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 ($h^2_{\mathrm{med}}/h^2_{\mathrm{SNP}}$) 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 $h^2_{\mathrm{med}}/h^2_{\mathrm{SNP}}$ 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, only weak eQTL variants, often in
59 regions distal to the gene's promoter[21], may remain and affect traits. In fact, eQTL-mediated
60 heritability was enriched in genes showing mutational constraint and those with lower *cis*-
61 heritability[16]. These weaker eQTL effects would require larger sample sizes to be detected with
62 statistical significance and to show colocalization[19]. Second, causal eQTL effects may be specific
63 to cell types that have not been assayed[17]. For example, immune cell eQTLs are highly cell-type
64 specific, and eQTL effects specific to some immune cell types may mediate immune disease risk[22].
65 Specificity of eQTL effects can also be limited to specific cell states[23,24]. Detecting cell-type or
66 cell-state specific eQTL effects requires the necessary gene expression data sets from the relevant
67 cell types and states[17], which has been a limiting resource for such analyses.
68 To investigate why disease-associated eQTL signals have been missing, we focused on
69 immune-mediate diseases (IMDs) as a model set of complex traits. We aimed to collect IMD-
70 associated loci that are expected to show eQTL signals in some cell type. Since active regulatory
71 elements coordinate target gene expression[25], we reasoned that variants that affect chromatin
72 phenotypes at regulatory elements, such as transcription factor (TF) binding[26–28] and chromatin
73 accessibility[29,30], have the potential to impact gene expression[31]. These chromatin phenotypes may
74 show detectable genetic effects even when an eQTL effect in the same cell type was not identified
75 in the locus[32]. For example, only about 20% of lymphoblastoid cell lines' (LCLs') PU.1 binding
76 QTLs (bQTLs) that colocalized with blood cell traits' association showed an eQTL effect for a
77 nearby gene in LCLs[33].
78 Here, we analyzed genetic and functional genomic (*i.e.*, ATAC-seq and RNA-seq) data in
79 LCLs. LCLs are derived from B lymphocytes, and their *cis*-regulatory elements were enriched for
80 variants associated with some IMDs[34,35]. We evaluated whether chromatin accessibility QTLs
81 (caQTLs) in LCLs potentially explain IMD associations using mediated heritability analysis[16] and
82 colocalization[36,37]. Then, we searched for disease-associated loci that were significant caQTLs, but
83 not eQTLs.
84 We examined whether the various potential reasons for missing eQTLs can account for
85 IMD-associated loci that are explained by caQTLs but not eQTLs. First, we explored the extent to
86 which eQTLs may have been missed because of limited statistical power. We compared *cis*-
87 heritability of colocalized caQTLs and eQTLs stratified by distance between the accessible region
88 and the transcription start site of the associated gene. We also investigated whether meta-analysis

89 of published LCL eQTL summary statistics, in order to effectively increase the sample size, can
90 uncover previously missed eQTLs. Second, we surveyed whether cell-type specificity of
91 regulatory variant effect may account for the missing regulation. We surveyed various immune
92 cell eQTL data to identify loci with which they colocalize even if LCL eQTLs did not colocalize
93 with those loci.

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

100 Results

101 Accessible chromatin in LCLs explains a significant proportion of immune-mediated 102 disease heritability

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

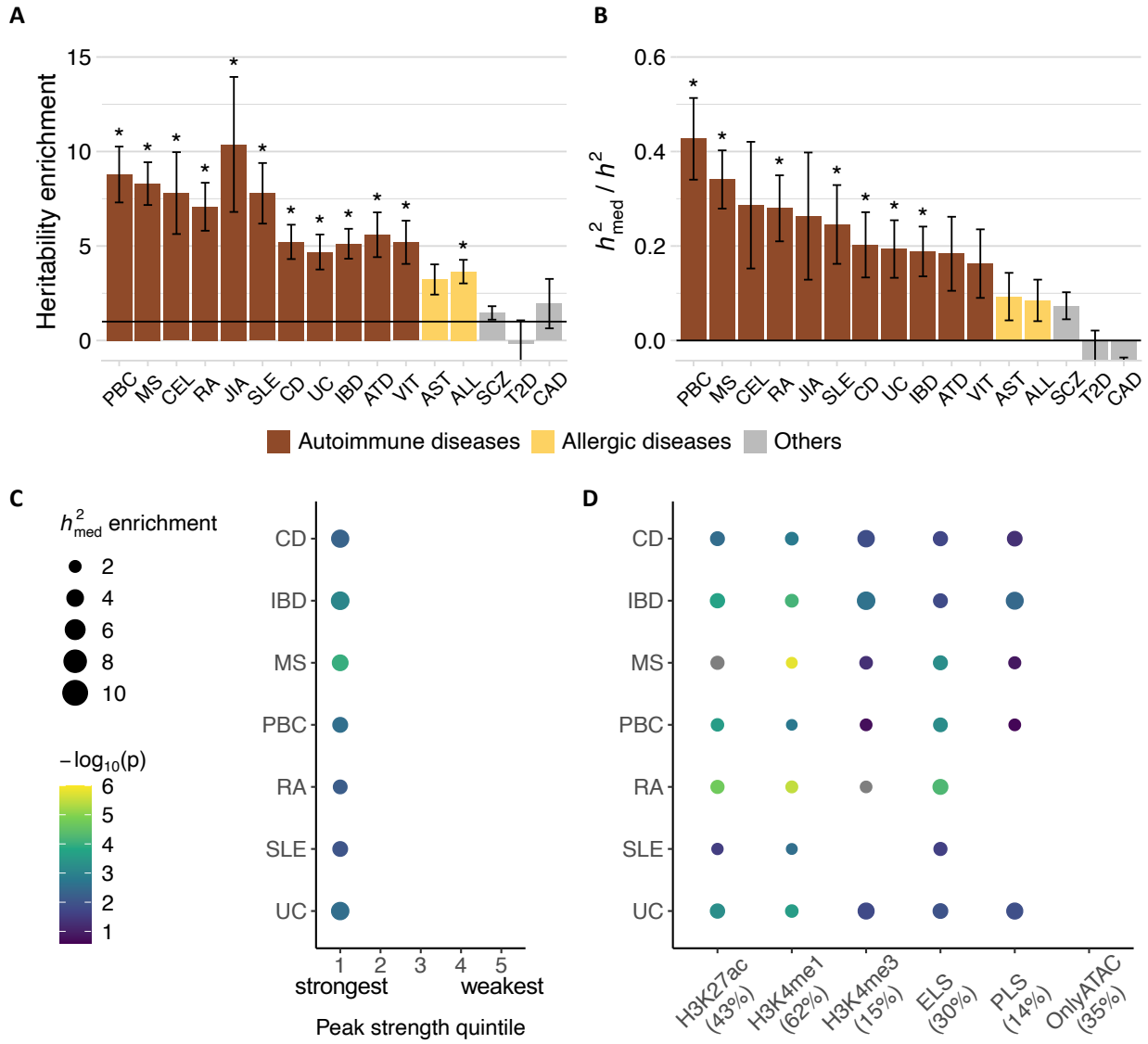


Figure 1. Immune disease heritability mediated by chromatin accessibility in LCLs. (**A-B**) Heritability enrichment in accessible regions in LCLs, based on (**A**) stratified LD score regression (S-LDSC) and (**B**) the proportion of heritability mediated by caQTLs in LCLs based on mediated expression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

