Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery

Hakhamanesh Mostafavi[*,1], Jeffrey P. Spence[1], Sahin Naqvi[1,2], Jonathan K. Pritchard[*,1,3]

[1] Department of Genetics, Stanford University, Stanford CA

[2] Department of Chemical and Systems Biology, Stanford University, Stanford CA

[3] Department of Biology, Stanford University, Stanford CA

[*] To whom correspondence should be addressed:

hmostafavi@stanford.edu, pritch@stanford.edu

May 8, 2022

Abstract

Most signals in genome-wide association studies (GWAS) of complex traits point to noncoding genetic variants with putative gene regulatory effects. However, currently identified expression quantitative trait loci (eQTLs) explain only a small fraction of GWAS signals. By analyzing GWAS hits for complex traits in the UK Biobank, and cis-eQTLs from the GTEx consortium, we show that these assays systematically discover different types of genes and variants: eQTLs cluster strongly near transcription start sites, while GWAS hits do not. Genes near GWAS hits are enriched in numerous functional annotations, are under strong selective constraint and have a complex regulatory landscape across different tissue/cell types, while genes near eQTLs are depleted of most functional annotations, show relaxed constraint, and have simpler regulatory landscapes. We describe a model to understand these observations, including how natural selection on complex traits hinders discovery of functionally-relevant eQTLs. Our results imply that GWAS and eQTL studies are systematically biased toward different types of variants, and support the use of complementary functional approaches alongside the next generation of eQTL studies.

1   Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants linked with a variety of human complex traits and diseases [1]. However, uncovering the functional mechanisms of these GWAS hits remains challenging, notably because ∼90% of trait-associated variants lie in noncoding regions [2]. GWAS signals are predominantly located in open chromatin regions in relevant cell types, and are enriched in gene regulatory elements and expression quantitative trait loci (eQTLs). These observations suggest that most GWAS hits are mediated by altering gene regulation of nearby genes [2–7].

Motivated by these observations, many studies have integrated GWAS findings with eQTL mapping to gain functional understanding of trait-associated variants [8–12]. However, despite extensive efforts to catalog eQTLs across diverse sets of biosamples, particularly those conducted by the GTEx consortium [13,14], most GWAS hits are not explained by currently known eQTLs [15–17]. For example, across a number of traits analyzed in the GTEx project, just 43% of GWAS hits and a median of 21% of GWAS hits per trait, colocalized with eQTLs (i.e., indicate the same genetic variant) [14]. Similarly, it has been estimated that only 11% of heritability for complex traits is mediated by gene expression in GTEx tissues [18].

Multiple potential explanations have been proposed for the limited overlap between GWAS hits and eQTLs. One is that some GWAS hits may only be eQTLs in specific contexts, e.g., during development [16, 19–22], in specific cell types [23–26], or in response to physiological stimuli such as immune responses [27–31]. These effects are expected to be absent, or hard to detect, in conventional eQTL assays using post-mortem adult whole tissues. Although hundreds of context-specific eQTLs have been identified, their contribution in explaining trait-associated variants has thus far been modest [17,31]. For example, eQTLs identified at multiple stages of induced pluripotent stem cell (iPSC) differentiation towards neuronal cell types added ∼10% more colocalizations with neurological trait loci beyond GTEx eQTLs [22]. Also in GTEx, cell type-specific eQTLs colocalized with ∼8% of GWAS hits [25]. While context-specific effects undoubtedly contribute to the limited colocalization, it remains to be seen how much of the gap can be resolved by deeper sampling of cell types and contexts.

A different type of explanation is that perhaps many trait-relevant eQTLs have not yet been discovered due to incomplete statistical power, even if mapping was performed in the correct contexts [18,32]. However, others have proposed that eQTL mapping is already "saturated" in well-studied tissues [16].

Alternatively, effects on complex traits could be driven by processes other than gene expression, such as splicing [33] and polyadenylation [34]. However, so far those other mechanisms explain fewer trait-associated variants than do eQTLs [14,31]. Lastly, with current sample sizes, eQTL studies are mainly powered to detect cis-eQTLs (affecting nearby genes), while many trait-relevant variants may act as trans-eQTLs (affecting genes elsewhere in the genome) [35–37]. However, standard models of gene regulation predict that trans-eQTLs should be mediated indirectly through cis effects on nearby genes, and thus such variants should in principle be discoverable as cis-eQTLs [16, 38]. Together, these observations suggest that most GWAS hits are indeed cis-eQTLs, but that many have not yet been discovered in eQTL mapping.

To better understand the lack of overlap between GWAS hits and eQTLs, we analyzed GWAS data for 44 complex traits in the UK Biobank, and eQTL data for 38 tissues in the GTEx data. We show that in fact GWAS hits and eQTLs differ systematically: GWAS hits lie at relatively long distances from transcription start sites (TSSs); they are enriched near genes with a wide variety

of functional annotations, such as transcription factors; they are under strong selective constraint; and they typically have complex regulatory landscapes across different tissues and cell types. In contrast, eQTLs are tightly clustered near the TSSs of genes that are typically depleted of most functional annotations, show relaxed selective constraint, and have simpler regulatory landscapes.

We close with a model of variant discovery in GWAS and eQTL assays to explain these observations. We show that even if genetic effects on complex traits were entirely mediated by gene expression, many GWAS hits would not be discovered as significant eQTLs (even in the correct causal contexts). One important reason is that natural selection at constrained genes has very different impacts on GWAS discovery than in eQTL discovery.

In summary, GWAS and eQTL mapping tend to maximize power for different types of variants, and current eQTL mapping has limited discovery at the most functionally-important variants. While further context-specific eQTL studies will help somewhat in explaining GWAS hits, we argue here that these efforts should be complemented by a variety of other functional approaches.

