HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2018 December 25.

Published in final edited form as:

Nat Genet. 2018 July ; 50(7): 1041–1047. doi:10.1038/s41588-018-0148-2.

Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits

Farhad Hormozdiari[1,2], **Steven Gazal**[1,2], **Bryce van de Geijn**[1,2], **Hilary Finucane**[1,2], **Chelsea J.-T. Ju**[3], **Po-Ru Loh**[2,4], **Armin Schoech**[1,2], **Yakir Reshef**[1,2], **Xuanyao Liu**[1,2], **Luke O'Connor**[1,5], **Alexander Gusev**[4,6], **Eleazar Eskin**[3,7], and **Alkes L. Price**[1,2]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[3]Department of Computer Science, University of California, Los Angeles, California, USA

[4]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

[5]Program in Bioinformatics and Integrative Genomics, Harvard Graduate School of Arts and Sciences, Boston, Massachusetts, USA

[6]Dana Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA

[7]Department of Human Genetics, University of California, Los Angeles, California, USA

Abstract

There is increasing evidence that many GWAS risk loci are molecular QTL (eQTL, hQTL, sQTL, and/or meQTL). Here, we introduce a new set of functional annotations based on causal posterior probabilities of fine-mapped molecular cis-QTL, using data from the GTEx and BLUEPRINT consortia. We show that these annotations are far more strongly enriched for heritability (e.g. 5.84x for eQTL; $P=1.19\times10^{-31}$) across 41 independent diseases and complex traits than annotations containing all significant molecular QTL (1.80x for eQTL). eQTL annotations that were obtained by meta-analyzing all GTEx tissues generally performed best, but tissue-specific eQTL annotations produced stronger enrichments for blood- and brain-related diseases and traits. Notably, eQTL annotations restricted to loss-of-function intolerant genes from ExAC were even more strongly enriched for heritability (17.06x; $P=1.20\times10^{-35}$). All molecular QTL except sQTL remained significantly enriched in a joint analysis, implying that each of these annotations is uniquely informative for disease and complex trait architectures.

Correspondence should be addressed to F.H. (Hormozdiari@hsph.harvard.edu) and A.L.P (aprice@hsph.harvard.edu).

Contributions

F.H. and A.L.P. designed experiments. F.H. performed experiments. F.H., S.G, B.V.G., H.K.F., C.J.J, P.-R.L., A.S., Y.K., X.L, L.O., A.G., E.E., analyzed data. F.H. and A.L.P. wrote the manuscript with assistance from all authors.

Competing interests

The authors declare no competing interests.

Introduction

Although Genome-wide association studies (GWAS) have been extremely successful in detecting thousands of risk loci for diseases and traits[1,2,3], our understanding of disease architecture is far from complete as most risk loci lie in non-coding regions of the genome[4,5,6,7,8,9]. Leveraging molecular phenotypes such as gene expression[10,11,12,13,14] or chromatin marks[15,16,17,18] can aid in understanding the disease architecture: in particular, previous studies have shown that cis-eQTL are enriched in GWAS loci as well as genome-wide heritability of several diseases[5,6,19,20], motivating further work on colocalization[21–23] and transcriptome-wide association studies[24–26]. Partitioning heritability using raw genotypes/phenotypes[27–31] or summary association statistics[32–34] can aid our understanding of disease architectures, but it is currently unclear how to best leverage molecular QTL from rich resources such as GTEx[12,14] and BLUEPRINT[18] using these methods.

Here, we introduce a new set of annotations constructed from eQTL, hQTL, sQTL, and meQTL data that are very strongly enriched for disease heritability across 41 independent diseases and complex traits. We construct these annotations by applying a fine-mapping method[35] (allowing for multiple causal variants at a locus) to compute causal posterior probabilities for each variant to be a causal cis-QTL. We show that our annotations are far more enriched for disease heritability than standard annotations. We further show that our eQTL annotations produce tissue-specific enrichments (despite high cis-genetic correlations of eQTL effect sizes across tissues[12,36], and produce much larger enrichments when restricted to loss-of-function intolerant genes from ExAC[37]. Finally, we quantify the extent to which annotations constructed from eQTL, hQTL, sQTL, and meQTL provide complementary information about disease.

Results

Overview of Methods

Our goal is to construct molecular QTL-based annotations that are maximally enriched for disease heritability. For a given molecular QTL data set, we construct a probabilistic (continuous-valued) annotation as follows. First, for each molecular phenotype (e.g. each gene) with at least one significant (FDR < 5%) cis-QTL (e.g. 1Mb from TSS), we compute the causal posterior probability (CPP) of each cis SNP in the fine-mapped 95% credible set, using our CAVIAR fine-mapping method[35] (see URLs). Then, for each SNP in the genome, we assign an annotation value based on the maximum value of CPP across all molecular phenotypes; SNPs that do not belong to any 95% credible set are assigned an annotation value of 0. We refer to this annotation as MaxCPP. For comparison purposes, we also construct three other molecular QTL-based annotations. First, we construct a binary annotation containing all SNPs that are a significant (FDR < 5%) cis-QTL for at least one molecular phenotype[19,20]; we refer to this annotation as AllcisQTL. Second, we construct a binary annotation containing all SNPs that belong to the 95% credible set (see above) for at least one molecular phenotype; we refer to this annotation as 95%CredibleSet. Third, we construct a binary annotation containing the most significant SNP for each molecular

phenotype with at least one significant (FDR < 5%) QTL. We refer to this annotation as TopcisQTL (see Online Methods).

We apply a previously developed method, stratified LD score regression (S-LDSC)[32,33], to partition disease heritability using functional annotations. We use two metrics to quantify an annotation's contribution to disease heritability: enrichment and standardized effect size ($\tau*$). Enrichment is defined as the proportion of heritability explained by SNPs in an annotation divided by the proportion of SNPs in the annotation[32]; here, we generalize this definition to probabilistic annotations such as MaxCPP. Standardized effect size ($\tau*$) is defined as the proportionate change in per-SNP heritability associated to a one standard deviation increase in the value of the annotation, conditional on other annotations included in the model[33]. Unlike enrichment, $\tau*$ quantifies effects that are unique to the focal annotation (see Online Methods).

We constructed MaxCPP and other annotations using eQTL data from the GTEx Consortium[12,14] and eQTL, hQTL, sQTL and meQTL data from the BLUEPRINT Consortium[18] (Table 1; see URLs). We included a broad set of 75 functional annotations from the baselineLD model (Supplementary Table 1) in most analyses. We have made our annotations and partitioned LD scores freely publicly available (see URLs).

Simulations

We performed a comprehensive set of simulations to assess whether S-LDSC produces unbiased estimates of an annotation's contribution to disease heritability for the AllcisQTL, 95%CredibleSet, TopcisQTL and MaxCPP annotations. We performed simulations using real genotypes from the UK Biobank, restricting to 749,024 SNPs on chromosome 1 (see Online Methods). In our main simulation, we simulated gene expression phenotypes for 500 individuals assuming that 10% of cis-variants for a gene are causal cis-eQTL (heritability=16%), simulated complex trait phenotypes for an independent set of 40,000 individuals assuming that the set of causal variants is exactly the set of causal eQTLs, with independent effect sizes (heritability=20%), and subsequently assumed that 10% of causal eQTL are missing from the data analyzed. We also performed secondary simulations under other genetic architectures and assumptions about missing data (see below). We estimated each annotation's contribution to complex trait heritability using S-LDSC. We performed 400 independent simulations, and averaged results across simulations.

Enrichment estimates and true enrichments for each annotation are displayed in Figure 1 and Supplementary Table 2. We determined that S-LDSC estimates are severely upward biased for the TopcisQTL annotation On the other hand, S-LDSC estimates were slightly conservative for the AllcisQTL, 95%CredibleSet and MaxCPP annotations. Thus, we restrict our analyses to these three annotations in our analyses of real phenotypes below Of these three annotations, MaxCPP had the highest enrichment (Figure 1 and Supplementary Table 2). For comparison purposes, we also computed estimates using GCTA[27,28] (see URLs), a method that has previously been applied to assess eQTL enrichment for complex traits[19,20]; we computed GCTA estimates for all annotations except MaxCPP, as GCTA is only applicable to binary annotations. We determined that GCTA estimates generally exhibited

greater bias than S-LDSC estimates (Figure 1 and Supplementary Tables 2-3). We obtained similar results for both S-LDSC and GCTA at other simulation parameters (Supplementary Table 3). We also obtained similar results for both S-LDSC and GCTA using an alternative simulation framework, drawn from our previous work[25], that directly uses simulated gene expression to generate complex trait phenotypes (see Online Methods and Supplementary Table 4).

All functional enrichment methods (GCTA[27,28], BOLT-REML[30], S-LDSC[32,33], and LDAK[31]) that we are currently aware of assume that causal disease effect sizes are independent and identically distributed (i.i.d) conditional on MAF, LD and function annotation values. However, this assumption may be violated for molecular QTL-based annotations when causal variants are sparse (see Online Methods for an example); in particular, this is a limitation of our S-LDSC method. Indeed, our simulations confirm these biases (Figure 1 and Supplementary Tables 2-4). We note that other functional enrichment methods are also subject to this limitation. Specifically, GCTA[27,28] is shown by our simulations to exhibit greater biases than S-LDSC (Figure 1 and Supplementary Tables 2-4). BOLT-REML[30] is a computationally efficient method that produces the same results as GCTA[27,28], and LDAK[31] has been shown in separate work to produce biased estimates in much simpler settings[38]. S-LDSC produces slightly conservative estimates across a comprehensive set of simulations for the AllcisQTL, 95%CredibleSet and MaxCPP annotations that we consider in our analyses of real phenotypes below.