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

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

135 To understand which features characterize accessible regions that mediate IMD heritability, we
136 estimated h^2_{med} enrichment (proportion of h^2_{med} / proportion of peaks)[16] in specific sets of
137 accessible regions. We found that peaks with a larger number of nonredundant sequencing reads
138 (*i.e.*, "stronger" peaks) in LCLs showed stronger h^2_{med} enrichment (**Figure 1C**) and thus likely
139 affect IMD-relevant gene expression more than "weaker" peaks do. This observation is consistent
140 with the "Activity-by-Contact" model[9], in which peaks with greater chromatin accessibility and
141 H3K27ac ChIP-seq signal are predicted to have proportional effects on target gene expression.
142 Next, we considered peaks with the active histone marks H3K27ac, H3K4me1, or
143 H3K4me3[28,52]. Consistent with prior observations that putative cell-type-specific regulatory
144 elements marked with H3K27ac and H3K4me1 are enriched for relevant disease associations[34,35],
145 we found that caQTLs with H3K27ac and H3K4me1 marks in LCLs were significantly enriched
146 for mediated IMD heritability (q-value < 0.05, MESC; **Figure 1D**). Strikingly, both peak sets
147 explained almost all of caQTL-mediated IMD heritability (**Supplementary Figure 1**). Peaks with
148 H3K4me3 marks, representative of promoters[53], also showed significant h^2_{med} enrichment for most

149 IMDs (q-value < 0.05, MESC). Peaks with promoter-like signatures (*i.e.*, H3K27ac and
150 H3K4me3)[54] and those with enhancer-like signatures (*i.e.*, H3K27ac, but no H3K4me3)[54] were
151 also enriched for all IMD heritability (q-value < 0.05, MESC). Conversely, peaks without any of
152 the three active histone marks were completely depleted of caQTL-mediated IMD heritability
153 (**Supplementary Figure 1**). These 'ATAC only' peaks were shorter, weaker, and further away
154 from the TSS compared to peaks with active histone marks (**Supplementary Figure 2**). Altogether,
155 these results indicate that peaks characterized as putatively active regulatory elements explain
156 nearly all of caQTL-mediated IMD heritability.

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

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

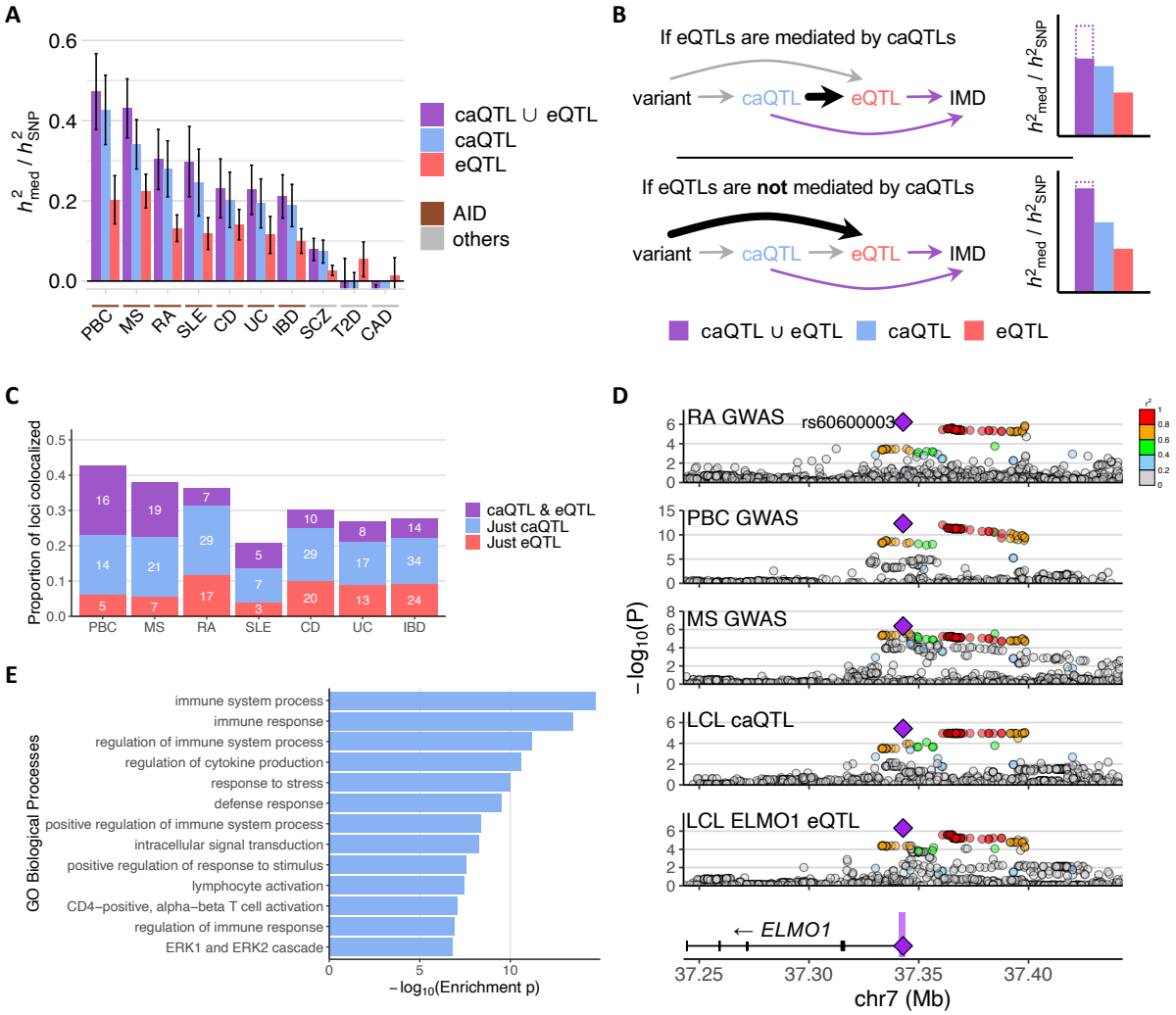


Figure 2. IMD heritability mediated by caQTLs and eQTLs. (**A**) h^2_{med} / h^2_{SNP} estimates of various IMDs for caQTLs, eQTLs, and their union. AID: autoimmune disease. Disease abbreviations along the *x*-axis are as in **Figure 1**. (**B**) Schema of the potential causal relationships between genetic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

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

170 size would potentially bias the estimates toward zero[16]. A possible explanation is that some IMD-
171 associated regulatory variants may show detectable effects on chromatin accessibility, but not on
172 gene expression, in LCLs at the current sample size (n = 373); such loci may account for the
173 missing regulation[15].

