Shared effect modeling reveals that a fraction of autoimmune disease associations are consistent with eQTLs in three immune cell types

Sung Chun[1,2,3], Alexandra Casparino[4], Nikolaos A Patsopoulos[3,5,6], Damien Croteau-Chonka[2,7], Benjamin A Raby[2,7], Philip L De Jager[2,3,5,6], Shamil R Sunyaev[1,2,3]*, and Chris Cotsapas[3,4,8]*

[1]Division of Genetics, Brigham and Women's Hospital, Boston MA, USA
[2]Department of Medicine, Harvard School of Medicine, Boston MA, USA
[3]Broad Institute of Harvard and MIT, Cambridge MA, USA
[4]Department of Neurology, Yale School of Medicine, New Haven CT, USA
[5]Department of Neurology, Brigham and Women's Hospital, Boston MA, USA
[6]Ann Romney Center for Neurological Diseases, Brigham and Women's Hospital, Boston MA, USA
[7]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston MA, USA
[8]Department of Genetics, Yale School of Medicine, New Haven CT, USA
* correspondence to SRS (ssunyaev@rics.bwh.harvard.edu) and CC (cotsapas@broadinstitute.org)

The majority of autoimmune disease risk effects identified by genome-wide association studies (GWAS) localize to open chromatin with gene regulatory activity. GWAS loci are also enriched for expression quantitative trait loci (eQTLs), suggesting that most disease risk variants exert their pathological effects by altering gene expression[1,2]. However, because causal variants are difficult to identify and *cis*-eQTLs occur frequently, it remains challenging to translate this bulk observation into specific instances of a disease risk variant driving changes to gene regulation. Here, we use a novel joint likelihood framework with higher resolution than previous methods to identify loci where disease risk and an eQTL are driven by a single, shared genetic effect as opposed to distinct effects in close proximity. We find that approximately 25% of autoimmune disease loci harbor an eQTL driven by the same genetic effect, but the majority of loci do not. Thus, we uncover a fraction of gene regulatory changes as strong mechanistic hypotheses for disease risk, but conclude that most risk mechanisms do not involve changes to basal gene expression.

The autoimmune and inflammatory diseases (AID) – multiple sclerosis, type 1 diabetes, inflammatory bowel disease, and more than 80 others – are highly heritable, complex diseases that cumulatively afflict 8% of the population[3,4]. Pathology is driven by loss of tolerance to self-antigens, resulting in either systemic or tissue-specific immune attack. They are extensively co-morbid[5,6] with extensive sharing of AID genetic risk[7-9], indicating that pathogenic mechanisms are also shared[10]. Disease mechanisms are still poorly understood, so current therapies control symptoms rather than root causes, often by suppression of major immune responses. Consortium-driven genetic mapping studies have identified hundreds of genomic regions mediating risk to several AID, and these communities have collaborated to develop the ImmunoChip custom genotyping array to deeply interrogate 185 of these loci[11].

These disease associations are primarily non-coding: lead GWAS SNPs are more likely to be associated with expression levels of neighboring genes than expected by chance[12], and the same lead SNPs are enriched in regulatory regions marked by chromatin accessibility and modification[1,13]. Fine-mapping reveals enrichment of AID-associated variants in enhancer elements specifically active in stimulated T cell subpopulations[14], and formal disease heritability partitioning analyses also show strong enrichment in such regions of gene regulatory potential[15,16]. Collectively, these strands of evidence suggest that the majority of disease risk is mediated by changes to gene regulation in specific cell subpopulations. However, these bulk overlap and enrichment analyses do not formally assess whether expression levels and disease risk are associated to the same underlying variant or are due to

independent effects in the same region[17,18]. Several methods have been developed to identify pathogenic genes within GWAS loci relying on eQTL co-localization[19-22]. However, these are limited in resolution to detect situations where causal variants for the disease trait and the eQTL are distinct but in linkage disequilibrium.

Here, we present an approach to test if a GWAS risk association and an eQTL are driven by the same underlying genetic effect, accounting for the LD between causal variants. Using data from ImmunoChip studies of seven AID comprising >180,000 samples in total (Table S2), we apply this approach to test if associations in 272 known risk loci are consistent with *cis*-eQTL for genes in each region, measured in three relevant immune cell populations: lymphoblastoid cell lines (LCLs), $CD4^+$ T cells and $CD14^+$ monocytes[23,24].

When associations to two traits in a locus – here, a disease trait and an eQTL – are driven by the same underlying causal variant, the joint evidence of association should be maximized at the markers in tightest LD with the (potentially unobserved) causal variant[18,25]. Here, we directly evaluate the joint likelihood that both trait associations are due to the same underlying causal variant (Figure S1), unlike previous approaches that look for similarities in the shape of the association curve over multiple markers[19,20,26,27]. We expect that when the underlying causal effect is shared, joint likelihood is maximized when we model the same causal variant in both traits; conversely, when the underlying causal variants are different, we expect maximum joint likelihood when we model their closest proxies. We empirically derive the null distribution of the joint likelihood ratio statistic by comparing disease associations to permuted eQTL data. The resulting *P* value is asymptotically conservative against the null of distinct causal variants, as the likelihoods of two competing models will be further contrasted with increasing sample and effect sizes (Supplementary Notes). We are thus able to directly evaluate whether associations in the same genomic locus for two traits, observed in different cohorts of individuals, are due to the same underlying causal variant.

To assess the performance of our method, we benchmarked it against *coloc*[19], a well-calibrated Bayesian framework that considers spatial similarities in association data across windows of markers in a locus. We simulated pairs of case-control cohorts with either the same or distinct causal variants driving association in each. Though both methods had excellent performance in simulations where the two distinct causal variants are in complete linkage equilibrium (AUC=0.99 for both when $r^2 < 0.5$), compared to *coloc* we found that our method maintained higher specificity even as the linkage disequilibrium between distinct causal variants became high (AUC = 0.92 compared to 0.70 for *coloc* when $0.7 < r^2 < 0.8$; Figure S3, Tables S3 and S4). In practice, our resolution becomes limited at very high LD levels ($r^2>0.8$), and we are unable to reliably distinguish between two causal variants in very high LD and a single causal variant associated to both traits. Thus, within these limits, we can accurately detect cases of shared genetic effects between two traits.

We first identified densely genotyped ImmunoChip loci showing strong association to each disease from publicly available summary data (immunobase.org; Table 1). These include associations to both Crohn disease and ulcerative colitis, collectively designated inflammatory bowel disease, and to each disease alone[28]. Due to the extensive LD and complex natural selection present in the Major Histocompatibility Locus, we excluded this region from consideration. We next identified genes in a 1Mb window centered on the most associated variant in each locus. Consistent with previous observations that eQTLs are frequently found in GWAS loci, we found that all loci but 11 had at least one gene with an eQTL ($p < 0.05$) in at least one of the three cell types, with most such effects common across all three tissues (Table 1). In total, we found 9,268 pairs of disease and eQTL associations across 261 ImmunoChip loci. We then tested each of these pairs with our joint association likelihood method to assess if the eQTLs appear driven by the same underlying effect as the disease associations. We find evidence for shared effects for only 57/9,268 pairs in 41/261 loci across all

diseases, with the proportion varying from 2/40 (5%) for type 1 diabetes loci to 4/10 (40%) for ulcerative colitis loci (false discovery rate < 5%; Tables 1 and 2). Of these 57 shared effects, 43 pass even the more stringent family-wise multiple testing correction (Bonferroni corrected P < 0.05). Thus, our analysis reveals that in the majority of AID loci, variants causally involved in disease phenotypes do not overlap variants responsible for eQTL signals. Overall, we find that only a small minority of tested disease-eQTL pairs in the same locus show any evidence of a shared association, whereas >75% show evidence of being driven by distinct genetic variants in the same locus (Figure 1).

