Leveraging molecular QTL 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. Ju³, Po-Ru Loh^{2,4}, Armin Schoech^{1,2}, Yakir Reshef^{2,5}, Xuanyao Liu^{1,2},
Luke O'Connor^{1,2}, Alexander Gusev^{4,6}, Eleazar Eskin^{3,7}, and Alkes L. Price^{*1,2}

¹Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA
02115, USA

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

³Department of Computer Science, University of California, Los Angeles, California 90095

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

⁵Department of Computer Science, Harvard University, Cambridge, MA 02138, USA

⁶Dana Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA

⁷Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

*Corresponding author: hormozdiari@hsph.harvard.edu, aprice@hsph.harvard.edu

Abstract

There is increasing evidence that many GWAS risk loci are molecular QTL for gene expression (eQTL), histone modification (hQTL), splicing (sQTL), and/or DNA methylation (meQTL). Here, we introduce a new set of functional annotations based on causal posterior probabilities (CPP) of fine-mapped molecular cis-QTL, using data from the GTEx and BLUEPRINT consortia. We show that these annotations are very strongly enriched for disease heritability across 41 independent diseases and complex traits (average $N=320K$): 5.84x for GTEx eQTL, and 5.44x for eQTL, 4.27-4.28x for hQTL (H3K27ac and H3K4me1), 3.61x for sQTL and 2.81x for meQTL in BLUEPRINT (all $P \leq 1.39e-10$), far higher than enrichments obtained using standard functional annotations that include all significant molecular cis-QTL (1.17-1.80x). eQTL annotations that were obtained by meta-analyzing all 44 GTEx tissues generally performed best, but tissue-specific blood eQTL annotations produced stronger enrichments for autoimmune diseases and blood cell traits and tissue-specific brain eQTL annotations produced stronger enrichments for brain-related diseases and traits, despite high cis-genetic correlations of eQTL effect sizes across tissues. Notably, eQTL annotations restricted to loss-of-function intolerant genes from ExAC were even more strongly enriched for disease heritability (17.09x; vs. 5.84x for all genes; $P = 4.90e-17$ for difference). All molecular QTL except sQTL remained significantly enriched for disease heritability in a joint analysis conditioned on each other and on a broad set of functional annotations from previous studies, implying that each of these annotations is uniquely informative for disease and complex trait architectures.

Introduction

Although Genome-wide association studies (GWAS) have been extremely successful in detecting thousands of risk loci for diseases and traits^{1–4}, our understanding of disease architecture is far from complete as most risk loci lie in non-coding regions of genome^{5–9,9,10}. Leveraging molecular phenotypes such as gene expression^{11–16} or chromatin marks^{17–20} can aid in understanding 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^{6,7,21,22}, motivating further work on colocalization^{23–28} and transcriptome-wide association studies^{29–31}. Partitioning heritability using raw genotypes/phenotypes^{32–36} or summary association statistics^{37–40} can aid our understanding of disease architectures, but it is currently unclear how to best leverage molecular QTL from rich resources such as GTEx^{13,16} and BLUEPRINT²⁰ 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 method⁴¹ (allowing for multiple causal variants at a locus) to compute causal posterior probabilities (CPP) for each variant to be a causal cis-QTL. We show that our annotations are far more enriched for disease heritability than standard annotations that include all significant molecular cis-QTL. We further show that our eQTL annotations produce tissue-specific enrichments (despite high cis-genetic correlations of eQTL effect sizes across tissues^{13,42}, and produce much larger enrichments when restricted to loss-of-function intolerant genes from ExAC⁴³. Finally, we quantify the extent to which annotations constructed from eQTL, hQTL, sQTL, and meQTL provide complementary information about disease, conditional on each other and on functional annotations from previous studies.

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 ($\text{FDR} < 5\%$) 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 method⁴¹ (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 ($\text{FDR} < 5\%$) cis-QTL for at least one molecular phenotype^{21,22}; 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 ($\text{FDR} < 5\%$) QTL. We refer to this annotation as TopcisQTL. Further details of each annotation are provided in the Online Methods section.

We apply a previously developed method, stratified LD score regression (S-LDSC)^{37,38}, 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 (τ^*). Enrichment is defined as the proportion of heritability explained by SNPs in an annotation divided by the proportion of SNPs in the annotation³⁷; here, we generalize this definition to probabilistic annotations such as MaxCPP. Standardized effect size (τ^*) is defined as the proportionate change in per-SNP heritability associated to a one standard deviation increase in the value of the annotation³⁸. Unlike enrichment, τ^* is conditional on other annotations included in the model. Further details of each metric are provided in the Online Methods section.

