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

Raehoon Jeong[1,2] and Martha L. Bulyk[1,2,3†]

[1] Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.

[2] Bioinformatics and Integrative Genomics Graduate Program, Harvard University, Cambridge, MA 02138, USA.

[3] Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.

† Correspondence: mlbulyk@genetics.med.harvard.edu

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.'

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 genome[3], and disease heritability is concentrated in putative regulatory regions[4]. 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 trait[7]. 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 variation[8]. 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 them[12].

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_{med}/h^2_{SNP}) of assayed gene expression was estimated to be only about 11% on average[16]. 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 introduced[15]. 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_{med}/h^2_{SNP} estimates for eQTLs are relatively low[16]. First, statistical power to detect disease-associated eQTLs may be insufficient[19]. 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

58 against their having large eQTL effects[20]. Consequently, the negative selection induces a
59 "flattening" effect, in which weak eQTL variants, often in regions distal to the gene's promoter[21]
60 may reach high enough frequency, whereas strong eQTL variants remain at low frequency[7]. In
61 fact, eQTL-mediated heritability was enriched in genes showing mutational constraint and those
62 with lower *cis*-heritability[16]. These weaker or low-allele-frequency eQTL effects would require
63 larger sample sizes to be detected with statistical significance and to show colocalization[19]. Second,
64 causal eQTL effects may be specific to cell types that have not been assayed[17]. For example,
65 immune cell eQTLs are highly cell-type specific, and eQTL effects specific to some immune cell
66 types may mediate immune disease risk[22]. Specificity of eQTL effects can also be limited to
67 specific cell states[23,24]. Detecting cell-type or cell-state specific eQTL effects requires the
68 necessary gene expression data sets from the relevant cell types and states[17], which has been a
69 limiting resource for such analyses.

70 To investigate why disease-associated eQTL signals have been missing, we focused on
71 immune-mediate diseases (IMDs) as a model set of complex traits. We aimed to collect IMD-
72 associated loci that are expected to show eQTL signals in some cell type. Since active regulatory
73 elements coordinate target gene expression[25], we reasoned that variants that affect chromatin
74 phenotypes at regulatory elements, such as transcription factor (TF) binding[26–28] and chromatin
75 accessibility[29,30], have the potential to impact gene expression[31]. These chromatin phenotypes may
76 show detectable genetic effects even when an eQTL effect in the same cell type was not identified
77 in the locus[32]. For example, only about 20% of lymphoblastoid cell lines' (LCLs') PU.1 binding
78 QTLs (bQTLs) that colocalized with blood cell traits' association showed an eQTL effect for a
79 nearby gene in LCLs[33].

80 Here, we analyzed genetic and functional genomic (*i.e.*, ATAC-seq and RNA-seq) data in
81 LCLs. LCLs are derived from B lymphocytes, and their *cis*-regulatory elements were enriched for
82 variants associated with some IMDs[34,35]. We evaluated whether chromatin accessibility QTLs
83 (caQTLs) in LCLs potentially explain IMD associations using mediated heritability analysis[16] and
84 colocalization[36,37]. Then, we searched for disease-associated loci that were significant caQTLs, but
85 not eQTLs.

86 We examined whether the various potential reasons for missing eQTLs can account for
87 IMD-associated loci that are explained by caQTLs but not eQTLs. First, we explored the extent to
88 which eQTLs may have been missed because of limited statistical power. We compared *cis*-

89 heritability of colocalized caQTLs and eQTLs stratified by distance between the accessible region
90 and the transcription start site of the associated gene. We also investigated whether meta-analysis
91 of published LCL eQTL summary statistics, in order to effectively increase the sample size, can
92 uncover previously missed eQTLs. Second, we surveyed whether cell-type specificity of
93 regulatory variant effect may account for the missing regulation. We surveyed various immune
94 cell eQTL data to identify loci with which they colocalize even if LCL eQTLs did not colocalize
95 with those loci.

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

102 Results

103 Accessible chromatin in LCLs explains a significant proportion of immune-mediated 104 disease heritability

105 We aimed to evaluate whether variants that alter chromatin accessibility in LCLs may explain
106 genetic associations to IMD. First, we verified whether accessible regions in LCLs are enriched
107 for IMD heritability. We reanalyzed 100 LCL ATAC-seq samples[30] to define accessible regions
108 in this cell type. With stratified LD score regression (S-LDSC)[38], we estimated their heritability
109 enrichment across 13 IMDs, including 11 autoimmune diseases — autoimmune thyroid disease
110 (ATD)[39], celiac disease (CEL)[40], Crohn's disease (CD)[41], inflammatory bowel disease (IBD)[41],
111 juvenile idiopathic arthritis (JIA)[42], multiple sclerosis (MS)[43], primary biliary cholangitis (PBC)[39],
112 rheumatoid arthritis (RA)[44], systemic lupus erythematosus (SLE)[45], ulcerative colitis (UC)[41], and
113 vitiligo (VIT)[46] — and 2 allergic diseases — allergy (ALL)[47] and asthma (AST)[47]. We also
114 analyzed GWAS data for 3 non-immune diseases — type 2 diabetes (T2D)[48], coronary artery
115 disease (CAD)[49] and schizophrenia (SCZ)[50] — for comparison. Single nucleotide polymorphisms
116 (SNPs) in accessible regions in LCLs were significantly enriched for IMD heritability ($p <$
117 0.003125 (Bonferroni-corrected), S-LDSC; **Figure 1A**) and there was no significant enrichment
118 for non-immune diseases ($p > 0.05$, S-LDSC). These results indicate that accessible regions in

119 LCLs harbor many variants specifically associated with IMDs, and therefore that LCLs share
120 IMD-associated accessible regions with those of the causal cell type(s).

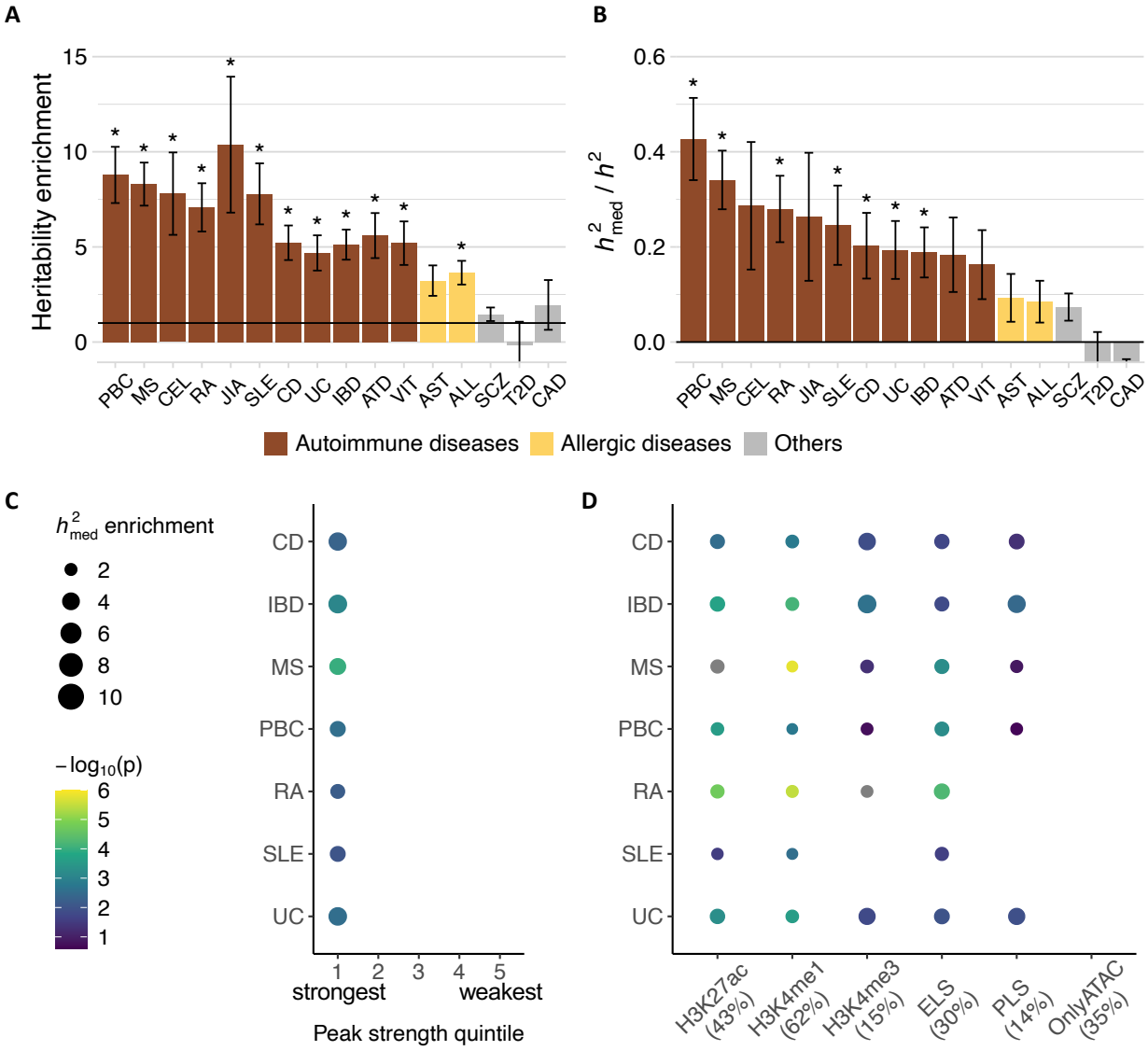


Figure 1. Immune disease heritability mediated by chromatin accessibility in LCLs. (**A-B**) Heritability enrichment in accessible regions in LCLs, based on (**A**) stratified LD score regression (S-LDSC) and (**B**) the proportion of heritability mediated by caQTLs in LCLs based on mediated expression score regression (MESC). 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. Stronger and more significant enrichment indicate that mediated heritability is concentrated in that subset. (**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.

