Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types

Sung Chun1,3, Alexandra Casparino4, Nikolaos A Patsopoulos3,5,6, Damien C Croteau-Chonka2,7, Benjamin A Raby2,7, Philip I De Jager2,3,5,6, Shamil R Sunyaev1,3,8 & Chris Cotsapas3,4,9

Most autoimmune-disease-risk effects identified by genome-wide association studies (GWAS) localize to open chromatin with gene-regulatory activity. GWAS loci are also enriched in expression quantitative trait loci (eQTLs), thus suggesting that most risk variants alter gene expression1,2. However, because causal variants are difficult to identify, and cis-eQTLs occur frequently, it remains challenging to identify specific instances of disease-relevant changes to gene regulation. Here, we used a novel joint likelihood framework with higher resolution than that of previous methods to identify loci where autoimmune-disease risk and an eQTL are driven by a single shared genetic effect. Using eQTLs from three major immune subpopulations, we found shared effects in only ~25% of the loci examined. Thus, we show that a fraction of gene-regulatory changes suggest strong mechanistic hypotheses for disease risk, but we conclude that most risk mechanisms are not likely to involve changes in basal gene expression.

The autoimmune and inflammatory diseases (AIDs) are heritable, complex diseases in which loss of tolerance to self-antigens results in either systemic or tissue-specific immune attack3,4. GWAS have identified hundreds of genomic regions mediating risk for several AIDs. These associations are primarily noncoding: lead GWAS SNPs are more likely to be associated with expression levels of neighboring genes than is expected by chance5,6, and the same lead SNPs are enriched in regulatory regions marked by chromatin accessibility and modification1,7. Fine-mapping has revealed enrichment of AID-associated variants in enhancer elements active in stimulated T-cell sub-populations8, and heritability is strongly enriched in such regulatory regions9,10. Collectively, these lines of evidence suggest that the majority of disease risk is mediated by changes to gene regulation in specific cell subpopulations.

However, these bulk analyses do not formally assess whether expression levels and disease risk can be attributed to a single underlying trait variant or to independent effects in a locus11,12. Though several methods have been developed to assess these alternatives by using eQTL data13–17, they have limited resolution to detect cases in which distinct disease and eQTL causal variants are in linkage disequilibrium. Here, we present an approach to test whether a GWAS risk association and an eQTL are driven by the same underlying genetic effect, thereby accounting for the linkage disequilibrium (LD) between causal variants. Using data from ImmunoChip studies of seven AIDs comprising >180,000 samples in total (Supplementary Table 1), we tested whether associations in 272 known risk loci were consistent with cis-eQTLs for genes in each region, in three relevant immune-cell populations: lymphoblastoid cell lines (LCLs), CD4+ T cells and CD14+ monocytes18,19.

When associations with two traits—here, disease trait and 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 causal variant12,20. Here, we directly evaluated this joint likelihood (Supplementary Fig. 1), in contrast to previous approaches that look for similarities in the shape of the association curve over multiple markers13,14,21–23. When the underlying causal effect is shared, joint likelihood is maximized when the same causal variant is modeled in both traits; in contrast, when the underlying causal variants are different, maximum joint likelihood is expected when their closest proxies are modeled. We empirically derived the null distribution of the joint likelihood ratio statistic by comparing disease associations with permuted eQTL data (Online Methods, Supplementary Fig. 2 and Supplementary Note). We thus directly evaluated whether two associations in the same locus, observed in different cohorts, were due to the same underlying effect.

To assess the performance of our method, we benchmarked it against three recently reported methods: coloc13, a well-calibrated...
Bayesian framework that considers spatial similarities in association data across sets of markers; gwas-pw24, which extends this idea to hierarchical priors and optimizes model parameters; and HEIDI/SMR 15, which applies Mendelian randomization between traits. We simulated 260 disease associations in ImmunoChip regions with at least one eQTL within 100 kb of the most-associated SNP. Only 55/260 (21%) of these associations showed evidence of a shared effect with an eQTL in that region. Thus, whereas eQTLs are abundant in disease-associated loci, they do not appear to be driven by the same causal variant as the disease association.

We identified 260 disease associations in ImmunoChip regions with at least one eQTL within 100 kb of the most-associated SNP. Only 55/260 (21%) of these associations showed evidence of a shared effect with an eQTL in that region. Thus, whereas eQTLs are abundant in disease-associated loci, they do not appear to be driven by the same causal variant as the disease association.

Bayesian framework that considers spatial similarities in association data across sets of markers; gwas-pw24, which extends this idea to hierarchical priors and optimizes model parameters; and HEIDI/SMR 15, which applies Mendelian randomization between traits. We simulated pairs of case–control cohorts with either the same or distinct causal variants driving association, and we found that our approach had the best overall performance (Supplementary Tables 2 and 3). When independent causal variants (i.e., not in LD) drove GWAS and

Bayesian framework that considers spatial similarities in association data across sets of markers; gwas-pw24, which extends this idea to hierarchical priors and optimizes model parameters; and HEIDI/SMR 15, which applies Mendelian randomization between traits. We simulated pairs of case–control cohorts with either the same or distinct causal variants driving association, and we found that our approach had the best overall performance (Supplementary Tables 2 and 3). When independent causal variants (i.e., not in LD) drove GWAS and