We constructed MaxCPP and other annotations using eQTL data from the GTEx Consortium^{13,16} and eQTL, hQTL, sQTL and meQTL data from the BLUEPRINT Consortium²⁰ (Table

1; see URLs). We included a broad set of 75 functional annotations from the baselineLD model (Table S1) in most analyses. We have made our annotations freely publicly available (see URLs).

Simulations

We performed 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. The S-LDSC model assumes that causal disease effect sizes are i.i.d. conditional on the value of the annotation, but this assumption may be violated for molecular QTL-based annotations. For example, 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).

We performed simulations using real genotypes from the UK Biobank, restricting to 749,024 SNPs on chromosome 1 (see Online Methods). We simulated gene expression phenotypes for 500 individuals assuming that each gene has exactly one causal eQTL (heritability=10%), and 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=50%). We constructed the AllcisQTL, 95%CredibleSet, TopcisQTL and MaxCPP annotations as described above, and 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 Table S2. We determined that S-LDSC estimates are severely upward biased for the TopcisQTL annotation, consistent with the above intuition. On the other hand, S-LDSC estimates were conser-

vative for the AllcisQTL, 95%CredibleSet and MaxCPP annotations; thus, we restrict our analyses to these three annotations in the remainder of the manuscript. Of these three annotations, MaxCPP had the highest enrichment (Figure 1 and Table S2) and highest τ^* (Table S3). We obtained similar results at other values of gene expression heritability and complex trait heritability (Table S4).

Fine-mapped eQTL are strongly enriched across 41 independent diseases and complex traits

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 (Table S5). We applied 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 Table S6). We meta-analyzed results across the 47 data sets, which were chosen to be independent (no highly genetically correlated traits in overlapping samples; see Online Methods) We computed enrichment and τ^* for each annotation, in an analysis that included 75 functional annotations (Table S1) from the baselineLD model³⁸; we note that enrichment has a simpler interpretation, whereas τ^* explicitly conditions on annotations from the baselineLD model and is directly comparable across annotations that span different proportions of SNPs. Results for whole blood, a widely studied tissue, are reported in Figure 2 and Table S7. We determined that the MaxCPP annotation had far higher values of enrichment (5.41x, s.e. = 0.82; $P = 7.68e-08$) and τ^* ($\tau^* = 0.23$, s.e. = 0.05; $P = 2.00e-05$) than AllcisQTL (enrichment = 1.51x and $\tau^* = 0.01$) or 95%CredibleSet (enrichment = 2.18x and $\tau^* = 0.07$). The MaxCPP annotation also had higher values of enrichment and τ^* in analyses that did not condition on the baselineLD model (Figure S1 and Table S8).

We investigated whether the τ^* for MaxCPP in each respective tissue varied with sample size, which determines the statistical power to detect eQTL. We observed a correlation ($R^2 = 0.69$, $P = 1.36e-12$) between sample size and τ^* (Figure 3 and Table S9). This implies that 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 (Meta-Tissue); we note that despite pervasive sample overlap, noise is largely uncorrelated across tissues (e.g. average $r=0.11$ between blood and brain-hippocampus expression despite high cis-genetic correlation of 0.675⁴²). We determined that Meta-Tissue annotations had slightly larger enrichments and much larger τ^* (due to larger annotation size) than annotations constructed from individual tissues (Figure 2, Table S7 and Table S10). As with individual tissues, the MaxCPP annotation had far higher values of enrichment (5.84x, s.e. = 0.40; $P = 1.19e-31$) and τ^* ($\tau^* = 0.52$, s.e. = 0.05; $P = 2.73e-27$) than AllCisQTL(enrichment = 1.80x and $\tau^* = 0.03$) or 95%CredibleSet (enrichment = 2.75x and $\tau^* = 0.17$). A histogram of MaxCPP annotation values for GTEx Meta-Tissue is provided in Figure S2, and correlations with baselineLD model annotations and their LD scores are provided in Figures S3 and S4.