121 Next, we applied mediated expression score regression (MESC)[16] to investigate the causal
122 relationship between caQTLs in LCLs and IMD associations (**Supplementary Figure 1A**).
123 Compared to S-LDSC analysis that tests for heritability enrichment of SNPs with some functional
124 annotation (*e.g.*, accessible regions), MESC analysis specifically estimates the heritability that is
125 mediated (*i.e.*, h^2_{med}) by the SNPs' *cis*-effects on a molecular phenotype (*e.g.*, caQTLs). We
126 estimated that caQTLs in LCLs mediate 16.3–42.7 % of autoimmune disease heritability and 8.5–
127 9.4 % of allergic disease heritability (**Figure 1B**). For non-immune diseases, the estimates were
128 lower and not significant (p > 0.003125 (Bonferroni-corrected), MESC). Interestingly,
129 schizophrenia showed a nominally significant proportion of caQTL-mediated heritability in LCLs
130 ($p < 0.05$, MESC), consistent with the hypothesis that B cells may play some role in schizophrenia
131 pathogenesis[50,51]. Our results indicate that LCLs are a valid cell type in which to search for caQTLs
132 that mediate genetic risk for 7 IMDs — CD, IBD, MS, PBC, RA, SLE, and UC — but not for
133 allergic diseases. In subsequent analyses, we focused on 7 IMDs – CD, IBD, MS, PBC, RA, SLE,
134 and UC – that showed significant caQTL-mediated heritability (p < 0.003125 (Bonferroni-
135 corrected), MESC).

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

138 To understand which features characterize accessible regions that mediate IMD heritability, we
139 estimated h^2_{med} enrichment (proportion of h^2_{med} / proportion of peaks)[16] in specific sets of
140 accessible regions. We found that peaks with a larger number of nonredundant sequencing reads
141 (*i.e.*, "stronger" peaks) in LCLs showed stronger h^2_{med} enrichment (**Figure 1C**) and thus likely
142 affect IMD-relevant gene expression more than "weaker" peaks do. This observation is consistent
143 with the "Activity-by-Contact" model[9], in which peaks with greater chromatin accessibility and
144 H3K27ac ChIP-seq signal are predicted to have proportional effects on target gene expression.
145 Next, we considered peaks with the active histone marks H3K27ac, H3K4me1, or
146 H3K4me3[28,52]. Consistent with prior observations that putative cell-type-specific regulatory
147 elements marked with H3K27ac and H3K4me1 are enriched for relevant disease associations[34,35],
148 we found that caQTLs with H3K27ac and H3K4me1 marks in LCLs were significantly enriched
149 for mediated IMD heritability (q-value < 0.05, MESC; **Figure 1D**). Strikingly, both peak sets
150 explained almost all of caQTL-mediated IMD heritability (**Supplementary Figure 2**). Peaks with

151 H3K4me3 marks, representative of promoters[53], also showed significant h^2_{med} enrichment for most
152 IMDs (q-value < 0.05, MESC). Peaks with promoter-like signatures (*i.e.*, H3K27ac and
153 H3K4me3)[54] and those with enhancer-like signatures (*i.e.*, H3K27ac, but no H3K4me3)[54] were
154 also enriched for all IMD heritability (q-value < 0.05, MESC). Conversely, peaks without any of
155 the three active histone marks were completely depleted of caQTL-mediated IMD heritability
156 (**Supplementary Figure 2**). These 'ATAC only' peaks were shorter, weaker, and further away
157 from the TSS compared to peaks with active histone marks (**Supplementary Figure 3**). Altogether,
158 these results indicate that peaks characterized as putatively active regulatory elements explain
159 nearly all of caQTL-mediated IMD heritability.

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

162 The model that gene regulatory activity explains a significant fraction of noncoding genetic
163 associations to IMDs is supported by our findings that caQTLs mediate a significant proportion of
164 IMD heritability and that those with active histone marks show strong h^2_{med} enrichment. This is in
165 contrast with relatively low average h^2_{med} / h^2_{SNP} estimates (~11%) previously having been
166 observed for eQTLs across 48 human tissues in GTEx and various human traits[55]. To directly
167 compare the proportion of IMD heritability mediated by caQTLs and eQTLs in the same cell type
168 (*i.e.*, LCLs), we additionally applied MESC to gene expression data from LCLs (*i.e.*, Geuvadis
169 data, **Supplementary Figure 1B**)[56].

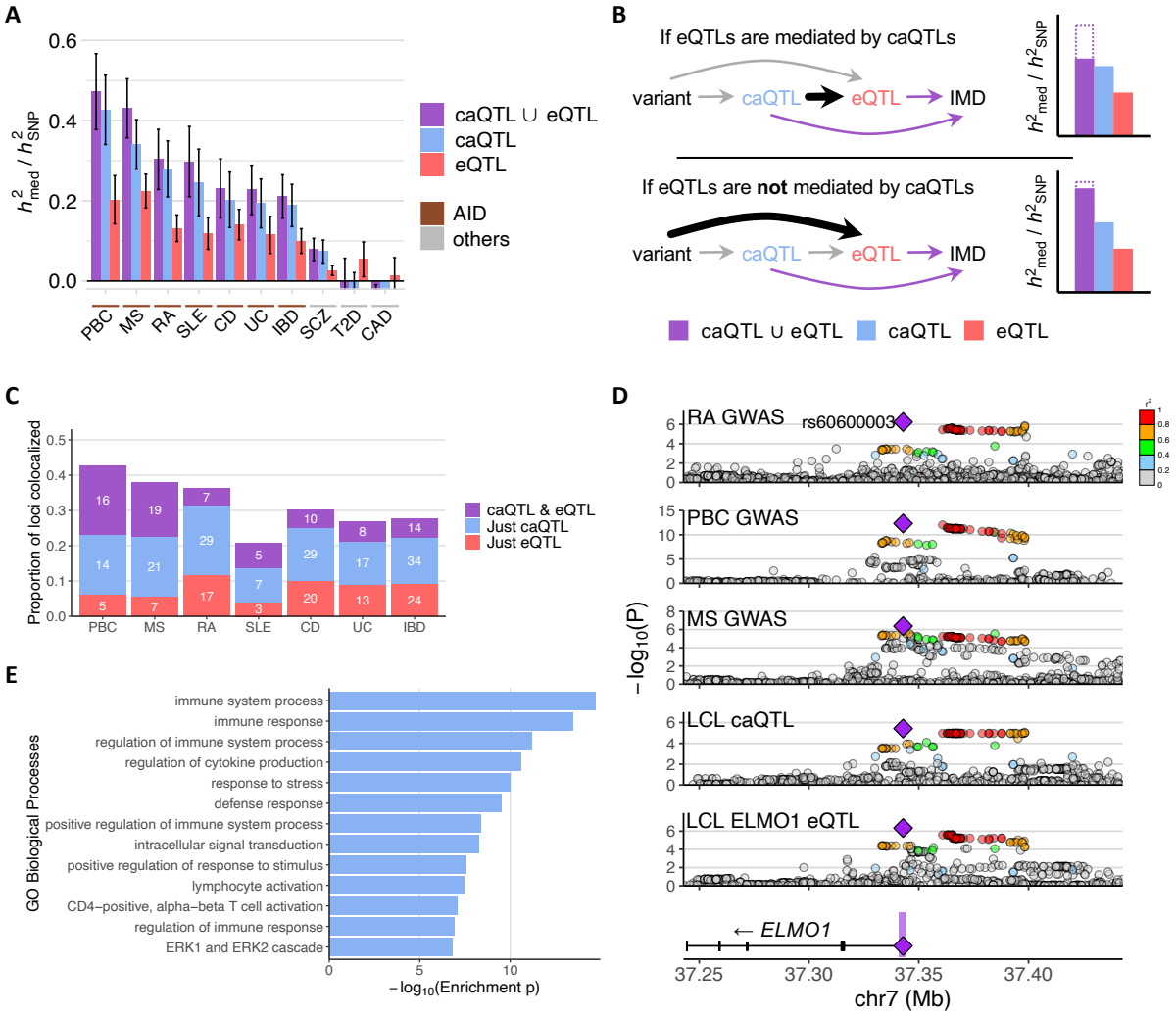


Figure 2. IMD heritability mediated by caQTLs and eQTLs. (**A**) h^2_{med} / h^2_{SNP} estimates of various IMDs for caQTLs, eQTLs, and their union. AID: autoimmune disease. Disease abbreviations along the *x*-axis are as in **Figure 1**. (**B**) Schema of the potential causal relationships between genetic 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.