Bayesian framework that considers spatial similarities in association data across sets of markers; gwas-pw24, which extends this idea to hierarchical priors and optimizes model parameters; and HEIDI/SMR 15, which applies Mendelian randomization between traits. We simulated pairs of case–control cohorts with either the same or distinct causal variants driving association, and we found that our approach had the best overall performance (Supplementary Tables 2 and 3). When independent causal variants (i.e., not in LD) drove GWAS and

Bayesian framework that considers spatial similarities in association data across sets of markers; gwas-pw24, which extends this idea to hierarchical priors and optimizes model parameters; and HEIDI/SMR 15, which applies Mendelian randomization between traits. We simulated pairs of case–control cohorts with either the same or distinct causal variants driving association, and we found that our approach had the best overall performance (Supplementary Tables 2 and 3). When independent causal variants (i.e., not in LD) drove GWAS and

Bayesian framework that considers spatial similarities in association data across sets of markers; gwas-pw24, which extends this idea to hierarchical priors and optimizes model parameters; and HEIDI/SMR 15, which applies Mendelian randomization between traits. We simulated pairs of case–control cohorts with either the same or distinct causal variants driving association, and we found that our approach had the best overall performance (Supplementary Tables 2 and 3). When independent causal variants (i.e., not in LD) drove GWAS and
Table 2  Fifty five loci contain eQTLs driven by the same variants as an association with at least one of seven diseases (continued)

| Disease | Lead SNPx | Gene   | CD4+ T cell | CD14+ monocytes | LCL |
|---------|-----------|--------|-------------|-----------------|-----|
| IBD     | rs2143178 | RPL3   | 4.1 × 10^{-4} | 5 × 10^{-4}     | 5.8 × 10^{-3} | 0.87 |
| IBD     | rs2266961 | UBE2L3 | 1.0 × 10^{-4} | 8 × 10^{-4}     | 1.0 × 10^{-9} | <10^{-6} |
| Crohn   | rs6752107 | SCARNA5 | 1.6 × 10^{-4} | 0.11            | 1.8 × 10^{-20} | <10^{-6} |
| Crohn   | rs71624119 | ANKRD55 | 2.0 × 10^{-10} | 2 × 10^{-9}     | –               | –               |
| Crohn   | rs71624119 | IL6ST  | 5.9 × 10^{-5} | 4 × 10^{-4}     | 4.9 × 10^{-1}  | 0.78 |
| Crohn   | rs3801810 | SKAP2  | 9.7 × 10^{-14} | <10^{-6}       | 7.8 × 10^{-5}  | 1.00 |
| UC      | rs2147905 | TNFRSF14 | 7.9 × 10^{-3} | 0.25           | –               | –               |
| UC      | rs11742304 | TTPP   | 1.0 × 10^{-4} | 0.003          | 4.0 × 10^{-8}  | 0.85 |
| UC      | rs7472814 | TNPO3  | 4.2 × 10^{-12} | 2 × 10^{-5}    | 6.8 × 10^{-8}  | 0.96 |
| UC      | rs11150589 | ITGA1 | 5.6 × 10^{-3} | 0.32           | –               | –               |
| UC      | rs889561  | NFA5   | 2.3 × 10^{-13} | <10^{-6}       | 5.0 × 10^{-3}  | 0.85 |
| UC      | rs889561  | ZFP90  | 5.9 × 10^{-23} | 1 × 10^{-6}    | –               | –               |
| UC      | rs6017342 | OSER1  | 2.8 × 10^{-3} | 0.003          | –               | –               |
| RA      | rs4681851 | FLNB   | 1.1 × 10^{-6} | 2.2 × 10^{-4}  | 3.1 × 10^{-3}  | 0.89 |
| RA      | rs71624119 | ANKRD55 | 2.0 × 10^{-10} | 9 × 10^{-6}    | –               | –               |
| RA      | rs71624119 | IL6ST  | 5.9 × 10^{-5} | 2 × 10^{-4}    | 4.9 × 10^{-4}  | 0.77 |
| RA      | rs3807306 | IRF5   | –               | 2.0 × 10^{-4}  | 0.97           | 6.7 × 10^{-20} | <10^{-6} |
| RA      | rs12936409 | NRI1D1 | 1.8 × 10^{-3} | 3 × 10^{-4}    | –               | –               |
| RA      | rs12936409 | NARA   | 1.5 × 10^{-3} | 5 × 10^{-4}    | –               | –               |
| CEL     | rs1359062 | RGS1   | –               | 1.6 × 10^{-21} | <10^{-6}       | –               |
| CEL     | rs1908422 | CD28   | 1.2 × 10^{-3} | 1.6 × 10^{-3}  | –               | –               |
| CEL     | rs2097282 | CCR2   | 3.0 × 10^{-5} | 6 × 10^{-5}    | 1.7 × 10^{-8}  | 1.00 |
| CEL     | rs7975879 | ELMO1  | 1.2 × 10^{-8} | 3 × 10^{-6}    | 4.0 × 10^{-4}  | 0.07 |
| CEL     | rs1893592 | UBASH3A | 4.5 × 10^{-14} | 4 × 10^{-6}    | –               | 5.2 × 10^{-6} | 1.00 |
| CEL     | rs4821124 | UBE2L3 | 1.0 × 10^{-4} | 1.3 × 10^{-3}  | 1.0 × 10^{-9}  | 3 × 10^{-5} |
| T1D     | rs12416116 | KLIN   | –               | –               | –               | 4.7 × 10^{-4} | 9 × 10^{-5} |
| T1D     | rs917911 | CLEC28 | 1.4 × 10^{-5} | 1.4 × 10^{-5}  | 6.8 × 10^{-5}  | 0.83 |
| T1D     | rs705705 | SUOX   | 9.3 × 10^{-6} | 6 × 10^{-6}    | 1.9 × 10^{-10} | 0.98 |
| T1D     | rs72727394 | RASGAP1 | 3.9 × 10^{-3} | 0.72           | –               | 3.7 × 10^{-10} | 9 × 10^{-4} |
| T1D     | rs7239671 | CD22G  | 2.1 × 10^{-3} | 0.83           | 9.7 × 10^{-3}  | 0.13 |
| T1D     | rs280497 | KEAP1  | –               | 6.7 × 10^{-3}  | 0.82           | 2.0 × 10^{-4} | 1.6 × 10^{-4} |
| T1D     | rs280497 | CDC37  | –               | 8.3 × 10^{-3}  | 0.85           | 1.4 × 10^{-3} | 1.0 × 10^{-3} |
| T1D     | rs6518530 | LINCO1424 | –               | –               | 9.0 × 10^{-4}  | 5 × 10^{-4} |