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

Although the 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 Meta-Tissue separately (conditional on the baselineLD model) and meta-analyzing results across the six diseases. We obtained higher estimates of τ^* (and higher or comparable estimates of enrichment) for blood than for Meta-Tissue or any other individual tissue (Table S11). We then analyzed MaxCPP annotations for blood and Meta-Tissue jointly (conditional on the baselineLD model). We obtained a significantly positive τ^* estimate for blood ($\tau^* = 0.91$, s.e. = 0.34; $P = 9.15e-03$) (Figure 4 and Table S12), implying that fine-mapped blood eQTL provides additional information about these six diseases conditional on fine-mapped Meta-Tissue eQTL. We repeated these analyses for the five blood cell traits. When analyzing MaxCPP annotations for blood and Meta-Tissue separately, we obtained higher estimates of τ^* (and higher or comparable estimates of enrichment) for blood than for Meta-Tissue or any other individual tissue (Table S13). When analyzing MaxCPP annotations for blood and Meta-Tissue jointly, we obtained a significantly positive τ^* estimate for blood ($\tau^* = 1.17$, s.e. = 0.24; $P = 1.77e-06$) (Figure 4 and Table S12), implying that fine-mapped blood eQTL provides additional information about these five traits conditional on fine-mapped Meta-Tissue eQTL.

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 Meta-Tissue separately, we obtained higher or comparable estimates of enrichment and τ^* for Brain+Nerve than for Meta-Tissue or any individual tissue (Table S14). When analyzing MaxCPP annotations for Brain+Nerve and Meta-Tissue jointly, we obtained a significantly positive τ^* estimate for Brain+Nerve ($\tau^* = 0.28$, s.e. = 0.07; $P = 9.81\text{e-}05$) (Figure 4 and Table S12), implying that fine-mapped Brain+Nerve eQTL provides additional information about these eight traits conditional on fine-mapped Meta-Tissue eQTL. 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 ($\text{FDR} < 5\%$) tissue-specific τ^* (Tables S15 and S16). 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^{13,42}.

Heritability enrichment of fine-mapped eQTL 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^{43–45}. To investigate the importance of non-coding common variants in these gene sets, for each gene set S we used GTEx 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 variants⁴³. 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 Table S17). In particular, MaxCPP(ExAC) was much more strongly enriched (17.06x, s.e. = 1.28; $P = 1.20\text{e-}35$) than MaxCPP(All Genes) (5.84x) ($P = 4.90\text{e-}17$ 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 eQTL⁴³. We then

analyzed MaxCPP(ExAC) and MaxCPP(All Genes) annotations jointly (conditional on the baselineLD model). We obtained a significantly positive τ^* for MaxCPP(ExAC) ($\tau^* = 0.42$, s.e. = 0.04; $P = 1.40\text{e-}23$; Figure 5B and Table S18), implying that MaxCPP(ExAC) provides additional information about disease heritability conditional on MaxCPP (All Genes). We observed that the effect size (τ^*) for MaxCPP(ExAC) conditional on MaxCPP (All Genes) and baselineLD is five times larger, and more statistically significant, than the τ^* of allSNP (ExAC) conditional on the baselineLD model (Table S24). 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 mutations⁴⁴ (Samocha); a set of 2,984 genes with strong selection against protein-truncating variants⁴⁵ (Cassa); a set of 1,878 genes predicted to be essential based on CRISPR experiments in a human cancer cell line⁴⁶ (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)⁴⁷. 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 Table S17). In addition, for each gene set except the Wang gene set, we obtained a significantly positive τ^* 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, Table S18 and Tables S19-S23). As with the ExAC gene set, the τ^* for MaxCPP(S) conditional on MaxCPP(All Genes) and the baselineLD model were substantially larger than the τ^* of allSNPs(ExAC) conditional on the baselineLD model and were often more statistically significant (Table S24), indicating that MaxCPP can increase power to identify enriched gene sets in which regulatory variants play an important role.

Fine-mapped QTL for a broad set of molecular traits are strongly enriched across 41 independent diseases and complex traits

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; Table S25) as well as a fixed-effect meta-analysis of molecular QTL effect sizes across the 3 cell types (Meta-Tissue). We again used S-LDSC