2   Results

Analysis overview.   For GWAS analyses, we used publicly available summary statistics for 44 traits in the UK Biobank (Table S1). For eQTL analyses, we used the GTEx v8 data for 38 tissues [14] (Table S1), focusing on cis-eQTLs linked to 18,332 protein-coding genes (Table S2). To make the GWAS hits and eQTLs more comparable, we used identical quality control and SNP selection pipelines for both datasets. We removed lead SNPs in strong LD with protein-altering variants, in order to focus on variants that most likely act through gene expression. This pipeline resulted in 22,119 GWAS hits across traits, and 118,996 eQTLs across all gene-tissue pairs (Tables S3,S4). See Methods for all details.

For each GWAS hit and each eQTL, we evaluated properties of the lead SNP with respect to various SNP and genic features. To study genic features, we linked each GWAS hit to the closest TSS among the same 18,332 protein-coding genes. Although we usually do not know the true causal genes for GWAS hits, recent work has shown that using the nearest gene is an extremely useful proxy [39]. In contrast, for eQTLs, we do of course know the relevant SNP-gene pairs. However to make the analyses comparable between GWAS and eQTLs, in most analyses we masked the true eGenes, and instead linked the eQTL SNP to the *nearest* gene in the same way that we did for GWAS hits. We refer to the genes assigned this way as "GWAS genes" and "eQTL genes" in the rest of this paper.

We observed several systematic differences between eQTL and GWAS lead SNPs (Figure S1): specifically, that eQTL SNPs have higher MAF (median 0.24 compared to 0.2), and are in more gene-dense regions (median 11 genes per Mb compared to 8) than GWAS lead SNPs. These differences could confound our comparisons if they covary with features of interest. Therefore, in all our analyses we included control SNPs, for both GWAS hits and eQTLs, that are matched for MAF, LD score and gene density.

Figure 1 provides an overview of our analysis pipeline and main observations. At the end of the paper, we present models to conceptualize these results.
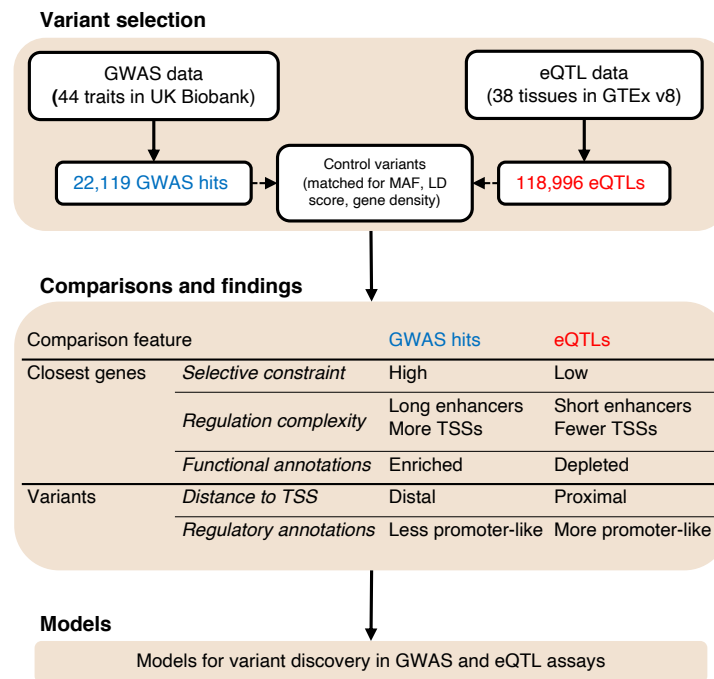
Figure 1: **Study workflow and key results.**
Overview of our pipeline for the analysis of GWAS hits and eQTLs. The list of traits and tissues can be found in Table S1. We compare GWAS hits and eQTLs with respect to a number of genic and SNP features. Summary of our main results is presented in this figure. We describe models for variant discovery in GWAS and eQTL assays to conceptualize these results. See Methods for details.

Selectively constrained genes are enriched in GWAS genes but depleted in eQTL genes.
Previous studies have suggested that the genetic architecture of most complex human traits is shaped by natural selection, such that mutations with large effect sizes are kept at lower frequencies than would be expected in the absence of selection [40–44]. Consistent with selection, SNP heritability for complex traits is enriched near genes depleted of loss-of-function (LoF) variants [45–47], as measured by the pLI score [45]. Although pLI is not an accurate quantitative measure of selection acting on a gene [48], genes with high pLI scores serve as reasonable candidates for being genes under strong selection.

Consistent with previous findings, we observed that GWAS genes are enriched for high-pLI genes (pLI > 0.9) (Figure 2A): 26% of genes linked with GWAS hits compared to 21% for genes linked to matched control SNPs. In contrast, eQTL genes are depleted of high-pLI genes: 12% of genes linked with eQTLs compared to 18% for matched SNPs. These results are in line with previous reports that genes without detectable eQTLs have relatively higher pLI [37, 45].

In this and most following analyses, we considered GWAS hits pooled across all 44 traits, and eQTLs pooled across all 38 tissues and all genes. We repeated our analysis for each trait/tissue separately, showing that trends in Figure 2A are not driven by a few outlier traits/tissues (Figure 2B). Notably, the depletion of high-pLI genes near eQTLs is replicated in all tissues.

We did, however, see a slight shift towards the null, i.e., less depletion of high-pLI genes, for tissues with larger number of eQTLs (Figure 2B). This is likely due to the inclusion of weaker eQTLs that reach significance detection levels in tissues with larger sample sizes, considering that the depletion of high-pLI genes is increasingly pronounced among top ranked eQTLs (Figure 2C). We observed the same trend for eGenes (true target genes, instead of the closest genes), suggesting

4

that these results are not driven by less accurate gene assignments for weaker eQTLs.