We found 77 instances of shared disease-eQTL effects in 55 loci (joint likelihood of shared association FDR <5%).

*xVariant with the minimum association P value with disease in the ImmunoChip summary statistics. **Minimum eQTL P value for any SNP within 100 kb of the lead SNP. Dashes (–) indicate that in the majority of AID loci, variants causally involved in disease association even when the underlying causal variants were in strong LD (area under the curve of 0.883 when 0.7 < r^2 < 0.8; Supplementary Figs. 3 and 4), whereas the other methods showed substantial false-positive rates, reporting distinct effects as being shared. We also found that our method was robust to within-continent levels of population structure (Supplementary Figs. 5 and 6) and remained robust when analysis was limited to a subset of SNPs for computational efficiency (Supplementary Fig. 7; color fared similarly, Supplementary Fig. 8). Our method also performed well when multiple independent causal variants affected one or both traits (Supplementary Figs. 9–11). In practice, our resolution became limited at high LD levels (r^2 > 0.8), thus causing the false-positive rate to increase dramatically. The resolution was also limited when the eQTL effect was very weak (P > 0.01; Supplementary Figs. 12–15). Thus, within these limits, we were able to accurately detect cases of shared genetic effects between two traits.

To dissect AID-risk loci, we first identified densely genotyped ImmunoChip loci showing genome-wide significant association, excluding the major histocompatibility locus, owing to the extensive LD structure in that region (URLs and Table 1). We next identified genes in a 1-Mb window centered on the most associated variant in each locus. In agreement with previous observations that eQTLs are frequently found in GWAS loci, we found that 260/272 loci had at least one gene with an eQTL (P < 0.01) in at least one cell type, and most such effects were common across all three cell types (Table 1). We tested whether any eQTLs in these loci appeared to be driven by the same underlying effect as the disease associations. We found evidence for shared effects for only 77/5,749 pairs in 55/260 (21%) loci across all diseases, and the proportion varied from 4/34 (12%) for rheumatoid arthritis loci to 6/10 (60%) for ulcerative colitis loci (false discovery rate (FDR) <5%; Tables 1 and 2). Of these 77 shared effects, 45 passed even the more stringent familywise multiple testing correction (Bonferroni-corrected P < 0.05). Thus, our analysis revealed that in the majority of AID loci, variants causally involved in disease phenotypes did not overlap with variants responsible for eQTL signals in the three broad cell populations analyzed, which represent the major arms of the immune lineages. Overall, we found that >75% of the tested disease-eQTL pairs appeared to be associated with distinct genetic variants in the same locus (Fig. 1).
Figure 1 Only a minority of disease associations share genetic effects with eQTLs across three immune-cell subpopulations. (a) We found strong evidence that approximately 75% of eQTLs are driven by distinct genetic effects (orange) for 260 disease-risk associations across 154 ImmunoChip regions. The proportion of shared effects (green) that we were able to detect was <25%, even for relatively strong eQTLs with nominal association $P < 10^{-5}$. We found 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 cell type (blue line) at more liberal significance thresholds remains constant after false-positive adjustment, thus 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 31–47% of multiple sclerosis associations and 30–45% of inflammatory bowel disease associations are consistent with eQTL effects. MS, multiple sclerosis; IBD, inflammatory bowel disease. Equivalent data for the other diseases are presented in Supplementary Figure 19.

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 (Supplementary Fig. 4), we remained concerned that this lack of overlap might have been due to low statistical power in the eQTL data, owing to the limited sample size of the cohorts. However, we found that even among the most strongly supported eQTLs (nominal $P < 10^{-5}$), <25% showed evidence of shared effects.