to assess each annotation’s contribution to disease heritability, meta-analyzed across 41 independent diseases and complex traits (47 data sets). We determined that for each QTL data set the MaxCPP annotation (enrichment = 2.06-6.02x and $\tau^* = 0.02$ -0.45) outperformed the AllcisQTL (enrichment = 1.17-1.56x and $\tau^* = 0.01$ -0.09) and 95%CredibleSet (enrichment = 1.44-2.28x and $\tau^* = -0.01$ -0.16) annotations (Table S26). A histogram of MaxCPP annotation values for each QTL data set is provided in Figure S5. 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; $P = 3.26 \times 10^{-16}$), H3K27ac (4.28x, s.e. = 0.37; $P = 2.59 \times 10^{-19}$), H3K4me1 (4.27x, s.e. = 0.36; $P = 1.29 \times 10^{-20}$), sQTL (3.61x, s.e. = 0.40; $P = 1.39 \times 10^{-10}$), and meQTL (2.81x, s.e. = 0.19; $P = 8.36 \times 10^{-22}$) (Figure 6A and Table S27). Meta-Tissue generally attained higher enrichments and τ^* than MaxCPP computed using each of the three immune cell types individually (Table S28), similar to our GTEx results (Tables S7 and S9). MaxCPP computed using Meta-Tissue also generally outperformed each of the three cell types in a meta-analysis across the six autoimmune diseases (Table S29) and a meta-analysis across the five blood cell traits (Table S30), in contrast to the stronger enrichments for tissue-specific GTEx blood eQTL annotations for blood cell traits (Tables S11 and S13).

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 τ^* remained statistically significant for all molecular QTL except sQTL (Figure 6B and Table S31); a joint analysis of just the five BLUEPRINT molecular QTL (conditional on the baselineLD model) produced similar findings (Table S32). 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 Figure S4), 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 very strongly enriched for disease heritability across 41 diseases and complex traits. 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 heritability of obsessive-compulsive disorder and Tourette Syndrome²¹ as well as type 2 diabetes²². Our findings provide additional motivation for colocalization studies^{23–27} to identify shared causal variants for eQTL and disease and transcriptome-wide association studies (TWAS)^{29–31} to identify genes whose predicted expression is associated to disease. Our fine-mapped eQTL annotations were able to detect tissue-specific enrichments for blood and brain related traits, despite high cis-genetic correlations^{13,42} of eQTL effect sizes across tissues and despite the fact that TWAS have generally concluded that their results “did not suggest tissue-specific enrichment”³¹.

We note that a previous study showed that cis-eQTL often lie close to the transcription start site (TSS) or transcription end site (TES)⁴⁸; we did not observe additional disease signal for cis-eQTL variants close to the TSS or TES in the GTEx or BLUEPRINT data sets (see Tables S33 and S34). 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.

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^{49–53}. We note that 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 eQTL⁵⁴, motivating further investigation in larger sQTL data sets. Finally, although we have used the CAVIAR method⁴¹ for fine-mapping molecular QTL, our results motivate a comparison of disease/trait enrichment using other QTL fine-mapping methods^{55–59}.

We note several limitations of our work. First, we restrict our analyses to common variants, as S-LDSC is not currently applicable to rare variants³⁷. Recent work has shown that rare variants can have substantial effects on gene expression⁶⁰, motivating ongoing work to extend 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. Third, we show that S-LDSC is generally unable to produce unbiased enrichment estimates for molecular QTL based annotations, which violate its model assumptions; S-LDSC produces slightly conservative enrichment estimates for the MaxCPP annotation that we focus on here, although 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; analyses of larger sample sizes and/or different tissues or contexts may produce larger enrichments in the future. Fifth, although we analyzed very large GWAS samples sizes (average N=320K), power increases only modestly with GWAS sample size at very large sample sizes, as the finite size of the genome is a stricter constraint⁶¹. Sixth, we performed a fixed-effect meta-analysis of molecular QTL effect sizes across tissues (Meta-Tissue), but performing fixed-effect meta-analysis with overlapping samples can limit power⁶²⁻⁶⁴. We believe that this is a minor issue in the current context, as noise is largely uncorrelated across tissues, but further investigation using random-effect meta-analysis is warranted. 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.

Acknowledgements

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

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 $\mathbf{a}$ 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), $0 \leq a_j \leq 1$.

MaxCPP annotation: Let $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_g)$ 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 $\mathbf{s}_i$ be the vector of marginal association statistics of gene i for all the cis-variants. Utilizing $\mathbf{s}_i$ 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 CAVIAR⁴¹. 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., $\theta_{ji}=1$ indicates that the SNP j is in the gene i 95%credible set and $\theta_{ji}=0$ 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: AllcisQTL annotation is a binary annotation, where any variant whose marginal association statistic for at least one gene passes the significance threshold ($\text{FDR} < 0.05$) 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 $\text{FDR}(j) < 0.05$ and $t_{ij} = 0$ otherwise). More formally, we have: $a_j = \max_i t_{ji}$.

95%CredibleSet annotation: 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 annotation: 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., $\gamma_{ji}=1$ if SNP j is the most significant SNP among all cis-variants for the gene i and $\gamma_{ji}=0$ otherwise). More formally, we have: $a_j = \max_i \gamma_{ji}$.