170 Across the 7 autoimmune diseases, the estimated proportion of heritability mediated by
171 eQTLs ($h^2_{med;\ eQTL}$ / h^2_{SNP}) ranged from 9% to 22% (**Figure 2A**). For all 7 diseases, we estimated
172 that eQTLs mediated less heritability than did caQTLs, even though the caQTLs' smaller sample

173 size would potentially bias the estimates toward zero[16]. A possible explanation is that some IMD-
174 associated regulatory variants may show detectable effects on chromatin accessibility, but not on
175 gene expression, in LCLs at the current sample size (n = 373); such loci may account for the
176 missing regulation.

177 We anticipated that IMD-associated variants that affect gene expression in *cis* do so by
178 modulating regulatory element's activity. Therefore, we investigated whether eQTL-mediated
179 IMD heritability is shared by caQTL-mediated signals (**Figure 2B**). We performed MESC on both
180 caQTLs and eQTLs together to estimate the amount of IMD heritability mediated by both
181 collectively (Methods). For the 7 IMDs, the combined $h^2_{\text{med; caQTL} \cup \text{eQTL}} / h^2_{\text{SNP}}$ was only slightly
182 higher (2.2–9.0 %) than the estimates for just caQTLs ($h^2_{\text{med; just caQTL}} / h^2_{\text{SNP}}$; **Figure 2A**),
183 suggesting that approximately 56–82% of eQTL-mediated heritability is shared with caQTL-
184 mediated heritability (*i.e.*, $h^2_{\text{med; caQTL} \cap \text{eQTL}} / h^2_{\text{med; eQTL}}$; **Supplementary Figure 4**). These
185 estimates are consistent with substantial sharing of caQTL- and eQTL-mediated IMD heritability.
186 Nevertheless, 9–27% of IMD heritability is explained just by caQTLs, while only 2–9% of IMD
187 heritability is explained just by eQTLs.

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

189 We applied colocalization analysis[37] to identify IMD-associated loci that share genetic signals with
190 caQTLs or eQTLs in LCLs. We selected candidate loci of 200-kb windows for each IMD with the
191 following conditions: (1) lead IMD association at $p < 10^{-6}$, (2) lead caQTL or eQTL association at
192 $p < 10^{-4}$, and (3) at least one variant simultaneously showed caQTL or eQTL χ^2 statistics greater
193 than 0.8 × lead χ^2 statistics for the caQTL or eQTL, respectively, and IMD association χ^2 statistics
194 greater than 0.8 × χ^2 statistics for the IMD lead variant in the locus. We applied gwas-pw[37] and
195 considered loci with posterior probability of colocalization (PPA3) > 0.98 to be colocalized[57].
196 Some loci colocalized with only either a caQTL or an eQTL, while others colocalized with both
197 (**Figure 2C** and **Supplementary Tables 1-2**).

198 We investigated which proteins might be interacting with the colocalized caQTL peaks
199 using Cistrome[58]. We tested for overlap of the colocalized caQTL peak regions with ChIP-seq
200 peaks detecting diverse proteins in immune-related cells (**Supplementary Table 3**). Consistent
201 with the enriched mediated heritability in accessible regions with active histone marks (**Figure 1D**
202 and **Supplementary Figure 1**), proteins that are most often detected at the colocalized accessible

203 regions are those related to RNA transcription (POL2RA and MED1), chromatin remodeling
204 (EP300, BRD4, SMARCA4, and MTA2) or immune cell transcription factors (IKZF1, RUNX3,
205 SPI1, RELA, RUNX1, and EBF1) (**Supplementary Figure 5**). Interestingly, TRIM28, which
206 functions as a repressor, was one of the most overlapping protein factors.

207 To confirm that the colocalized genes are relevant to IMD, we tested for their enrichment
208 of Gene Ontology (GO) annotation terms for specific biological processes[59]. Considering all genes
209 within 500 kb of the IMD GWAS lead variants at colocalized loci as background, the genes that
210 showed eQTL colocalization for any IMD were enriched for various immune responses and
211 signaling processes, such as "positive regulation of immune system process" and "regulation of
212 lymphocyte activation" (**Supplementary Figure 6**), indicating that the colocalized genes in LCLs
213 are involved in immune function. For example, *IL6R* and *IL12A* encode direct or indirect targets
214 of approved drugs – Tocilizumab and Ustekinumab – for autoimmune diseases like RA[60] and CD[61].
215 These two genes showed colocalization with both caQTLs and eQTLs in CD and PBC GWAS,
216 respectively (**Supplementary Figure 7A-B**). Increased *IL6R* expression was associated with
217 higher risk for CD, and increased *IL12A* expression was associated with lower risk for PBC and
218 SLE. The former observation is in line with Tocilizumab, a monoclonal antibody to IL-6 receptor,
219 showing efficacy in CD patients[60], although it is not pursued for approval because of potential side
220 effects[62]. Interestingly, *IL6R* and *IL12A* eQTLs did not colocalize with the association signals of
221 RA and CD, respectively, which are the diseases for which these drugs are approved
222 (**Supplementary Figure 7C-D**). Moreover, *ELMO1*, which previously had not been associated
223 with autoimmune diseases, showed eQTL colocalization with RA, PBC, and MS association
224 signals (**Figure 2D)**. In all three, decreased *ELMO1* expression was associated with increased
225 disease risk. In mice, *Elmo1* was required for polarization and migration of B and T lymphocytes[63].

226 Across the IMDs, there were many loci that colocalized with a caQTL but not with an
227 eQTL (**Figure 2C**). These "caQTL only" loci showed enrichment for immune response genes in
228 *cis* compared to all accessible regions in LCLs[64,65], even though the colocalized eQTLs were
229 enriched for immune response genes as well (**Figure 2E** and **Supplementary Figure 6**), indicating
230 that IMD-relevant genes without eQTL colocalization in Geuvadis LCL data[56] are likely found in
231 these loci.

232 **Distance to transcription start sites affect eQTLs but not caQTLs**

233 Why might there be loci with caQTL colocalization only, despite the caQTL data having fewer
234 samples than the eQTL data (100 vs. 373)? Limited statistical power can prevent some eQTLs
235 from being detected and showing significant colocalization[19]. As hypothesized by Mostafavi and
236 colleagues, disease-relevant eQTLs may be weaker and more distal[18]. To understand the extent to
237 which this effect may result in many loci showing colocalization only with caQTLs, we compared
238 the *cis*-heritability (h^2_{cis}) of caQTLs and eQTLs depending on the distance from the ATAC peak
239 to the transcription start site (TSS) of the gene (*i.e.*, peak-to-TSS distance). We considered all
240 caQTLs and eQTLs regardless of disease association.

241 We identified pairs of caQTLs and eQTLs that colocalized with each other[37], which implies
242 that the regulatory variant modulating chromatin accessibility also affects gene expression. Then,
243 the distance between the ATAC peak and the TSS of the eQTL gene (*i.e.*, eGene) is the distance
244 between a regulatory element and its target gene's TSS. We stratified the pairs into peak-to-TSS
245 distance quintiles and compared the eQTL h^2_{cis} distribution of the first quintile (*i.e.*, closest pairs)
246 with that of the later quintiles. We observed that eQTL h^2_{cis} distribution decreased with increasing
247 distance of the paired ATAC peaks from the TSS ($p = 1.0\times10^{-4}$, 2.1×10^{-10}, 1.6×10^{-15}, and $1.7\times10^{-}$
248 ^{20}, respectively, one-sided Wilcoxon rank-sum test; **Figure 3A**), consistent with the negative
249 relationship between promoter-enhancer genomic distance and impact on gene expression[66,67].
250 This result also explains why discovered eQTLs are concentrated near the promoter, where the
251 variants are more likely to show stronger effects[68]. In contrast, caQTL h^2_{cis} distribution was similar
252 across peak-to-TSS distances ($p > 0.05$, one-sided Wilcoxon rank-sum test; **Figure 3A**). For all
253 distance quintiles, caQTL h^2_{cis} was significantly higher than that of the paired eQTLs ($p = 4.0\times10^{-}$
254 ^{23}, 1.3×10^{-24}, 1.5×10^{-19}, 1.1×10^{-28}, and 2.4×10^{-49}, respectively, one-sided paired Wilcoxon rank-
255 sum test; **Figure 3A**), and the contrast between them was greater at more distant quintiles,
256 suggesting that the statistical power to detect and colocalize eQTLs is increasingly lower than that
257 for caQTLs for regulatory effects far from the TSS.

258 Overall, caQTL h^2_{cis} had a significant positive effect ($p = 6.6\times10^{-11}$, linear regression) and
259 peak-to-TSS distance had a negative effect ($p = 2.4\times10^{-23}$, linear regression) on eQTL h^2_{cis} (**Figure**
260 **3B**). Thus, for regulatory variants that showed both caQTL and eQTL signals, those with larger
261 effects on chromatin accessibility tended to exhibit larger effects on gene expression, but their
262 eQTL effects diminished with increasing distance from TSSs.