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 with multiple sclerosis (MS) risk (top; shading denotes strength of LD with the most associated variant, rs10783847). This association is consistent with eQTLs for METTL21B in CD4+ T cells (middle) and CD14+ monocytes (bottom, both shaded by LD with rs10783847), but not with eQTL data for any other genes in the region (upper gene track; black boxes denote 31 genes with eQTL data available in addition to METTL21B (red); gray denotes genes that were not reliably detected in our data or do not have eQTL association $P < 0.01$ in the region). (b) Joint likelihood $P$ values for 32 candidate genes analyzed for this multiple sclerosis association peak in three cell types. Those with FDR <5% are shown in red. (c) Association $P$ values for multiple sclerosis risk (x axis) and eQTLs (y axis) are strongly correlated for both CD4+ T cells (top) and CD14+ monocytes (bottom). (d) Similarly, eQTL association Z statistics scale linearly with LD (r, x axis) with rs10783847, in agreement with a model of a single causal variant driving both disease association and eQTL.
with disease associations. In contrast, we found strong evidence for distinct effects for the majority of disease–eQTL pairs, and only a subset of comparisons were ambiguous, thus suggesting that our method is adequately powered to detect shared effects where they exist (Fig. 1a and Supplementary Figs. 16–18). To assess whether power, rather than eQTLs, affected the total number of loci that could be resolved, we looked more deeply at our significance-threshold settings. We found that thresholds that were more liberal did not increase the number of true-positive results after adjustment for false-positive rate, thus indicating that most loci did not contain any gene with an eQTL consistent with the disease association (Fig. 1b and Supplementary Fig. 19). Cumulatively, our results demonstrated that only a minority of AID-risk effects drive eQTLs in the three cell populations tested, which were drawn from diverse lineages of the immune system.

We next focused on the subset of 77 disease–eQTL pairs in 55 loci where we were able to detect strong evidence of a shared effect (Table 2). We found that 59/77 (77%) of the effects were restricted to one cell population, thus indicating that tissue-specific eQTLs are important components of the molecular underpinnings of disease (Supplementary Figs. 20 and 21). The remaining 18 effects were detected in multiple cell populations; for example, the multiple sclerosis association at rs10783847 on chromosome 12 was consistent with eQTLs for the transcript of methyltransferase-like 21B (METTL21B) in both CD4+ T cells and CD14+ monocytes but not for the remaining 31 genes in the immediate locus (Fig. 2). Although METTL21B is expressed in LCLs, there was no evidence of an eQTL in this cell population within 1 Mb from rs10783847. Similarly, for the multiple sclerosis association at rs1966115 on chromosome 8 and eQTLs for ZC2H1CA, and for the inflammatory bowel disease association at rs55770741 on chromosome 5 and eQTLs for ERAP2, we detected a shared effect in all three cell populations. In several cases, we found tissue-specific shared effects despite strong eQTLs for the same gene in other tissues: for ZFP90 and ulcerative colitis risk at rs889561 on chromosome 16, we also found shared effects in CD4+ and CD14+ but not LCLs, where we observed a ZFP90 eQTL at \( P = 0.005 \) with a low likelihood of a shared effect with GWAS (joint likelihood \( P = 0.85 \)). Instead, we found 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 were consistent with disease associations, thus suggesting that the diseaserelevant eQTLs were tissue specific.

We also found cases in which an eQTL was consistent with associations with multiple diseases. The ankyrin-repeat domain 55 (ANKRD55)
transcript on chromosome 5 had an eQTL in CD4⁺ T cells that exhibited shared associations with multiple sclerosis, Crohn disease and rheumatoid arthritis (Fig. 3), and all observations were significant after Bonferroni correction. We also found weaker evidence for shared effects across all three diseases and an eQTL for interleukin 6 signal transducer (IL6ST) in CD4⁺ T cells, which passed the FDR threshold but not Bonferroni correction (Supplementary Fig. 22). Similarly, a CD4⁺ eQTL for ELMO1 on chromosome 7 was consistent with associations with both celiac disease and multiple sclerosis (Supplementary Fig. 23), a CD14⁺ eQTL for RGS1 on chromosome 1 was consistent with associations with both celiac disease and multiple sclerosis (Supplementary Fig. 24), and three other eQTLs were consistent with associations with multiple diseases (Supplementary Figs. 25–27). In all cases, these were the only genome-wide significant disease associations reported in these loci. Because we considered each disease association independently, these results indicated that the same underlying risk variants drive risk for multiple diseases in these loci by altering gene expression, in agreement with observations of shared underlying risk variants driving risk for multiple diseases in these loci25.

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

Previous efforts to detect shared effects between traits in specific loci have relied on conditional analyses26 or have indirectly leveraged linkage disequilibrium to determine whether the shapes of a association peaks in the region are similar13,22,23,27. In contrast, our method directly evaluates whether the data support a shared effect through joint likelihood estimation. Through this direct evaluation, cases in which two associations are proximal can be resolved with higher resolution (Supplementary Figs. 3 and 4, and Supplementary Tables 2 and 3). Because our method is general, it may be useful in other contexts, such as establishing whether shared heritability between diseases is driven by the same underlying causal effects28.

More broadly, our results raise the question of how causal disease variants alter cell function and consequently induce risk, given the strong enrichment of disease-risk signals in gene-regulatory regions1, particularly gene enhancers8. We suggest that although gene-regulatory regions bearing risk variants are accessible in multiple immune-cell subpopulations, they may control gene expression in either a tissue-specific or condition-specific manner. These gene-regulatory events may be restricted to very specific cell populations, and easily accessible subsets — such as those analyzed here — may not adequately capture these events. Our results therefore reinforce the view that the appropriate cell-type and physiological conditions must be sought out in order to capture the pathologically relevant gene-regulatory changes driving disease risk.