We sought to explain this lack of overlap between disease associations and eQTLs, despite their frequent co-occurrence in the same loci. In particular, although our method showed good performance in simulated data (Figure S4), we remained concerned that this lack of overlap may be due to low statistical power in the eQTL data, which come from cohorts of limited sample size. However, we find that even amongst the most strongly supported eQTLs ($p < 10^{-5}$), <25% show evidence of shared effects with disease associations. Conversely, we find strong evidence for distinct effects for the majority of disease-eQTL pairs, with only a subset of comparisons being ambiguous, suggesting that our method is adequately powered to detect shared effects where they exist (Figures 1a and S11-13). To assess whether power affects the total number of loci, rather than eQTL, that can be resolved, we looked more deeply at our significance threshold settings. We find that more liberal thresholds do not increase the number of true positive results after adjusting for false positive rate, indicating that most loci do not contain *any* gene with an eQTL consistent with the disease association (Figures 1b and S14). Cumulatively, our results demonstrate that only a minority of AID risk effects drive eQTLs in the three cell populations we tested, which are drawn from diverse lineages of the immune system.

We next focused on the subset of 57 disease/eQTL pairs in 41 loci where we could detect strong evidence of a shared effect (Table 2). We find that 51/57 (89%) of effects are restricted to one cell population, indicating that tissue-specific eQTLs are important components of the molecular underpinnings of disease (Figures S5 and S6). The remaining six effects are detected in multiple cell populations; for example, the multiple sclerosis association at rs10783847 on chromosome 12 is consistent with eQTLs for the transcript of methyltransferase-like 21B (*METTL21B*) in both $CD4^+$ T cells and $CD14^+$ monocytes, but not with eQTLs for the remaining 38 genes in the immediate locus (Figure 2). Although *METTL21B* is expressed in LCLs, there is no evidence of an eQTL in this tissue within 1Mb from rs10783847. Similarly, for the multiple sclerosis association at rs1966115 on chromosome 8 and eQTLs for *ZC2HC1A*, and for the inflammatory bowel disease association at rs55770741 on chromosome 5 and eQTLs for *ERAP2*, we detect a shared effect in all three cell populations. In several cases we find tissue-specific shared effects despite strong eQTLs for the same gene in other tissues: for *TUFM* and inflammatory bowel disease risk at rs12448902 on chromosome 16, we find shared effects in $CD4^+$ and $CD14^+$ but not LCLs, where we see a *TUFM* eQTL at p = 0.01 (joint likelihood P = 0.97). For *ZFP90* and ulcerative colitis risk at rs889561 on chromosome 16, we also find shared effects in $CD4^+$ and $CD14^+$ but not LCLs, where we observe a *ZFP90* eQTL at p = 0.005 that has a low likelihood of shared effect with GWAS (joint likelihood P = 0.95). Instead, we find evidence of sharing between disease risk and an eQTL for *NFAT5* in LCLs. Thus, despite the presence of eQTLs for a gene in multiple tissues, not all these effects are consistent with disease associations suggesting that disease-relevant eQTLs are tissue specific.

Among our findings are cases where an eQTL is consistent with associations to multiple diseases. For example, the ankyrin repeat domain 55 (*ANKRD55*) transcript encoded on chromosome 5 has an eQTL in $CD4^+$ T cells that is shared with proximal associations to multiple sclerosis, Crohn disease and rheumatoid arthritis (Figure 3, all observations are significant after Bonferroni correction). We also find weaker evidence for shared effects between all three diseases and an eQTL for interleukin 6 signal transducer (*IL6ST*) in $CD4^+$ T cells, which passes the false discovery rate threshold but not the more stringent Bonferroni correction (Figure S7). Similarly, a $CD4^+$ eQTL for *ELMO1* on chromosome 7 is

consistent with associations to both celiac disease and multiple sclerosis (Figure S8), a $CD14^+$ eQTL for *RGS1* on chromosome 1 is consistent with associations to both celiac disease and multiple sclerosis (Figure S9), and a $CD14^+$ eQTL for *UBE2L3* on chromosome 22 is consistent with associations to both celiac disease and inflammatory bowel disease (Figure S10). In all cases, these are the only genome-wide significant disease associations reported in these loci. As we consider each disease association independently, these results indicate that the same underlying risk variants drive risk to multiple diseases in these loci by altering gene expression, consistent with observations of shared effects across diseases[7].

Overall, our results suggest that some autoimmune and inflammatory disease loci are consistent with eQTLs acting in specific immune cell subpopulations, which form strong mechanistic hypotheses for the molecular mechanisms driving disease risk. However, these only account for a small fraction of eQTLs present in disease risk loci; this suggests that abundant caution must be exercised before inferring pathological relevance for an observed eQTL simply due to proximity to a disease association. Strong evidence of a shared genetic effect should therefore be established prior to embarking on time-consuming and costly experimental dissection of such effects.

Previous efforts to detect shared effects between traits in specific loci rely on conditional analyses[29] or indirectly leverage linkage disequilibrium to test if the shape of association peaks in the region are similar[19,26,27,30]. In contrast, we directly evaluate whether the data support a shared effect through joint likelihood estimation. Through this direct evaluation, we are able to resolve cases where the associations are proximal with higher resolution (Figure S3, Tables S3 and S4). As our method is general, we suggest it may be useful in other contexts, for example in establishing if the shared heritability between diseases is driven by the same underlying causal effects[31].

More broadly, our results raise the question of how causal disease variants alter cell function to induce risk, given the strong enrichment of disease risk signal on regions of chromatin accessibility with gene regulatory potential[1], and gene enhancers in particular[14]. We suggest that although gene regulatory regions harboring risk variants are accessible in multiple immune cell subpopulations, they may control gene expression in either a tissue-specific or condition-specific manner, which is not manifest in all cell populations. Our results therefore reinforce the view that we must seek the appropriate cell type and physiological conditions in order to capture the pathologically relevant gene regulatory changes driving disease risk.

Methods

Simulated dataset. We selected two loci with previously known associations to base our positive and negative simulations: CD58 on chromosome 1, rs667309 associated to MS risk[32], and ATG16L1 on chromosome 2, rs2241880 associated to IBD[33]. In each, we used Hapgen[34] and phased base haplotypes of the locus ($2n$=112, downloaded from http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) to simulate up to ten cohorts, each with 1,000 cases and 1,000 controls, and up to 90 cohorts with other causal variants (Table S1). We chose these other variants to have a range of LD to rs667309 or rs2241880, so we could assess the LD resolution of our joint likelihood method. In each cohort, we simulated genotypes for all variants within 2Mb of rs667309 or rs2241880. To make the cohorts comparable to each other, we adjusted the effect sizes of causal SNPs at different MAF to maintain full power in each simulated cohort.