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

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

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

195 To confirm that the colocalized genes are relevant to IMD, we tested for their enrichment
196 of Gene Ontology (GO) annotation terms for specific biological processes[58]. Considering all genes
197 within 500 kb of the IMD GWAS lead variants at colocalized loci as background, the genes that
198 showed eQTL colocalization for any IMD were enriched for various immune responses and
199 signaling processes, such as "positive regulation of immune system process" and "regulation of

200 lymphocyte activation" (**Supplementary Figure 4**), indicating that the colocalized genes in LCLs
201 are involved in immune function. For example, *IL6R* and *IL12A* encode direct or indirect targets
202 of approved drugs – Tocilizumab and Ustekinumab – for autoimmune diseases like RA[59] and CD[60].
203 These two genes showed colocalization with both caQTLs and eQTLs in CD and PBC GWAS,
204 respectively (**Supplementary Figure 5A-B**). Increased *IL6R* expression was associated with
205 higher risk for CD, and increased *IL12A* expression was associated with lower risk for PBC and
206 SLE. The former observation is in line with Tocilizumab, a monoclonal antibody to IL-6 receptor,
207 showing efficacy in CD patients[59], although it is not pursued for approval because of potential side
208 effects[61]. Interestingly, *IL6R* and *IL12A* eQTLs did not colocalize with the association signals of
209 RA and CD, respectively, which are the diseases for which these drugs are approved
210 (**Supplementary Figure 5C-D**). Moreover, *ELMO1*, which previously had not been associated
211 with autoimmune diseases, showed eQTL colocalization with RA, PBC, and MS association
212 signals (**Figure 2D)**. In all three, decreased *ELMO1* expression was associated with increased
213 disease risk. In mice, *Elmo1* was required for polarization and migration of B and T lymphocytes[62].
214 Across the IMDs, there were many loci that colocalized with a caQTL but not with an
215 eQTL (**Figure 2C**). These "caQTL only" loci showed enrichment for immune response genes in
216 *cis* compared to all accessible regions in LCLs[63,64], even though the colocalized eQTLs were
217 enriched for immune response genes as well (**Figure 2E** and **Supplementary Figure 4**), indicating
218 that IMD-relevant genes without eQTL colocalization in Geuvadis LCL data[56] are likely found in
219 these loci.

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

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

229 We identified pairs of caQTLs and eQTLs that colocalized with each other[37], which implies
230 that the regulatory variant modulating chromatin accessibility also affects gene expression. Then,
231 the distance between the ATAC peak and the TSS of the eQTL gene (*i.e.*, eGene) is the distance
232 between a regulatory element and its target gene's TSS. We stratified the pairs into peak-to-TSS
233 distance quintiles and compared the eQTL h^2_{cis} distribution of the first quintile (*i.e.*, closest pairs)
234 with that of the later quintiles. We observed that eQTL h^2_{cis} distribution decreased with increasing
235 distance of the paired ATAC peaks from the TSS ($p = 1.0\times10^{-4}$, 2.1×10^{-10}, 1.6×10^{-15}, and $1.7\times10^{-}$
236 ^{20}, respectively, one-sided Wilcoxon rank-sum test; **Figure 3A**), consistent with the negative
237 relationship between promoter-enhancer genomic distance and impact on gene expression[65,66].
238 This result also explains why discovered eQTLs are concentrated near the promoter, where the
239 variants are more likely to show stronger effects[67]. In contrast, caQTL h^2_{cis} distribution was similar
240 across peak-to-TSS distances ($p > 0.05$, one-sided Wilcoxon rank-sum test; **Figure 3A**). For all
241 distance quintiles, caQTL h^2_{cis} was significantly higher than that of the paired eQTLs ($p = 4.0\times10^{-}$
242 ^{23}, 1.3×10^{-24}, 1.5×10^{-19}, 1.1×10^{-28}, and 2.4×10^{-49}, respectively, one-sided paired Wilcoxon rank-
243 sum test; **Figure 3A**), and the contrast between them was greater at more distant quintiles,
244 suggesting that the statistical power to detect and colocalize eQTLs is increasingly lower than that
245 for caQTLs for regulatory effects far from the TSS.
246 Overall, caQTL h^2_{cis} had a significant positive effect ($p = 6.6\times10^{-11}$, linear regression) and
247 peak-to-TSS distance had a negative effect ($p = 2.4\times10^{-23}$, linear regression) on eQTL h^2_{cis} (**Figure**
248 **3B**). Thus, for regulatory variants that showed both caQTL and eQTL signals, those with larger
249 effects on chromatin accessibility tended to exhibit larger effects on gene expression, but their
250 eQTL effects diminished with increasing distance from TSSs.
251 Next, we investigated how caQTL-mediated IMD heritability is distributed with respect to
252 TSS. If caQTLs beyond the typical *cis*-eQTL window of 1 megabase (Mb) around the genes' TSS
253 explain some proportion of IMD heritability, then *cis*-eQTL analyses might require a wider
254 window to detect disease-associated eQTLs. Across the 7 diseases, caQTLs within 500 kb of the
255 TSS of expressed genes explained almost all of the caQTL-mediated IMD heritability (92–100%;
256 **Figure 3C**), indicating that regulatory variants are most likely within 500 kb of the target gene's
257 TSS and supporting the use of a 1 Mb window for *cis*-eQTL analyses. Depending on the disease,
258 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 $h^2_{\text{med; caQTL}}$ 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.