Enrichment and effect size (τ^*) metrics

We use two metrics to measure the importance of an annotation in the context of diseases and complex traits: Enrichment and effect size (τ^*) of annotation. We use S-LDSC^{37,38} to compute enrichment and effect size (τ^*). Let a_{cj} indicate the annotation value of the SNP j for the annotation c . S-LDSC^{37,38} assumes that the variance of each SNP is a linear additive contribution to each annotation:

$$\text{Var}(\beta_j) = \sum_c a_{cj} \tau_c \quad (1)$$

where τ_c is the contribution of the annotation c to per-SNP heritability. S-LDSC^{37,38} estimates τ_c using the following equation:

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

where N is the GWAS sample size and $\ell(j, c)$ is the LD score of the SNP j for the annotation c . S-LDSC computes the LD scores as follow: $\ell(j, c) = \sum_k a_{ck} r_{jk}^2$ 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.³⁸ defined τ_{c^*} for

an annotation as the per-standardized annotation effect size:

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

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:

$$\text{Enrichment} = \frac{\%h_g^2(c)}{\%\text{SNP}(c)} = \frac{\frac{h_g^2(c)}{h_g^2}}{\frac{\sum_j a_{jc}}{M}} \quad (4)$$

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

$$h_g^2(c) = \sum_j a_{jc} \text{Var}(\beta_j) = \sum_j a_{jc} \left(\sum_c a_{jc} \tau_c \right) \quad (5)$$

Although enrichment has a simpler interpretation than τ_{c*} , τ_{c*} is computed by conditioning on other annotations in the model. Thus, τ_{c*} captures the signal that is unique to the annotation c and is not captured by other annotations in the model. For example, two annotations that are highly correlated will have similar enrichment, while τ_{c*} for the annotation that better captures true causal variants will be larger and more statistically significant.

Simulation framework

In our simulation, 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. We implanted one causal variant per each gene for all the 4,493 genes on chromosome 1, and assumed that the set of causal trait variants was exactly the same set of 4,493 variants.

We set the total heritability for each gene and trait phenotype to be σ_{eQTL}^2 and σ_{GWAS}^2 , respectively. We simulated causal trait effect sizes using polygenetic model, $\beta_i \sim N(0, \frac{\sigma_{GWAS}^2}{n_c})$, where β_i is the true effect size of i -th causal variant and n_c is the number of causal variants. In the case of gene expression, we simulated 4,493 phenotypes to represent our simulated gene expression data and we set n_c to one for each gene. In the case of trait phenotypes, we set n_c to 4,493 as the number of causal variant is equal to the number of genes in chromosome 1. In our simulated data sets, we set σ_{GWAS}^2 to 0.5 and considered different values of σ_{eQTL}^2 to test our results on different input parameters. We set σ_{eQTL}^2 to 0.05, 0.1, and 0.2 resulting in three different simulated data sets. 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 software⁶⁵ (see URLs). In the case of simulated trait phenotypes, we computed the association of each variant with the simulated trait phenotypes. In the case of simulated gene expression phenotypes, we computed the marginal statistics for all variants within 1Mb of the TSS. We generated the four annotations as described above. We used CAVIAR⁴¹ to generate the 95% Credible set and MaxCPP annotations. We utilized European samples from the 1000 Genomes Project (1000G)⁶⁶ (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.

After obtaining the four annotations, we ran S-LDSC³⁷ to generate the LD score of each variant in each annotation using the same procedure described in the previous studies^{37,38}. Regression SNPs, which are used by S-LDSC to estimate τ from marginal association statistics, were obtained from the HapMap Project phase 3⁶⁷. These SNPs are considered as well-imputed SNPs. SNPs with marginal association statistics larger than 80 or larger than $0.001N$ 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 1000G⁶⁶. Heritability SNPs, which are used to estimate $sd(c)$ and h_g^2 , were defined as common variants ($MAF \geq 0.05$) 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 τ^* for each simulation. Then, we compared the S-LDSC estimated enrichment and true enrichment for each annotation. All results are averaged across 400 simulations.