The pairs of simulated cohorts were used either as positive controls if they simulate the same causal variant and negative controls otherwise, to assess the LD resolution limits of our method. Cohort pairs simulating distinct causal variants in high LD (r^2 > 0.8) were excluded. For each pair, one cohort was

treated as the primary trait (i.e. a disease GWAS) and the other as a secondary trait (i.e. eQTL). Due to the nature of the coalescent forward simulation model, 7.7% of simulated cohorts showed peak association to a SNP in only moderate LD with the specified causal variant (r^2 < 0.8). We kept these cohorts as secondary traits only, to better capture the vagaries of resolution limits inherent in the small sample size of eQTL studies. In total, we generated 1,629 cohort pairs for positive controls simulating the same causal variant and 6,106, 530, and 160 cohort pairs for negative controls for which distinct causal variants are separated by r^2 of 0 – 0.5, 0.5 – 0.7, and 0.7 – 0.8, respectively.

Disease GWAS datasets. We downloaded association summary statistics for type 1 diabetes (T1D), rheumatoid arthritis (RA), celiac disease (CEL), multiple sclerosis (MS), inflammatory bowel disease (IBD), Crohn's disease (Crohn), and ulcerative colitis (UC) from ImmunoBase (immunobase.org; Table S2). For MS, we used the association statistics derived from the combined cohort of discovery and validation samples[8] in order to maximize the sample size and genetic resolution. For IBD, Crohn, and UC, summary data are from European subset of a trans-ethnic association stud[28]. All association data are solely based on ImmunoChip samples and do not include imputed genotypes. To address population structure, we limited our analyses to European subjects only with the exception of RA[35], which includes 620 Punjab individuals out of a total of 27,345. T1D summary statistics are from the meta-analysis between case/control association and affected sib-pair analysis.

As our method works best on dense genotype data, we restricted our analyses to the 188 loci genotyped at high density on ImmunoChip. We excluded the Major Histocompatibility Complex (MHC) locus, due to the complex landscape of selection and resulting complex LD patterns. For each disease, we sought the largest published genetic mapping study and identified genome-wide significant associations reported in the 188 ImmunoChip loci. We note that these reports may contain additional samples, so the associations may not be genome-wide significant in the ImmunoChip studies alone. We also excluded any secondary associations after conditioning on initial results, as these are inconsistently reported across diseases. If multiple independent associations are reported within the same ImmunoChip region for any disease, we divide the region at the mid-point between the reported markers and select lead SNPs in each sub-interval separately.

eQTL dataset. We examined eQTLs in lymphoblastoid cell lines[23] (LCLs) and primary $CD4^+$ T cells and $CD14^+$ monocytes[24] obtained from healthy donors (Table S2). For LCLs, we obtained imputed genotypes and normalized RNAseq in RPKM for 278 non-Finnish European donors in the Geuvadis project. We removed SNPs with minor allele frequency (< 5%), high probability of Hardy-Weinberg disequilibrium ($P_{HW} < 10^{-5}$), or high genotype missing rate (>5%). We removed pseudogenes and transcripts without assigned gene symbols from the expression data, and calculated association statistics by linear regression of genotype on expression levels, including three population principal components to control for structure[36,37]. For $CD4^+$ and $CD14^+$, we regressed normalized expression levels for European Americans (n=213 and 211, respectively) on similarly QCed imputed allele dosages. For all cell types, we generated adaptive permutation statistics from 10^3 up to 10^6 iterations[36], using all covariates.

Joint likelihood mapping (JLIM). To test the hypothesis that association signals for two traits are driven by the same causal variant, we contrasted the joint likelihood of observed association statistics under the assumption of same compared to distinct causal variant. Due to limited genetic resolution, distinct causal variants were defined by separation in LD space by $r^2 < \theta$ from each other. The limit of genetic resolution θ is a user-specified parameter and was set to 0.8 in this study. We assumed that at most one causal variant was present in the locus for each trait and that no samples overlap between the traits. We designed the joint likelihood mapping (JLIM) statistic Λ in an asymmetrical fashion, requiring only summary-level statistics for one trait (primary trait) but genotype-level data for the other (secondary trait). Specifically, Λ was defined as the sum of log likelihood that the causal variant

underlying secondary trait is more likely to be same as than distinct from the variant underlying primary trait, as integrated over a set of likely causal variants under a GWAS peak of primary trait:

$$\Lambda = \sum_{i \in \mathcal{N}_\theta^1(m^*)} L_1(i) \cdot \log \frac{L_1(i)L_2(i)}{\max\limits_{j \notin \mathcal{N}_\theta^2(i)} L_1(i)L_2(j)}$$

$$= \sum_{i \in \mathcal{N}_\theta^1(m^*)} L_1(i) \cdot (\log\, L_2(i) \,-\, \max_{j \notin \mathcal{N}_\theta^2(i)} \{\log\, L_2(j)\})$$

where m^* is the most associated SNP for primary trait, $L_1(i)$ and $L_2(i)$ are the likelihood of SNP *i* being causally associated with primary and secondary traits, respectively, and $\mathcal{N}_\theta^1(i)$ and $\mathcal{N}_\theta^2(i)$ are the sets of SNPs within LD neighborhood around SNP *i*, as defined by $\{\, SNP\ j \mid r_{i,j}^2 > \theta \,\}$. We derived $\mathcal{N}_\theta^1$ from the reference LD panel and $\mathcal{N}_\theta^2$ directly from the genotypes of secondary trait cohort. We used disease outcome as primary trait, leveraging the larger sample size and dense genotyping, and gene expression as secondary trait, taking advantage of the availability of individual genotype data.

We calculated the likelihood of causal association by approximating the local LD structure with pairwise correlation as previously described[38,39]. Briefly, when SNP *c* is the only causal variant in the locus with non-centrality λ_c, the association static z_i of non-causal SNP *i* follows a normal distribution $N(r_{i,c}\ \lambda_c, 1)$, where $r_{i,c}$ is LD between SNPs *i* and *c* measured in pairwise Pearson correlation of genotypes. In general, when association statistics $\boldsymbol{Z} = (z_1, z_2, \dots, z_M)^T$ are provided for all *M* SNPs in the analysis window, the likelihood of SNP *i* being the causal variant with non-centrality λ_i is:

$$L(\boldsymbol{Z};\ \lambda_i \neq 0) = \phi_{MVN}(\boldsymbol{Z}; mean = \Sigma(\lambda_i \circ C), var = \Sigma)$$

where ϕ_{MVN} is the multivariate normal density function, *C* is an incident vector with $C_k = 1$ if and only if $k = i$, Σ is a *M* x *M* local LD matrix defined by pairwise Pearson correlation between genotypes, and $\circ$ is element-wise multiplication [Kichaev]. Since we do not know the true non-centrality of causal variant, we estimated the profile likelihood, which simplifies to a closed form:

$$\log L\left(\boldsymbol{Z};\ \lambda_i^{mle}\right) = \frac{1}{2}\left(-\boldsymbol{Z}^T\Sigma^{-1}\boldsymbol{Z} + z_i^2\right) - \frac{1}{2}\log((2\pi)^M\ |\Sigma|)$$

with $\lambda_i^{mle} = z_i$. Thus, given association statistics for primary and secondary traits, $\boldsymbol{Z} = (z_1, z_2, \dots, z_m)^T$ and $\boldsymbol{W} = (w_1, w_2, \dots, w_m)^T$, the test statistic Λ simplifies to:

$$\Lambda = \sum_{i \in \mathcal{N}_\theta^1(m^*)} e^{\frac{1}{2}(z_i^2 - z_{m^*}^2)} \cdot (w_i^2 - \max_{j \notin \mathcal{N}_\theta^2(i)} w_j^2)$$