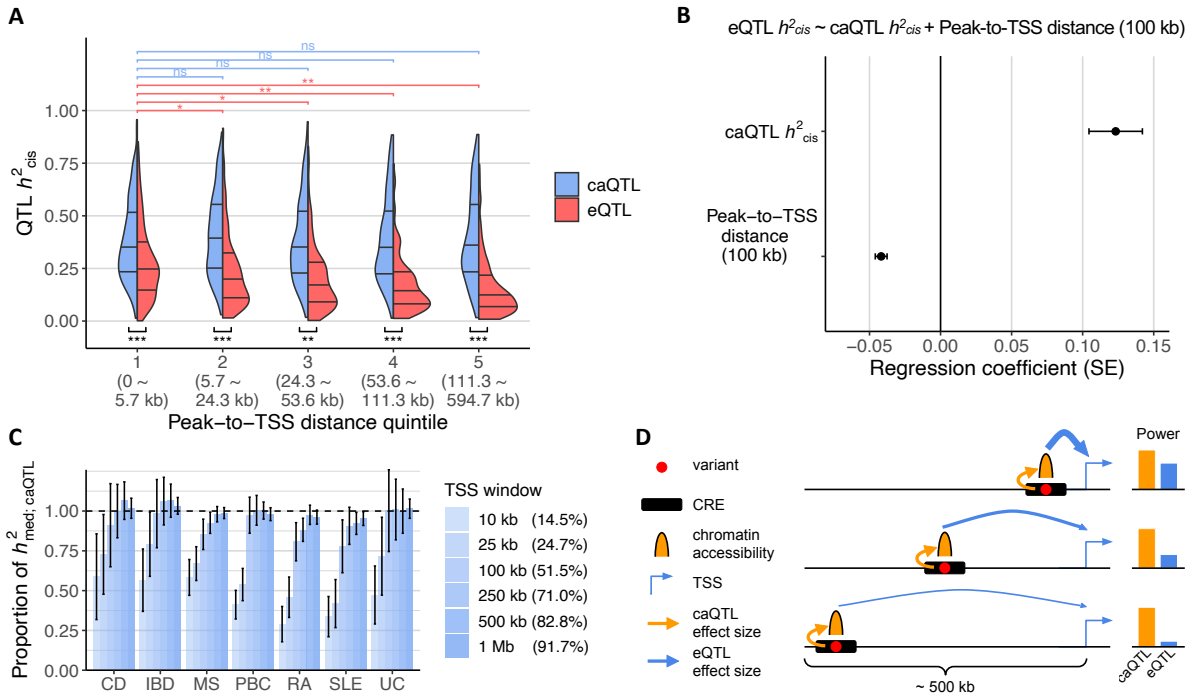


Figure 3. Effect of peak-to-TSS distance on *cis*-heritability of caQTLs and eQTLs and IMD heritability. (**A**) Distribution of *cis*-heritability (h^2_{cis}) of caQTLs and eQTLs by peak-to-TSS distance quintiles. The ranges of peak-to-TSS distance are shown in parentheses. The comparisons shown on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

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

268 For genetic association studies, increasing the sample size is a way to increase statistical power.
269 Therefore, we meta-analyzed four LCL eQTL summary statistics[55,56,68,69], leading to a total sample
270 size of 1,128 individuals. We performed colocalization analysis using the meta-analyzed summary
271 statistics to evaluate whether effectively increasing the sample sizes would uncover more disease-
272 associated eQTLs in LCLs, especially in loci where a caQTL already showed IMD colocalization.
273 Up to 6 additional loci showing eQTL colocalization were thus detected for each IMD (**Figure 4A**
274 and **Supplementary Table 3**). For example, *CIITA* is the class II major histocompatibility
275 complex transactivator, which causes severe immunodeficiency if dysfunctional[70]. The *CIITA*
276 locus is associated with IBD right below the genome-wide significance level (rs10445003, p =
277 7.5×10^{-8}), and it colocalized with a caQTL signal, but initially not with any eQTL (**Supplementary**
278 **Figure 6A**). However, the meta-analyzed statistics showed a stronger association to *CIITA*
279 expression ($p = 6.7\times10^{-8}$) than without meta-analysis ($p = 5.4\times10^{-4}$) and exhibited a significant
280 colocalization. Interestingly, two of the causal CD genes that previously lacked colocalized
281 eQTLs[15], *CARD9* and *ATG16L1*, showed significant colocalization in the meta-analyzed LCL
282 eQTL data (**Supplementary Figure 6B-C**).

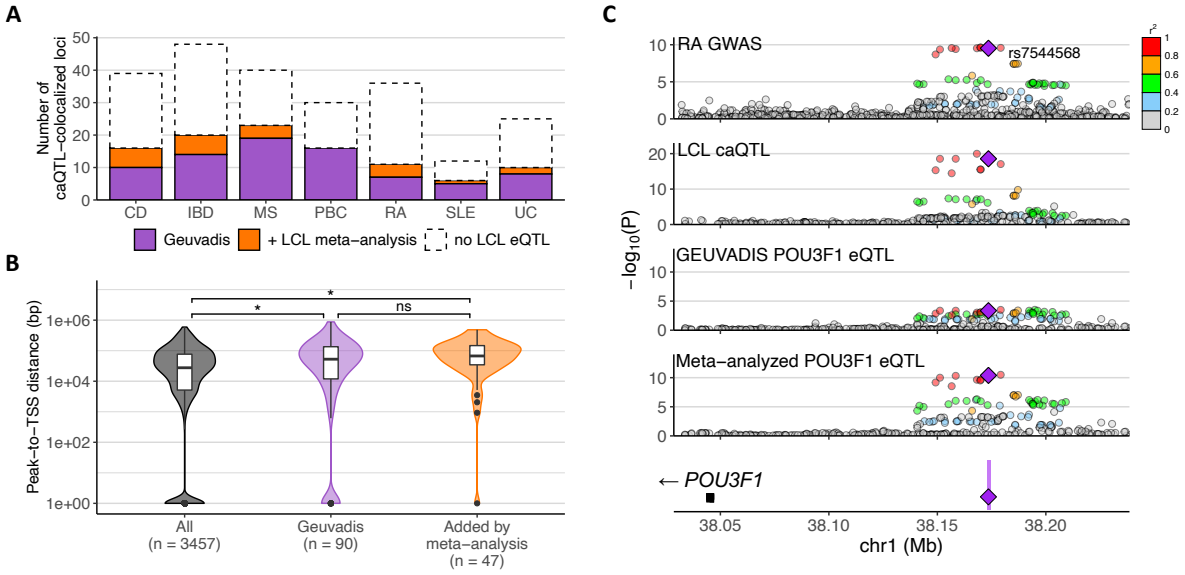


Figure 4. Additional colocalization of caQTL-colocalized IMD loci with meta-analyzed LCL eQTL data. (**A**) Number of caQTL-colocalized IMD loci that showed eQTL colocalization in LCLs. Disease abbreviations along the *x*-axis are as in **Figure 1**. (**B**) Distribution of peak-to-TSS distance of all 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[71], 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.

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

297 Despite uncovering more eQTL colocalizations through meta-analysis, more than 40% of the
298 caQTL-colocalized loci nevertheless showed no colocalization with an eQTL in LCLs (**Figure**
299 **4B**). We investigated whether these "caQTL only" peaks might be inactive regulatory elements.
300 We quantified the active histone mark levels for H3K27ac, H3K4me1, and H3K4me3 at
301 colocalized caQTL peaks in LCLs and then compared their levels between the "caQTL and eQTL"
302 and "caQTL only" loci. On average, H3K27ac marks were stronger at "caQTL and eQTL" peaks,
303 supporting that the corresponding regulatory elements might be more active (**Supplementary**
304 **Figure 7**). H3K4me3 marks were detected more often at "caQTL and eQTL" peaks, leading to
305 stronger average signal. In contrast, H3K4me1 levels were highly similar between the two sets of
306 peaks. Although "caQTL only" peaks generally showed lower levels of active histone marks,
307 several individual "caQTL only" peaks showed comparable levels. These peaks could be inactive
308 *cis*-regulatory elements in LCLs that affect gene expression in a different cellular context.
309 Therefore, we next examined whether those caQTLs appear as eQTLs in other immune cell types.

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

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

316 25–42% of the caQTL-colocalized IMD loci that were missing eQTL colocalizations in
317 LCLs showed eQTL colocalizations in at least one immune cell type (**Figure 5A** and
318 **Supplementary Table 5**). The overlap of LCL caQTLs with non-LCL immune cell eQTLs was
319 greater than expected by chance for 5 of the 7 IMDs ($p < 0.00714$ (Bonferroni-corrected), Fisher's
320 exact test, for CD, IBD, PBC, RA, and UC; **Supplementary Table 6**), suggesting that the caQTLs
321 found in LCLs may also show regulatory function in those immune cell types. Comparing across
322 the datasets, we found that the number of loci with eQTL colocalization varied depending on the
323 cell type, but that the effect of the sample size was more profound ($r^2 = 0.60–0.79$; **Figure 5B** and
324 **Supplementary Table 7**). We meta-analyzed eQTL data of 3 immune cell types with multiple

325 sources – naïve CD4$^+$ T cell, monocyte, and memory regulatory T cell (Treg) – and this also
326 increased the number of loci with significant eQTL colocalization (**Supplementary Table 8**).
327 Altogether, these results suggest that although generating eQTL data in more cell types and cell
328 states uncovers context-specific eQTLs, increasing the sample size should also be a priority to
329 ensure sufficient statistical power.