263 Next, we investigated how caQTL-mediated IMD heritability is distributed with respect to
264 TSS. If caQTLs beyond the typical *cis*-eQTL window of 1 megabase (Mb) around the genes' TSS
265 explain some proportion of IMD heritability, then *cis*-eQTL analyses might require a wider
266 window to detect disease-associated eQTLs. Across the 7 diseases, caQTLs within 500 kb of the
267 TSS of expressed genes explained almost all of the caQTL-mediated IMD heritability (92–100%;
268 **Figure 3C**), indicating that regulatory variants are most likely within 500 kb of the target gene's
269 TSS and supporting the use of a 1 Mb window for *cis*-eQTL analyses. Depending on the disease,
270 41–66% of the caQTL-mediated IMD heritability was detected in distal peaks further than 10 kb
271 from the TSS of expressed genes, further supporting the analysis of regulatory variants beyond
272 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 $h^2_{\text{med; caQTL}}$ are distributed mostly within 500 kb of genes' TSS, the IMD loci 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.

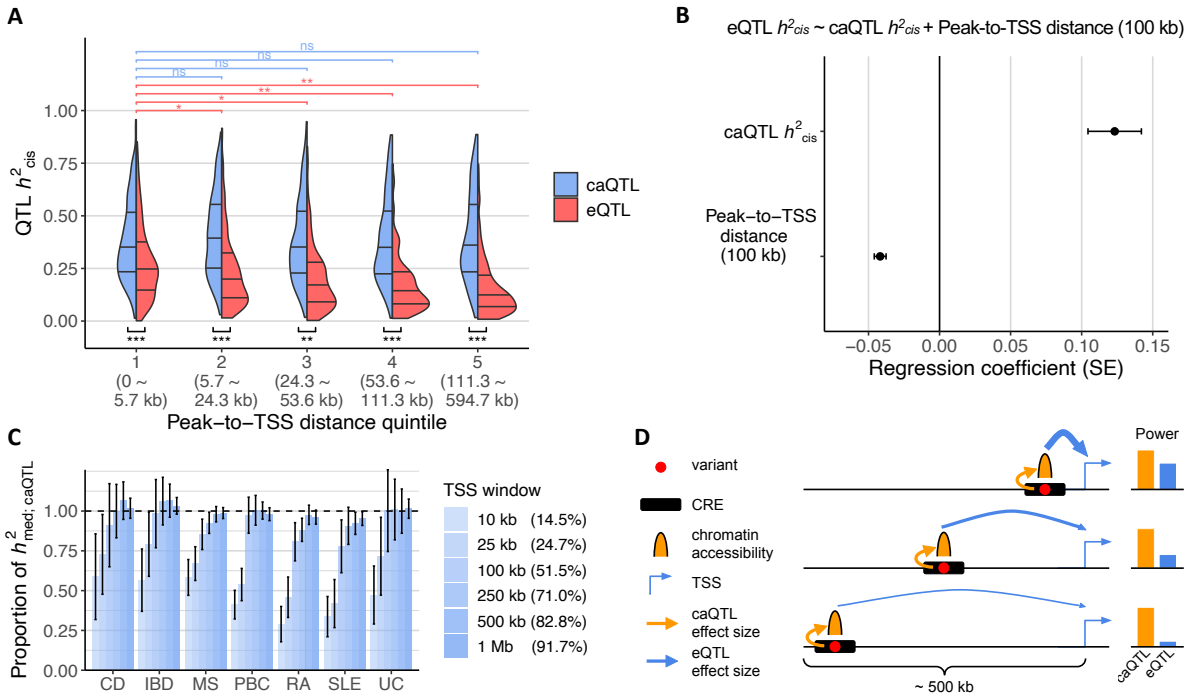


Figure 3. Effect of peak-to-TSS distance on *cis*-heritability of caQTLs and eQTLs and IMD heritability. (**A**) Distribution of *cis*-heritability (h^2_{cis}) of caQTLs and eQTLs by peak-to-TSS distance quintiles. The ranges of peak-to-TSS distance are shown in parentheses. The comparisons shown on 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.

279 **Increasing the sample size reveals some eQTL colocalization**

280 For genetic association studies, increasing the sample size is a way to increase statistical power.
281 Therefore, we meta-analyzed four LCL eQTL summary statistics[55,56,69,70], leading to a total sample
282 size of 1,128 individuals. We performed colocalization analysis using the meta-analyzed summary
283 statistics to evaluate whether effectively increasing the sample sizes would uncover more disease-
284 associated eQTLs in LCLs, especially in loci where a caQTL already showed IMD colocalization.
285 Up to 6 additional loci showing eQTL colocalization were thus detected for each IMD (**Figure 4A**
286 and **Supplementary Table 4**). For example, *CIITA* is the class II major histocompatibility
287 complex transactivator, which causes severe immunodeficiency if dysfunctional[71]. The *CIITA*
288 locus is associated with IBD right below the genome-wide significance level (rs10445003, p =
289 7.5×10^{-8}), and it colocalized with a caQTL signal, but initially not with any eQTL (**Supplementary**
290 **Figure 8A**). However, the meta-analyzed statistics showed a stronger association to *CIITA*
291 expression ($p = 6.7\times10^{-8}$) than without meta-analysis ($p = 5.4\times10^{-4}$) and exhibited a significant
292 colocalization. Interestingly, two of the causal CD genes that previously lacked colocalized
293 eQTLs[15], *CARD9* and *ATG16L1*, showed significant colocalization in the meta-analyzed LCL
294 eQTL data (**Supplementary Figure 8B-C**).

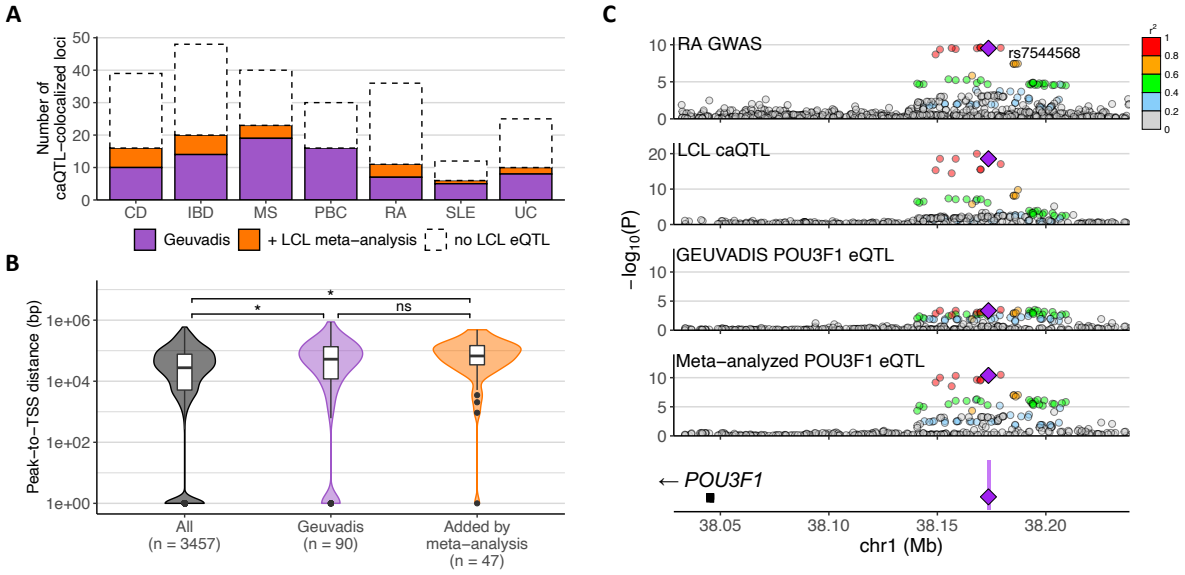


Figure 4. Additional colocalization of caQTL-colocalized IMD loci with meta-analyzed LCL eQTL data. (**A**) Number of caQTL-colocalized IMD loci that showed eQTL colocalization in LCLs. Disease abbreviations along the *x*-axis are as in **Figure 1**. (**B**) Distribution of peak-to-TSS distance of all 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.

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 $p = 5.9\times10^{-6}$ for the meta-analyzed data; **Figure 4B**). For example, a RA-associated locus near *POU3F1*, a neuronal transcription factor that is also induced by interferon[72], 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.

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

309 Despite uncovering more eQTL colocalizations through meta-analysis, more than 40% of the
310 caQTL-colocalized loci nevertheless showed no colocalization with an eQTL in LCLs (**Figure**
311 **4B**). We investigated whether these "caQTL only" peaks might be inactive regulatory elements.
312 We quantified the active histone mark levels for H3K27ac, H3K4me1, and H3K4me3 at
313 colocalized caQTL peaks in LCLs and then compared their levels between the "caQTL and eQTL"
314 and "caQTL only" loci. On average, H3K27ac marks were stronger at "caQTL and eQTL" peaks,
315 supporting that the corresponding regulatory elements might be more active (**Supplementary**
316 **Figure 9**). H3K4me3 marks were detected more often at "caQTL and eQTL" peaks, leading to
317 stronger average signal. In contrast, H3K4me1 levels were highly similar between the two sets of
318 peaks. Although "caQTL only" peaks generally showed lower levels of active histone marks,
319 several individual "caQTL only" peaks showed comparable levels. These peaks could be inactive
320 *cis*-regulatory elements in LCLs that affect gene expression in a different cellular context.
321 Therefore, we next examined whether those caQTLs appear as eQTLs in other immune cell types.