Fine-mapped eQTL are enriched for disease heritability

We used the GTEx eQTL data set (Table 1) to construct AllcisQTL, 95%CredibleSet and MaxCPP annotations. We constructed annotations using each of the 44 tissues (Supplementary Table 5). We applied S-LDSC to assess each annotation's contribution to disease heritability for each of 41 independent diseases and complex trait data sets (average N=320K); for six traits we analyzed two different data sets, leading to a total of 47 data sets analyzed (see Supplementary Table 6). We meta-analyzed results across the 47 data sets, which were chosen to be independent (see Online Methods). We computed enrichment and $\tau*$ for each annotation, in analyses that included 75 functional annotations (Supplementary Table 1) from the baselineLD model[33]. Results for whole blood, a widely studied tissue, are reported in Figure 2 and Supplementary Table 7. We determined that the MaxCPP annotation had far higher values of enrichment and $\tau*$ than AllcisQTL or 95%CredibleSet; the enrichment estimates remain much higher for MaxCPP even after accounting for the fact that S-LDSC generally produces more conservative estimates for AllcisQTL and 95%CredibleSet than for MaxCPP in our simulations (1.10-1.58x and 0.76-1.23x more conservative respectively, across all simulations; Figure 2 and Supplementary Table 7). The MaxCPP annotation also had higher values of enrichment and $\tau*$ in analyses that did not condition on the baselineLD model (Supplementary Figure 1 and Supplementary Table 8).

We investigated whether the $\tau*$ for MaxCPP in each respective tissue varied with sample size. We observed a correlation (R^2 = 0.69, P = 1.36e-12) between sample size and $\tau*$ (Figure 3 and Supplementary Table 9). We observed a similar pattern in simulations (Supplementary Figure 2). This suggests that the correlation between sample size and $\tau*$ in

GTEx data is related to statistical power and not because tissue-specific eQTL from tissues with larger sample size are more relevant for the 41 traits analyzed. Thus, annotations constructed from tissues with larger sample sizes are more informative for disease/trait architectures.

To maximize sample size, we performed a fixed-effect meta-analysis of eQTL effect sizes across the 44 tissues (FE-Meta-Tissue)(see Online Methods). We determined that FE-Meta-Tissue annotations had slightly larger enrichments and much larger $\tau*$ (due to larger annotation size) than annotations constructed from individual tissues (Figure 2, Supplementary Table 7 and Supplementary Table 10). As with individual tissues, the MaxCPP annotation had far higher values of enrichment (5.84x, s.e. = 0.40; P = 1.19e-31) and $\tau*$ ($\tau*$ = 0.52, s.e. = 0.05; P = 2.73e-27) than AllCisQTL (enrichment = 1.80x and $\tau*$ = 0.03) or 95%CredibleSet (enrichment = 2.75x and $\tau*$ = 0.17). It is particularly notable that the $\tau*$ for GTEx FE-Meta-Tissue MaxCPP (conditional on the baselineLD model) is much larger than the $\tau*$ values of the 6 continuous annotations that we introduced in our previous work[33] (Supplementary Table 11). A histogram of MaxCPP annotation values for GTEx FE-Meta-Tissue is provided in Supplementary Figure 3, and correlations with baselineLD model annotations and their LD scores are provided in Supplementary Figures 4 and 5.

Tissue-specific fine-mapped eQTL enriched for heritability

Although the FE-Meta-Tissue MaxCPP annotation outperformed each of the 44 tissue-specific MaxCPP annotations in the meta-analysis across 41 traits (Figure 2 and Figure 3), this was not the case for every trait. We examined six autoimmune diseases, five blood cell traits, and eight brain-related diseases and traits in detail (see Online Methods). We first analyzed the six autoimmune diseases, analyzing MaxCPP annotations for blood and FE-Meta-Tissue separately (conditional on the baselineLD model) and meta-analyzing results across the six diseases. We obtained higher estimates of $\tau*$ (and higher or comparable estimates of enrichment) for blood than for FE-Meta-Tissue or any other individual tissue (Supplementary Table 12). We then analyzed MaxCPP annotations for blood and FE-Meta-Tissue jointly (conditional on the baselineLD model). We obtained a significantly positive $\tau*$ estimate for blood ($\tau*$ = 0.91, s.e. = 0.34; P = 9.15e-03) (Figure 4 and Supplementary Table 13), implying that fine-mapped blood eQTL provides additional information about these six diseases conditional on fine-mapped FE-Meta-Tissue eQTL. We repeated these analyses for the five blood cell traits. When analyzing MaxCPP annotations for blood and FE-Meta-Tissue separately, we obtained higher estimates of $\tau*$ (and higher or comparable estimates of enrichment) for blood than for FE-Meta-Tissue or any other individual tissue (Supplementary Table 14). When analyzing MaxCPP annotations for blood and FE-Meta-Tissue jointly, we obtained a significantly positive $\tau*$ estimate for blood ($\tau*$ = 1.17, s.e. = 0.24; P = 1.77e-06) (Figure 4 and Supplementary Table 13), implying that fine-mapped blood eQTL provides additional information about these five traits conditional on fine-mapped FE-Meta-Tissue eQTL. The correlations of blood MaxCPP annotation values with baselineLD model annotations and their LD scores are provided in Supplementary Figures 6 and 7.

We then analyzed the eight brain-related diseases and traits. We performed a fixed-effect meta-analysis of eQTL effect sizes across the 10 brain tissues and one nerve tissue (Brain +Nerve). When analyzing MaxCPP annotations for Brain+Nerve and FE-Meta-Tissue separately, we obtained higher or comparable estimates of enrichment and $\tau*$ for Brain +Nerve than for FE-Meta-Tissue or any individual tissue (Supplementary Table 15). When analyzing MaxCPP annotations for Brain+Nerve and FE-Meta-Tissue jointly, we obtained a significantly positive $\tau*$ estimate for Brain+Nerve ($\tau*$ = 0.28, s.e. = 0.07; P = 9.81e-05) (Figure 4 and Supplementary Table 13), implying that fine-mapped Brain+Nerve eQTL provides additional information about these eight traits conditional on fine-mapped FE-Meta-Tissue eQTL. The correlations of Brain+Nerve MaxCPP annotation values with baselineLD model annotations and their LD scores are provided in Supplementary Figures 6 and 7. We repeated these analyses for each trait and each tissue separately, but determined that only three blood cell traits (white blood count, red blood cell distribution width, and eosinophil count traits), in conjunction with MaxCPP for blood, attained a significantly positive (FDR <5%) tissue-specific $\tau*$ (Supplementary Tables 16 and 17). Overall, these results demonstrate that tissue-specific eQTL effects on steady-state expression can be significant for diseases and complex traits, despite the well-documented high cis-genetic correlations of eQTL effect sizes across tissues[12,36].

MaxCPP signal is concentrated in disease-relevant gene sets

Recent studies have identified gene sets that are depleted for coding variants and enriched for *de novo* coding mutations impacting disease[37,39,40]. To investigate the importance of non-coding common variants in these gene sets, for each gene set S we used GTEx FE-Meta-Tissue to construct a new annotation MaxCPP(S), defined as the maximum CPP restricted to genes in S. For comparison purposes, we also constructed an annotation allSNP(S), defined as the set of all SNPs within 100Kb of genes in S.

We first analyzed the ExAC gene set, consisting of 3,230 genes that are strongly depleted for protein-truncating variants[37]. We determined that MaxCPP(ExAC) was very strongly enriched in an analysis conditional on the baselineLD model, meta-analyzed across 41 independent traits (see Figure 5a and Supplementary Table 18). In particular, MaxCPP(ExAC) was much more strongly enriched (17.06x, s.e. = 1.28; P = 1.20e-35) than MaxCPP(All Genes) (5.84x) (P = 4.90e-17 for difference). This implies that eQTL for these 3,230 genes have a disproportionately strong impact on disease heritability, consistent with the fact that these genes are depleted for eQTL[37]. We then analyzed MaxCPP(ExAC) and MaxCPP(All Genes) annotations jointly (conditional on the baselineLD model). We obtained a significantly positive $\tau*$ for MaxCPP(ExAC) ($\tau*$ = 0.41, s.e. = 0.04; P = 1.40e-23; Figure 5b and Supplementary Table 19), implying that MaxCPP(ExAC) provides additional information about disease heritability conditional on MaxCPP (All Genes). We observed that the effect size ($\tau*$) for MaxCPP(ExAC) conditional on MaxCPP (All Genes) and baselineLD is five times larger, and more statistically significant, than the $\tau*$ of allSNP (ExAC) conditional on the baselineLD model (Supplementary Table 20). Thus, MaxCPP can increase power to identify enriched gene sets.

We analyzed four additional gene sets S: a set of 1,003 genes that are strongly depleted for missense mutations[40] (Samocha); a set of 2,984 genes with strong selection against protein-truncating variants[39] (Cassa); a set of 1,878 genes predicted to be essential based on CRISPR experiments in a human cancer cell line[41] (Wang); and a set of 11,983 genes with evidence of allelic heterogeneity in analyses of GTEx gene expression data using our previously developed methods (AH)[42]. For each of these gene sets, MaxCPP(S) was strongly enriched in analyses conditional on the baselineLD model, meta-analyzed across 41 independent traits (Figure 5a and Supplementary Table 18). In addition, for each gene set except the Wang gene set, we obtained a significantly positive $\tau*$ for MaxCPP(S) (after correcting for five gene sets tested) when analyzing MaxCPP(S) and MaxCPP(All Genes) jointly (conditional on the baselineLD model) (Figure 5b, Supplementary Table 19 and Supplementary Tables 21-25). As with the ExAC gene set, the $\tau*$ for MaxCPP(S) conditional on MaxCPP(All Genes) and the baselineLD model were substantially larger than the $\tau*$ of allSNPs (ExAC) conditional on the baselineLD model and were often more statistically significant (Supplementary Table 20), indicating that MaxCPP can increase power to identify enriched gene sets in which regulatory variants play an important role.