330 We investigated the potential reasons why IMD loci that colocalized with caQTLs in LCLs
331 showed eQTLs not in LCLs but in other immune cells. First, LCL caQTLs may correspond to gene
332 regulatory elements that exert their effects on gene expression in a different cellular context. For
333 instance, monocyte H3K27ac levels in the "caQTL only" loci where monocyte eQTLs colocalized
334 were higher than those with no monocyte eQTLs (**Supplementary Figure 8**). Second, some
335 examples, such as for *TNFSF15*, were due to cell-type specific gene expression (**Figure 6**): despite
336 a significant colocalization of IBD with a caQTL in LCLs in the *TNFSF15* locus, disease-
337 associated eQTL signal was detected only in monocytes (**Figure 6A**). Tumor necrosis factor-like
338 cytokine 1A (TL1A), the protein encoded by *TNFSF15*, is secreted by monocytes and many other
339 cells to activate helper T cells, regulatory T cells, and B cells[75]. *TNFSF15* expression was low in
340 LCLs (mean transcript per million [TPM] = 0.30), but higher in monocytes (mean TPM = 2.23).
341 Of the genes that showed exclusively monocyte eQTL colocalization, those with low expression
342 (mean TPM < 1) in LCLs generally showed higher expression in monocytes (**Figure 6B-C**).

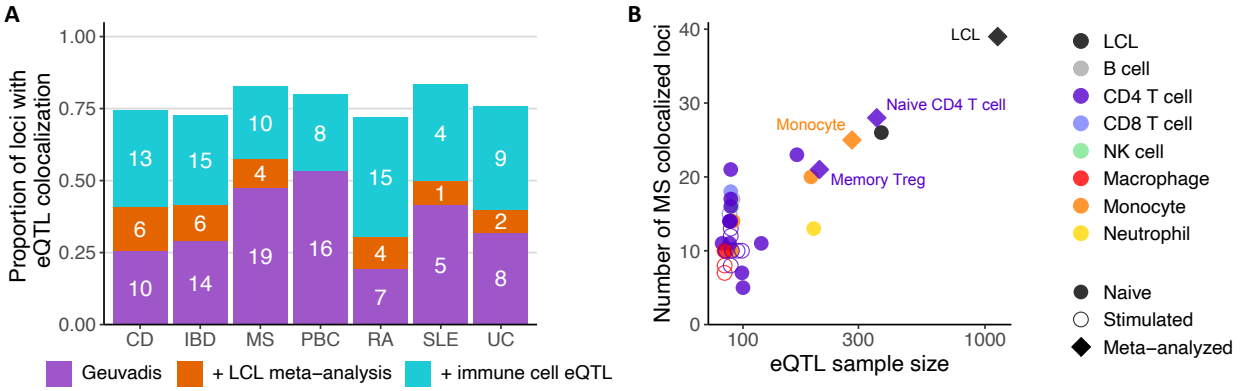
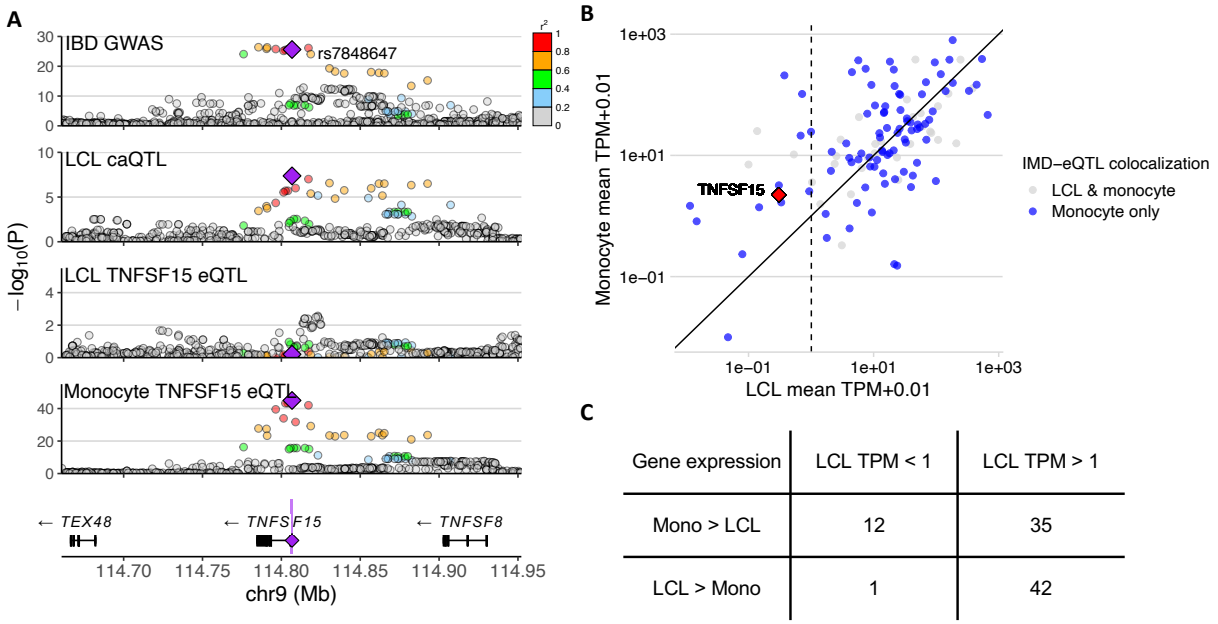


Figure 5. Added utility of various immune cell eQTL data. (**A**) Number of loci that additionally colocalized with eQTLs by LCL meta-analysis (orange) and immune cell data (cyan) compared to the original analysis with Geuvadis LCL eQTL data (purple). The height of the bar is the proportion of loci with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

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

354 the potential reasons for the missing regulation will guide future data generation projects to
355 elucidate disease-associated loci. To determine why some loci might lack colocalized eQTLs, we
356 focused on chromatin accessibility, which is a molecular phenotype affected by regulatory variants
357 more directly than are eQTLs. We approached this question with mediated heritability analysis[16]
358 and colocalization analysis[36,37].