The p-value of joint likelihood is estimated by permuting phenotypes of secondary traits as under the trivial null hypothesis that that there is no casual variant for secondary trait in the locus ("H_0"). With respect to the more likely null that distinct causal variants underlie association signals of two traits ("H_2"), we can show that asymptotically as the non-centrality of causal variant increases, p-values estimated from H_0 behave conservatively with respect to H_2 (Supplementary Notes):

$$P_{JLIM} = P(\Lambda \geq l|H_0) \geq P(\Lambda \geq l|H_2)$$

Thus, with large enough sample or effect sizes, joint likelihood test against H_0 will also reject H_2 in favor of alternative hypothesis of shared causal variant ("H_1"). Further, to evaluate whether this property holds for practical non-centrality values, we examined the negative controls simulating H_2 in ATG16L1 and CD58 loci, specifically, if P_{JLIM} was highly shifted toward 1.0 (Figure S2) and larger than empirically estimated false positive rates as expected (Table S4).

For both simulated and real GWAS data, we applied JLIM to SNPs with data for both primary and secondary traits, present in the reference LD panels, and within 100kb of the most associated marker to disease ("lead SNP"). In ImmunoChip data, the analysis windows were further confined by the boundaries of the fine-mapping intervals. We compared each lead SNP to eQTL data for all genes with a transcription start sites (TSS) up to 1Mb from the lead SNP, and an eQTL association $p < 0.05$ for at least one SNP in the analysis window. To minimize computational burden, we did not consider SNPs associated with neither disease or eQTL (association p > 0.1 to both). For the reference LD panel, we used the base haplotypes of Hapgen simulation for simulated datasets, and non-Finnish European samples (*n*=404) of the 1000 Genomes Project (phase 3, release 2013/05/02) for ImmunoChip loci.

We corrected for multiple tests using false discovery rate (FDR) levels and Bonferroni correction. The FDR was calculated separately for specific disease and cell type combination as:

$$FDR(p) = \frac{p\ N}{\#\{P_{JLIM} \leq p\}}$$

where *p* is a JLIM p-value cut-off, and *N* is the number of all tested disease lead SNP-eQTL candidate gene combinations. The FDR was calculated for each cell type since the distribution of JLIM p-values can vary depending on the disease relevance of cell type. To provide a list of higher confidence hits in each disease, we also applied the Bonferroni correction to nominal JLIM p-values for the number of tests across all three cell types.

Bayesian *coloc*. We ran Bayesian *coloc*[19] using default parameter settings on both simulated and real data as described for JLIM. For simulated data, we used colocalization prior p_{12} values of 10^{-5} or 10^{-6}, which are default values for higher sensitivity and higher specificity, respectively. The beta and variance of beta were used for all SNPs in the analysis window in case/control mode. We calculated accuracy as the area under the receiver operator curve (ROC; Figure S3).

As ImmunoChip data is only available as summary statistics, we used the minor allele frequencies from non-Finnish Europeans from the 1000 Genomes Project, and quantitative beta and variance of beta calculated on eQTL association data, and a colocalization prior $p_{12} = 10^{-6}$. We did not consider the type 1 diabetes data, where case/control sample size is limited after excluding affected sib pair data.

Estimating the number of disease GWAS loci with consistent eQTL effects. We expect JLIM p-values to follow a bimodal distribution with modes close to zero and one when the data support a model of shared or distinct causal effects, respectively. Conversely, under the null model of no *cis*-eQTL association, we expect a uniform p-value distribution. We can thus estimate the proportion disease-eQTL pairs belonging to the null π_0 , same $\pi_1$and distinct π_2 causal variant models from the observed p-value distribution[40] (Figures S11-13). To assess if the strength of the eQTL association influences the likelihood of identifying a shared causal variant, we calculate these proportions for subsets of trait pairs defined by minimum eQTL p-value. In each bin, we identified the limits of the uniform portion of the distribution γ_1 and γ_2 and estimate π_0, π_1, and π_2 as:

$$\pi_0 = \frac{\#\{\gamma_1 < P_{JLIM} < \gamma_2\}}{(\gamma_2 - \gamma_1)N}$$

$$\pi_1 = \frac{\#\{P_{JLIM} \leq \gamma_1\}}{N} - \gamma_1\pi_0$$

$$\pi_2 = \frac{\#\{P_{JLIM} \geq \gamma_2\}}{N} - (1 - \gamma_2)\pi_0$$

To estimate the number of disease GWAS loci that can be explained by consistent effect of same causal variant on disease and eQTL (denoted by $\mathcal{C}$ below), we incrementally relaxed the p-value cut-offs of JLIM and examined the trends of the number of disease loci with at least one JLIM hit and subtracted the expected number of false positive loci (Figures 1 and S14). Specifically, at each JLIM p-value cutoff p_i, we successively calculated $\mathcal{C}(p_i)$:

$$\mathcal{C}(p_i) = \mathcal{C}(p_{i-1}) + |\mathcal{D}(p_i) \cap \mathcal{D}(p_{i-1})^c| - \sum_{d \in \mathcal{D}(p_{i-1})^c} \mathcal{E}(d, p_i) + \sum_{d \in \mathcal{D}(p_{i-1})^c} \mathcal{E}(d, p_{i-1})$$

where $p_{i-1} < p_i$ with $p_0 = 0$, $\mathcal{D}(p)$ is the set of disease GWAS loci with at least one eQTL gene in any cell type passing the JLIM p-value cut-off p, and $\mathcal{E}(d, p)$ is the probability that disease GWAS locus d has a false positive eQTL gene passing the JLIM p-value cutoff p. We estimated the lower and upper bounds of $\mathcal{E}(d, p)$ using the Monte Carlo method by randomly selecting false positive eQTL genes within the locus d at rates of $(1 - \pi_1) \cdot lb$ or $(1 - \pi_1) \cdot ub$ over 1,000 iterations. The lb and ub are the lower and upper bounds of false positive rate of JLIM against true null. Note that π_1 and lb depend on the cell type and strength of eQTL association.

As the true null is mixture of two nulls, H_0 and H_2, the false positive rate of JLIM against true null $P(\Lambda \geq l | H_0 \cup H_2)$ can be bounded by using the following decomposition:

$$P(\Lambda \geq l | H_0 \cup H_2) = P(\Lambda \geq l | H_0) \frac{P(H_0)}{P(H_0) + P(H_2)} + P(\Lambda \geq l | H_2) \frac{P(H_2)}{P(H_0) + P(H_2)}$$

While the false positive rate under distinct null $P(\Lambda \geq l | H_2)$ is difficult to estimate, it is non-negative by definition and asymptotically bounded by permutation p-value $P(\Lambda \geq l | H_0)$, i.e. P_{JLIM}, as the non-centrality of causal variant increases. Therefore, we took:

$$ub = P_{JLIM}$$

$$lb = P_{JLIM} \frac{\pi_0}{\pi_0 + \pi_2} = P_{JLIM} \frac{\pi_0}{1 - \pi_1}$$

and estimated the bounds of locus-level false positive rates $\mathcal{E}(d, p)$ and number of disease loci with consistent effects $\mathcal{C}(p_i)$.