322 **Various immune cell types exhibit eQTL colocalization where LCLs did not**

323 We downloaded eQTL summary statistics generated from 26 naïve and stimulated immune cell
324 types[22,23,73–75] to search for eQTLs that may correspond to the remaining, IMD-colocalized caQTLs
325 (**Supplementary Table 5**). The profiled cell types range from B cells and monocytes to subtypes
326 of T cells, as well as stimulated T cells and macrophages. We tested for colocalization of these
327 eQTLs to IMD associations.

328 25–42% of the caQTL-colocalized IMD loci that were missing eQTL colocalizations in
329 LCLs showed eQTL colocalizations in at least one immune cell type (**Figure 5A** and
330 **Supplementary Table 6**). The overlap of LCL caQTLs with non-LCL immune cell eQTLs was
331 greater than expected by chance for 5 of the 7 IMDs ($p < 0.00714$ (Bonferroni-corrected), Fisher's
332 exact test, for CD, IBD, PBC, RA, and UC; **Supplementary Table 7**), suggesting that the caQTLs
333 found in LCLs may also show regulatory function in those immune cell types. Comparing across
334 the datasets, we found that the number of loci with eQTL colocalization varied depending on the
335 cell type, but that the effect of the sample size was more profound ($r^2 = 0.60–0.79$; **Figure 5B** and
336 **Supplementary Table 8**). We meta-analyzed eQTL data of 3 immune cell types with multiple

337 sources – naïve CD4$^+$ T cell, monocyte, and memory regulatory T cell (Treg) – and this also
338 increased the number of loci with significant eQTL colocalization (**Supplementary Table 9**).
339 Altogether, these results suggest that although generating eQTL data in more cell types and cell
340 states uncovers context-specific eQTLs, increasing the sample size should also be a priority to
341 ensure sufficient statistical power.

342 We investigated the potential reasons why IMD loci that colocalized with caQTLs in LCLs
343 showed eQTLs not in LCLs but in other immune cells. First, LCL caQTLs may correspond to gene
344 regulatory elements that exert their effects on gene expression in a different cellular context. For
345 instance, monocyte H3K27ac levels in the "caQTL only" loci where monocyte eQTLs colocalized
346 were higher than those with no monocyte eQTLs (**Supplementary Figure 10**). Second, some
347 examples, such as for *TNFSF15*, were due to cell-type specific gene expression (**Figure 6**): despite
348 a significant colocalization of IBD with a caQTL in LCLs in the *TNFSF15* locus, disease-
349 associated eQTL signal was detected only in monocytes (**Figure 6A**). Tumor necrosis factor-like
350 cytokine 1A (TL1A), the protein encoded by *TNFSF15*, is secreted by monocytes and many other
351 cells to activate helper T cells, regulatory T cells, and B cells[76]. *TNFSF15* expression was low in
352 LCLs (mean transcript per million [TPM] = 0.30), but higher in monocytes (mean TPM = 2.23).
353 Of the genes that showed exclusively monocyte eQTL colocalization, those with low expression
354 (mean TPM < 1) in LCLs generally showed higher expression in monocytes (**Figure 6B-C**).

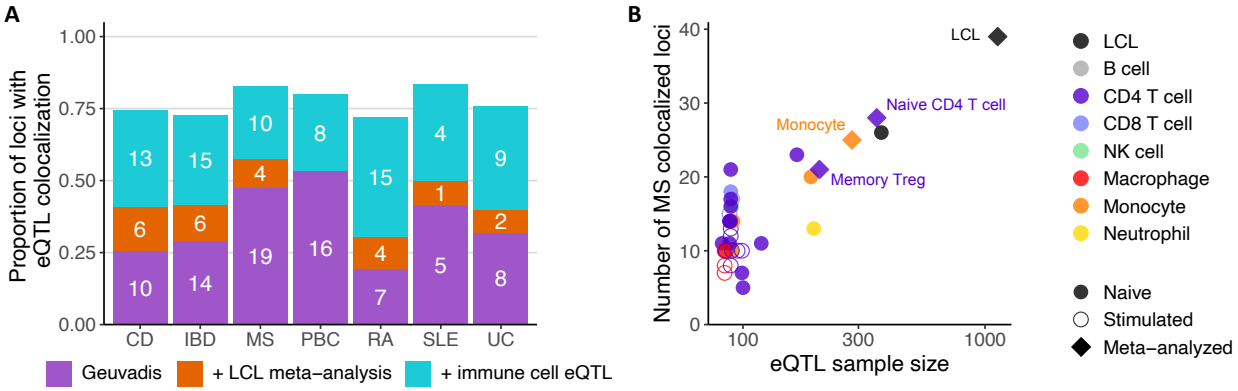
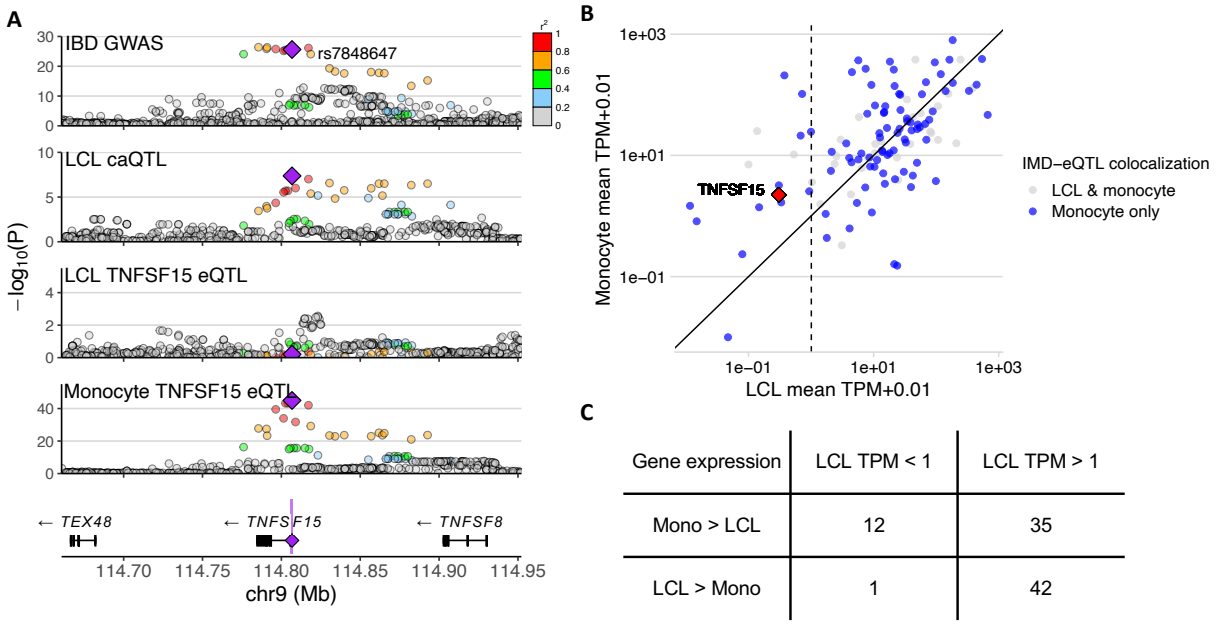


Figure 5. Added utility of various immune cell eQTL data. (**A**) Number of loci that additionally colocalized with eQTLs by LCL meta-analysis (orange) and immune cell data (cyan) compared to the original analysis with Geuvadis LCL eQTL data (purple). The height of the bar is the proportion of loci with 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.



Gene expression	LCL TPM < 1	LCL TPM > 1
Mono > LCL	12	35
LCL > Mono	1	42

Figure 6. IMD loci that colocalized with an eQTL in monocytes, but not in LCLs, even though they colocalized 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).

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

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

363 **Discussion**

364 A lack of the link between many noncoding GWAS loci to the associated variants' gene regulatory
365 effects have posed challenges in understanding their genetic mechanism[15]. Various hypotheses

366 have been presented, including disease genes showing more complex gene regulation[77], context
367 specificity of gene regulation[17], and a combination of both, due to selective constraints against
368 damaging eQTLs[18]. A better evaluation of the potential reasons for the missing regulation will
369 guide future data generation projects to elucidate disease-associated loci. To determine why some
370 loci might lack colocalized eQTLs, we focused on chromatin accessibility, which is a molecular
371 phenotype affected by regulatory variants more directly than are eQTLs. We approached this
372 question with mediated heritability analysis[16] and colocalization analysis[36,37].

373 We found that caQTLs in LCLs mediate a significant proportion of heritability for many
374 autoimmune diseases. In contrast, LCLs did not appear to be an effective cell type to model gene
375 regulatory effects in allergic diseases. The h^2_{med} / h^2_{SNP} estimates for caQTLs were higher than
376 those of eQTLs in most autoimmune diseases, even though the smaller sample size of caQTL data
377 (*i.e.*, 100 vs. 373) could bias caQTLs' estimates toward zero[16]. We also showed that disease-
378 associated chromatin accessibility effects often share the genetic signal with gene expression
379 effects but that there are also many loci without an eQTL detected in LCLs.