Fine-mapped molecular QTL are enriched for heritability

We analyzed five molecular QTL from the BLUEPRINT data set (Table 1), including eQTL, hQTL (H3K27ac and H3K4me1), sQTL and meQTL. In each case, we constructed AllcisQTL, 95%CredibleSet and MaxCPP annotations using each of the three immune cell types (CD14+ monocytes, CD16+ neutrophils, and naive CD4+ T cells; Supplementary Table 26) as well as a fixed-effect meta-analysis of molecular QTL effect sizes across the 3 cell types (FE-Meta-Tissue). We determined that for each QTL data set the MaxCPP annotation outperformed the AllcisQTL and 95%CredibleSet annotations (Supplementary Table 27). A histogram of MaxCPP annotation values for each QTL data set is provided in Supplementary Figure 8. MaxCPP for each molecular QTL was significantly enriched in an analysis conditional on the baselineLD model, meta-analyzed across the 41 traits: eQTL (5.44x, s.e. = 0.55; P = 3.26e-16), H3K27ac (4.28x, s.e. = 0.37; P = 2.59e-19), H3K4me1 (4.27x, s.e. = 0.36; P = 1.29e-20), sQTL (3.61x, s.e. = 0.40; P =1.39e-10), and meQTL (2.81x, s.e. = 0.19; P = 8.36e-22) (Figure 6a and Supplementary Table 28); the enrichment for BLUEPRINT eQTL was almost as large as the enrichment for GTEx eQTL (5.84x), despite the much smaller total sample size of FE-Meta-Tissue in BLUEPRINT. This implies that BLUEPRINT sample sizes, though small, are adequately powered for eQTL detection. Consistent with this finding, we observed a high replication rate between cis-QTL in GTEx and BLUEPRINT (see Supplementary Table 29), confirming that GTEx FE-Meta-Tissue provides increased power relative to GTEx blood (Supplementary Table 28). BLUEPRINT FE-Meta-Tissue generally attained higher enrichments and $\tau*$ than MaxCPP computed using each of the three immune cell types individually (Supplementary Table 30), similar to our GTEx results (Supplementary Tables 7 and 10). MaxCPP computed using FE-Meta-Tissue also generally outperformed each of the three cell types in a meta-analysis across the six autoimmune diseases (Supplementary Table 31) and a meta-analysis across the five blood cell traits (Supplementary Table 32), in contrast to the stronger enrichments for tissue-specific GTEx blood eQTL annotations for blood cell traits (Supplementary Tables 12 and

14). FE-Meta-Tissue generally attained higher enrichments and $\tau*$ than MaxCPP computed using each of the three immune cell types individually (Supplementary Table 30), similar to our GTEx results (Supplementary Tables 7 and 10). MaxCPP computed using FE-Meta-Tissue also generally outperformed each of the three cell types in a meta-analysis across the six autoimmune diseases (Supplementary Table 31) and a meta-analysis across the five blood cell traits (Supplementary Table 32), in contrast to tissue-specific results in GTEx (Supplementary Tables 12 and 14).

Finally, we jointly analyzed MaxCPP annotations for GTEx eQTL and each of the five BLUEPRINT molecular QTL (conditional on the baselineLD model). The purpose of this analysis was to determine whether each of these molecular QTL provides independent information about disease and complex trait architectures. We determined that $\tau*$ remained statistically significant for all molecular QTL except sQTL (Figure 6b and Supplementary Table 33); a joint analysis of just the five BLUEPRINT molecular QTL (conditional on the baselineLD model) produced similar findings (Supplementary Table 34). LD scores of the sQTL annotation had the highest correlation with LD scores of the GTEx eQTL and BLUEPRINT eQTL annotations ($R = 0.56 - 0.57$; see Supplementary Figure 5), implying that much of the informativeness of sQTL in this analysis is captured by eQTL. However, eQTL, hQTL (H3K27ac and H3K4me1) and meQTL are each uniquely informative for disease and complex trait architectures.

Discussion

We have shown that annotations constructed using fine-mapped posterior probabilities for several different molecular QTL are strongly enriched for disease heritability. These results improve upon two previous studies that made key contributions in showing that annotations constructed using all significant cis-eQTL were significantly enriched for trait heritability[19,20]. Our findings provide additional motivation for colocalization studies[21–23] and transcriptome-wide association studies (TWAS)[24–26]. Our fine-mapped eQTL annotations were able to detect tissue-specific enrichments for blood and brain related traits, despite high cis-genetic correlations[12,36] of eQTL effect sizes across tissues and despite the fact that TWAS have generally concluded that their results "did not suggest tissue-specific enrichment"[26].

We note that a previous study showed that cis-eQTL often lie close to the transcription start site (TSS) or transcription end site (TES)[43], motivating us to investigate the orthogonal question of whether cis-eQTL that lie near the TSS/TES produce more disease signal than cis-eQTL that do not lie near the TSS/TES; we did not observe such an effect in the GTEx or BLUEPRINT data sets (see Supplementary Tables 35 and 36). Notably, our eQTL annotations produced particularly large enrichments when restricted to disease-relevant gene sets such as loss-of-function intolerant genes from ExAC, highlighting the potential to increase signal in analyses of gene sets harboring regulatory signals by prioritizing fine-mapped cis-eQTL. Our eQTL annotations may also prove useful in future analyses of gene pathways.

We also detected strong enrichments using annotations based on other molecular QTL, with eQTL, hQTL and meQTL all providing complementary information about disease, conditional on each other and on functional annotations from previous studies. These results motivate applying colocalization and TWAS methods to other molecular QTL; it may also be possible to prioritize other molecular QTL in gene set analyses by connecting regulatory regions to genes[44,45]. Although annotations constructed from sQTL were not conditionally significant in our analysis, previous work has shown that sQTL can contain information that is independent from eQTL[46], motivating further investigation in larger sQTL data sets.

We note several limitations of our work. First, we restrict our analyses to common variants, as S-LDSC is not currently applicable to rare variants[47]. Recent work has shown that rare variants can have substantial effects on gene expression[50], motivating ongoing work to extending S-LDSC to rare variants. Second, the CAVIAR fine-mapping method allows up to six causal variants per locus; this may limit power at loci that harbor more than six causal variants, although this would not lead to spurious signals. We determined that our results were very similar when modifying CAVIAR to allow up to three causal variants per locus (Supplementary Figures 9 and 10 and Supplementary Table 37), suggesting that modeling only six causal variants per locus is unlikely to greatly impact our results. Third, we show that S-LDSC is generally unable to produce unbiased enrichment estimates for molecular QTL based annotations when causal variants are sparse, due to violations of model assumptions (which also impact other functional enrichment methods, including GCTA[27,28], BOLT-REML[30] and LDAK[31], which also assume that causal disease effect sizes are i.i.d conditional on MAF, LD and functional annotation values); S-LDSC produces slightly conservative enrichment estimates for the MaxCPP annotation that we focus on here; however, the S-LDSC estimates should not be viewed as rigorous lower bounds, because our simulations do not include all possible genetic architectures. We caution that the TopcisQTL annotation produces large upward biases and should be avoided (Figure 1). Fourth, our results are a function of the molecular QTL sample size (Figure 3) and set of tissues; although current molecular QTL sample sizes are clearly informative, analyses of larger sample sizes and/or different tissues or contexts may produce larger enrichments in the future. Fifth, we performed a fixed-effect meta-analysis of molecular QTL effect sizes across tissues (FE-Meta-Tissue) that does not account for overlapping samples and heterogeneity across tissues in eQTL effect sizes, which could in principle limit our power[48]. However, noise is largely uncorrelated across tissues (despite pervasive sample overlap), and recently developed random-effect cross-tissue eQTL meta-analysis methods[48,49] are not applicable in the current setting (see Online Methods). Sixth, our approach cannot distinguish causal mediation from horizontal pleiotropy (i.e. independent effects on molecular QTL and disease), thus our molecular QTL enrichment results should not be viewed as a quantification of mediated effects. Despite these limitations, our results indicate that fine-mapped QTL annotations are strongly enriched for disease heritability and can help elucidate the genetic architecture of diseases and complex traits.

Online Methods

Molecular QTL-based annotations

We construct four annotations for any given QTL data set using the observed marginal association statistics. The four annotations are MaxCPP, AllcisQTL, 95%CredibleSet, and TopcisQTL. Each annotation is a vector that assigns a value to each SNP. Let a indicate our annotation for one QTL data set where a_j indicates the value assigned to SNP j. For binary annotations (AllcisQTL, 95%CredibleSet, and TopcisQTL) $a_j \in \{0, 1\}$, and $a_j = 0$ indicates that SNP j is not included in the annotation, while $a_j = 1$ indicates that SNP j is included in the annotation. For continuous probabilistic annotations (MaxCPP), $0 \leq a_j \leq 1$.

Let $S = (s_1, s_2, \cdots s_g)$ indicate an $(m \times g)$ matrix of the observed marginal association statistics obtained for each QTL data set, where m is the number of SNPs and g is the number of eGenes (e.g., genes that have at least one significant cis-eQTL). Let s_i be the vector of marginal association statistics of gene i for all the cis-variants. Utilizing s_i and the LD structure, we can compute the causal posterior probability (CPP) for each variant. CPP is the probability that a variant is causal. Let α_{ji} be the posterior probability that the SNP j is causal for the gene i. We obtain the CPP values from CAVIAR[35]. In addition to the CPP values, CAVIAR provides a 95%credible set that contains all of the causal variants with probability at least 95%. Let θ_{ji} indicate whether SNP j is in the 95%credible set for the gene i (i.e., θ_{ji}=1 indicates that the SNP j is in the gene i 95%credible set and θ_{ji}=0 otherwise). We construct the MaxCPP annotation for SNP j by computing the maximum value of CPP over all genes where SNP j is in the 95%credible set of the gene i. More formally, we have: $a_j = \max_i \alpha_{ji}$ where the maximum is over genes i with $\theta_{ij} = 1$.

AllcisQTL annotation is a binary annotation, where any variant whose marginal association statistic for at least one gene passes the significance threshold (FDR < 0.05) has annotation value 1, and each other variant has annotation value 0. Let t_{ji} indicate whether the SNP j is statistically significant for the gene i ($t_{ji} = 1$ when FDR(j) < 0.05 and $t_{ij} = 0$ otherwise). More formally, we have: $a_j = \max_i t_{ji}$.

95%CredibleSet is a binary annotation, any variant that is in a 95%credible set of at least one gene has annotation value 1 and each other variant has annotation value 0. More formally, we have: $a_j = \max_i \theta_{ji}$.

TopcisQTL is a binary annotation where any variant that is the most significant variant for at least one gene has annotation value 1 and each other variant has annotation value 0. Let γ_{ji} indicate whether the SNP j is the most significant SNP for the gene i (i.e., γ_{ji}=1 if SNP j is the most significant SNP among all cis-variants for the gene i and γ_{ji}=0 otherwise). More formally, we have: $a_j = \max_i \gamma_{ji}$.