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 which summary association statistics were computed using BOLT-LMM^{35,68} (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 (Table S6). The 47 data sets included 6 traits that were duplicated in two different data sets (genetic correlation of at least 0.9, Table S6). Thus, we analyzed 41 independent diseases and complex traits. We ran S-LDSC using the same procedures described in previous studies^{37,38} (see above). Most analyses included the baselineLD model v1.1, which is identical to the baselineLD model as previously described³⁸ except that we fixed an error in the promoter annotation (inherited from previous studies^{34,37}); we determined that fixing this error did not affect our results. 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 LDSC⁶⁹. The meta-analyzed values of enrichment and τ^* across the 47 data sets were computed using a random-effect meta-analysis, as implemented in the *rmeta* R package.

Blood and brain related diseases and complex traits

Autoimmune diseases: We analyzed six autoimmune diseases: Crohn’s disease⁷⁰, Rheumatoid arthritis (ref.⁷¹ and UK Biobank), Ulcerative colitis⁷⁰, Lupus⁷², Celiac⁷³, and all autoimmune and inflammatory diseases in UK Biobank).

Blood cell traits: 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.

Brain-related diseases and complex traits: We analyzed eight independent brain-related diseases and complex traits: Age at menarche⁷⁴, BMI (ref.⁷⁵ and UK Biobank), Bipolar Disorder⁷⁶, Depressive symptoms³, Neuroticism (UK Biobank), Schizophrenia¹, Smoking Status (ref.⁷⁷ and UK Biobank), and Year of education (ref.⁷⁸ and UK Biobank). These traits are a subset of traits from Table S6 that were reported to be brain-enriched^{37,61}.

URLs

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

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

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

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

QTL-based annotations: <https://data.broadinstitute.org/alkesgroup/LDSCORE/>

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/UKB>

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

References

1. Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* *511*, 421–427.
2. Price, A. L., Spencer, C. C. A., and Donnelly, P. (2015). Progress and promise in understanding the genetic basis of common diseases. *Proceedings of the Royal Society B: Biological Sciences* *282*, 20151684.
3. Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S. F. W., et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* *533*, 539–542.
4. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics* *101*, 5–22.
5. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* *106*, 9362–9367.
6. Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics* *6*, e1000888.
7. Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., and Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics* *6*, e1000895.
8. Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* *337*, 1190–1195.

9. Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B. E., Klein, R. J., Han, B., and Raychaudhuri, S. (2015). Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *The American Journal of Human Genetics* *97*, 139–152.
10. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* *518*, 317–330.
11. Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.-H., et al. (2014). Heritability and genomics of gene expression in peripheral blood. *Nature Genetics* *46*, 430–437.
12. Zhang, X., Joehanes, R., Chen, B. H., Huan, T., Ying, S., Munson, P. J., Johnson, A. D., Levy, D., and O'Donnell, C. J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nature Genetics* *47*, 345–352.
13. Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., et al. (2015). The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* *348*, 648–660.
14. Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics* *48*, 1094–1100.
15. Zhernakova, D. V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van t Hof, P., Mei, H., van Dijk, F., Westra, H.-J., et al. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics* *49*, 139–145.
16. Aguet, F., Brown, A. A., Castel, S., Davis, J. R., Mohammadi, P., Segre, A. V., Zappala, Z., Abell, N. S., Fresard, L., Gamazon, E. R., et al. (2016). Local genetic effects on gene expression across 44 human tissues. *bioRxiv*.

17. McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science* *342*, 747–749.
18. Waszak, S. M., Delaneau, O., Gschwind, A. R., Kilpinen, H., Raghav, S. K., Witwicki, R. M., Orioli, A., Wiederkehr, M., Panousis, N. I., Yurovsky, A., et al. (2015). Population variation and genetic control of modular chromatin architecture in humans. *Cell* *162*, 1039–1050.
19. Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., Greenside, P., Srivas, R., Phanstiel, D. H., Pekowska, A., et al. (2015). Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* *162*, 1051–1065.
20. Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* *167*, 1398–1414.e24.
21. Davis, L. K., Yu, D., Keenan, C. L., Gamazon, E. R., Konkashbaev, A. I., Derks, E. M., Neale, B. M., Yang, J., Lee, S. H., Evans, P., et al. (2013). Partitioning the heritability of tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genetics* *9*, e1003864.
22. Torres, J. M., Gamazon, E. R., Parra, E. J., Below, J. E., Valladares-Salgado, A., Wachter, N., Cruz, M., Hanis, C. L., and Cox, N. J. (2014). Cross-tissue and tissue-specific eQTLs: Partitioning the heritability of a complex trait. *The American Journal of Human Genetics* *95*, 521–534.
23. He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., and Li, H. (2013). Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS. *The American Journal of Human Genetics* *92*, 667–680.
24. Hu, X., Kim, H., Raj, T., Brennan, P. J., Trynka, G., Teslovich, N., Slowikowski, K., Chen, W.-M., Onengut, S., Baecher-Allan, C., et al. (2014). Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in CD4 effector memory t cells. *PLoS Genetics* *10*, e1004404.

25. Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics* *10*, e1004383.
26. Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics* *99*, 1245–1260.
27. Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., Jager, P. L. D., Sunyaev, S. R., and Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nature Genetics* *49*, 600–605.
28. Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLOS Genetics* *13*, e1006646.
29. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* *47*, 1091–1098.
30. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* *48*, 245–252.
31. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *The American Journal of Human Genetics* *100*, 473–487.
32. Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* *43*, 519–525.

33. Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., Keller, M. C., Visscher, P. M., and Wray, N. R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics* *44*, 247–250.
34. Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics* *95*, 535–552.
35. Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., et al. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* *47*, 284–290.
36. Won, H.-H., Natarajan, P., Dobbyn, A., Jordan, D. M., Roussos, P., Lage, K., Raychaudhuri, S., Stahl, E., and Do, R. (2015). Disproportionate contributions of select genomic compartments and cell types to genetic risk for coronary artery disease. *PLOS Genetics* *11*, e1005622.
37. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* *47*, 1228–1235.
38. Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B. M., Gusev, A., et al. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics* *49*, 1421–1427.
39. Pasaniuc, B. and Price, A. L. (2016). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* *18*, 117–127.
40. Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell* *169*, 1177–1186.

41. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* *198*, 497–508.
42. Liu, X., Finucane, H. K., Gusev, A., Bhatia, G., Gazal, S., O'Connor, L., Bulik-Sullivan, B., Wright, F. A., Sullivan, P. F., Neale, B. M., et al. (2017). Functional architectures of local and distal regulation of gene expression in multiple human tissues. *The American Journal of Human Genetics* *100*, 605–616.
43. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., et al. (2016). Analysis of protein-coding genetic variation in 60, 706 humans. *Nature* *536*, 285–291.
44. Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics* *46*, 944–950.
45. Cassa, C. A., Weghorn, D., Balick, D. J., Jordan, D. M., Nusinow, D., Samocha, K. E., O'Donnell-Luria, A., MacArthur, D. G., Daly, M. J., Beier, D. R., et al. (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nature Genetics* *49*, 806–810.
46. Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., and Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science* *350*, 1096–1101.
47. Hormozdiari, F., Zhu, A., Kichaev, G., Segre, A. V., Ju, C. J.-T., Joo, J. W., Won, H., Sankararaman, S., Pasaniuc, B., Shiffman, S., et al. (2017 (In press)). Widespread allelic heterogeneity in complex traits. *The American Journal of Human Genetics*.
48. Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics* *4*, e1000214.
49. Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., et al. (2016). Lineage-specific genome

- architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* *167*, 1369–1384.e19.
50. Shooshtari, P., Huang, H., and Cotsapas, C. (2017). Integrative genetic and epigenetic analysis uncovers regulatory mechanisms of autoimmune disease. *The American Journal of Human Genetics* *101*, 75–86.
51. Gusev, A., Mancuso, N., Finucane, H. K., Reshef, Y., Song, L., Safi, A., Oh, E., , McCarroll, S., Neale, B., et al. (2016). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *bioRxiv*.
52. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., Reshef, Y., Song, L., Safi, A., of the Psychiatric Genomics Consortium, S. W. G., McCarroll, S., et al. (2016). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature Genetics* (In Press).
53. Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature Genetics*.
54. Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., and Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* *352*, 600–604.
55. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *Plos Genetics*.
56. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. *Bioinformatics* *31*, i206–i213.
57. Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A., and Schaid, D. J. (2015). Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics* pp. genetics–115.

58. Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* *32*, 1493–1501.
59. Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics* *98*, 1114 – 1129.
60. Li, X., Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C., Zappala, Z., Strober, B. J., Scott, A. J., Ganna, A., et al. (2016). The impact of rare variation on gene expression across tissues. *bioRxiv*.
61. Finucane, H., Reshef, Y., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shores, N., et al. (2017). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *bioRxiv*.
62. Han, B. and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics* *88*, 586–598.
63. Sul, J. H., Han, B., Ye, C., Choi, T., and Eskin, E. (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genetics* *9*, e1003491.
64. Han, B., Duong, D., Sul, J. H., de Bakker, P. I. W., Eskin, E., and Raychaudhuri, S. (2016). A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Human Molecular Genetics* *25*, 1857–1866.
65. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* *81*, 559–575.
66. Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti,