URLs. Immunobase, http://www.immunobase.org/; PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

S.R.S. and C.S.C. were supported by NIH awards R01-MH101244-04, R01-GM105857-03, R01-GM078598-09 and U01-HG009881-01.

AUTHOR CONTRIBUTIONS

S.C. designed and performed research and authored the manuscript; A.C. performed research; N.A.P., D.C.-C., B.A.R. and P.L.D.J. contributed data and approved the manuscript; S.R.S. and C.C. designed and performed research and authored the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

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. Nature 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 autoimmunity diseases coordinating committee. (American Autoimmune Related Diseases Association (AARD) and National Coalition of Autoimmune Patient Groups (NCAPG), 2011).
5. Trynka, G. et al. Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. Am. J. Hum. Genet. 97, 139–152 (2015).
6. Nicolae, D.L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 6, e1000888 (2010).
7. Trynka, G. et al. Chromatic marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. 45, 124–130 (2013).
8. Farh, K.K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015).
9. 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).
10. Finucane, H.K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47, 1228–1235 (2015).
11. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. 14, 483–495 (2013).
12. Nica, A.C. & Dermitzakis, E.T. Using gene expression to investigate the genetic basis of complex traits. Nature 497, 138–147 (2013).
13. Nica, D.L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 6, e1000888 (2010).
14. Guo, H. et al. Candidate causal regulatory effects by integration of expression QTL and GWAS. Hum. Mol. Genet. 24, 3305–3313 (2015).
15. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. 48, 481–487 (2016).
16. 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).
17. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet. 99, 1245–1260 (2016).
18. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511 (2013).
19. Raj, T. et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science 344, 519–523 (2014).
20. Nica, A.C. et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 7, e1002003 (2011).
21. Ferretti, L.M. et al. Statistical colocalization of genetic-risk variants for related autoimmune diseases in the context of common controls. Nat. Genet. 47, 839–846 (2015).
22. Malik, R. et al. Shared genetic basis for migraine and ischemic stroke: a genome-wide analysis of common variants. Neurology 84, 2132–2145 (2015).
23. Winship, B.S. et al. Genetic analysis for a shared biological basis between migraine and coronary artery disease. Neuroil Genet. 1, e10 (2015).
24. Pickrell, J.K. et al. Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. 48, 709–717 (2016).
25. Cotsapas, C. et al. Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet. 7, e1002554 (2011).
26. Nica, A.C. et al. Candidate causal regulatory effects by integration of QTLs with complex trait genetic associations. PLoS Genet. 6, e1000895 (2010).
27. 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).
28. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47, 1236–1241 (2015).
29. Jodis, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124 (2012).
online methods
Simulated data set. We randomly sampled 97 genomic loci of length 200 kb across the genome to base our simulations. We excluded subtelomeric/cen tromeric regions, sex chromosomes and regions of sparse genetic-map coverage. In each locus, we simulated disease (20,000 cases and 20,000 controls each) and eQTL (250 individuals each) cohorts by using HapGen2 (ref. 30) and phased haplotypes from the CEU population (2n = 198) of the 1000 Genomes Project\textsuperscript{11}. For disease cohorts, we set the variant nearest the center of the interval as causal, with an odds ratio (OR) of 1.1 for each minor-allele copy, and simulated five replicate cohorts of cases and controls in each locus. In each locus, we then simulated three different genetic models for the eQTL cohort: no causal variant (H,\textsubscript{0}), the same causal variant as disease (H,\textsubscript{1}) and distinct causal variants between disease and eQTL (H,\textsubscript{2}). For H,\textsubscript{2}, we selected eQTL causal variants within 50 kb of the disease-causing variant and with differing levels of LD between the two causal variants (\(r^2\) of 0–0.4, 0.4–0.5, 0.5–0.6, 0.6–0.7, 0.7–0.8 or 0.8–0.9 in CEU). We selected all disease and eQTL causal variants to have a minor-allele frequency (MAF) >10% in CEU. We generated genotypes for the eQTL cohorts with HapGen2 with a null effect size, then simulated a quantitative phenotype with the allelic-mean-difference model implemented in GCTA\textsuperscript{23} with effect sizes of 0.05, 0.1 or 0.2 in cis-heritability (\(h^2\)). In each locus, we generated five replicate eQTL cohorts for H,\textsubscript{0} and H,\textsubscript{1} each; for H,\textsubscript{2}, we generated a single cohort with up to five distinct causal variants per locus.
We used PLINK (URLs) to calculate the genetic association with disease and expression phenotypes in logistic and linear-regression models, respectively, after filtering out SNPs with MAF <5% in each cohort. We rejected cohorts showing weak maximum association signals (association \(P > 10^{-5}\) for disease cohorts and association \(P > 0.01\) for eQTL cohorts). In addition, as expected from the coalescent forward-simulation model on which HapGen2 is based, a fraction of our simulated cohorts showed maximal association with a SNP in low LD with the causal variant that we had specified (\(r^2 < 0.8\) measured in sample).
We kept these cohorts as gene-expression traits only, to better capture the vagaries of resolution limits inherent in the small sample sizes of eQTL studies, but we excluded the disease cohorts. Overall, we rejected 20% of disease cohorts and 11% of eQTL cohorts and generated a total of 5,680, 829 and 4,666 disease-eQTL comparisons under H,\textsubscript{0}, H,\textsubscript{1} and H,\textsubscript{2}, respectively.
To test the effect of mild population mismatch, we also generated a second set of eQTL cohorts, this time using base haplotypes of all non-Finnish Europeans from the 1000 Genomes Project (CEU \& GBR \+ TSI \+ IBS, \(2n = 808\)).
To explore the scenarios in which multiple independent causal variants in a locus affect the same phenotype, we generated another set of simulated disease-eQTL pairs, assuming two causal variants for disease, eQTL or both traits in the genetic background of 48 genomic loci. We set the disease-causing variant at the center of the locus as the reference SNP and added a second causal variant, varying the LD with the reference SNP (\(r^2\) bins of 0–0.4, 0.4–0.5, 0.5–0.6, 0.6–0.7, 0.7–0.8 or 0.8–0.9). For disease cohorts, we set the OR of the central risk variant to 1.1, as in the original simulations, and we set the OR of the second causal variant to 1.1, 1.049 or 1.024 (i.e., 1.0, 0.5\times or 0.1x the effect size of the central variant on a log scale). For expression phenotypes, the total \(r^2\) of 0.1 was split between two independent causal variants. The relative effect size of the second causal variant was scaled to 1.0, 0.5\times or 0.1x relative to the central causal SNP, without standardizing the genotypes. For each combination of causal variants, we generated two replicate cohorts for disease and a single cohort for eQTL. Again, we rejected cohorts in which we observed the strongest association at a variant in low LD with the specified causal variants (\(r^2 < 0.8\)).
Disease GWAS data set. We downloaded association summary statistics for type 1 diabetes (T1D), rheumatoid arthritis (RA), celiac disease (CEL), multiple sclerosis (MS), inflammatory bowel disease (IBD), Crohn disease (Crohn) and ulcerative colitis (UC) from ImmunoBase (URLs; Supplementary Table 1). For MS, we used the association statistics derived from the combined cohort of discovery and validation samples\textsuperscript{31} to maximize the sample size and genetic resolution. For IBD, Crohn, and UC, summary data were from the European subset of the most recently available transethnic association study\textsuperscript{29}. All association data were based on only ImmunoChip samples and did not include imputed genotypes. To address population structure, we limited our analyses to European subjects, except in the case of RA, which included 620 Punjabi individuals out of a total of 27,345. T1D summary statistics are from the meta-analysis between case–control association and affected sibling-pair analysis.
Because our method works best on dense genotype data, we restricted our analyses to the 188 loci genotyped at high density on ImmunoChips. We excluded the major histocompatibility complex locus, owing to the complex landscape of selection and the resulting complex LD patterns. For each disease, we identified the largest published genetic mapping study and identified genome-wide significant associations reported in the 188 ImmunoChip loci. However, these reports may have contained additional samples, so the associations may not have been significant genome-wide in the ImmunoChip studies alone. We also excluded any secondary associations after conditioning on initial results, because these results were inconsistently reported across diseases. If multiple independent associations were reported within the same ImmunoChip region for any disease, we divided the region at the midpoint between the reported markers and selected lead SNPs in each subinterval separately.