Bibiliography

1. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337,** 1190–1195 (2012).
2. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28,** 1045–1048 (2010).
3. Walsh, S. J. & Rau, L. M. Autoimmune diseases: a leading cause of death among young and middle-aged women in the United States. *Am J Public Health* **90,** 1463–1466 (2000).
4. Autoimmune Diseases Coordinating Committee. *Report of the Autoimmune Diseases Coordinating Committee*. (American Autoimmune Related Diseases Association (AARDA) & National Coalition of Autoimmune Patient Groups (NCAPG), 2011).
5. Eaton, W. W., Rose, N. R., Kalaydjian, A., Pedersen, M. G. & Mortensen, P. B. Epidemiology of autoimmune diseases in Denmark. *Journal of Autoimmunity* **29,** 1–9 (2007).
6. Criswell, L. A. *et al.* Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. *AJHG* **76,** 561–571 (2005).
7. Cotsapas, C. *et al.* Pervasive Sharing of Genetic Effects in Autoimmune Disease. **7,** e1002254 (2011).
8. International Multiple Sclerosis Genetics Consortium (IMSGC) *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics* **45,** 1353–1360 (2013).
9. Fortune, M. D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nature Genetics* (2015). doi:10.1038/ng.3330
10. Cotsapas, C. & Hafler, D. A. Immune-mediated disease genetics: the shared basis of pathogenesis. *Trends Immunol.* **34,** 22–26 (2013).
11. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13,** 101 (2011).
12. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genetics* **6,** e1000888 (2010).
13. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics* **45,** 124–130 (2013).
14. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).
15. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–552 (2014).
16. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47,** 1228–1235 (2015).
17. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14,** 483–495 (2013).
18. Nica, A. C. & Dermitzakis, E. T. Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet.* **17,** R129–34 (2008).
19. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics* **10,** e1004383 (2014).
20. Guo, H. *et al.* Integration of disease association and eQTL data using a Bayesian colocalisation

approach highlights six candidate causal genes in immune-mediated diseases. *Hum. Mol. Genet.* **24,** 3305–3313 (2015).

21. Zhu, Z., Zhang, F., Hu, H., Bakshi, A. & Robinson, M. R. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* (2016).
22. He, X. *et al.* Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet.* **92,** 667–680 (2013).
23. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
24. Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344,** 519–523 (2014).
25. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genetics* **7,** e1002003 (2011).
26. Malik, R. *et al.* Shared genetic basis for migraine and ischemic stroke: A genome-wide analysis of common variants. *Neurology* **84,** 2132–2145 (2015).
27. Winsvold, B. S. *et al.* Genetic analysis for a shared biological basis between migraine and coronary artery disease. *Neurol Genet* **1,** e10 (2015).
28. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
29. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics* **6,** e1000895 (2010).
30. Wallace, C. *et al.* Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Hum. Mol. Genet.* **21,** 2815–2824 (2012).
31. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47,** 1236–1241 (2015).
32. de Jager, P. L. *et al.* The role of the CD58 locus in multiple sclerosis. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 5264–5269 (2009).
33. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* **40,** 955–962 (2008).
34. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27,** 2304–2305 (2011).
35. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2013).
36. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *AJHG* **81,** 559–575 (2007).
37. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38,** 904–909 (2006).
38. Kichaev, G. *et al.* Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genetics* **10,** e1004722 (2014).
39. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B. & Eskin, E. Identification of causal genes for complex traits. *Bioinformatics* **31,** i206–i213 (2015).
40. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100,** 9440–9445 (2003).

Table 1. Only a minority of disease associations share causal variants with eQTLs across three immune cell subpopulations. We identified 261 disease associations in ImmunoChip regions with at least one eQTL within 100kb of the most associated SNP. Only 41/261 (16%) of these associations show evidence of a shared effect with an eQTL in that region. Thus, while eQTLs are abundant in disease-associated loci, they do not appear to be driven by the same causal variant as the disease association. [a]We only consider associations reported at genome-wide significant levels and overlapping genomic regions densely genotyped on ImmunoChip, excluding conditional peaks and MHC loci (see Methods). [b]eQTLs are selected if there is nominal association (eQTL $p < 0.05$) to at least one SNP within 100kb of the most associated SNP to disease, and a transcription start site of the gene within 1Mb of that SNP. [c]Number of loci where disease association is consistent with a shared effect for at least one eQTL (FDR < 5%).

	Number of loci								
	Densely	eQTL present [b]				Driven by same effect [c]			
Disease	genotyped [a]	CD4$^+$	CD14$^+$	LCL	Total	CD4$^+$	CD14$^+$	LCL	Total
MS	59	54	55	55	56	8	3	6	12
IBD	69	69	69	68	69	6	9	1	12
Crohn	19	18	18	18	18	2	1	0	3
UC	10	10	9	10	10	2	1	3	4
T1D	47	39	40	36	40	2	0	0	2
RA	34	34	34	34	34	2	0	1	3
CEL	34	34	34	34	34	3	2	0	5
Overall	272	258	259	255	261	25	16	11	41

Table 2. Forty one loci harbor eQTLs driven by the same variants as an association to at least one of seven diseases. We find 57 instances of shared disease-eQTL effects in 41 loci (joint likelihood of shared association FDR < 5%). [a]Variant with the minimum association *p* value to disease in the ImmunoChip summary statistics. [b]Minimum eQTL *p* value for any SNP within 100kb of the lead SNP. Dashes (-) indicate genes that are either not detected or with minimum eQTL P > 0.05 in that cell type. [c]Highlighted in bold are disease-eQTL pairs with false discovery rate < 5%. Asterisk (*) marks eQTL genes passing Bonferroni correction.

