL compare to the number of gene expressions used in eQTL?

- Details about RNA expression data: In the method section, it states that raw data (ERP001942) was accessed, and in data availability, processed data (E-GEUV-1) was used. These need to be consistent.

How many samples were used (the text states 373, but how was it reduced from the original 465, and the total genotype is said to be 493 samples while ATAC has n=100; what are the 20 others?), and it mentions European samples, but does this exclude YRI?

(2) Experimental results determining which TFs might bind to the representative signals of caQTL are required.

(3) It is stated that caQTL is less tissue-specific compared to eQTL; would caQTL performed with ATAC-seq results from different cell types, yield similar results?

<https://doi.org/10.7554/eLife.98289.1.sa0>