eQTL data set. We examined eQTLs in LCLs and primary CD4\textsuperscript{+} T cells and CD14\textsuperscript{+} monocytes obtained from healthy donors\textsuperscript{18,19} (Supplementary Table 1). For LCLs, we obtained imputed genotype and normalized RNA-seq data in RPKM for 278 non-Finnish European donors in the Groudens project. We removed SNPs with a minor-allele frequency <5%, a high probability of Hardy–Weinberg disequilibrium (\(P_{\text{HW}} < 10^{-3}\)) or a 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\textsuperscript{34,35}. For CD4\textsuperscript{+} and CD14\textsuperscript{+}, we regressed normalized expression levels for European Americans (\(n = 213\) and 211, respectively) on similarly quality-controlled imputed allele dosages. For all cell types, we generated adaptive permutation statistics from \(10^3\) to \(10^6\) iterations, using all covariates\textsuperscript{34}.

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 with distinct causal variants. Owing 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, \(\theta\), 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 samples of two-trait association were not overlapping. We designed the joint likelihood mapping (JLIM) statistic, \(A\), in an asymmetrical fashion, requiring only summary-level statistics for one trait (primary trait) but genotype-level data for the other (secondary trait). Specifically, \(A\) was defined as the sum of the log likelihood that the causal variant underlying a secondary trait is more likely to be same as, rather than distinct from, the variant underlying primary trait, as integrated over a set of likely causal variants under a GWAS peak of a primary trait:

\[
A = \sum_{i \in N_{q1}^d | m^*} \frac{L_i(1) \log L_i(1)}{\max_{j \in N_{q1}^d(0)} L_j(1)} \frac{L_i(0) L_j(0)}{L_i(0) L_j(0)} = \sum_{i \in N_{q1}^d | m^*} L_i(1) \times \frac{\log L_2(j)}{\max_{j \in N_{q1}^d(0)} \log L_2(j)}
\]

where \(m^*\) is the most associated SNP for a primary trait, \(L_i(1)\) and \(L_i(0)\) are the likelihood of SNP \(i\) being causally associated with primary and secondary traits, respectively, and \(N_{q1}^d\) and \(N_{q1}^d(0)\) are the sets of SNPs within the LD neighborhood around SNP \(i\), as defined by \(\{SNP_j | q_1^2 > \theta\}\). We derived \(N_{q1}^d\) from the reference LD panel and \(N_{q1}^d(0)\) directly from the genotypes of the secondary-trait cohort. We used disease outcome as primary trait, leveraging the larger sample size and dense genotyping, and gene expression as a secondary trait, taking advantage of the availability of individual genotype data.
The likelihood of causal association was calculated by approximating the local LD structure with pairwise correlation, similarly to methodology described in Kichaev et al. and Hormozdiari et al.