			CD4$^+$ T cell		CD14$^+$ monocytes		LCL	
Disease	**Lead SNP [a]**	**Gene**	**eQTL P [b]**	**JLIM P [c]**	**eQTL P [b]**	**JLIM P [c]**	**eQTL P [b]**	**JLIM P [c]**
MS	rs12749591	*PRKCZ*	-	-	-	-	2.4×10^{-4}	**5.0×10^{-4}**
MS	rs35967351	*SLAMF7*	2.3×10^{-3}	0.27	4.1×10^{-3}	0.94	5.4×10^{-8}	**$< 10^{-6}$ ***
MS	rs35967351	*NHLH1*	8.2×10^{-5}	**3.0×10^{-6} ***	-	-	8.9×10^{-3}	0.96
MS	rs1359062	*RGS1*	0.042	0.031	1.6×10^{-21}	**$< 10^{-6}$ ***	-	-
MS	rs9989735	*SP140*	7.5×10^{-13}	1.00	7.0×10^{-3}	0.58	1.3×10^{-9}	**1.0×10^{-6} ***
MS	rs71624119	*ANKRD55*	2.0×10^{-10}	**1.5×10^{-5} ***	-	-	-	-
MS	rs71624119	*IL6ST*	5.9×10^{-5}	**4.0×10^{-4}**	4.9×10^{-4}	0.98	-	-
MS	rs917116	*JAZF1*	6.2×10^{-16}	**$< 10^{-6}$ ***	3.3×10^{-6}	0.90	1.1×10^{-3}	1.00
MS	rs60600003	*ELMO1*	1.2×10^{-8}	**8.0×10^{-6} ***	1.5×10^{-4}	8.0×10^{-3}	-	-
MS	rs1966115	*ZC2HC1A*	4.5×10^{-12}	**$< 10^{-6}$ ***	3.4×10^{-40}	**$< 10^{-6}$ ***	3.4×10^{-30}	**$< 10^{-6}$ ***
MS	rs1966115	*PKIA*	1.2×10^{-15}	1.00	0.020	0.71	1.1×10^{-9}	**3.0×10^{-5}**
MS	rs10783847	*METTL21B*	8.8×10^{-21}	**$< 10^{-6}$ ***	2.0×10^{-21}	**$< 10^{-6}$ ***	-	-
MS	rs7132277	*CDK2AP1*	2.1×10^{-4}	0.95	1.3×10^{-8}	0.018	6.5×10^{-13}	**$< 10^{-6}$ ***
MS	rs12946510	*GSDMB*	4.1×10^{-17}	**$< 10^{-6}$ ***	-	-	2.8×10^{-4}	0.92
MS	rs12946510	*ORMDL3*	5.7×10^{-13}	**$< 10^{-6}$ ***	0.025	0.63	3.1×10^{-26}	**1.0×10^{-6} ***
MS	rs17785991	*SLC9A8*	2.2×10^{-6}	**2.0×10^{-4}**	6.3×10^{-9}	1.00	5.2×10^{-3}	0.95
IBD	rs13001325	*IL18R1*	7.2×10^{-11}	**7.0×10^{-6} ***	0.022	0.71	0.026	0.78
IBD	rs3749171	*GPR35*	7.6×10^{-3}	0.92	9.5×10^{-8}	**3.0×10^{-4}**	2.1×10^{-3}	0.76
IBD	rs55770741	*ERAP2*	2.2×10^{-60}	**$< 10^{-6}$ ***	1.9×10^{-57}	**$< 10^{-6}$ ***	1.1×10^{-105}	**$< 10^{-6}$ ***
IBD	rs17622378	*SHROOM1*	0.043	0.62	1.6×10^{-3}	**2.0×10^{-4}**	0.017	0.79
IBD	rs17622378	*KIF3A*	4.0×10^{-5}	**1.3×10^{-4}**	0.017	0.30	-	-
IBD	rs444210	*RNASET2*	9.3×10^{-51}	**$< 10^{-6}$ ***	3.9×10^{-14}	1.00	9.0×10^{-8}	1.00

IBD	rs7848647	*TNFSF15*	1.3×10^{-3}	0.017	1.3×10^{-16}	**2.0×10^{-6} ***	0.042	0.087
IBD	rs34779708	*CUL2*	2.7×10^{-7}	**3.0×10^{-5}**	0.034	0.79	4.1×10^{-3}	0.76
IBD	rs2590348	*CISD1*	-	-	9.0×10^{-13}	**$< 10^{-6}$ ***	0.045	0.22
IBD	rs12448902	*TUFM*	1.0×10^{-20}	**$< 10^{-6}$ ***	5.6×10^{-28}	**$< 10^{-6}$ ***	0.011	0.97
IBD	rs9808651	*ETS2*	-	-	1.4×10^{-7}	**3.0×10^{-4}**	-	-
IBD	rs4456788	*ICOSLG*	2.9×10^{-6}	1.00	1.4×10^{-6}	**5.0×10^{-5}**	6.8×10^{-4}	0.97
IBD	rs2266961	*UBE2L3*	1.0×10^{-4}	4.0×10^{-3}	1.0×10^{-9}	**$< 10^{-6}$ ***	-	-
Crohn	rs6752107	*scaRNA5*	1.6×10^{-4}	0.12	1.8×10^{-20}	**$< 10^{-6}$ ***	-	-
Crohn	rs71624119	*ANKRD55*	2.0×10^{-10}	**2.0×10^{-5} ***	0.027	0.67	-	-
Crohn	rs71624119	*IL6ST*	5.9×10^{-5}	**5.0×10^{-4}**	4.9×10^{-4}	0.89	-	-
Crohn	rs3801810	*SKAP2*	9.7×10^{-14}	**1.0×10^{-6} ***	7.8×10^{-5}	1.00	-	-
UC	rs2147905	*TNFRSF14*	7.9×10^{-3}	0.43	-	-	2.5×10^{-8}	**3.0×10^{-5} ***
UC	rs11742304	*TPPP*	1.0×10^{-4}	1.00	0.012	0.70	1.5×10^{-7}	**1.1×10^{-5} ***
UC	rs11150589	*ITGAL*	4.2×10^{-12}	**1.0×10^{-5} ***	6.8×10^{-4}	0.99	0.025	0.64
UC	rs889561	*NFAT5*	5.6×10^{-3}	0.90	-	-	7.7×10^{-3}	**2.0×10^{-4}**
UC	rs889561	*ZFP90*	5.9×10^{-23}	**$< 10^{-6}$ ***	2.3×10^{-19}	**$< 10^{-6}$ ***	5.0×10^{-3}	0.95
RA	rs4681851	*FLNB*	1.1×10^{-6}	**2.0×10^{-4}**	3.1×10^{-3}	0.99	1.3×10^{-3}	0.90
RA	rs71624119	*ANKRD55*	2.0×10^{-10}	**2.0×10^{-5} ***	-	-	-	-
RA	rs71624119	*IL6ST*	5.9×10^{-5}	**3.0×10^{-4}**	4.9×10^{-4}	0.93	-	-
RA	rs3807306	*IRF5*	-	-	2.0×10^{-4}	1.00	6.7×10^{-20}	**1.0×10^{-6} ***
CEL	rs1359062	*RGS1*	0.042	0.025	1.6×10^{-21}	**$< 10^{-6}$ ***	-	-
CEL	rs2097282	*CCR2*	3.0×10^{-5}	**5.0×10^{-5} ***	1.7×10^{-8}	1.00	0.013	0.79
CEL	rs79758729	*ELMO1*	1.2×10^{-8}	**1.0×10^{-5} ***	4.0×10^{-4}	0.22	-	-
CEL	rs1893592	*UBASH3A*	4.5×10^{-14}	**3.0×10^{-6} ***	0.011	0.82	5.2×10^{-6}	1.00
CEL	rs4821124	*UBE2L3*	1.0×10^{-4}	9.0×10^{-3}	1.0×10^{-9}	**2.0×10^{-5} ***	-	-
T1D	rs917911	*CLEC2B*	1.4×10^{-5}	**3.0×10^{-5} ***	6.8×10^{-5}	0.94	1.6×10^{-3}	0.84
T1D	rs705705	*SUOX*	9.3×10^{-6}	**1.1×10^{-5} ***	1.9×10^{-10}	0.98	9.3×10^{-3}	0.085