Together, these results show that although a substantial fraction of GWAS hits likely regulate genes that are under strong selection, most discovered eQTLs are not linked with such genes. This suggests that eQTLs with large effects on constrained genes are purged by selection, and is consistent with recent work showing that the fraction of trait heritability estimated to be mediated via gene expression is mostly dominated by genes with low cis-heritability for expression levels [18, 46].
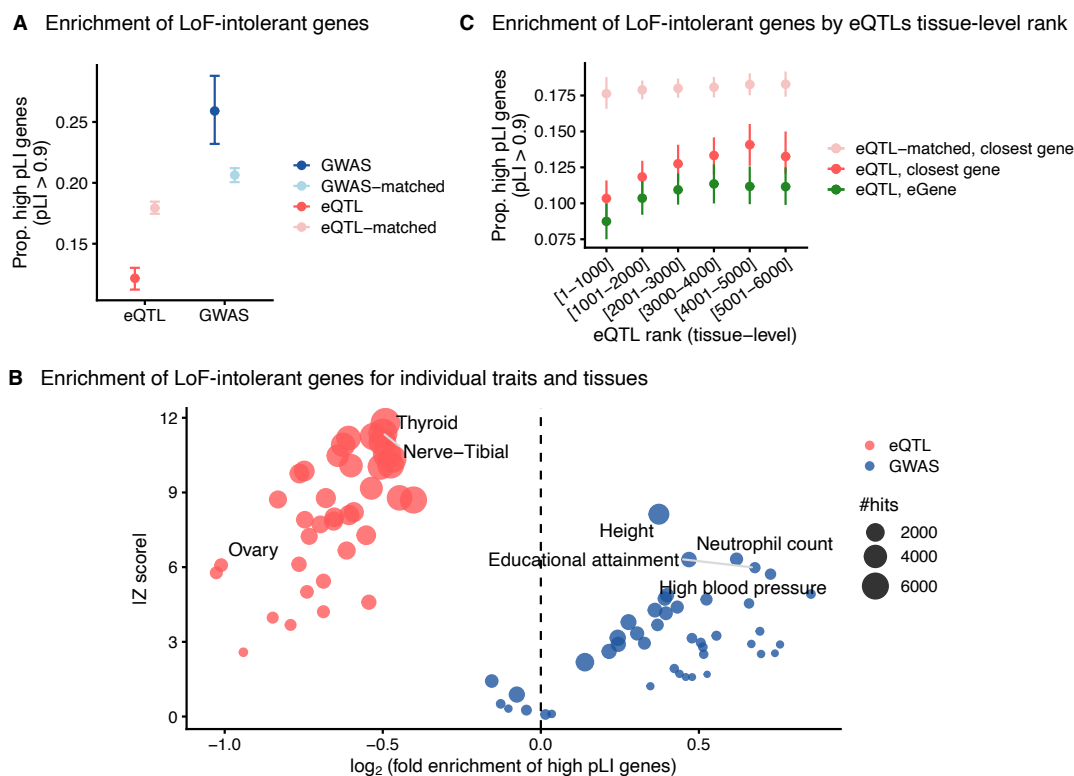

Figure 2: **GWAS and eQTL genes are under different selective constraints.** *A) Fraction of genes with high pLI (pLI > 0.9, a measure of selective constraint) among genes nearest to GWAS hits (blue), eQTLs (red), and control SNPs matched for MAF, LD score and gene density. B) Enrichment of high-pLI genes in GWAS genes for individual traits and eQTL genes for individual tissues. Enrichment values (on the x-axis) and Z scores (on the y-axis) were computed based on values observed in 1000 sampling iterations of control SNPs. C) Fraction of genes with high pLI among eQTL genes (closest gene, red), eGenes (green) and nearest genes to control SNPs (light red) versus eQTL tissue-level rank. For groupings on the x-axis, we first bin ranked eQTLs by association p-values in groups of 1000 eQTLs, and then pooled eQTLs across tissues by the ranked bins. In panels (A) and (C), for GWAS hits and eQTLs, error bars show 95% confidence intervals as determined by quantile bootstrapping. For control SNPs, points and error bars show mean values and 95% confidence intervals in 1000 sampling iterations (see Methods).*

GWAS genes have more complex regulatory landscapes than eQTL genes. As we showed in Figure 2, genes near GWAS hits and eQTLs differ markedly in terms of their levels of selective constraint. Recently, Wang and Goldstein demonstrated that genes near GWAS hits and eGenes also differ with respect to features of their linked enhancer domains [49]. To explore this further, we considered the transcriptional regulatory landscapes of genes, defined based on the variation in enhancer activity and transcription start site (TSS) usage across diverse tissues and cell types. Many enhancers regulate gene expression in a context-specific manner [50]. Similarly, alternative

TSSs could direct context-specific expression, and may also lead to different isoforms [51, 52].

We used the enhancer-gene links in the Roadmap Epigenomics Consortium dataset inferred by Liu et al. [53]. For each gene, we computed (i) the number of active tissue/cell types (out of 111), in which the gene is linked to > 1 enhancer elements, and (ii) the total length of linked enhancers per active tissue/cell type (Figure 3A). We considered these two features jointly using a logistic regression model to distinguish GWAS hits and eQTLs from random SNPs, controlling for MAF, LD score and gene density (Methods). In this framework, compared to genes linked with random SNPs, GWAS genes have longer enhancer regions per active tissue/cell type, while eQTLs have shorter enhancers (Figure 3B), consistent with Wang and Goldstein [49].