Enrichment and effect size ($\tau*$) metrics

We use two metrics to measure the importance of an annotation in the context of diseases and complex traits: Enrichment and standardized effect size ($\tau*$) of annotation. We use S-LDSC[32,33] to compute enrichment and standardized effect size ($\tau*$). Let a_{cj} indicate the annotation value of the SNP j for the annotation c. S-LDSC[32,33] assumes that the variance of each SNP is a linear additive contribution to each annotation:

$$\mathrm{Var}\left(\beta_j\right) = \sum_c a_{cj}\tau_c \quad (1)$$

where τ_c is the contribution of the annotation c to per-SNP heritability. S-LDSC[32,33] estimates τ_c using the following equation:

$$E[\chi_j^2] = N\sum_c l(j,c)\tau_c + 1 \quad (2)$$

where N is the GWAS sample size and $l(j,c)$ is the LD score of the SNP j for the annotation c. S-LDSC computes the LD scores as follow: $l(j,c) = \sum_k a_{ck} r_{jk}^2$ where r_{jk} is the genetic correlation between the SNPs j and k.

Since τ_c depends on the trait heritability and the size of the annotation, ref.[33] defined τ_c* for an annotation as the standardized annotation effect size:

$$\tau_c* = \frac{\tau_c \mathrm{sd}(c)}{h_g^2/M} \quad (3)$$

where sd(c) is the standard deviation of the annotation c, h_g^2 is the SNP-heritability, and M is the number of variants used to compute h_g^2. In our experiments, M is equal to 5,961,159 (see below).

The enrichment of an annotation is defined as the fraction of heritability captured by the annotation divide by the fraction of SNPs in that annotation. We extend the definition of enrichment to continuous probabilistic annotations with values between 0 and 1:

$$\mathrm{Enrichment} = \frac{\%h_g^2(c)}{\%\mathrm{SNP}(c)} = \frac{\frac{h_g^2(c)}{h_g^2}}{\frac{\sum_j a_{jc}}{M}} \quad (3)$$

where $h_g^2(c)$ is the heritability captured by the c-th annotation. We can compute this quantity as follows:

$$h_g^2(c) = \sum_j a_{jc} \mathrm{Var}(\beta_j) = \sum_j a_{jc} (\sum_c a_{jc} \tau_c) \quad (4)$$

Although both enrichment and $\tau*$ are computed using a model that includes all annotations, $\tau*$ quantifies effects that are unique to the focal annotation (after conditioning on all other annotations in the model), whereas enrichment quantifies effects that are either unique and/or non-unique to the focal annotation. For example, consider a model that includes two annotations, where the first annotation is a highly disease-informative functional annotation and the second annotation is the first annotation plus a random set of SNPs. Only the first annotation will have significant $\tau*$, but both annotations will be significantly enriched. We confirmed via simulation that, under a generative model in which only the baselineLD and GTEx FE-Meta-Tissue MaxCPP annotations directly influence trait heritability, $\tau*$ estimates for the GTEx-Whole-Blood MaxCPP annotations are equal to 0 on average, with a correctly calibrated null distribution of P-values for nonzero $\tau*$ (Supplementary Figure 11).

We computed the statistical significance level (p-value) of enrichment for each annotation via block-jackknife, as described in our previous studies[32,33,50]. We computed the statistical significance (p-value) of standardized effect size ($\tau*$) for each annotation by assuming that $\frac{\tau*}{se(\tau*)}$ follows a normal distribution with mean 0 and variance of 1 ($\frac{\tau*}{se(\tau*)} \sim N(0,1)$)[33].

Simulation framework

Main simulation framework—We simulated both gene expression and trait phenotypes. We utilized UK Biobank genotypes from chromosome 1, which consists of 749,024 variants, for our simulation. We used 40,000 individuals to generate the trait phenotypes and a non-overlapping set of 500 individuals to generate gene expression phenotypes. Let σ_{ge}^2 and σ_t^2 indicate the total heritability of gene expression and trait phenotypes, respectively. We simulated causal trait effect sizes using a polygenetic model, $\beta_i \sim N(0, \frac{\sigma_t^2}{n_t})$, where β_i is the causal (true) effect size of the i-th causal variant and n_t is the number of causal variants for the trait. Similarly, we simulated causal gene expression effect sizes using a polygenetic model, $\beta'_{ji} \sim N(0, \frac{\sigma_{ge}^2}{n_{ge}})$, where β'_{ji} is the causal (true) effect size of the i-th causal variant on gene expression of gene j and n_{ge} is the number of causal variants. We use the following model to simulate the gene expression and traits:

$$y = X\beta + e \quad (5)$$
$$g_j = X'\beta'_j + e_j,$$

where e is environmental and measurement noise, y is the simulated trait phenotypes, g_j is the gene expression of gene j, and e_j is the environmental and measurement noise for gene j. In the case of gene expression, we simulated 1,860 phenotypes to represent our simulated gene expression data as 1,860 genes lies on chromosome one. We simulated three different data sets where n_{ge} is set to 1 causal variant, 10 causal variants, or 10% of cis-variants that are causal for each gene. The default value of n_{ge} is set to 10% of cis-variants. In the case of trait phenotypes, we set n_t to be the union of causal variants for all genes on chromosome 1 ($n_t = \cup_{i=1}^{1,860} C_{ge}(i)$ where $C_{ge}(i)$ is the set of causal variants for gene i). In our simulated data sets, we set σ_t^2 to 0.2 and considered different values of σ_{ge}^2 to test our results on different input parameters. We set σ_{ge}^2 to 0.1, 0.16, and 0.2 resulting in different simulated data sets. The default value of 0.16 is used for σ_{ge}^2. For each simulated data set, we performed 400 simulations.

After simulating the gene expression and trait phenotypes, we obtained marginal association statistics for each variant using linear regression implemented in the PLINK software[51] (see URLs). In the case of simulated trait phenotypes, we computed the association of each variant with the simulated trait phenotypes ($y \sim x_i$). In the case of simulated gene expression phenotypes, we computed the marginal statistics for all variants within 1Mb of the TSS ($g_j \sim x'_i$). In some simulations, we assumed that a subset of the causal variants are missing (not measured). Let n_m indicate the fraction of causal eQTL that are missing. In our simulated datasets, we set n_m to 0% (no causal eQTL is missing), 5% (5% of causal eQTL are missing), 10% (10% of causal eQTL are missing), or 50% (half of the causal eQTL are missing). The default value of 10% is used for n_m.

Alternative simulation framework—We also considered an alternative simulation framework, drawn from our previous work[25] that directly uses simulated gene expression to generate complex trait phenotypes. We utilized UK Biobank genotypes from chromosome 1, which consists of 749,024 variants, for our simulation. We used 40,000 individuals to generate the trait phenotypes and gene expression phenotypes. We assume that the trait phenotype is a mixture of direct genotype effects (effect not medicate through gene expression) and gene expression effects. Let σ_{dt}^2 indicate the phenotypic variance explained directly by genotypes, σ_{ge}^2 indicate the gene expression variance explained directly by genotypes, and σ_{gt}^2 indicate the phenotypic variance explained by gene expression. We simulated causal gene expression effect sizes using a polygenetic model, $\beta'_{ji} \sim N(0, \frac{\sigma_{ge}^2}{n_{ge}})$, where β'_{ji} is the causal (true) effect size of the i-th causal variant on the gene expression of gene j and n_{ge} is the number of causal variants. We simulated causal trait effect sizes using a

polygenetic model, $\beta_i \sim N(0, \frac{\sigma^2_{dy}}{n_{ct}})$, where β_i is the causal (true) effect size of i-th causal variant and n_{ct} is the number of causal variants for simulated trait phenotypes. As above, we simulated expression values for the 1,860 genes on chromosome 1. We simulated the effect size of gene expression on traits using polygenetic model, $\gamma_k \sim N(0, \frac{\sigma^2_{gt}}{1,860})$, where σ^2_{gt} is set to $\frac{0.1}{180}$[25]. We use the following model to simulate the gene expression and traits:

$$g_j = X'\beta'_i + e_j \qquad (6)$$
$$y = X\beta + \sum_{k=1}^{1,860} g_k\gamma_j + e\,.$$

After simulating the gene expression and trait phenotypes, we obtained marginal association statistics for each variant using linear regression implemented in the PLINK software[52] (see URLs). In the case of simulated trait phenotypes, we computed the association of each variant with the simulated trait phenotypes ($y \sim x_i$). In the case of simulated gene expression phenotypes, we computed the marginal statistics for all variants within 1Mb of the TSS ($g_j \sim x''_i$) where X'' is the subset of X genotype matrix restricted to 500 individuals. We assumed that 10% of the causal variants are missing (not measured), i.e we set n_m to 10%.

We generated the four annotations as described above. We used CAVIAR[35] to generate the 95%Credible set and MaxCPP annotations. We utilized European samples from the 1000 Genomes Project (1000G)[52] (see URLs) to estimate the LD structure required as input to CAVIAR. We applied CAVIAR under a setting where we allowed up to six causal variants for each gene. We observed that the results for cases where we allowed up to six or three causal variants for each gene are not statistically different (Supplementary Figures 9 and 10 and Supplementary Table 37). We note that the 95% credible set is not guaranteed to be unique; in this work we use a single 95% credible set for each gene for three reasons: First, use of a single 95% credible set is consistent with the output of all existing fine-mapping methods of which we are currently aware. Second, computing all 95% credible sets for each gene is computationally costly. Third, taking the union of all 95% credible sets might reduce the enrichment and $\tau*$, whereas the goal of this paper is to construct annotations with highest possible enrichment and $\tau*$.

After obtaining the four annotations, we ran S-LDSC[32] to generate the LD score of each variant in each annotation using the same procedure described in the previous studies[32,33]. Regression SNPs, which are used by S-LDSC to estimate τ from marginal association statistics, were obtained from the HapMap Project phase 3[53]. These SNPs are considered as well-imputed SNPs. SNPs with marginal association statistics larger than 80 or larger than 0.001 N and SNPs that are in the major histocompatibility complex (MHC) region were excluding from all the analyses. Reference SNPs, which are used to compute LD scores,

were defined using the European samples in 1000G[52]. Heritability SNPs, which are used to estimate sd(c) and h_g^2, were defined as common variants (MAF ≥ 0.05) in the set of reference SNPs. Using the LD score for each annotation and the marginal statistics obtained from the trait phenotypes, we computed the enrichment and $\tau*$ for each simulation. Then, we compared the S-LDSC estimated enrichment and true enrichment for each annotation. All results are averaged across 400 simulations.