Briefly, when SNP \( c \) is the only causal variant in the locus with noncentrality \( \lambda_c \), the association statistic \( z_i \) of noncausal SNP \( i \) follows a normal distribution \( N(\lambda_i, \lambda_c, 1) \), where \( n_c \), is the LD between SNPs \( i \) and \( c \), measured in pairwise Pearson correlation of genotypes. In general, when association statistics \( Z = (z_1, z_2, \ldots, z_N)^T \) are provided for all \( M \) SNPs in the analysis window, the likelihood of SNP \( i \) being the causal variant with noncentrality \( \lambda_i \) is:

\[
L(Z; \lambda_i = 0) = \phi_{MVN}(Z; \text{mean} = \Sigma(\lambda_i^C), \text{var} = \Sigma)
\]

where \( \phi_{MVN} \) is the multivariate normal density function, \( C \) is an incident vector with \( C_q = 1 \) if only if \( k = i, \Sigma \) is a \( M \times M \) local LD matrix defined by pairwise Pearson correlation between genotypes, and \( ^T \) is element-wise multiplication. Because the true noncentrality of the causal variant was unknown, we estimated the profile likelihood, which simplifies to a closed form:

\[
\log L(Z; \lambda_i^{MLE}) = -\frac{1}{2}(Z^T \Sigma^{-1}Z + z_i^2) - \frac{1}{2} \log((2\pi)^M | \Sigma|)
\]

with \( \lambda_i^{MLE} = z_i \). Thus, given association statistics for primary and secondary traits, \( Z = (z_1, z_2, \ldots, z_M)^T \) and \( W = (w_1, w_2, \ldots, w_M)^T \), the test statistic, \( A \), simplifies to:

\[
A = \sum_{i=1}^{M} \frac{e^{2(z_i^2 - z_i^{2,\text{MLE}})}}{\max_{j \in \mathbb{N}_0^M} w_j^2}
\]

The \( \Lambda \) value of joint likelihood was estimated by permuting phenotypes of secondary traits as under the trivial null hypothesis that there is no causal variant for a secondary trait in the locus \( (H_0) \). With respect to the more likely null hypothesis that distinct causal variants underlie association signals of two traits \( (H_1) \), we can show that, asymptotically, as the noncentrality of a causal variant increases, \( P \) values estimated from \( H_0 \) behave conservatively with respect to \( H_2 \) (Supplementary Note):

\[
P_{\text{JLIM}} = P(\Lambda \geq |H_0|) \geq P(\Lambda \geq |H_2|)
\]

Thus, with sufficiently large sample or effect sizes, joint likelihood tests against \( H_0 \) will also reject \( H_2 \) in favor of the alternative hypothesis of the shared causal variant \( (H_1) \). Further, to evaluate whether this property holds for practical noncentrality values, we examined our negative controls simulating \( H_2 \), specifically, whether \( P_{\text{JLIM}} \) was highly shifted toward 1.0 (Supplementary Fig. 2) and whether it was similar to or larger than empirically estimated false-positive rates, as expected (Supplementary Table 3; \( P < 0.05 \)).

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 100 kb of the marker most associated with disease (i.e., the lead SNP). In ImmunoChip data, the analysis windows were further confined on the basis of the boundaries of the dense genotyping intervals. We compared each lead SNP with eQTL data for all genes with transcription start sites up to 1 Mb from the lead SNP, and an eQTL association \( P < 0.01 \) for at least one SNP in the analysis window. To minimize computational burden, we did not consider SNPs associated with either disease or eQTL (association \( P > 0.1 \) for both). For the reference LD panel, we used the base haplotypes of HapGen simulation for simulated data sets, and we used 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 on the basis of FDR levels and Bonferroni correction. The FDR was calculated separately for specific disease and cell-type combinations as:

\[
\text{FDR}(p) = \frac{pN}{#(P_{\text{JLIM}} \leq p)}
\]

where \( p \) is a JLIM \( P \)-value cutoff, and \( N \) is the number of all tested disease lead SNP–eQTL candidate-gene combinations. The FDR was calculated for each cell type, because the distribution of JLIM \( P \)-values varied depending on the disease relevance of cell type. To provide a list of higher-confidence hits in each disease, we also applied Bonferroni correction to nominal JLIM \( P \)-values for the number of tests across all three cell types.