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

366 By focusing on disease-associated caQTL that lacked significant eQTLs, we explored how
367 additional colocalized eQTL effects could be uncovered. First, increasing the sample sizes for the
368 eQTL statistics via meta-analysis demonstrated that more eQTL colocalizations can be detected
369 with increased statistical power and robustness[19]. These results are consistent with the hypothesis
370 that disease-associated eQTLs are typically weaker and distal due to negative selection against
371 large expression changes for causal genes[18]. Second, many caQTLs in LCLs without eQTLs in
372 LCLs showed eQTL colocalization in other immune cell types. Context specificity of eQTLs has
373 been widely considered to be the primary explanation for the difficulty of pinpointing disease-
374 associated eQTLs[17,22,23,73]. Our observation that many IMD-colocalized caQTLs in LCLs show
375 eQTLs in other immune cell types suggests that caQTL effects may be shared across cell types
376 whereas eQTL effects are more context-specific[23]. eQTL data for the specific cellular contexts
377 need to be generated in future studies to uncover more such examples. Our results also showed
378 that having a large sample size in eQTL data is likely a key factor for discoveries of more disease-
379 associated eQTLs.

380 **Limitations of the study**

381 The h^2_{med} / h^2_{SNP} estimates can be biased because of insufficient sample size of the QTL data. Thus,
382 the proportion of mediated IMD heritability could change based on the specific caQTL and eQTL
383 data. Colocalization analysis tests whether the QTL and GWAS data likely share genetic signal,
384 and such shared signal could arise from either causal mediation or pleiotropy. Therefore, further

experiments are needed to establish causality of colocalized 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 ERP110508[30]. The available files were cram alignment files mapped to the b37 reference genome, so we extracted unique read pairs using SAMtools[76] and bamtofastq command from bedtools[77]. 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[78] 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 WASP[80].

To represent peaks across the samples, we subsampled 3 million read pairs from each and pooled them. Then, we used MACS2[81] 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[82] 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[83]. 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.

418 **LCL samples' genotype preparation**

419 To utilize the genotype calls of the highest quality, we downloaded the high-coverage 1000
420 Genomes (1kG) Project data[84]. There was a total of 493 samples that had either ATAC-seq or
421 RNA-seq data in this study. Of those, 479 samples had genotypes derived from the high-coverage
422 whole-genome sequencing data. 14 samples had genotypes derived from microarrays[85]. The
423 samples that needed imputation are listed in **Supplementary Table 9**. We lifted over the
424 microarray data based on the hg19 reference genome to GRCh38 and filtered for variants present
425 in the high-coverage 1kG data. We first imputed microarray genotype data using the TOPMed
426 imputation server[86] and extracted SNPs with imputation $R^2 \geq 0.5$ and imputed the rest of the
427 variants, including short indels, using high-coverage data with Beagle5.2[87,88]. After keeping only
428 variants with DR2 $\geq$ 0.7, we merged the imputed genotypes with the high-coverage 1kG genotypes.

429 **Curation of IMD GWAS data**

430 We downloaded GWAS summary statistics for 13 IMDs, including 11 autoimmune diseases
431 (autoimmune thyroid disease (ATD)[39], celiac disease (CEL)[40], Crohn's disease (CD)[41],
432 inflammatory bowel disease (IBD)[41], juvenile idiopathic arthritis (JIA)[42], multiple sclerosis (MS)[43],
433 primary biliary cholangitis (PBC)[39], rheumatoid arthritis (RA)[44], systemic lupus erythematosus
434 (SLE)[45], ulcerative colitis (UC)[41], and vitiligo (VIT)[46]) and 2 allergic diseases (allergy (ALL)[47]
435 and asthma (AST)[47]). For each disease, we searched for more recent studies with larger sample
436 sizes and prioritized those with genome-wide statistics, rather than those with only Immunochip
437 variants[89]. To compare with non-immune diseases, we also downloaded summary statistics for 3
438 non-immune diseases (type 2 diabetes (T2D)[48], coronary artery disease (CAD)[49] and schizophrenia
439 (SCZ)[50]). In this study, we analyzed only those GWAS summary statistics derived from cohorts
440 of individuals with European ancestries.

441 Since the LCL samples' genotypes are based on the GRCh38 reference genome, we lifted
442 over any GWAS data based on b37 reference genome to the GRCh38 genomic coordinates. Briefly,
443 we formatted the summary statistics as bed files and used the liftOver tool[90] to convert them to
444 GRCh38 genomic coordinates. Then, to ensure that the reference alleles match the sequences of
445 the GRCh38 reference genome, we used gwas2vcf tool[91].

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

447 First, we quantified the chromatin accessibility levels at ATAC-seq peaks identified earlier. We
448 counted the number of read fragments overlapping each peak using featureCounts[92]. For each
449 sample, the read counts were normalized for library size using trimmed mean of M-values[93] so that
450 the values are comparable across the samples. Then, the phenotype values were further normalized
451 to follow a standard normal distribution across the samples, using quantile normalization. Peaks
452 with counts per million (CPM) < 0.8 or counts < 10 for more than 20% of the samples were
453 discarded.

454 Next, we performed a principal component analysis (PCA) on the phenotype matrix to
455 derive potential latent covariates. We selected the number of principal components (PC) to
456 incorporate in the regression model based on the Buja and Eyuboglu algorithm[94] that is
457 implemented in PCAForQTL[95]. Ultimately, we accounted for sex, library size, 3 genotype PCs
458 and 13 phenotype PCs in the QTL analysis. We performed genetic association tests on variants
459 within 200 kilobases (kb) of the peak using tensorQTL[96]. We discarded variants with minor allele
460 frequency less than 5%.

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

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

468 **Mediated heritability estimation of QTLs**

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

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

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

477 To evaluate whether certain peak sets are enriched for mediated heritability, we utilized
478 the gene set analysis functionality. For peak strength, we considered the 95% percentile CPM value
479 of each peak and stratified the peaks into quintiles. For histone marks, we first generated histone
480 ChIP-seq peaks using data from Delaneau *et al*.[52]. Similar to calling ATAC-seq peaks, we sampled
481 3 million reads per sample and merged them before applying MACS2[81]. We downloaded 3 control
482 ChIP-seq data from ENCODE to use as input (File IDs: ENCFF066RCS, ENCFF159XTB, and
483 ENCFF850RIE)[79]. Then, we curated sets of ATAC-seq peaks that overlapped H3K27ac,
484 H3K4me1, and H3K4me3 ChIP-seq peaks by at least 1 bp. ATAC-seq peaks that overlapped
485 H3K27ac but not H3K4me3 regions were labeled as 'enhancer-like signature', while those that
486 overlapped H3K27ac and H3K4me3 regions were labeled as 'promoter-like signature'. The
487 estimand of interest is the proportion of mediated heritability explained by the peak set (peak set
488 h^2_{med} / total h^2_{med}).

489 *Comparison with eQTL-mediated heritability.* First, we estimated h^2_{med} / h^2_{SNP} of eQTLs (*i.e.*,
490 $h^2_{med;\ eQTL}$ / h^2_{SNP}) the same way as we did for that of caQTLs. Then, we also estimated h^2_{med} / h^2_{SNP}
491 of caQTLs and eQTLs together (*i.e.*, $h^2_{med;\ caQTL \cup eQTL}$) with MESC meta-analysis[16]. caQTLs and
492 eQTLs were also stratified as separate sets to account for potential differences in the relationship
493 of QTL *cis*-heritability and h^2_{med}. This meta-analyzed estimate is effectively the amount of
494 heritability mediated by either caQTLs or eQTLs in LCLs. This estimate reveals the overall
495 relationship between heritability mediated by caQTLs and eQTLs. For instance, the estimate of
496 the heritability mediated exclusively by caQTLs would be $h^2_{med;\ just\ caQTL} = h^2_{med;\ caQTL \cup eQTL} - h^2_{med;}$
497 $_{caQTL}$. The estimate of mediated heritability shared by caQTLs and eQTLs is $h^2_{med;\ caQTL \cap eQTL} =$
498 $(h^2_{med;\ caQTL} + h^2_{med;\ eQTL}) - h^2_{med;\ caQTL \cup eQTL}$.