In addition to S-LDSC, we used GCTA[27,28] to compute the enrichment of each binary annotation. We first computed the GRM for each annotation using the set of all variants in that annotation. We then used the –reml option in GCTA to estimate the heritability explained by each annotation.

Concrete example for S-LDSC bias estimates of TopcisQTL

Let x be a SNP in the TopcisQTL annotation, let y be a SNP not in the TopcisQTL annotation that is in LD with x, and let z be a random SNP not in the TopcisQTL annotation. Then, in expectation, y has a larger *causal* disease effect size than z (violating S-LDSC model assumptions), because it is possible that y is a causal molecular QTL (tagged by x, which may be more significant due to statistical chance), and that y is also causal for disease. The fact that SNPs in LD with the TopcisQTL annotation have larger causal disease effect sizes may cause TopcisQTL enrichment to be overestimated by S-LDSC, which attributes higher χ^2 statistics for such SNPs entirely to tagging of causal TopcisQTL enrichment. On the other hand, enrichments of more inclusive annotations may be underestimated by S-LDSC, because SNPs in the annotation with high LD to other SNPs in the annotation are expected to have lower causal disease effect sizes than random SNPs in the annotation (again violating S-LDSC model assumptions).

Set of 41 independent diseases and complex traits

We initially considered 34 GWAS summary association statistic data sets that are publicly available and 55 UK Biobank traits (see URLs) for which summary association statistics were computed using BOLT-LMM[30,54] (see URLs; up to N=459K European-ancestry samples). We restricted our analyses to 47 data sets with z-scores of total SNP heritability at least 6 (Supplementary Table 6). The 47 data sets included 6 traits that were duplicated in two different data sets (genetic correlation of at least 0.9). Thus, we analyzed 41 independent diseases and complex traits. We ran S-LDSC using the same procedures described in previous studies[32,33] (see above). All analyses that included the baselineLD model are based on baselineLD model v1.1, which is identical to the baselineLD model as previously described[33] except that we fixed an error in the promoter annotation (inherited from previous studies[32,33]); we determined that fixing this error did not affect our results (see Supplementary Table 38). The z-score of total SNP heritability was computed using S-LDSC with the baselineLD model, and the genetic correlation between pairs of traits was computed using cross-trait LDSC[55]. The meta-analyzed values of enrichment and $\tau*$ across the 47 data sets were computed using a random-effect meta-analysis, as implemented in the rmeta R package.

Fixed-effect meta-analysis of eQTL effect sizes (FE-Meta-Tissue)

Given a set of effect sizes for SNP i ($\beta_1, \beta_2, \cdots \beta_t$) for t tissues, where β_j is the eQTL effect size for tissue j, we used fixed-effect meta-analysis (FE-Meta-Tissue) to compute inverse-variance weighted meta-analysis z-scores z_{FE} as follows:

$$z_{FE} = \frac{\sum_{j=1}^{t} w_j \beta_j}{\sqrt{\sum_{j=1}^{t} w_j}} \quad (7)$$

where $w_j = \frac{1}{se(\beta_j)^2}$ and $se(\beta_j)$ is the standard error of β_j. We note that equation (7) is equivalent to computing a weighted average of z-scores.

Our use of FE-Meta-Tissue has two limitations. First, expression levels in two tissues in the same individual are not independent (sample overlap). Second, true effect sizes in two tissues may be different (heterogeneity). We discuss each limitation in turn.

Regarding the sample overlap limitation, we determined that noise is largely uncorrelated across tissues. For example, the correlation of normalized gene expression (read count) between whole blood and brain hippocampus is 0.11. Furthermore, the genetic correlation between these two tissues is ~0.67[36] and the heritability explained by cis-eQTL is ~0.12. This implies that the bulk of gene expression correlation is due to genetic correlation. Thus, the noise (environmental and measurement) in expression levels in two tissues in the same individual is close to independent.

Regarding the heterogeneity limitation, we determined that recently developed random-effect cross-tissue eQTL meta-analysis methods[48,49,56,57] are not applicable to our problem. The Meta-Tissue method[40] is computationally intractable for data sets as large as GTEx (number of tissues and sample size). Other existing methods[49,56,57], which are Bayesian methods, do not produce summary statistics (e.g. z-scores) that are required to compute the MaxCPP annotation. Thus, these methods are not applicable to our work. Previous studies[12,36] have shown that eQTL effects are highly correlated across tissues, suggesting that our fixed-effect meta-analysis approach is likely to be fairly close to optimal.

We note that the above limitations pertain only to power and not to false positives in our setting, which involves building eQTL annotations to apply to independent disease data. Our results (Figure 2 and Supplementary Tables 7 and 27) show that we have improved our results by utilizing FE-Meta-Tissue. Furthermore, utilizing FE-Meta-Tissue increases replication rates for both eQTL and hQTL (Supplementary Table 29).

Blood and brain related diseases and complex traits

We analyzed six autoimmune diseases: Crohn's disease[58], Rheumatoid arthritis (ref.[59] and UK Biobank), Ulcerative colitis[58], Lupus[60], Celiac[61], and all autoimmune and inflammatory diseases in UK Biobank).

We analyzed five blood cell traits: white blood cell count, red blood cell count, platelet count, eosinophil count, and red blood cell distribution width. All of these data sets were obtained from UK Biobank.

We analyzed eight independent brain-related diseases and complex traits: Age at menarche[62], BMI (ref.[63] and UK Biobank), Bipolar Disorder[64], Depressive symptoms[2], Neuroticism (UK Biobank), Schizophrenia[1], Smoking Status (ref.[65] and UK Biobank), and Year of education (ref.[66] and UK Biobank). These traits are a subset of traits from Supplementary Table 6 that were reported to be brain-enriched[32,50].

Data availability

The S-LDSC software, baselineLD, and MaxCPP QTL-based annotations, and a tutorial on how to use S-LDSC with QTL-based annotations are available online (see URLs). A Life Sciences Reporting Summary is available.

URLs

CAVIAR: http://genetics.cs.ucla.edu/caviar/

GTEx (Release v6, dbGaP Accession phs000424.v6.p1): http://www.gtexportal.org.

GCTA: cnsgenomics.com/software/gcta/

BLUEPRINT: ftp://ftp.ebi.ac.uk/pub/databases/blueprint/blueprint_Epivar/qtl_as/

baselineLD annotations: https://data.broadinstitute.org/alkesgroup/LDSCORE/

MaxCPP QTL-based annotations and partitioned LD scores: https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_QTL/

95%Causal Set QTL-based annotations and partitioned LD scores: https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_QTL/

1000 Genomes Project Phase 3 data: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502

PLINK software: https://www.cog-genomics.org/plink2

BOLT-LMM software: https://data.broadinstitute.org/alkesgroup/BOLT-LMM

BOLT-LMM summary statistics for UK Biobank traits: https://data.broadinstitute.org/alkesgroup/UKBB

UK Biobank: http://www.ukbiobank.ac.uk/

UK Biobank Genotyping and QC Documentation: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Soumya Raychaudhuri, Noah Zaitlen, Bogdan Pasaniuc, Michel Nivard, Jae-Hoon Sul, and Fereydoun Hormozdiari for helpful discussions. This research was funded by NIH grants U01 HG009379, R01 MH101244, R01 MH109978, T32 DK110919, and R01 MH107649. This research was conducted using the UK Biobank Resource under Application 16549.

References

1. Schizophrenia Working Group of the Psychiatric Genomics Consortium. et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–427. [PubMed: 25056061]
2. Okbay A, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016; 533:539–542. [PubMed: 27225129]
3. 10 Years of GWAS Discovery: Biology, Function, and Translation. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/28686856. (Accessed: 24th March 2018)
4. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009; 106:9362–9367. [PubMed: 19474294]
5. Nicolae DL, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010; 6:e1000888. [PubMed: 20369019]
6. Nica AC, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010; 6:e1000895. [PubMed: 20369022]
7. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337:1190–1195. [PubMed: 22955828]
8. 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. 2015; 97:139–152. [PubMed: 26140449]
9. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
10. Wright FA, et al. Heritability and genomics of gene expression in peripheral blood. Nat Genet. 2014; 46:430–437. [PubMed: 24728292]
11. Zhang X, et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat Genet. 2015; 47:345–352. [PubMed: 25685889]
12. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–660. [PubMed: 25954001]
13. Zhernakova DV, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017; 49:139–145. [PubMed: 27918533]
14. GTEx Consortium. et al. Genetic effects on gene expression across human tissues. Nature. 2017; 550:204–213. [PubMed: 29022597]
15. McVicker G, et al. Identification of genetic variants that affect histone modifications in human cells. Science. 2013; 342:747–749. [PubMed: 24136359]
16. Waszak SM, et al. Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell. 2015; 162:1039–1050. [PubMed: 26300124]
17. Grubert F, et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell. 2015; 162:1051–1065. [PubMed: 26300125]
18. Chen L, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 2016; 167:1398–1414.e24. [PubMed: 27863251]
19. Davis LK, et al. Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. PLoS Genet. 2013; 9:e1003864. [PubMed: 24204291]
20. Torres JM, et al. Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. Am J Hum Genet. 2014; 95:521–534. [PubMed: 25439722]