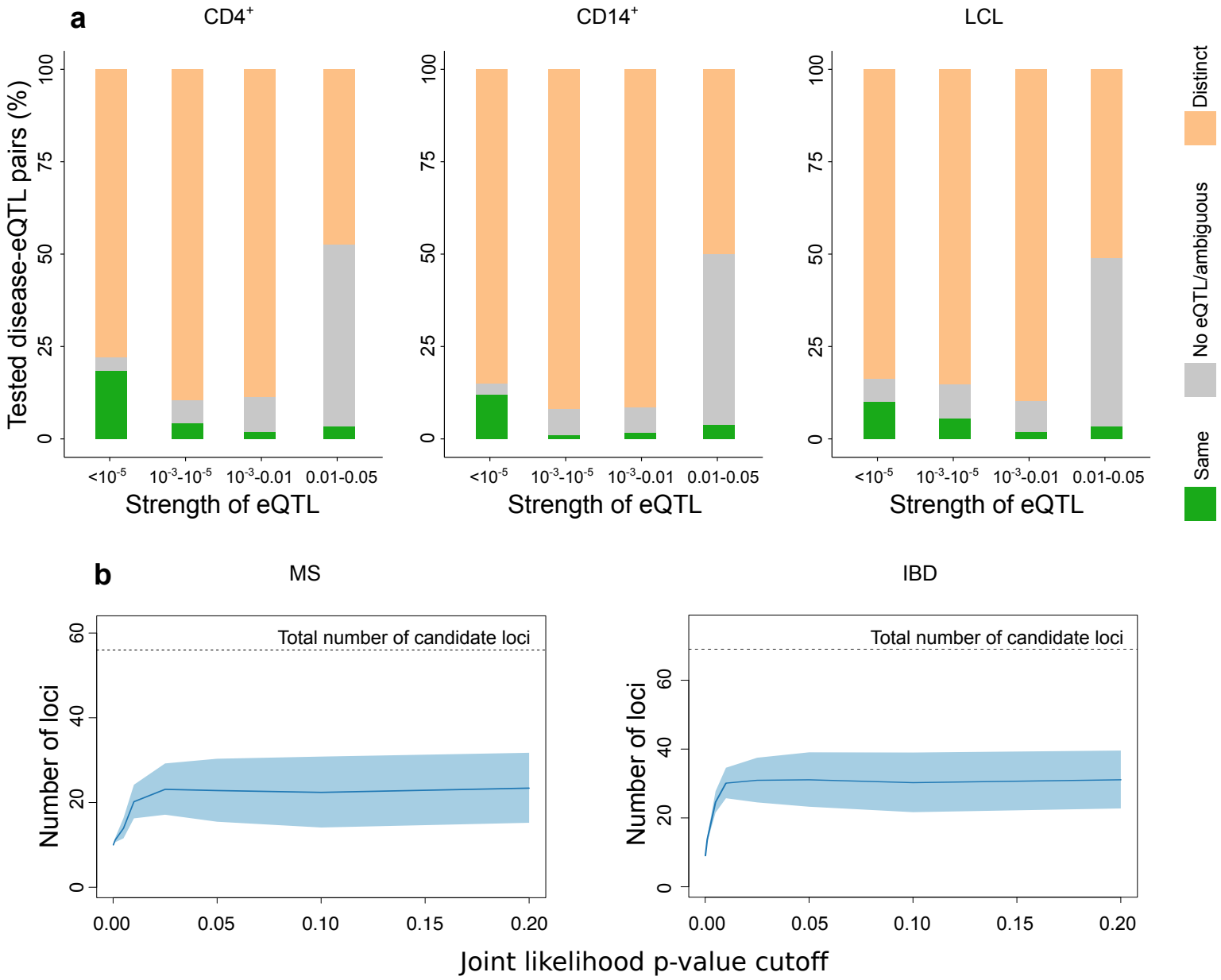

Figure 1. Only a minority of disease associations share causal variants with eQTLs across three immune cell subpopulations. (a) We find strong evidence that approximately 75% of eQTLs are driven by distinct causal variants (orange) to 261 disease risk associations across 155 ImmunoChip regions. The strength of eQTL association does not influence the proportion of shared effects (green) we are able to detect, suggesting this lack of overlap is not due to lack of power. We find no compelling evidence for either shared or distinct associations for a small proportion of disease-eQTL pairs (gray). **(b)** The median number of loci with at least one shared effect eQTL in any tissue (blue line) at more liberal significance thresholds remains constant after false positive adjustment, further supporting this conclusion. The shaded area represents the lower and upper expectation bounds for disease-eQTL pairs driven by the same causal variant. Only 31-57% of multiple sclerosis associations and 37-57% of inflammatory bowel disease associations are consistent with eQTL effects. Equivalent data for the other diseases are presented in Figure S14.