499 **Colocalization analyses**

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

504 lead χ^2 statistics for the caQTL or eQTL and IMD association χ^2 statistics greater than $0.8 \times \chi^2$
505 statistics for the IMD lead variant in the locus. Then, we applied gwas-pw[37] on the variants within
506 100 kb of the lead variant. We considered loci with posterior probability of colocalization (PP3) >
507 0.98 to be colocalized[57].

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

515 **Enrichment of biological processes**

516 To test whether colocalized genes are likely relevant to autoimmune diseases, we surveyed which
517 Biological Process Gene Ontology (GO) terms were overrepresented compared to all the genes within
518 500 kb of each IMD association signal. Enrichment of biological processes was evaluated using
519 Protein Analysis Through Evolutionary Relationships (PANTHER)[97]. The foreground list
520 comprised all of the genes whose eQTL signal colocalized with one of the 7 autoimmune diseases
521 (CD, IBD, MS, PBC, RA, SLE, and UC). The background list of genes was all of the genes within
522 500 kb of each IMD lead variant for which we observed colocalization. Moreover, we tested
523 whether colocalized caQTLs without eQTLs are closer to genes related to immune processes than
524 expected by chance. For this, we used Genomic Regions Enrichment of Annotations Tool
525 (GREAT)[63]. The foreground list comprised ATAC-seq peaks at IMD loci showing colocalization
526 only with caQTLs and not with eQTLs. The background list is all ATAC-seq peaks identified in
527 LCLs that we tested for caQTL association.

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

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

533 MESC analysis uses REML implemented in GCTA[98] to estimate QTL *cis*-heritability. We
534 referred to its output and compared the *cis*-heritability of caQTLs and eQTLs based on peak-to-
535 TSS distance. To visualize the distribution of *cis*-heritability estimates, we grouped pairs of
536 colocalized caQTLs and eQTLs based on peak-to-TSS distance quintiles.

537 **LCL eQTL meta-analysis**

538 We downloaded LCL eQTL data from 3 studies through eQTL Catalogue release 6[99]. The sample
539 sizes were 190, 147, and 418 for GENCORD[68], GTEx[55], and TwinsUK[69], respectively. We meta-
540 analyzed the summary statistics using the inverse variance weighted fixed effects model. If a
541 variant was missing in a subset of the studies, then only the available statistics were meta-analyzed.
542 We used these meta-analyzed statistics to perform colocalization the same way as earlier
543 colocalization analyses.

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

545 We downloaded eQTL data for various immune cell types from the eQTL Catalogue release 6[99].
546 The source studies from the eQTL Catalogue include BLUEPRINT[72], DICE[22], Alasoo *et al.*[23] and
547 Bossini-Castillo *et al*.[74]. The represented immune cell types include T cell subtypes, B cells,
548 neutrophils, and monocytes. We also separately downloaded data for $CD4^+$ T cells with and
549 without stimulation[73]. The selection of candidate loci and colocalization analysis on them followed
550 the same procedure as that for other QTLs.

551 We also evaluated whether meta-analysis of eQTL data can increase the number of loci
552 with eQTL colocalization. We meta-analyzed eQTL data for 3 immune cell types with multiple
553 sources – naïve $CD4^+$ T cell, monocyte, and memory regulatory T cell (Treg) – using the inverse
554 variance weighted fixed effects model. Specifically, we meta-analyzed naïve $CD4^+$ T cell eQTL
555 summary statistics from Soskic *et al.*[73], BLUEPRINT[72], and DICE[22]. We meta-analyzed monocyte
556 eQTL summary statistics from BLUEPRINT[72] and DICE[22]. We meta-analyzed memory Treg
557 eQTL summary statistics from Bossini-Castillo *et al*.[74] and DICE[22]. We used these meta-analyzed
558 statistics to perform colocalization the same way as earlier colocalization analyses.

559

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 Atlas[83] (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 catalog[100] (https://www.ebi.ac.uk/gwas/). eQTL summary statistics are downloaded either from the eQTL Catalogue[99] (https://www.ebi.ac.uk/eqtl/) or the source publication.

Acknowledgments

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.

578 **References**

579 1. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J.*
580 *Hum. Genet.* **101**, 5–22 (2017).
581 2. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
582 3. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in
583 regulatory DNA. *Science* **337**, 1190–1195 (2012).
584 4. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11
585 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
586 5. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance
587 discovery from GWAS. *PLoS Genet.* **6**, (2010).
588 6. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic
589 architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
590 7. O'Connor, L. J. *et al.* Extreme Polygenicity of Complex Traits Is Explained by Negative
591 Selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).
592 8. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic
593 to Omnigenic. *Cell* **169**, 1177–1186 (2017).
594 9. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–
595 243 (2021).
596 10. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes
597 underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).
598 11. Morris, J. A. *et al.* Discovery of target genes and pathways at GWAS loci by pooled single-cell
599 CRISPR screens. *Science* **380**, eadh7699 (2023).
600 12. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human
601 genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
602 13. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-
603 disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
604 14. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci.
605 *Genome Biol.* **22**, 49 (2021).
606 15. Connally, N. J. *et al.* The missing link between genetic association and regulatory function. *Elife*
607 **11**, (2022).
608 16. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease
609 mediated by assayed gene expression levels. *Nat. Genet.* 1–8 (2020) doi:10.1038/s41588-020-
610 0625-2.

611 17. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.*
612 **37**, 109–124 (2021).
613 18. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of
614 genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
615 19. Hukku, A. *et al.* Probabilistic Colocalization of Genetic Variants from Complex and Molecular
616 Traits: Promise and Limitations. *Am. J. Hum. Genet.* **108**, 25–35 (2021).
617 20. Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K. Evidence for weak selective
618 constraint on human gene expression. *Genetics* **211**, 757–772 (2019).
619 21. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent
620 manner. *Science* **325**, 1246–1250 (2009).
621 22. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene
622 Expression. *Cell* **175**, 1701–1715 (2018).
623 23. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for
624 enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
625 24. Nathan, A. *et al.* Single-cell eQTL models reveal dynamic T cell state dependence of disease loci.
626 *Nature* **606**, 120–128 (2022).
627 25. Field, A. & Adelman, K. Evaluating Enhancer Function and Transcription. *Annu. Rev. Biochem.*
628 **89**, 213–234 (2020).
629 26. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–
630 235 (2010).
631 27. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin
632 structure, and transcription. *Science* **342**, 744–747 (2013).
633 28. Waszak, S. M. *et al.* Population Variation and Genetic Control of Modular Chromatin Architecture
634 in Humans. *Cell* **162**, 1039–1050 (2015).
635 29. Degner, J. F. *et al.* DNase-I sensitivity QTLs are a major determinant of human expression
636 variation. *Nature* **482**, 390–394 (2012).
637 30. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal
638 interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
639 31. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat.*
640 *Rev. Genet.* **16**, 197–212 (2015).
641 32. Wu, Y. *et al.* Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large
642 fraction of GWAS signals shared with molecular phenotypes. *Cell Genomics* 100344 (2023)
643 doi:10.1016/j.xgen.2023.100344.
644 33. Jeong, R. & Bulyk, M. L. Blood cell traits' GWAS loci colocalization with variation in PU.1