21. Hu X, et al. Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in CD4+ effector memory T cells. PLoS Genet. 2014; 10:e1004404. [PubMed: 24968232]
22. Giambartolomei C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014; 10:e1004383. [PubMed: 24830394]
23. Hormozdiari F, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016; 99:1245–1260. [PubMed: 27866706]
24. Gamazon ER, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015; 47:1091–1098. [PubMed: 26258848]
25. Gusev A, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016; 48:245–252. [PubMed: 26854917]
26. Mancuso N, et al. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. Am J Hum Genet. 2017; 100:473–487. [PubMed: 28238358]
27. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88:76–82. [PubMed: 21167468]
28. Lee SH, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet. 2012; 44:247–250. [PubMed: 22344220]
29. Gusev A, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014; 95:535–552. [PubMed: 25439723]
30. Loh PR, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet. 2015; 47:1385–1392. [PubMed: 26523775]
31. Speed D, et al. Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017; 49:986–992. [PubMed: 28530675]
32. Finucane HK, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015; 47:1228–1235. [PubMed: 26414678]
33. Gazal S, et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat Genet. 2017; 49:1421–1427. [PubMed: 28892061]
34. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017; 18:117–127. [PubMed: 27840428]
35. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014; 198:497–508. [PubMed: 25104515]
36. Liu X, et al. Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. Am J Hum Genet. 2017; 100:605–616. [PubMed: 28343628]
37. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–291. [PubMed: 27535533]
38. Gazal S, Finucane H, Price A. Reconciling S-LDSC and LDAK functional enrichment estimates. BioRxiv. 2018
39. Cassa CA, et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nat Genet. 2017; 49:806–810. [PubMed: 28369035]
40. Samocha KE, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014; 46:944–950. [PubMed: 25086666]
41. Wang T, et al. Identification and characterization of essential genes in the human genome. Science. 2015; 350:1096–1101. [PubMed: 26472758]
42. Hormozdiari F, et al. Widespread Allelic Heterogeneity in Complex Traits. Am J Hum Genet. 2017; 100:789–802. [PubMed: 28475861]
43. Veyrieras JB, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008; 4:e1000214. [PubMed: 18846210]
44. Javierre BM, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016; 167:1369–1384.e19. [PubMed: 27863249]
45. Mumbach MR, et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet. 2017; 49:1602–1612. [PubMed: 28945252]

46. Li YI, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016; 352:600–604. [PubMed: 27126046]
47. Li X, et al. The impact of rare variation on gene expression across tissues. Nature. 2017; 550:239–243. [PubMed: 29022581]
48. Sul JH, Han B, Ye C, Choi T, Eskin E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. PLoS Genet. 2013; 9:e1003491. [PubMed: 23785294]
49. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet. 2013; 9:e1003486. [PubMed: 23671422]
50. Finucane H, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet. In press.
51. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]
52. 1000 Genomes Project Consortium. et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
53. International HapMap 3 Consortium. et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]
54. Loh P-R, et al. Mixed model association for biobank-scale data sets. Nat Genet. In press.
55. Bulik-Sullivan B, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015; 47:1236–1241. [PubMed: 26414676]
56. Urbut SM, et al. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. bioRxiv. 2016
57. Park Y, et al. Multi-tissue polygenic models for transcriptome-wide association studies. bioRxiv. 2017
58. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–124. [PubMed: 23128233]
59. Okada Y, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014; 506:376–381. [PubMed: 24390342]
60. Bentham J, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. 2015; 47:1457–1464. [PubMed: 26502338]
61. Dubois PCA, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet. 2010; 42:295–302. [PubMed: 20190752]
62. Day FR, et al. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. Nat Genet. 2017; 49:834–841. [PubMed: 28436984]
63. Speliotes EK, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010; 42:937–948. [PubMed: 20935630]
64. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet. 2011; 43:977–983. [PubMed: 21926972]
65. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010; 42:441–447. [PubMed: 20418890]
66. Rietveld CA, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science. 2013; 340:1467–1471. [PubMed: 23722424]

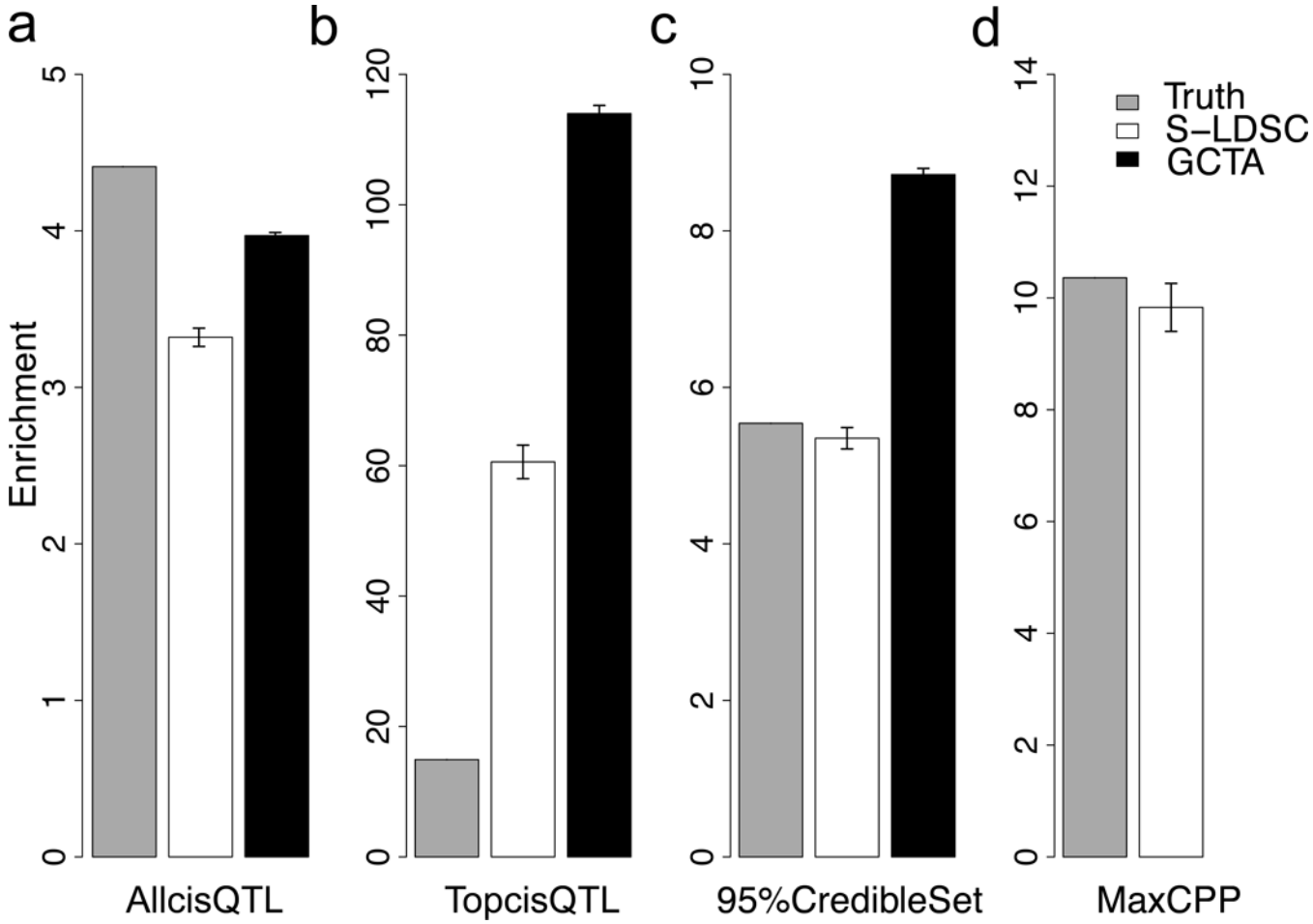

Figure 1. S-LDSC and GCTA estimate for TopcisQTL are upward biased in simulations

Panels (a), (b), (c), and (d) illustrate the true enrichment and S-LDSC and GCTA enrichment estimates for AllcisQTL, 95%CredibleSet, TopcisQTL, and MaxCPP annotations, respectively. GCTA is not applicable to continuous annotations (MaxCPP). The Y-axis indicates the mean of enrichment and error bars represent 95% confidence intervals that are computed for 400 simulations. Numerical results are reported in Supplementary Table 2. See Supplementary Table 3 and Supplementary Table 4 for additional simulation scenarios.