In contrast, both types of genes have enhancer activity across more tissue/cell types relative to genes linked with random SNPs (Figure 3B), though presumably for unrelated reasons: genes with broad enhancer activity are more likely to harbor regulatory variants in finite sets of tissue/cell types sampled in eQTL datasets. On the other hand, such genes are possibly involved in multiple biological pathways, thus more likely to contribute to complex traits. These results are replicated when using enhancer-gene links from the activity-by-contact model [54] (Figure S2).

We then considered another regulatory feature: how many different TSSs are used for a given gene across a diverse set of cell types. To this end, we computed TSS counts in the FANTOM project [55] (Methods). On average, GWAS genes harbor more TSSs than eQTL genes (6.4 versus 4.4; Figure 3C), and this difference is not explained by the difference observed for matched SNPs (Figure 3C).

We hypothesized that the regulatory landscape of a given gene, in part, corresponds with its regulatory function in a broader gene regulatory network, such that expression of genes with more regulatory domains is more likely to covary with the expression of other genes in the network (e.g., being a target of multiple regulators). Moreover, genes with many downstream regulatory connections in the network, are expected to be important contributors to heritability [35–37]. Motivated by these considerations, we analyzed co-expression networks inferred by Saha et al. for GTEx v6 tissues [56]. For each gene, we constructed a connectedness measure based on the gene's number of neighbors in tissue-specific networks (Methods). Focusing on 9,422 genes with at least one connection across tissues, we found that genes with more connections are progressively enriched near GWAS hits (Figure 3D), consistent with previous reports that trait heritability is enriched near such genes [57, 58]. In contrast, genes with more connections are progressively depleted for eQTLs (Figure 3D).

Taken together, these observations suggest that targets of GWAS hits are often genes with complex regulatory architecture, in a sense that they harbor regulatory mechanisms to control and diversify gene expression across different contexts and possibly in a context-specific manner, which could correspond with a functional role in gene regulatory networks. These kinds of genes are depleted of eQTLs.
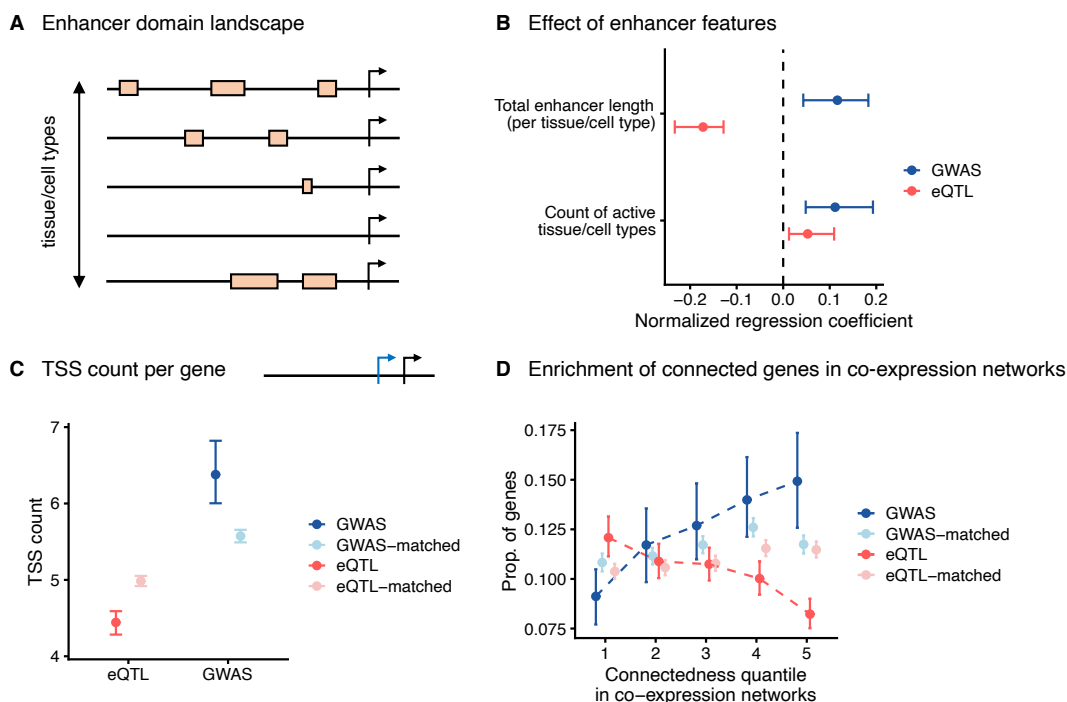
Figure 3: **GWAS and eQTL genes have different transcriptional regulatory landscapes.**
A) Schematic of enhancer activity across a number of tissue/cell types. Based on enhancer-gene links inferred from the Roadmap dataset [53], for a given gene, we computed (i) the number of tissue/cell types in which a gene has an enhancer, and (ii) the average total enhancer length (in base pairs) across tissue/cell types with enhancer activity. B) Logistic regression coefficients corresponding with the two enhancer features described in (A) for predicting GWAS hits (blue) and eQTLs (red) versus random SNPs after adjusting for confounders (see Methods). C) Mean count of TSSs per gene across cell types in the FANTOM project [55] for GWAS and eQTL genes. D) Fraction of GWAS and eQTL genes in gene bins ranked by connectedness values computed based on co-expression networks from Saha et al. [56]. In panels (B-D), for GWAS hits and eQTLs, error bars show 95% confidence intervals as determined by quantile bootstrapping. For matched SNPs (for MAF, LD score and gene density, shown in light blue and red colors), points and error bars show mean values and 95% confidence intervals in 1000 sampling iterations.
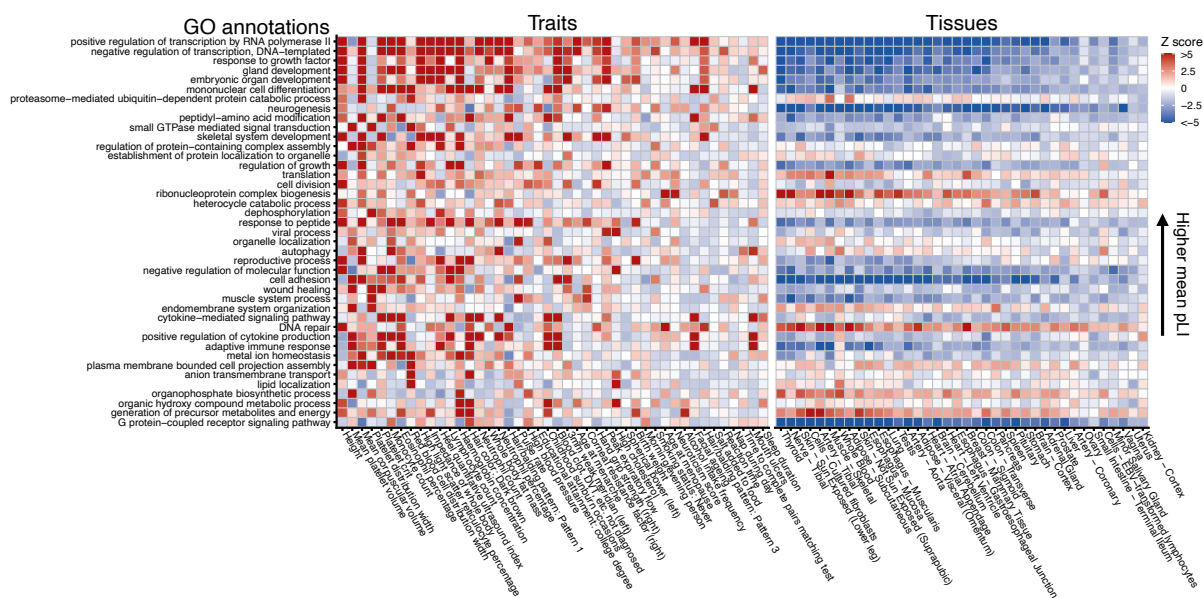
Many Gene Ontology annotations are enriched among GWAS genes, but depleted among eQTL genes. Differences in the selective constraint and regulatory landscapes of genes likely reflect the functional roles of different types of genes. To gain insight into the role of gene functions in shaping the differences between GWAS and eQTL genes, we analyzed 577 Gene Ontology (GO) biological process terms. For each category, we computed the enrichment of GWAS and eQTL genes in that category compared to control SNPs across all traits and tissues (Methods). For data visualization, we focused here on 41 categories that are broadly unrelated, prioritizing terms that are informative of GWAS/eQTL SNPs (Table S5; Methods).