645 genomic occupancy prioritizes causal noncoding regulatory variants. *Cell genomics* **3**, 100327
646 (2023).
647 34. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes.
648 *Nature* **518**, 317–329 (2015).
649 35. Farh, K. K. H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants.
650 *Nature* **518**, 337–343 (2015).
651 36. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association
652 studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
653 37. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits.
654 *Nat. Genet.* **48**, 709–717 (2016).
655 38. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide
656 association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
657 39. Cordell, H. J. *et al.* An international genome-wide meta-analysis of primary biliary cholangitis:
658 Novel risk loci and candidate drugs. *J. Hepatol.* **75**, 572–581 (2021).
659 40. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene
660 expression. *Nat. Genet.* **42**, 295–302 (2010).
661 41. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple
662 integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
663 42. López-Isac, E. *et al.* Combined genetic analysis of juvenile idiopathic arthritis clinical subtypes
664 identifies novel risk loci, target genes and key regulatory mechanisms. *Ann. Rheum. Dis.* **80**, 321–
665 328 (2021).
666 43. Patsopoulos, N. A. *et al.* Multiple sclerosis genomic map implicates peripheral immune cells and
667 microglia in susceptibility. *Science* **365**, (2019).
668 44. Ishigaki, K. *et al.* Multi-ancestry genome-wide association analyses identify novel genetic
669 mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).
670 45. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive
671 immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464
672 (2015).
673 46. Jin, Y. *et al.* Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci
674 and highlight key pathways and regulatory variants. *Nat. Genet.* **48**, 1418–1424 (2016).
675 47. Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for
676 biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
677 48. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture
678 and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).

679 49. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for
680 coronary artery disease. *Nat. Genet.* **43**, 333–8 (2011).
681 50. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from
682 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–7 (2014).
683 51. van Mierlo, H. C., Broen, J. C. A., Kahn, R. S. & de Witte, L. D. B-cells and schizophrenia: A
684 promising link or a finding lost in translation? *Brain. Behav. Immun.* **81**, 52–62 (2019).
685 52. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene
686 expression. *Science* **364**, (2019).
687 53. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters
688 and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
689 54. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and
690 mouse genomes. *Nature* **583**, 699–710 (2020).
691 55. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues.
692 *Science* **369**, 1318–1330 (2020).
693 56. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in
694 humans. *Nature* **501**, 506–511 (2013).
695 57. Kundu, K. *et al.* Genetic associations at regulatory phenotypes improve fine-mapping of causal
696 variants for twelve immune-mediated diseases. *Nat. Genet.* **44**, 0–23 (2022).
697 58. Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein*
698 *Sci.* **31**, 8–22 (2022).
699 59. Sanmartí, R., Ruiz-Esquide, V., Bastida, C. & Soy, D. Tocilizumab in the treatment of adult
700 rheumatoid arthritis. *Immunotherapy* **10**, 447–464 (2018).
701 60. Khanna, R. & Feagan, B. G. Ustekinumab for the treatment of Crohn's disease. *Immunotherapy* **5**,
702 803–15 (2013).
703 61. Monemi, S. *et al.* Incidence of Gastrointestinal Perforations in Patients with Rheumatoid Arthritis
704 Treated with Tocilizumab from Clinical Trial, Postmarketing, and Real-World Data Sources.
705 *Rheumatol. Ther.* **3**, 337–352 (2016).
706 62. Stevenson, C. *et al.* Essential role of Elmo1 in Dock2-dependent lymphocyte migration. *J.*
707 *Immunol.* **192**, 6062–70 (2014).
708 63. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat.*
709 *Biotechnol.* **28**, 495–501 (2010).
710 64. Tanigawa, Y., Dyer, E. S. & Bejerano, G. WhichTF is functionally important in your open
711 chromatin data? *PLoS Comput. Biol.* **18**, e1010378 (2022).
712 65. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of

713 CRISPR perturbations. *Nat. Genet.* **51**, (2019).
714 66. Zuin, J. *et al.* Nonlinear control of transcription through enhancer-promoter interactions. *Nature*
715 **604**, 571–577 (2022).
716 67. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and
717 polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, (2021).
718 68. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic
719 variation in gene regulation. *Elife* **2013**, 1–18 (2013).
720 69. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence
721 analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
722 70. Dziembowska, M. *et al.* Three novel mutations of the CIITA gene in MHC class II-deficient
723 patients with a severe immunodeficiency. *Immunogenetics* **53**, 821–9 (2002).
724 71. Hofmann, E. *et al.* Octamer-binding factor 6 (Oct-6/Pou3f1) is induced by interferon and
725 contributes to dsRNA-mediated transcriptional responses. *BMC Cell Biol.* **11**, 61 (2010).
726 72. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune
727 Cells. *Cell* **167**, 1398-1414.e24 (2016).
728 73. Soskic, B. *et al.* Immune disease risk variants regulate gene expression dynamics during CD4+ T
729 cell activation. *Nat. Genet.* **54**, 817–826 (2022).
730 74. Bossini-Castillo, L. *et al.* Immune disease variants modulate gene expression in regulatory CD4+
731 T cells. *Cell Genomics* **2**, 100117 (2022).
732 75. Xu, W.-D., Li, R. & Huang, A.-F. Role of TL1A in Inflammatory Autoimmune Diseases: A
733 Comprehensive Review. *Front. Immunol.* **13**, 891328 (2022).
734 76. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
735 77. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
736 features. *Bioinformatics* **26**, 841–2 (2010).
737 78. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–
738 359 (2012).
739 79. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**,
740 57–74 (2012).
741 80. Van De Geijn, B., Mcvicker, G., Gilad, Y. & Pritchard, J. K. WASP: Allele-specific software for
742 robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
743 81. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
744 82. Banovich, N. E. *et al.* Impact of regulatory variation across human iPSCs and differentiated cells.
745 *Genome Res.* **28**, 122–131 (2018).
746 83. Papatheodorou, I. *et al.* Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.*

48, D77–D83 (2020).

84. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).

85. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

86. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

87. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).

88. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).

89. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).

90. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

91. Lyon, M. S. *et al.* The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biol.* **22**, 32 (2021).

92. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–30 (2014).

93. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

94. Buja, A. & Eyuboglu, N. Remarks on Parallel Analysis. *Multivariate Behav. Res.* **27**, 509–40 (1992).

95. Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol.* **23**, 210 (2022).

96. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).

97. 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).

98. 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).

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

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