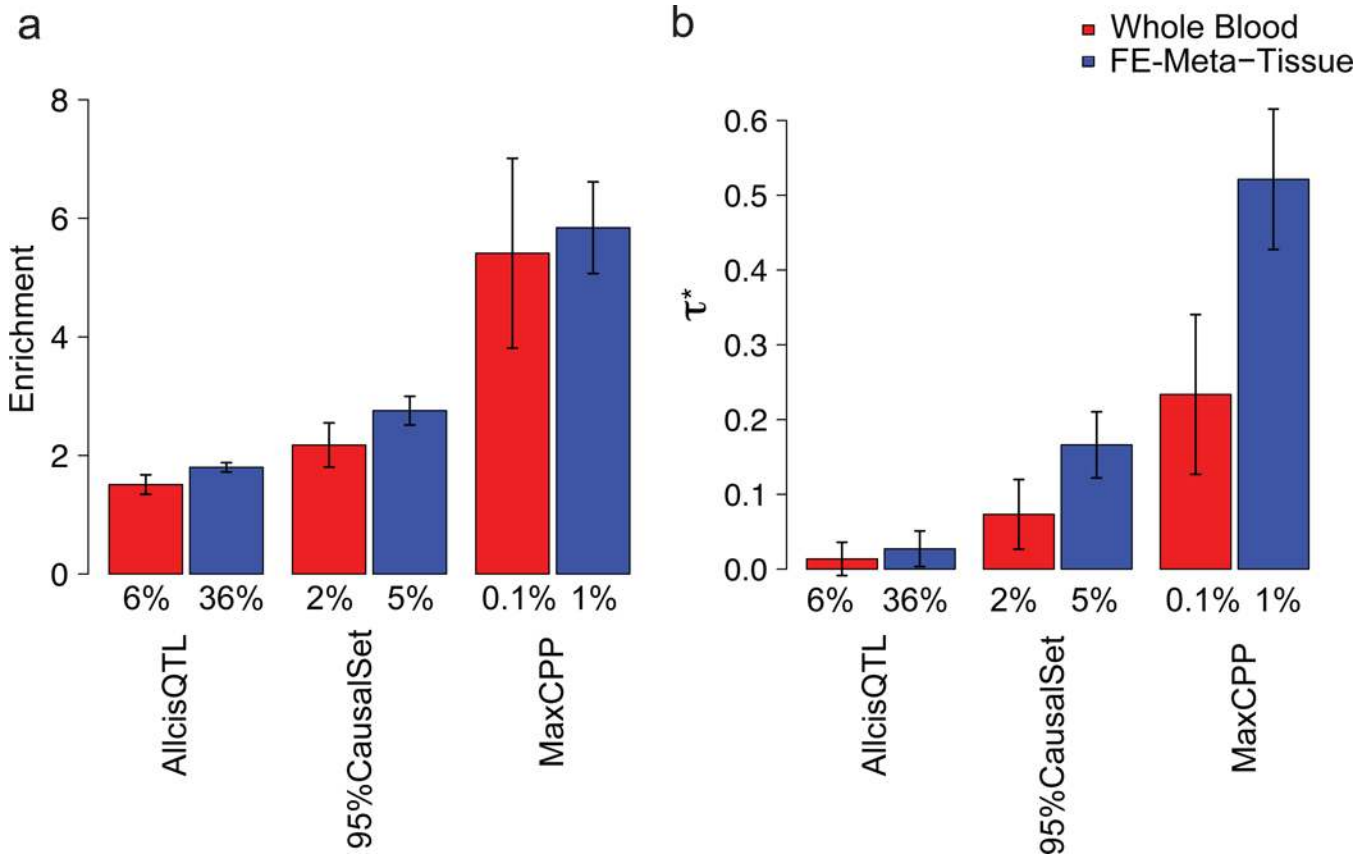

Figure 2. Fine-mapped eQTL are strongly enriched for disease/trait heritability

(a) Meta-analysis results across 41 traits of enrichment for whole blood and FE-Meta-Tissue from the GTEx data set conditioning on the baselineLD model. (b) Meta-analysis results across 41 traits of $\tau*$ for whole blood and FE-Meta-Tissue conditioning on the baselineLD model. In each panel, we report results for AllcisQTL, 95%CredibleSet, and MaxCPP. Error bars represent 95% confidence intervals. The % value under each bar indicates the proportion of SNPs in each annotation; for probabilistic annotations (MaxCPP), this is defined as the average value of the annotation. Numerical results are reported in Supplementary Table 7.

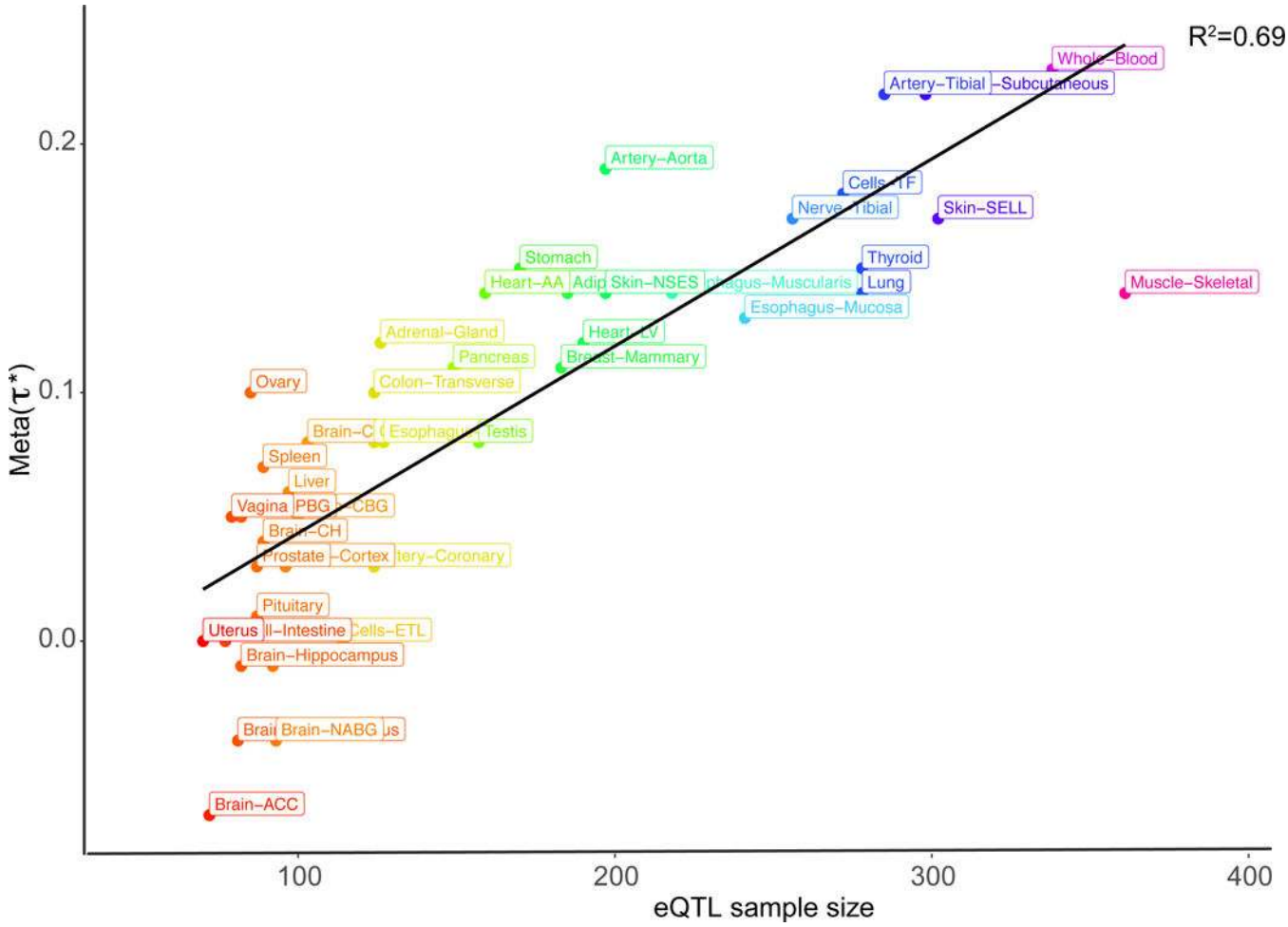

Figure 3. Relationship between eQTL sample size and the annotation effect size ($\tau*$)

For each tissue, we plot the $\tau*$ of the MaxCPP annotation, meta-analyzed across 41 traits, against the eQTL sample size. Numerical results are reported in Supplementary Table 10. For visualization purposes, we use the following abbreviations: Adipose Visceral Omentum (Adipose-Visceral), Brain Anterior cingulate cortex BA24 (Brain-ACC), Brain Caudate basal ganglia (Brain-CBG), Brain Cerebellar Hemisphere(Brain-CH), Brain Cerebellar Hemisphere (Brain-CH), Brain Frontal Cortex BA9 (Brain-FC), Brain Nucleus accumbens basal ganglia (Brain-NABG), and Brain Putamen basal ganglia (Brain-PBG), Cells EBV transformed lymphocytes (Cells-CETL), Cells Transformed fibroblasts (Cells-TF), Esophagus Gastroesophageal Junction (Esophagus-GJ), Heart Atrial Appendage (Heart-AA), Heart Left Ventricle (Heart-LV), Skin Not Sun Exposed Suprapubic (Skin-NSES), Skin Sun Exposed Lower leg (Skin-SELL), and Small Intestine Terminal Ileum (Small-Intestine).

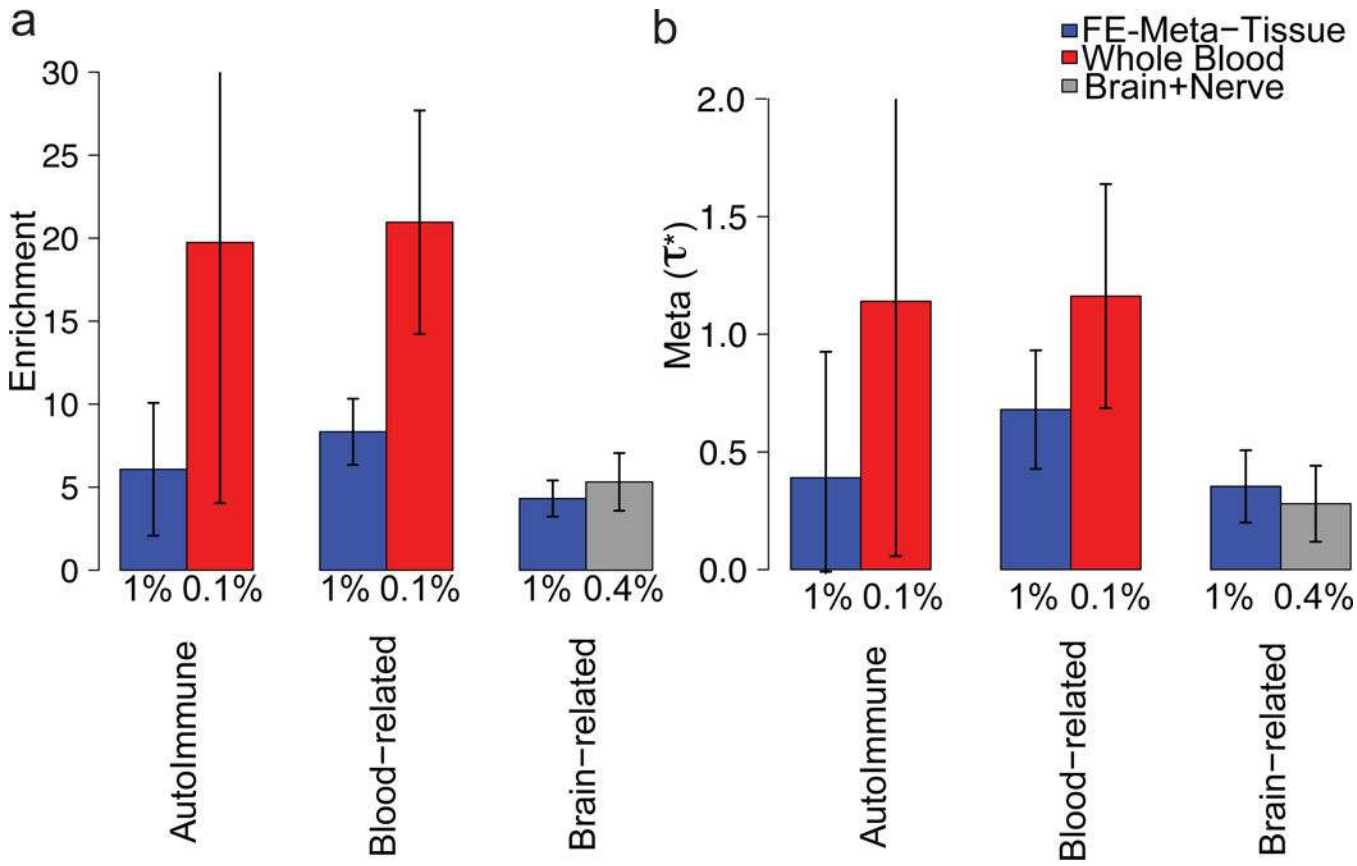

Figure 4. Tissue-specific fine-mapped eQTL enrichments for blood and brain related traits

Meta-analysis results of (a) enrichment and (b) $\tau *$ of FE-Meta-Tissue and tissue-specific MaxCPP annotations, conditional on each other and the baselineLD model, across six independent autoimmune diseases, five blood cell traits, and eight brain-related traits, respectively. The Y-axis is the meta-analyzed value and error bars represent 95% confidence intervals. The % value under each bar indicates the proportion of SNPs in each annotation, defined as the average value of the annotation. Numerical results are reported in Supplementary Table 13.