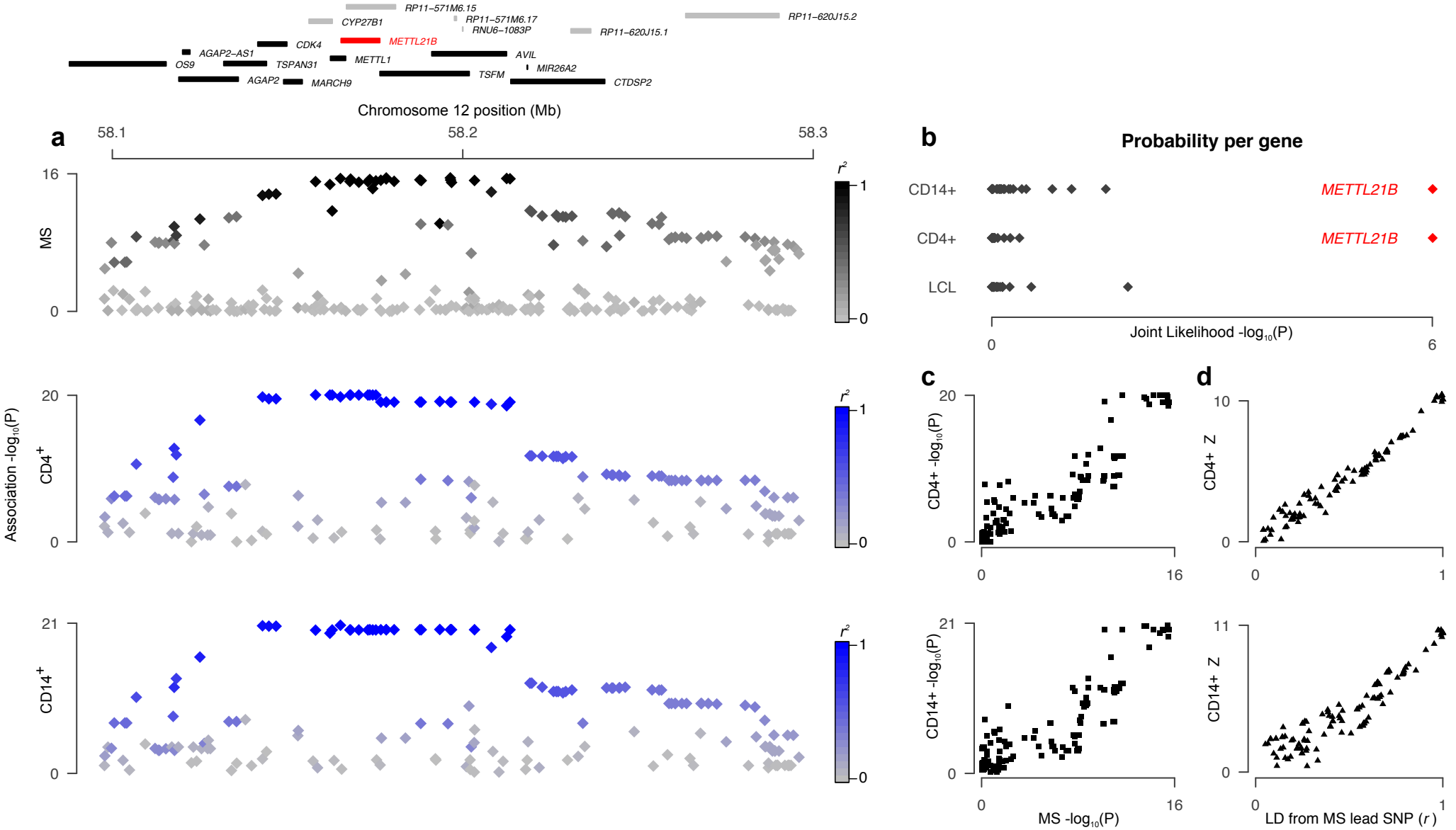

Figure 2. A multiple sclerosis association on chromosome 12 is consistent with eQTLs for *METTL21B* in both CD4⁺ T cells and CD14⁺ monocytes. (a) A genome-wide significant association to multiple sclerosis risk (upper panel; shading denotes strength of LD to the most associated variant rs10783847). This association is consistent with eQTLs for *METTL21B* in CD4⁺ T cells (middle panel) and CD14⁺ monocytes (lower panel, both shaded by LD to rs10783847), but not to eQTL data for any other genes in the region (upper gene track: black boxes denote 38 genes with eQTL data available in addition to *METTL21B* (red); gray denotes genes which are not reliably detected in our data or do not have eQTL p < 0.05 in the region). **(b)** Joint likelihood p-values for 39 candidate genes analyzed for this MS association peak in three cell types. Those with FDR < 5% are shown in red. **(c)** Association p-values for MS risk (x-axis) and eQTLs (y-axis) are strongly correlated for both CD4⁺ T cells (middle panel) and CD14⁺ monocytes (lower panel). **(d)** Similarly, eQTL association Z statistics scale linearly with LD (***r***, x axis) to rs10783847, consistent with a model of a single causal variant driving both disease association and eQTL.

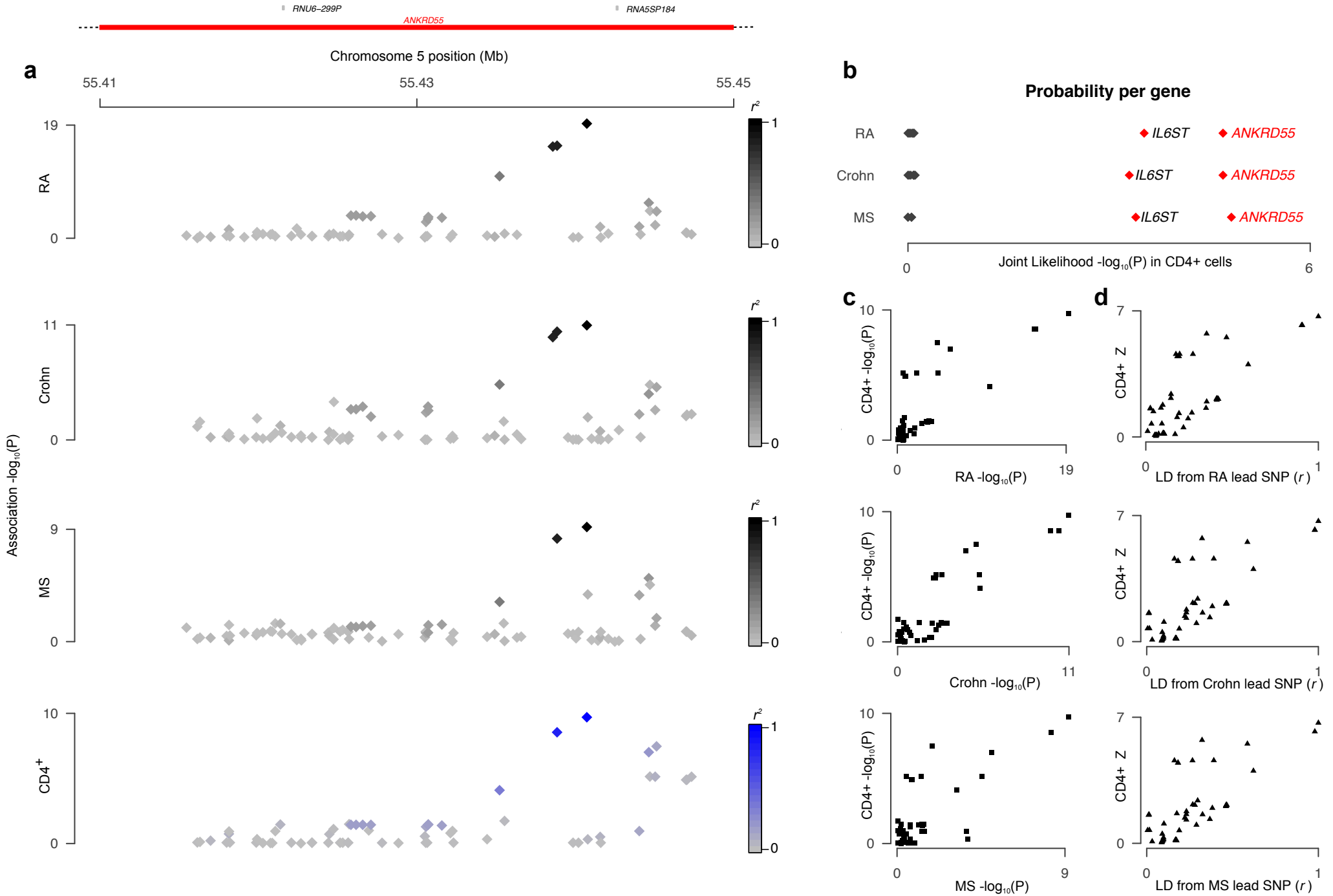

Figure 3. Associations to multiple sclerosis, Crohn disease and rheumatoid arthritis (RA) on chromosome 5 are consistent with an eQTL for *ANKRD55* in CD4$^+$ T cells. (a) Genome-wide significant associations to all three diseases (upper panels) and eQTL data for *ANKRD55* (lower panel; shading in all panels proportional to LD to the most associated variant rs71624119). Due to the variable density of ImmunoChip data, the analysis window is small and only overlaps the coding region of *ANKRD55*, though we test eQTLs for 11 genes with a transcriptional start site within 1Mb of the the association. **(b)** Joint likelihood p-values for nine candidate genes analyzed for this locus in CD4$^+$ T cells. Those with FDR < 5% are shown in red. **(c)** Association p-values for each disease (x axis) are strongly correlated to those for the *ANKRD55* eQTL in CD4$^+$ cells (y axis). **(d)** Similarly, eQTL association *Z* statistics scale linearly with LD (***r***, x axis) to rs71624119 for all three diseases, consistent with a model of a single causal variant driving all disease associations and the eQTL.