Notably, we found that GO categories are pervasively enriched among GWAS genes for many traits (Figure 4A). For some categories, the enriched categories highlight evidently relevant traits: e.g., *skeletal system development* for height, *lipid localization* for high cholesterol, and *adaptive immune response* for a number of blood and immune-related traits. That said, many categories, such as *embryonic organ development*, show enrichment across multiple traits.
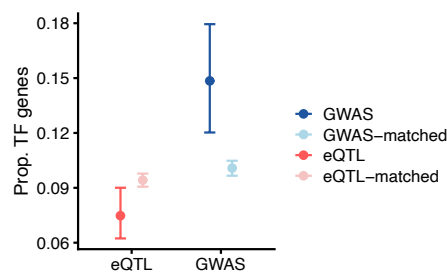
In contrast, many functional categories show clear depletion of eQTLs in all tissues (Figure 4A). Consistent with our earlier results on pLI, depletion is most prominent for categories with high aver-

age pLI, suggesting that selection purges eQTLs for genes with evolutionarily important functions. In contrast, certain categories, such as *DNA repair*, are enriched for eQTL genes, suggesting that they may be more tolerant of expression changes.
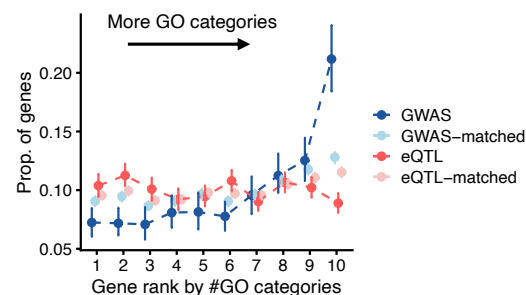

Figure 4: **Diverse categories of functional genes are enriched among GWAS genes, but not among eQTL genes.** *A) Enrichment of genes in 41 Gene Ontology (GO) annotations among GWAS and eQTL genes (see Methods for details of GO category selections). Traits and tissues (x-axis) are sorted by hit count (decreasing from left to right), and GO annotations (y-axis) are sorted by the mean pLI value of genes within categories. For each trait- or tissue-GO category pair we computed enrichment Z scores based on 1000 sampling iterations of SNPs matched for MAF, LD score, and gene density (see Methods). The color map represents enrichment (red) or depletion (blue) of a given gene set among GWAS or eQTL genes. See Table S6 for enrichment and Z score values. B) Fraction of GWAS and eQTL genes that are transcription factors. C) Fraction of GWAS and eQTL genes (y-axis) that are in different bins of genes ranked by the counts of GO categories they belong to (x-axis). In panels (B) and (C), for GWAS hits (blue) and eQTLs (red), error bars show 95% confidence intervals as determined by quantile bootstrapping. For matched SNPs (for MAF, LD score and gene density, shown in light blue and red colors), points and error bars show mean values and 95% confidence intervals in 1000 sampling iterations.*

We wondered whether transcription factors (TFs) would also show the same pattern, as TF dosages often play essential roles in development and cellular functions. Indeed, two categories related to transcriptional regulation, *positive regulation of transcription by RNA polymerase II* and *negative regulation of transcription, DNA-templated*, are particularly enriched in GWAS genes and

8

depleted in eQTL genes (Figure 4A). Focusing specifically on TFs, we found that they are enriched in GWAS genes (15% compared to 10% for matched SNPs), and depleted in eQTL genes (7% compared to 9% for matched SNPs; Figure 4B), in line with previous reports that TFs are underrepresented among eGenes [59]. TFs also contribute substantially to several developmental categories (Figure S3A) that are enriched in GWAS genes and depleted in eQTL genes.