380 By focusing on disease-associated caQTL that lacked significant eQTLs, we explored how
381 additional colocalized eQTL effects could be uncovered. First, increasing the sample sizes for the
382 eQTL statistics via meta-analysis demonstrated that more eQTL colocalizations can be detected
383 with increased statistical power and robustness[19]. These results are consistent with the hypothesis
384 that disease-associated eQTLs are typically weaker and distal due to negative selection against
385 large expression changes for causal genes[18]. Second, many caQTLs in LCLs without eQTLs in
386 LCLs showed eQTL colocalization in other immune cell types. Context specificity of eQTLs has
387 been widely considered to be the primary explanation for the difficulty of pinpointing disease-
388 associated eQTLs[17,22,23,74]. Our observation that many IMD-colocalized caQTLs in LCLs show
389 eQTLs in other immune cell types suggests that caQTL effects may be shared across cell types
390 whereas eQTL effects are more context-specific[23]. If a shared set of transcription factors are
391 expressed in similar yet distinct cell types, like the immune cells, genetic variants affecting their
392 DNA binding would affect chromatin accessibility similarly. On the other hand, regulation of gene
393 expression is a result of multiple regulatory elements, each binding multiple transcription factors[78],
394 so the measured effect of a regulatory variant on gene expression likely depends much more on
395 the cellular context. In such cases, eQTL data for the specific cellular contexts need to be generated
396 in future studies to uncover genetic signal shared with a complex trait or disease. All in all, our

397 results suggest that both increasing the sample size and generating gene expression data from more
398 relevant cellular contexts would be useful strategies for discovering more disease-associated
399 eQTLs.

400 Finally, we demonstrated that caQTLs can reveal the regulatory variant effect of disease-
401 associated variants that may have been difficult to detect with eQTLs, particularly in TSS-distal
402 regions. Although caQTLs cannot directly identify the target gene or the causal cellular context,
403 we anticipate that integrated analyses can improve the power to detect weaker eQTL signals, as
404 multi-trait GWAS analyses have shown[79]. Moreover, integrating multiple molecular QTL data,
405 like transcription factor binding QTLs and histone mark QTLs, may highlight the regulatory
406 elements associated with the GWAS phenotype, which may ultimately contribute to identifying
407 the causal gene[33]. Therefore, we anticipate that generation of data across the various molecular
408 phenotypes upstream of gene expression for QTL analyses will be informative.

409 **Limitations of the study**

410 The h^2_{med} / h^2_{SNP} estimates can be biased because of insufficient sample size of the QTL data. Thus,
411 the proportion of mediated IMD heritability could change based on the specific caQTL and eQTL
412 data. Colocalization analysis tests whether the QTL and GWAS data likely share genetic signal,
413 and such shared signal could arise from either causal mediation or pleiotropy. Therefore, further
414 experiments are needed to establish causality of colocalized eQTL genes. Lastly, we hypothesized
415 that caQTL effects are often shared across cell types, but we did not have access to relevant data
416 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 ERP110508[30]. The available files were cram alignment files mapped to the b37 reference genome, so we extracted unique read pairs using SAMtools[80] and bamtofastq command from bedtools[81]. 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 2[82] 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)[83] and potential PCR duplicates using scripts from WASP[84].

To represent peaks across the samples, we subsampled 3 million read pairs from each and pooled them. Then, we used MACS2[85] 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) samples[86] 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 samples[56] from the Expression Atlas[87]. These data consisted of TPM values of genes as processed by the Expression Atlas. 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.

447 **LCL samples' genotype preparation**

448 To utilize the genotype calls of the highest quality, we downloaded the high-coverage 1000
449 Genomes (1kG) Project data[88]. Of the European samples with ATAC-seq or RNA-seq data, 14
450 samples had genotypes derived from microarrays[89], and the remaining samples had genotypes
451 derived from the high-coverage whole-genome sequencing data. The samples with only
452 microarray-based genotypes that needed imputation are listed in **Supplementary Table 10**. We
453 lifted over the microarray data based on the hg19 reference genome to GRCh38 and filtered for
454 variants present in the high-coverage 1kG data. We first imputed microarray genotype data using
455 the TOPMed imputation server[90] and extracted SNPs with imputation $R^2 \geq 0.5$ and imputed the
456 rest of the variants, including short indels, using high-coverage data with Beagle5.2[91,92]. After
457 keeping only variants with DR2 ≥ 0.7, we merged the imputed genotypes with the high-coverage
458 1kG genotypes.

459 **Curation of IMD GWAS data**

460 We downloaded GWAS summary statistics for 13 IMDs, including 11 autoimmune diseases
461 (autoimmune thyroid disease (ATD)[39], celiac disease (CEL)[40], Crohn's disease (CD)[41],
462 inflammatory bowel disease (IBD)[41], juvenile idiopathic arthritis (JIA)[42], multiple sclerosis (MS)[43],
463 primary biliary cholangitis (PBC)[39], rheumatoid arthritis (RA)[44], systemic lupus erythematosus
464 (SLE)[45], ulcerative colitis (UC)[41], and vitiligo (VIT)[46]) and 2 allergic diseases (allergy (ALL)[47]
465 and asthma (AST)[47]). For each disease, we searched for more recent studies with larger sample
466 sizes and prioritized those with genome-wide statistics, rather than those with only Immunochip
467 variants[93]. To compare with non-immune diseases, we also downloaded summary statistics for 3
468 non-immune diseases (type 2 diabetes (T2D)[48], coronary artery disease (CAD)[49] and schizophrenia
469 (SCZ)[50]). In this study, we analyzed only those GWAS summary statistics derived from cohorts
470 of individuals with European ancestries.

471 Since the LCL samples' genotypes are based on the GRCh38 reference genome, we lifted
472 over any GWAS data based on b37 reference genome to the GRCh38 genomic coordinates. Briefly,
473 we formatted the summary statistics as bed files and used the liftOver tool[94] to convert them to
474 GRCh38 genomic coordinates. Then, to ensure that the reference alleles match the sequences of
475 the GRCh38 reference genome, we used gwas2vcf tool[95].

476 **Chromatin accessibility QTLs in lymphoblastoid cell lines**

477 First, we quantified the chromatin accessibility levels at ATAC-seq peaks identified earlier. We
478 counted the number of read fragments overlapping each peak using featureCounts[96]. For each
479 sample, the read counts were normalized for library size using trimmed mean of M-values[97] so that
480 the values are comparable across the samples. Then, the phenotype values were further normalized
481 to follow a standard normal distribution across the samples, using quantile normalization. Peaks
482 with counts per million (CPM) < 0.8 or counts < 10 for more than 20% of the samples were
483 discarded.

484 Next, we performed a principal component analysis (PCA) on the phenotype matrix to
485 derive potential latent covariates. We selected the number of principal components (PC) to
486 incorporate in the regression model based on the Buja and Eyuboglu algorithm[98] that is
487 implemented in PCAForQTL[99]. Ultimately, we accounted for sex, library size, 3 genotype PCs
488 and 13 phenotype PCs in the QTL analysis. We performed genetic association tests on variants
489 within 200 kilobases (kb) of the peak using tensorQTL[100]. We discarded variants with minor allele
490 frequency less than 5%.

491 **IMD heritability enrichment in accessible regions of LCLs**

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

498 **Mediated heritability estimation of QTLs**

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

501 *caQTL-mediated heritability*. We estimated the 'expression scores' for chromatin accessibility
502 in LCLs using individual genotypes and phenotypes. We analyzed the same set of peaks and
503 accounted for the same covariates as we did for the QTL analysis above. For mediated heritability

504 estimation, we accounted for baselineLD v2.2 annotation in hg38 without the QTL annotations, as
505 they could be redundant. The estimand of interest is the proportion of heritability mediated by
506 caQTLs (h^2_{med} / h^2_{SNP}).

507 To evaluate whether certain peak sets are enriched for mediated heritability, we utilized
508 the gene set analysis functionality. For peak strength, we considered the 95% percentile CPM value
509 of each peak and stratified the peaks into quintiles. For histone marks, we first generated histone
510 ChIP-seq peaks using data from Delaneau *et al*.[52]. Similar to calling ATAC-seq peaks, we sampled
511 3 million reads per sample and merged them before applying MACS2[85]. We downloaded 3 control
512 ChIP-seq data from ENCODE to use as input (File IDs: ENCFF066RCS, ENCFF159XTB, and
513 ENCFF850RIE)[83]. Then, we curated sets of ATAC-seq peaks that overlapped H3K27ac,
514 H3K4me1, and H3K4me3 ChIP-seq peaks by at least 1 bp. ATAC-seq peaks that overlapped
515 H3K27ac but not H3K4me3 regions were labeled as 'enhancer-like signature', while those that
516 overlapped H3K27ac and H3K4me3 regions were labeled as 'promoter-like signature'. The
517 estimand of interest is the proportion of mediated heritability explained by the peak set (peak set
518 h^2_{med} / total h^2_{med}).