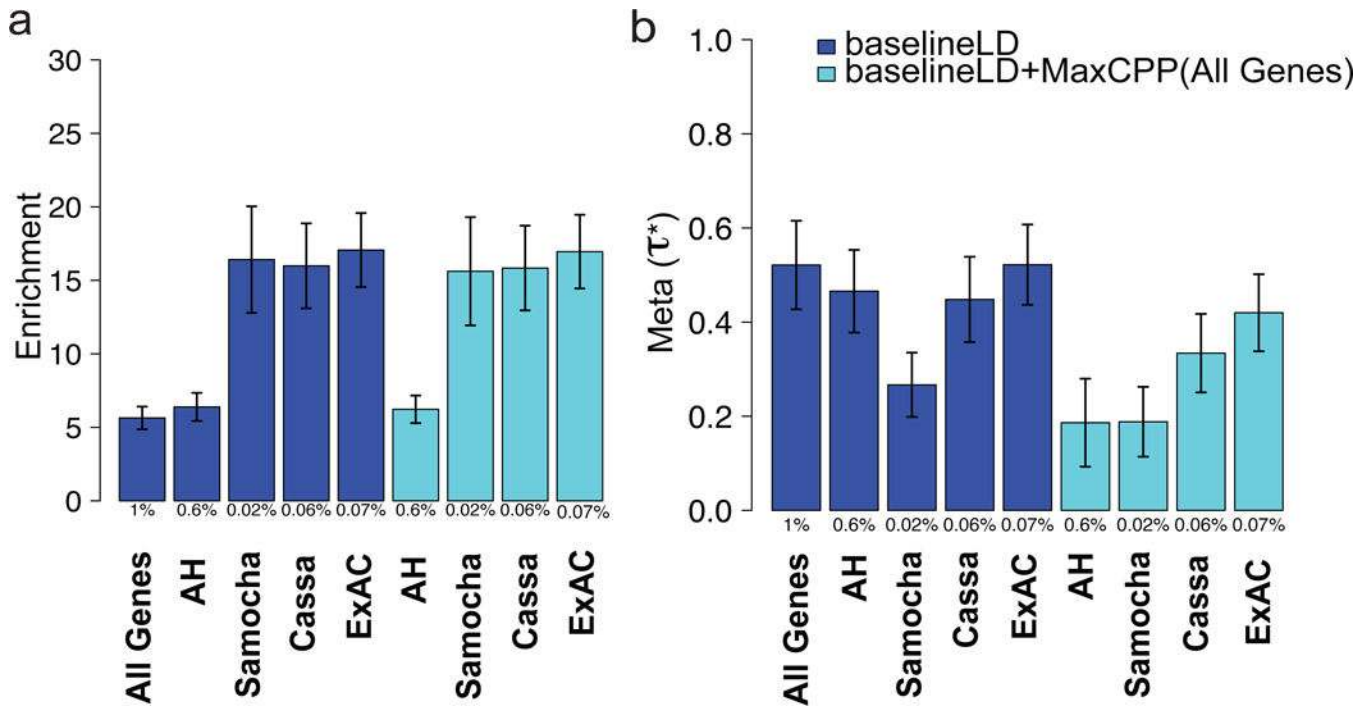

Figure 5. Heritability enrichment of fine-mapped eQTL is concentrated in disease-relevant gene sets

Meta-analysis results of (a) enrichment and (b) $\tau*$ of MaxCPP(S) for various gene sets S. We report results conditional on the baselineLD model (dark blue) and results conditional on both the baselineLD model and MaxCPP(All Genes) (light blue), meta-analyzed across 41 traits. As expected, $\tau*$ estimates are reduced by conditioning on MaxCPP(All Genes), but enrichment estimates are not affected. The Y-axis is the meta-analyzed value and error bars represent 95% confidence intervals that are computed over 41 traits. The % value under each bar indicates the proportion of SNPs in each annotation, defined as the average value of the annotation. Numerical *results are reported in* Supplementary Table 18.

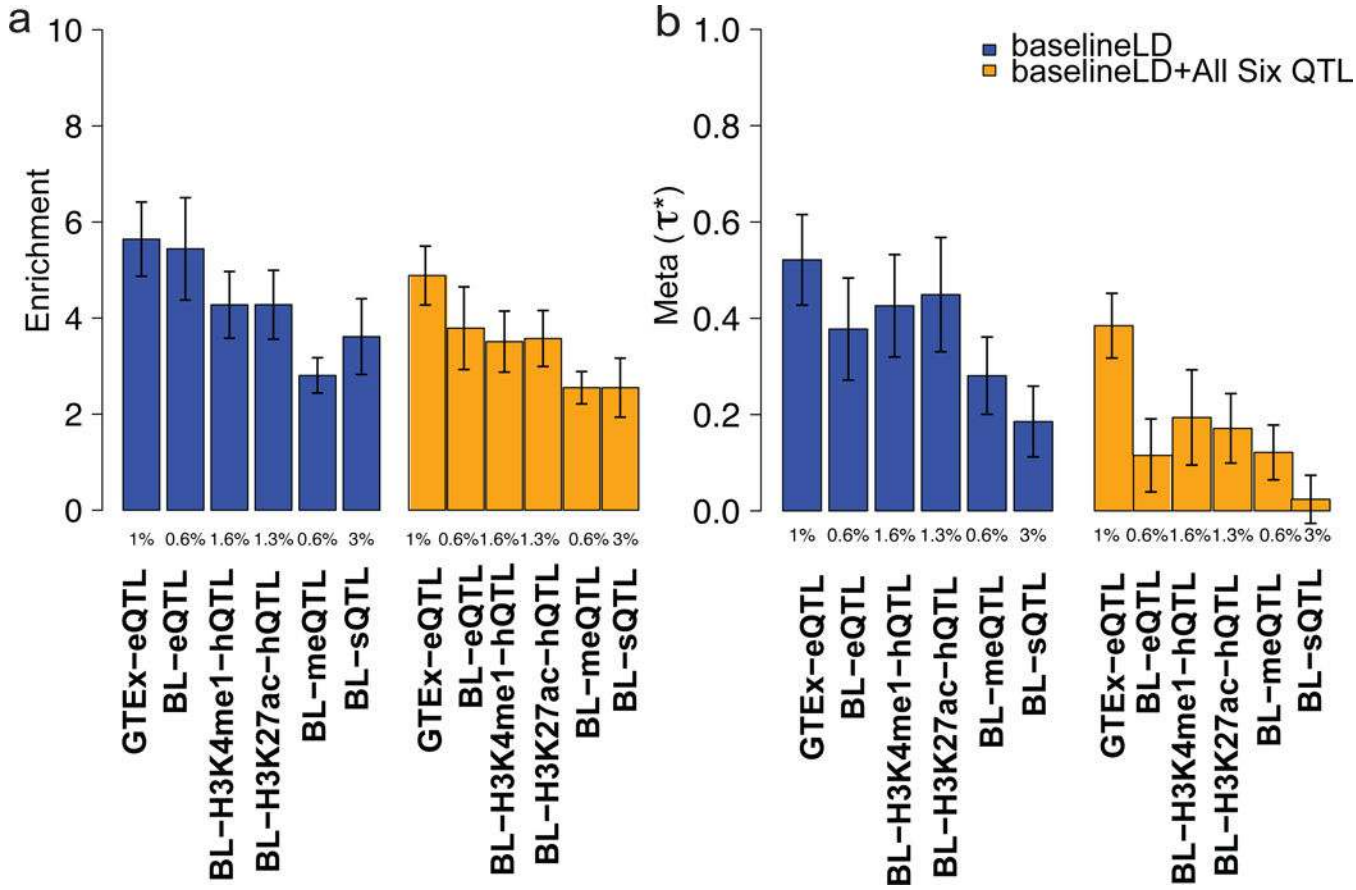

Figure 6. Fine-mapped eQTL, hQTL, sQTL, and meQTL annotations are enriched for disease/trait heritability

Meta-analysis results of (a) enrichment and (b) $\tau *$ of MaxCPP for various molecular QTL from GTEx and BLUEPRINT (BL). We report results conditional on the baselineLD model (dark blue) and results conditional on both the baselineLD model and MaxCPP for all six molecular QTL (orange), meta-analyzed across 41 traits. As expected, $\tau *$ estimates are reduced by conditioning on MaxCPP for all molecular QTL, but enrichment estimates are not affected. The Y-axis is the meta-analyzed value and error bars represent 95% confidence intervals that are computed over 41 traits. The % value under each bar indicates the proportion of SNPs in each annotation, defined as the average value of the annotation. Numerical results are reported in Supplementary Table 28 and Supplementary Table 33.

Table 1
List of molecular QTL data sets analyzed

GTEx includes eQTL for a wide range of tissues. BLUEPRINT includes eQTL, two hQTL, sQTL and meQTL for 3 blood cell types. Sample sizes for each tissue are provided in Supplementary Table 5 (GTEx) and Supplementary Table 26 (BLUEPRINT).

Dataset	QTL Type	Number of Tissues	N (per tissue)	N (total)
GTEx	eQTL	44	70-361	7014
BLUEPRINT	eQTL	3	169-194	555
BLUEPRINT	hQTL (H3K27ac)	3	143-174	479
BLUEPRINT	hQTL (H3K4me1)	3	104-173	449
BLUEPRINT	sQTL	3	169-194	555
BLUEPRINT	meQTL	3	132-197	525