Benchmark comparison. We used our simulations to compare the performance of our method (here abbreviated as JLIM, for joint likelihood mapping) with three existing methods: Bayesian colocal, gwas-pw and SMR/HEIDI. We ran colocal (version 2.3.1) with default parameter settings with the colocalization prior \( p_{\text{HEIDI}} < 0.05 \). We followed the authors’ recommendation to use beta and variance of beta for the case–control cohorts as summary statistics. For quantitative trait cohorts, we also provided in-sample minor-allele frequencies. We applied gwas-pw (version 0.21) with default parameters. All simulated disease–eQTL pairs were combined into a single batch and analyzed together so that gwas-pw could optimize the model parameters. We ran SMR/HEIDI (version 0.64) with default parameters, except that the \( P \)-value threshold to select the top associated eQTL (pseud-smr, default \( 5 \times 10^{-8} \)) was relaxed to 0.01 to enable use of the test on simulated disease–eQTL pairs with weak eQTL association. Tests producing significant heterogeneity by HEIDI (\( p_{\text{HEIDI}} < 0.05 \)) were called negative regardless of \( \text{PMEM} \) values, because they were likely to contain distinct causal variants between disease and eQTL. For colocal and gwas-pw, predictions were made on the basis of the reported posterior probability of colocalization (PP4 and PP3, respectively) although posteriors have been reported for other competing models. For overall performance comparison, we evaluated the area under the receiver operator curve (ROC).

Bayesian colocal on real data. Because ImmunoChip data are available only as summary statistics, we ran colocal with the \( P \) values of association for disease cohorts, and with quantitative beta and variance of beta calculated for eQTL association data cohorts. We also provided to coloc the minor-allele frequencies of non-Finnish Europeans from the 1000 Genomes Project. The colocalization prior \( p_{\text{JLIM}} \) was set to \( 10^{-6} \), and the prediction was made at PP4 \( \geq 0.75 \) for higher confidence (Supplementary Table 4). We did not consider the type 1 diabetes data, in which the case–control sample size was limited after exclusion of affected sib-pair data.

Estimating the number of disease GWAS loci with consistent eQTL effects. We expected JLIM \( P \)-values to follow a bimodal distribution with modes close to zero and one when the data supported a model of shared or distinct causal effects, respectively. In contrast, under the null model of no cis-eQTL association, we expected a uniform \( P \)-value distribution. We thus estimated the proportion of disease–eQTL pairs belonging to the null \( \pi_0 \), same \( \pi_1 \) and distinct \( \pi_2 \) causal-variant models from the observed \( P \)-value distribution (Supplementary Figs. 16–18). To assess whether the strength of the eQTL association influenced the likelihood of identifying a shared causal variant, we calculated these proportions for subsets of trait pairs defined by a minimum eQTL \( P \)-value. In each bin, we identified the limits of the uniform portion of the distribution \( \gamma_1 \) and \( \gamma_2 \) and estimated \( \pi_0 \), \( \pi_1 \) and \( \pi_2 \) as:

\[
\pi_0 = \frac{#(\gamma_1 \leq P_{\text{JLIM}} \leq \gamma_2)}{(\gamma_2 - \gamma_1)N}
\]

\[
\pi_1 = \frac{#(P_{\text{JLIM}} \leq \gamma_1)}{N} - \gamma_1 \pi_0
\]

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

To estimate the number of disease GWAS loci that could be explained by a consistent effect of the same causal variant on disease and eQTL (denoted by \( \mathcal{C} \) below), we incrementally relaxed the \( P \)-value cutoffs 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 (Fig. 1 and Supplementary Fig. 19). Specifically, at each JLIM P-value cutoff $p_0$, we successively calculated $\mathcal{E}(p_1)$:

$$\mathcal{E}(p_1) = \mathcal{E}(p_{l-1}) + |D(p_1) \cap D(p_{l-1})| - \sum_{d \in D(p_{l-1})} \epsilon(d, p_1) + \sum_{d \in D(p_{l-1})} \epsilon(d, p_{l-1})$$

where $p_{l-1} < p_l$ with $p_0 = 0$. $D(p)$ is the set of disease GWAS loci with at least one eQTL gene in any cell type passing the JLIM P-value cutoff $p$; and $\epsilon(d, p)$ is the probability that the 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 $\epsilon(d, p)$ by using the Monte Carlo method by randomly selecting false-positive eQTL genes within the locus $d$ at rates of $(1 - \pi_1) \times lb$ or $(1 - \pi_1) \times ub$ over 1,000 iterations. The $lb$ and $ub$ are the lower and upper bounds of the false-positive rate of JLIM against the true null. Notably, $\pi_1$ and $lb$ depend on the cell type and the strength of the eQTL association.

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

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

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

$$ub = P_{\text{JLIM}}$$

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

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

**Code availability.** The current implementation of JLIM is available from the laboratories of C.C. and S.R.S. at http://www.github.com/cotsapaslab/jlim/ and http://genetics.bwh.harvard.edu/wiki/sunyaevlab/jlim/, respectively.

**Data availability.** The publicly available 1000 Genomes genotype data were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. The publicly available eUVADIS LCL eQTL data were accessed via European Bioinformatics Institute ArrayExpress accession number E-GEUV-1. Gene-expression data for CD4$^+$ T cell and CD14$^+$ monocytes were accessed via National Center for Biotechnology Information Gene Expression Omnibus accession number GSE56035. ImmunoChip GWAS summary statistics are available at http://www.immunobase.org/.

30. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).
31. Abecasis, G.R. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
32. 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).
33. International Multiple Sclerosis Genetics Consortium (IMSGC). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
34. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
35. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
36. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
37. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
38. Chen, X. et al. Dominant genetic variation and missing heritability for human complex traits: insights from twin versus genome-wide common SNP models. *Am. J. Hum. Genet.* **97**, 708–714 (2015).
39. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).