519 *Comparison with eQTL-mediated heritability.* First, we estimated h^2_{med} / h^2_{SNP} of eQTLs (*i.e.*,
520 $h^2_{med;\ eQTL}$ / h^2_{SNP}) the same way as we did for that of caQTLs. Then, we also estimated h^2_{med} / h^2_{SNP}
521 of caQTLs and eQTLs together (*i.e.*, $h^2_{med;\ caQTL \cup eQTL}$) with MESC meta-analysis[16]. caQTLs and
522 eQTLs were also stratified as separate sets to account for potential differences in the relationship
523 of QTL *cis*-heritability and h^2_{med}. This meta-analyzed estimate is effectively the amount of
524 heritability mediated by either caQTLs or eQTLs in LCLs. This estimate reveals the overall
525 relationship between heritability mediated by caQTLs and eQTLs. For instance, the estimate of
526 the heritability mediated exclusively by caQTLs would be $h^2_{med;\ just\ caQTL} = h^2_{med;\ caQTL \cup eQTL} - h^2_{med;}$
527 $_{caQTL}$. The estimate of mediated heritability shared by caQTLs and eQTLs is $h^2_{med;\ caQTL \cap eQTL} =$
528 $(h^2_{med;\ caQTL} + h^2_{med;\ eQTL}) - h^2_{med;\ caQTL \cup eQTL}$.

529 **Colocalization analyses**

530 *caQTL and eQTL colocalization with IMD GWAS.* First, we selected candidate colocalization
531 loci by filtering for overlapping 'significant' variants. The candidate loci met the following
532 conditions: (1) lead IMD association at $p < 10^{-6}$, (2) lead caQTL or eQTL association at $p < 10^{-4}$,
533 and (3) at least one variant simultaneously showed caQTL or eQTL χ^2 statistics greater than 0.8 ×

534 lead χ^2 statistics for the caQTL or eQTL and IMD association χ^2 statistics greater than $0.8 \times \chi^2$
535 statistics for the IMD lead variant in the locus. Then, we applied gwas-pw[37] on the variants within
536 100 kb of the lead variant. We considered loci with posterior probability of colocalization (PP3) >
537 0.98 to be colocalized[57].

538 *Colocalization of caQTL with eQTL*. We performed a colocalization analysis of caQTLs and
539 eQTLs to curate a set of loci where the same genetic signal likely explains both associations. The
540 pairs of caQTLs and eQTLs reveal the distance between the regulatory element and the target
541 gene's TSS. We selected candidate colocalization loci with: (1) IMD association at $p < 10^{-6}$, (2)
542 QTL association at $p < 10^{-4}$, and (3) the two lead variants showed LD $r^2 > 0.8$. We applied gwas-
543 pw[37] on the variants within 200 kb of the tested caQTL peak. We considered loci with posterior
544 probability of colocalization (PP3) > 0.98 to be colocalized[57].

545 **Overlap of protein factor ChIP-seq and colocalized caQTL peaks**

546 We searched for any protein factors detected at the colocalized caQTL peaks using the Cistrome
547 database[58], which we accessed on July 24, 2024. We considered only ChIP-seq data from immune
548 cell types, progenitors and stem cells that can differentiate into immune cells. For each ChIP-seq
549 peak, we searched for those that show overlap of more than 50% with one of 305 caQTL peaks
550 that colocalized with IMD GWAS signals (**Supplementary Table 1**). Each protein factor detected
551 in the caQTL peaks and the cell type used to generate the ChIP-seq data are listed in
552 **Supplementary Table 3**.

553 **Enrichment of biological processes**

554 To test whether colocalized genes are likely relevant to autoimmune diseases, we surveyed which
555 Biological Process Gene Ontology (GO) terms were overrepresented compared to all the genes within
556 500 kb of each IMD association signal. Enrichment of biological processes was evaluated using
557 Protein Analysis Through Evolutionary Relationships (PANTHER)[101]. The foreground list
558 comprised all of the genes whose eQTL signal colocalized with one of the 7 autoimmune diseases
559 (CD, IBD, MS, PBC, RA, SLE, and UC). The background list of genes was all of the genes within
560 500 kb of each IMD lead variant for which we observed colocalization. Moreover, we tested
561 whether colocalized caQTLs without eQTLs are closer to genes related to immune processes than
562 expected by chance. For this, we used Genomic Regions Enrichment of Annotations Tool

563 (GREAT)[64]. The foreground list comprised ATAC-seq peaks at IMD loci showing colocalization
564 only with caQTLs and not with eQTLs. The background list is all ATAC-seq peaks identified in
565 LCLs that we tested for caQTL association.

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

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

571 MESC analysis uses REML implemented in GCTA[102] to estimate QTL *cis*-heritability. We
572 referred to its output and compared the *cis*-heritability of caQTLs and eQTLs based on peak-to-
573 TSS distance. To visualize the distribution of *cis*-heritability estimates, we grouped pairs of
574 colocalized caQTLs and eQTLs based on peak-to-TSS distance quintiles.

575 **LCL eQTL meta-analysis**

576 We downloaded LCL eQTL data from 3 studies through eQTL Catalogue release 6[103]. The sample
577 sizes were 190, 147, and 418 for GENCORD[69], GTEx[55], and TwinsUK[70], respectively. We meta-
578 analyzed the summary statistics using the inverse variance weighted fixed effects model. If a
579 variant was missing in a subset of the studies, then only the available statistics were meta-analyzed.
580 We used these meta-analyzed statistics to perform colocalization the same way as earlier
581 colocalization analyses.

582 **Colocalization analysis with immune cell eQTL data**

583 We downloaded eQTL data for various immune cell types from the eQTL Catalogue release 6[103].
584 The source studies from the eQTL Catalogue include BLUEPRINT[73], DICE[22], Alasoo *et al.*[23] and
585 Bossini-Castillo *et al.*[75]. The represented immune cell types include T cell subtypes, B cells,
586 neutrophils, and monocytes. We also separately downloaded data for $CD4^+$ T cells with and
587 without stimulation[74]. The selection of candidate loci and colocalization analysis on them followed
588 the same procedure as that for other QTLs.

589 We also evaluated whether meta-analysis of eQTL data can increase the number of loci
590 with eQTL colocalization. We meta-analyzed eQTL data for 3 immune cell types with multiple

591 sources – naïve $CD4^+$ T cell, monocyte, and memory regulatory T cell (Treg) – using the inverse
592 variance weighted fixed effects model. Specifically, we meta-analyzed naïve $CD4^+$ T cell eQTL
593 summary statistics from Soskic *et al.*[74], BLUEPRINT[73], and DICE[22]. We meta-analyzed monocyte
594 eQTL summary statistics from BLUEPRINT[73] and DICE[22]. We meta-analyzed memory Treg
595 eQTL summary statistics from Bossini-Castillo *et al*.[75] and DICE[22]. We used these meta-analyzed
596 statistics to perform colocalization the same way as earlier colocalization analyses.

597

598 **Data availability**

599 Processed data and code for generating the figures presented in the manuscript are available at
600 https://github.com/BulykLab/IMD-colocalization-manuscript-figures. Genotype data are
601 downloaded from the 1000 Genomes Project (https://www.internationalgenome.org/data-
602 portal/data-collection/30x-grch38). LCL ATAC-seq data are downloaded from ENA
603 (https://www.ebi.ac.uk/ena/browser/view/PRJEB28318). LCL RNA-seq count matrix is
604 downloaded from Expression Atlas[87] (https://www.ebi.ac.uk/gxa/experiments/E-GEUV-
605 1/Results). GWAS data for the studies listed in the Methods section are downloaded from the
606 GWAS catalog[104] (https://www.ebi.ac.uk/gwas/). eQTL summary statistics are downloaded
607 either from the eQTL Catalogue[103] (https://www.ebi.ac.uk/eqtl/) or the source publication.

608 **Acknowledgments**

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

611 **Author contributions**

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

614 **Declaration of interests**

615 The authors declare no competing interests.

616 **References**

617 1. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J.*
618 *Hum. Genet.* **101**, 5–22 (2017).
619 2. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
620 3. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in
621 regulatory DNA. *Science* **337**, 1190–1195 (2012).
622 4. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11
623 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
624 5. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance
625 discovery from GWAS. *PLoS Genet.* **6**, (2010).
626 6. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic
627 architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
628 7. O'Connor, L. J. *et al.* Extreme Polygenicity of Complex Traits Is Explained by Negative
629 Selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).
630 8. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic
631 to Omnigenic. *Cell* **169**, 1177–1186 (2017).
632 9. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–
633 243 (2021).
634 10. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes
635 underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).
636 11. Morris, J. A. *et al.* Discovery of target genes and pathways at GWAS loci by pooled single-cell
637 CRISPR screens. *Science* **380**, eadh7699 (2023).
638 12. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human
639 genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
640 13. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-
641 disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
642 14. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci.
643 *Genome Biol.* **22**, 49 (2021).
644 15. Connally, N. J. *et al.* The missing link between genetic association and regulatory function. *Elife*
645 **11**, (2022).
646 16. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease
647 mediated by assayed gene expression levels. *Nat. Genet.* 1–8 (2020) doi:10.1038/s41588-020-
648 0625-2.