Lastly, we were curious to understand how it is that *most* functional categories can show enrichment for complex traits (Figure 4A). We hypothesized that this is in part because many genes are involved in multiple categories, and such multi-functional genes would be particularly enriched among GWAS genes. Indeed, genes belonging to more GO categories are progressively enriched in GWAS genes, while being modestly depleted for eQTL genes (Figure 4C). Multi-functional genes tend to be highly connected in protein-protein interaction networks (Figure S4A). In line with this observation, we found that genes with the most interactions are enriched for GWAS hits (Figure S4B).

In summary, our GO analysis shows that a diverse range of biological process categories are enriched near GWAS hits; moreover, multi-functional genes are especially enriched for GWAS hits. In contrast, most annotated biological processes are underrepresented near eQTLs.

GWAS hits are further from transcription start sites (TSSs) than eQTLs. Our analysis so far has focused on differences between GWAS hits and eQTLs with respect to the properties of the target genes. We also found important differences between the SNP-level context of GWAS and eQTL hits.

It's well known that eQTLs tend to be tightly clustered near TSSs [60–62]. We observed similar results (Figure 5A): 43% of eQTLs lie within 10Kb of the nearest TSS, compared to 23% for matched SNPs. However, in sharp contrast, GWAS hits are only modestly enriched near TSSs (Figure 5A): 22% lie within 10Kb of the nearest TSS, compared to 17% for matched SNPs. GWAS hits typically lie at greater distances from the nearest TSS (median 36Kb) compared to eQTLs (median 13Kb).

Consistent with these trends, although both GWAS hits and eQTLs are enriched in genomic regions annotated with promoter and enhancer activity by the ENCODE and FANTOM projects (Figure 5B), the relative enrichment in promoter domains versus enhancer domains is much stronger for eQTLs than GWAS hits. Also, eQTLs enrichment in both promoter and enhancer elements sharply increases as selective constraint on the target gene relaxes, particularly exceeding GWAS enrichments for the least constrained genes (Figure 5C; see Methods for details on linking regulatory regions to target genes). These results suggest that a promoter versus enhancer hypothesis (i.e., eQTLs being promoter variants, while GWAS hits being enhancer variants) does not fully explain the differences in clustering near TSSs in Figure 5A.
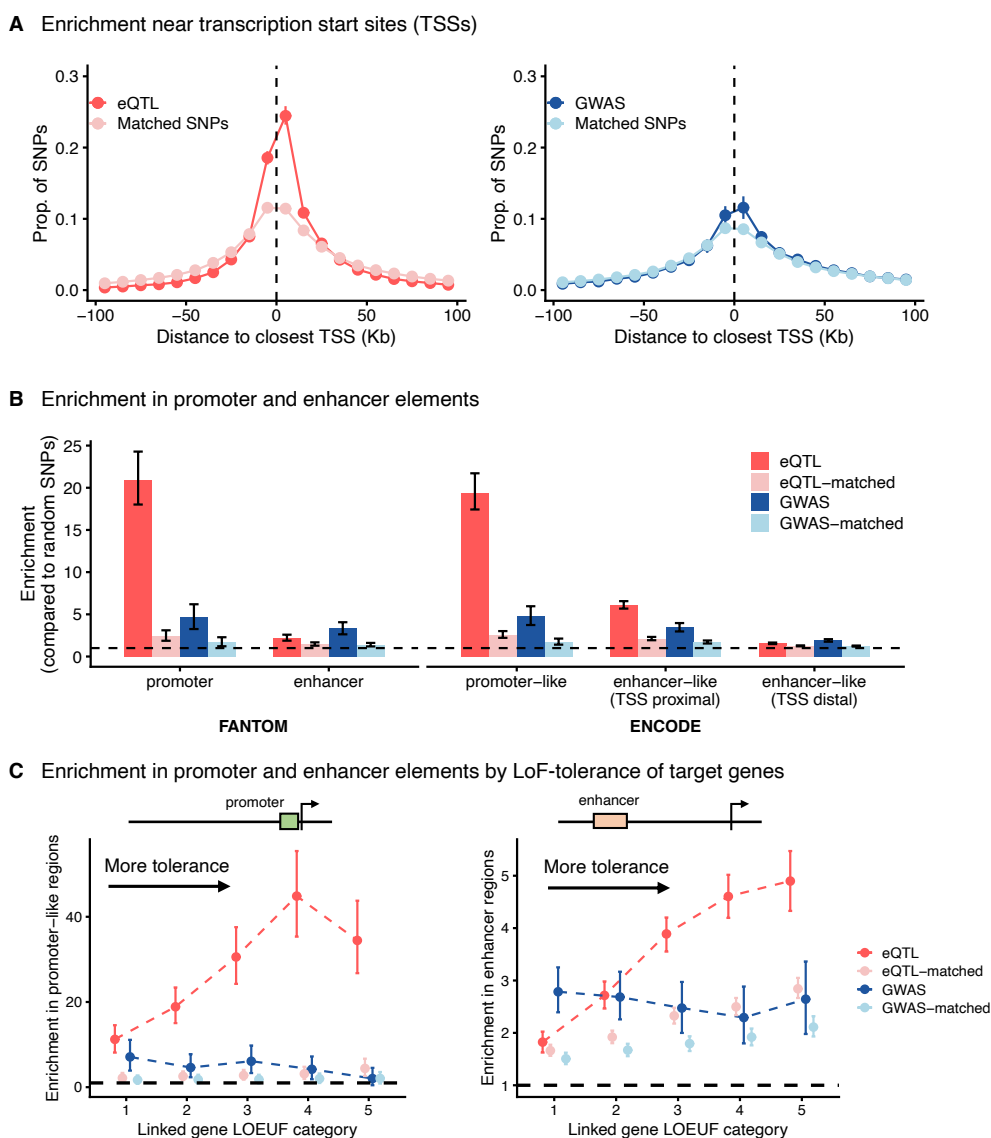
Figure 5: **GWAS hits are less enriched at transcription start sites than are eQTLs.** *A) Distance of eQTLs (left panel) and GWAS hits (right panel) to the nearest TSS. Points show fraction of SNPs in 10Kb bins. SNPs more than 100Kb away from their closest TSS are not shown for clarity.* **B)** *Enrichment of eQTLs and GWAS hits in promoter and enhancer elements annotated in the FANTOM project (left), and in promoter-like, TSS proximal enhancer-like, and TSS distal enhancer-like elements annotated in the ENCODE project (right). For each annotation, the enrichment value is computed as the fraction of SNPs in the annotation divided by the fraction of all SNPs in the annotation.* **C)** *Enrichment of eQTLs and GWAS hits in ENCODE promoter elements (left panel) and Roadmap enhancer elements (right panel) as a function of the quintile of their target gene LOEUF score (a measure of selective constraint, [63]). For promoter elements, linking to genes was done taking the closest TSS within 1Kb. For enhancers, enhancer-gene links from Liu et al. were used [53]. In panels (B) and (C), the black dashed lines mark the value of 1 on the y-axis. In all panels, values corresponding to SNPs matched for MAF, LD score, and gene density are also presented. For GWAS hits and eQTLs, error bars show 95% confidence intervals as determined by quantile bootstrapping. For matched SNPs, error bars show mean values and 95% confidence intervals in 1000 sampling iterations.*