- A., Clark, A. G., Collins, F. S., Vega, F. M. D. L., Donnelly, P., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* *467*, 1061–1073.
67. Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* *467*, 52–58.
68. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., and Price, A. L. (2017). Mixed model association for biobank-scale data sets. *bioRxiv*.
69. Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* *47*, 1236–1241.
70. Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., et al. (2012). Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* *491*, 119–124.
71. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al. (2013). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* *506*, 376–381.
72. Bentham, J., Morris, D. L., Graham, D. S. C., Pinder, C. L., Tomblinson, P., Behrens, T. W., Martín, J., Fairfax, B. P., Knight, J. C., Chen, L., et al. (2015). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics* *47*, 1457–1464.
73. Dubois, P. C. A., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G. A. R., Ádány, R., Aromaa, A., et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics* *42*, 295–302.
74. Day, F. R., Thompson, D. J., Helgason, H., Chasman, D. I., Finucane, H., Sulem, P., Ruth, K. S., Whalen, S., Sarkar, A. K., Albrecht, E., et al. (2017). Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nature Genetics* *49*, 834–841.

75. Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* *42*, 937–948.
76. Sklar, P., Ripke, S., Scott, L. J., Andreassen, O. A., Cichon, S., Craddock, N., Edenberg, H. J., Nurnberger, J. I., Rietschel, M., Blackwood, D., et al. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near *odx4*. *Nature Genetics* *43*, 977.
77. Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N., Ardissino, D., Bernardinelli, L., Mannucci, P. M., Mauri, F., Merlini, P. A., et al. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics* *42*, 441–447.
78. Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* *340*, 1467–1471.
79. Consortium, E. P. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature* *489*, 57–74.
80. Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., et al. (2012). Integrative annotation of chromatin elements from ENCODE data. *41*, 827–841.
81. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S. (2012). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics* *45*, 124–130.
82. Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., and Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell* *155*, 934–947.

83. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* *478*, 476–482.
84. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* *507*, 455–461.
85. Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using *gerp++*. *PLOS Computational Biology* *6*, 1–13.

Tables

Dataset	QTL Type	Number of Tissues	<i>N</i> (per tissue)	<i>N</i> (total)
GTE _x	eQTL	44	70-361	7014
BLUEPRINT	eQTL	3	169-194	555
BLUEPRINT	hQTL (H3K27 _{ac})	3	143-174	479
BLUEPRINT	hQTL (H3K4 _{me1})	3	104-173	449
BLUEPRINT	sQTL	3	169-194	555
BLUEPRINT	meQTL	3	132-197	525

Table 1. List of molecular QTL data sets analyzed. GTE_x 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 Table S5 (GTE_x) and Table S25 (BLUEPRINT).

Figures

Figure 1. S-LDSC estimates for TopcisQTL are upward biased in simulations. Panels (A), (B), (C), and (D) illustrate the true enrichment and S-LDSC enrichment estimates for AllcisQTL, 95%CredibleSet, TopcisQTL, and MaxCPP annotations, respectively. Error bars represent 95% confidence intervals. Numerical results are reported in Table S2.

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 Meta-Tissue from the GTEx data set conditioning on the baselineLD model. (B) Meta-analysis results across 41 traits of τ^* for whole blood and 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 Table S7.

Figure 3. Relationship between eQTL sample size and the annotation effect size (τ^*). For each tissue, we plot the τ^* of the MaxCPP annotation, meta-analyzed across 41 traits, against the eQTL sample size. Numerical results are reported in Table S9. 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) τ^* of 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. 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 Table S12.

Figure 5. Heritability enrichment of fine-mapped eQTL is concentrated in disease-relevant gene sets. Meta-analysis results of (A) enrichment and (B) τ^* 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, τ^* estimates are reduced by conditioning on MaxCPP(All Genes), but enrichment estimates are not affected. 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 Table S17.

Figure 6. Fine-mapped eQTL, hQTL, sQTL, and meQTL annotations are enriched for disease/trait heritability. Meta-analysis results of (A) enrichment and (B) τ^* 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, τ^* estimates are reduced by conditioning on MaxCPP for all molecular QTL, but enrichment estimates are not affected. 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 Table S27 and Table S31.

A

Enrichment

AllcisQTL**B** **95%CredibleSet****C** **TopcisQTL****D** **MaxCPP**

■ True Enr.
□ S-LDSC

A

B

A

B

A

Enrichment

BMeta (τ^*)