649 17. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.*
650 **37**, 109–124 (2021).
651 18. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of
652 genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
653 19. Hukku, A. *et al.* Probabilistic Colocalization of Genetic Variants from Complex and Molecular
654 Traits: Promise and Limitations. *Am. J. Hum. Genet.* **108**, 25–35 (2021).
655 20. Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K. Evidence for weak selective
656 constraint on human gene expression. *Genetics* **211**, 757–772 (2019).
657 21. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent
658 manner. *Science* **325**, 1246–1250 (2009).
659 22. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene
660 Expression. *Cell* **175**, 1701–1715 (2018).
661 23. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for
662 enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
663 24. Nathan, A. *et al.* Single-cell eQTL models reveal dynamic T cell state dependence of disease loci.
664 *Nature* **606**, 120–128 (2022).
665 25. Field, A. & Adelman, K. Evaluating Enhancer Function and Transcription. *Annu. Rev. Biochem.*
666 **89**, 213–234 (2020).
667 26. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–
668 235 (2010).
669 27. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin
670 structure, and transcription. *Science* **342**, 744–747 (2013).
671 28. Waszak, S. M. *et al.* Population Variation and Genetic Control of Modular Chromatin Architecture
672 in Humans. *Cell* **162**, 1039–1050 (2015).
673 29. Degner, J. F. *et al.* DNase-I sensitivity QTLs are a major determinant of human expression
674 variation. *Nature* **482**, 390–394 (2012).
675 30. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal
676 interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
677 31. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat.*
678 *Rev. Genet.* **16**, 197–212 (2015).
679 32. Wu, Y. *et al.* Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large
680 fraction of GWAS signals shared with molecular phenotypes. *Cell Genomics* 100344 (2023)
681 doi:10.1016/j.xgen.2023.100344.
682 33. Jeong, R. & Bulyk, M. L. Blood cell traits' GWAS loci colocalization with variation in PU.1

683 genomic occupancy prioritizes causal noncoding regulatory variants. *Cell genomics* **3**, 100327
684 (2023).
685 34. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes.
686 *Nature* **518**, 317–329 (2015).
687 35. Farh, K. K. H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants.
688 *Nature* **518**, 337–343 (2015).
689 36. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association
690 studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
691 37. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits.
692 *Nat. Genet.* **48**, 709–717 (2016).
693 38. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide
694 association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
695 39. Cordell, H. J. *et al.* An international genome-wide meta-analysis of primary biliary cholangitis:
696 Novel risk loci and candidate drugs. *J. Hepatol.* **75**, 572–581 (2021).
697 40. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene
698 expression. *Nat. Genet.* **42**, 295–302 (2010).
699 41. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple
700 integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
701 42. López-Isac, E. *et al.* Combined genetic analysis of juvenile idiopathic arthritis clinical subtypes
702 identifies novel risk loci, target genes and key regulatory mechanisms. *Ann. Rheum. Dis.* **80**, 321–
703 328 (2021).
704 43. Patsopoulos, N. A. *et al.* Multiple sclerosis genomic map implicates peripheral immune cells and
705 microglia in susceptibility. *Science* **365**, (2019).
706 44. Ishigaki, K. *et al.* Multi-ancestry genome-wide association analyses identify novel genetic
707 mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).
708 45. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive
709 immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464
710 (2015).
711 46. Jin, Y. *et al.* Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci
712 and highlight key pathways and regulatory variants. *Nat. Genet.* **48**, 1418–1424 (2016).
713 47. Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for
714 biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
715 48. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture
716 and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).

717 49. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for
718 coronary artery disease. *Nat. Genet.* **43**, 333–8 (2011).
719 50. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from
720 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–7 (2014).
721 51. van Mierlo, H. C., Broen, J. C. A., Kahn, R. S. & de Witte, L. D. B-cells and schizophrenia: A
722 promising link or a finding lost in translation? *Brain. Behav. Immun.* **81**, 52–62 (2019).
723 52. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene
724 expression. *Science* **364**, (2019).
725 53. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters
726 and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
727 54. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and
728 mouse genomes. *Nature* **583**, 699–710 (2020).
729 55. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues.
730 *Science* **369**, 1318–1330 (2020).
731 56. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in
732 humans. *Nature* **501**, 506–511 (2013).
733 57. Kundu, K. *et al.* Genetic associations at regulatory phenotypes improve fine-mapping of causal
734 variants for twelve immune-mediated diseases. *Nat. Genet.* **44**, 0–23 (2022).
735 58. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*
736 **12**, R83 (2011).
737 59. Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein*
738 *Sci.* **31**, 8–22 (2022).
739 60. Sanmartí, R., Ruiz-Esquide, V., Bastida, C. & Soy, D. Tocilizumab in the treatment of adult
740 rheumatoid arthritis. *Immunotherapy* **10**, 447–464 (2018).
741 61. Khanna, R. & Feagan, B. G. Ustekinumab for the treatment of Crohn's disease. *Immunotherapy* **5**,
742 803–15 (2013).
743 62. Monemi, S. *et al.* Incidence of Gastrointestinal Perforations in Patients with Rheumatoid Arthritis
744 Treated with Tocilizumab from Clinical Trial, Postmarketing, and Real-World Data Sources.
745 *Rheumatol. Ther.* **3**, 337–352 (2016).
746 63. Stevenson, C. *et al.* Essential role of Elmo1 in Dock2-dependent lymphocyte migration. *J.*
747 *Immunol.* **192**, 6062–70 (2014).
748 64. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat.*
749 *Biotechnol.* **28**, 495–501 (2010).
750 65. Tanigawa, Y., Dyer, E. S. & Bejerano, G. WhichTF is functionally important in your open

751 chromatin data? *PLoS Comput. Biol.* **18**, e1010378 (2022).
752 66. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of
753 CRISPR perturbations. *Nat. Genet.* **51**, (2019).
754 67. Zuin, J. *et al.* Nonlinear control of transcription through enhancer-promoter interactions. *Nature*
755 **604**, 571–577 (2022).
756 68. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and
757 polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, (2021).
758 69. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic
759 variation in gene regulation. *Elife* **2013**, 1–18 (2013).
760 70. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence
761 analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
762 71. Dziembowska, M. *et al.* Three novel mutations of the CIITA gene in MHC class II-deficient
763 patients with a severe immunodeficiency. *Immunogenetics* **53**, 821–9 (2002).
764 72. Hofmann, E. *et al.* Octamer-binding factor 6 (Oct-6/Pou3f1) is induced by interferon and
765 contributes to dsRNA-mediated transcriptional responses. *BMC Cell Biol.* **11**, 61 (2010).
766 73. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune
767 Cells. *Cell* **167**, 1398-1414.e24 (2016).
768 74. Soskic, B. *et al.* Immune disease risk variants regulate gene expression dynamics during CD4+ T
769 cell activation. *Nat. Genet.* **54**, 817–826 (2022).
770 75. Bossini-Castillo, L. *et al.* Immune disease variants modulate gene expression in regulatory CD4+
771 T cells. *Cell Genomics* **2**, 100117 (2022).
772 76. Xu, W.-D., Li, R. & Huang, A.-F. Role of TL1A in Inflammatory Autoimmune Diseases: A
773 Comprehensive Review. *Front. Immunol.* **13**, 891328 (2022).
774 77. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene
775 Discovery in Complex Disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).
776 78. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* **83**,
777 373–392 (2023).
778 79. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG.
779 *Nat. Genet.* **50**, 229–237 (2018).
780 80. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
781 81. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
782 features. *Bioinformatics* **26**, 841–2 (2010).
783 82. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–
784 359 (2012).

785 83. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**,
786 57–74 (2012).
787 84. Van De Geijn, B., Mcvicker, G., Gilad, Y. & Pritchard, J. K. WASP: Allele-specific software for
788 robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
789 85. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
790 86. Banovich, N. E. *et al.* Impact of regulatory variation across human iPSCs and differentiated cells.
791 *Genome Res.* **28**, 122–131 (2018).
792 87. Papatheodorou, I. *et al.* Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.*
793 **48**, D77–D83 (2020).
794 88. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000
795 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).
796 89. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
797 90. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–
798 1287 (2016).
799 91. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-
800 Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
801 92. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale
802 sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
803 93. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101
804 (2011).
805 94. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
806 95. Lyon, M. S. *et al.* The variant call format provides efficient and robust storage of GWAS summary
807 statistics. *Genome Biol.* **22**, 32 (2021).
808 96. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for
809 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–30 (2014).
810 97. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression
811 analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
812 98. Buja, A. & Eyuboglu, N. Remarks on Parallel Analysis. *Multivariate Behav. Res.* **27**, 509–40
813 (1992).
814 99. Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference
815 methods for molecular QTL mapping. *Genome Biol.* **23**, 210 (2022).
816 100. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs.
817 *Genome Biol.* **20**, 228 (2019).
818 101. Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the

PANTHER classification system (v.14.0). *Nat. Protoc.* **14**, 703–721 (2019).

102. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

103. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).

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