3   A model for variant discovery in GWAS and eQTL assays.

In summary, we have shown that GWAS hits are systematically different from eQTL hits in important ways, including that:

- GWAS hits are enriched near selectively constrained genes (high pLI); eQTL hits are depleted.

- GWAS hits are enriched at genes with complex regulatory landscapes across different cell types, and at genes with more co-expression partners; eQTL hits are relatively depleted at all these features.

- GWAS hits are enriched across a wide variety of GO categories and among transcription factors; eQTL hits are depleted from many GO categories, and depleted among transcription factors.

- GWAS hits show only slight promoter enrichment; eQTL hits are strongly enriched at promoters.

These stark differences between GWAS hits and eQTLs may seem surprising, especially given the expectation that most noncoding GWAS hits should be eQTLs in some cell type. (Other mechanisms such as splicing do contribute, but likely in smaller proportions.) We now describe a model that predicts these qualitative patterns and shows that incomplete colocalization should be expected.

Single cell-type model. We start by examining the case where there is only a single relevant cell type for the phenotype of interest, and eQTL mapping has been conducted in this cell type. This would be most relevant to studies where specific cell types can be studied by cell-sorting or single-cell sequencing.

For simplicity, we assume that *all* genetic effects on phenotypes are mediated via cis-effects on gene expression:

$$\text{genotype} \xrightarrow{\beta} \text{gene expression} \xrightarrow{\gamma} \text{phenotype},$$

where β is the per-allele effect size of genotype on expression, and γ is the effect size of a unit change in expression on the phenotype. The net phenotypic effect of the variant on phenotype is therefore $\beta\gamma$.

Whether a variant is discovered as an eQTL or a GWAS hit depends on how much of the variance it explains in gene expression and in phenotype; this is given by $2p(1-p)\beta^2$ and $2p(1-p)\beta^2\gamma^2$, respectively, where p is the allele frequency. In expectation, the variants that we discover satisfy the conditions

$$
\begin{aligned}
2p(1-p)\beta^2 &> c^*_{\text{eQTL}} \quad &&\text{[for eQTLs]} \\
2p(1-p)\beta^2\gamma^2 &> c^*_{\text{GWAS}} \quad &&\text{[for GWAS]}
\end{aligned}
$$

where c^* is a study-dependent discovery threshold. (Specifically, $c^* \approx \chi_c^2 V_y / n$, where χ_c^2 is the relevant chi-square critical-value for significance, V_y is the total variance in the trait, and n is sample size. Equivalently, in expectation, the discovered variants satisfy the condition $h_{SNP}^2 > \chi_c^2/n$, where h_{SNP}^2 is the fraction-of-variance explained by the tested variant; see Methods.)

Consequently, the power in any given study depends on the specific parameters of the study: the fraction-of-variance explained by GWAS hits is usually much smaller than for cis-eQTLs ($\sim 10^3$-fold smaller for GWAS), but this is roughly balanced by the much-larger sample sizes typical in GWAS

($\sim10^3$-fold larger; Supplementary Note). Given these two different approaches to variant discovery, the key question is: **Should we expect GWAS and eQTL mapping to find the same hits?**

GWAS and eQTL mapping have power in different parts of the parameter space. We can address the question of GWAS-eQTL overlap in terms of what parts of the parameter space have appreciable power for each assay (Figure 6A). For eQTLs, we discover variants provided that $2p(1-p)\beta^2$ is large enough; so conditional on p the discovery region is given by vertical contours, as shown in red. For GWAS, we need the product $\beta^2\gamma^2$ to be large enough, and so the discovery region is given by a curved blue line. Some hits are discovered by both assays because both β^2 and γ^2 are large enough (purple). Most importantly, some hits are discovered in GWAS only (blue region) because $\beta^2\gamma^2$ is large, but are not detected as eQTLs because their effects on expression alone (β^2) are small. If we change the sample size n for either study, this shifts the positions of the discovery contours, but not their shapes.
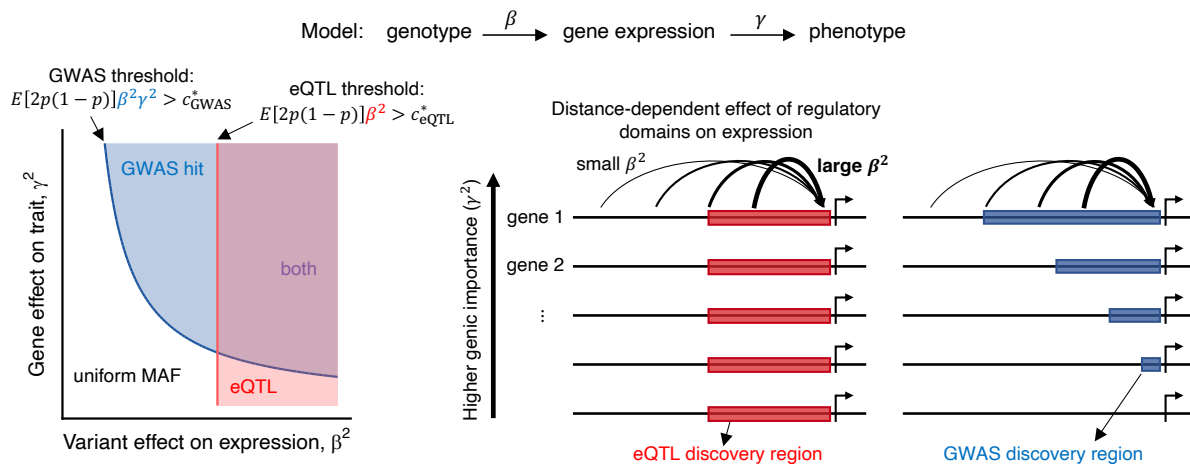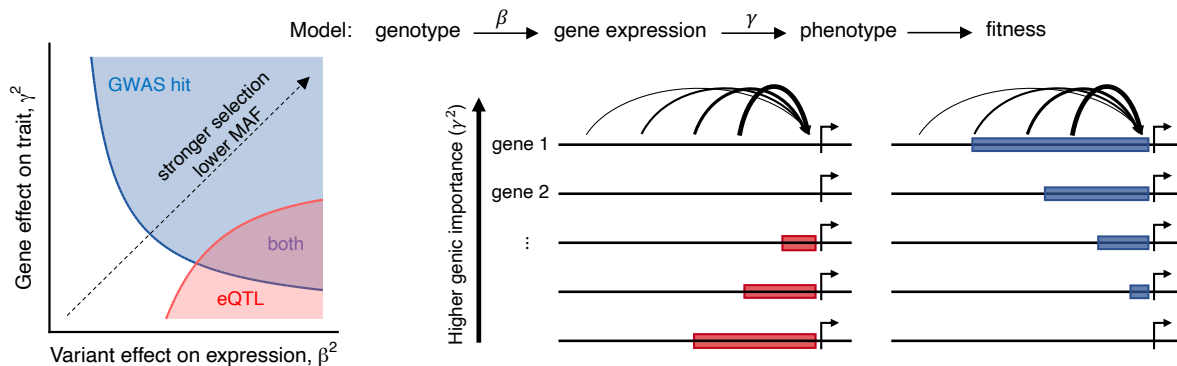

Figure 6: **A model for variant discovery in GWAS and eQTL assays.** *A) Case of a neutrally evolving phenotype. (Left panel) Variant discovery in the space defined by variant effect on expression, β, and genic effect on phenotype, γ. Shading colors represent parameter space for the discovery of GWAS hit only (blue), eQTL only (red), and both types (purple), determined conditional on $E[2p(1-p)]$ which is independent of the effect sizes in the neutral case. (Right panel) schematic of variant discovery mapped to cis-regulatory domains as a function of genic contribution to the phenotype. B) Case of a phenotype under selection. Same as panel (A), but now the the discovery regions in the left panel are determined based on how $E[2p(1-p)]$ covaries with effect sizes under a model of selection (see Methods for details).*

Selection further shrinks the region of overlap. So far we have described a neutral model, but many papers have shown selection against variants that affect complex traits [40–44]. In this scenario, selection keeps variants with large phenotypic effects at low frequencies. We derived how the discovery regions change under selection by modeling a negative correlation between $\beta^2\gamma^2$ and $E[2p(1-p)]$ (Methods; Figure 6B).

Given that selection strength is inversely related to the magnitude of the phenotypic effect, this does not systematically change the expected rankings of variants discovered in GWAS compared to the neutral scenario (Figure 6B), although it leads to a more uniform distribution of heritability across variants with intermediate and large effects (Figure S5A, a phenomenon referred to as "flattening" [40,44]). However, for eQTLs with a given magnitude of regulatory effect, β^2, selection is stronger on variants acting on genes with larger γ^2 values (Figure S5B). This lowers discovery power for regulatory variants linked with phenotypically important genes.

In summary, selection affects the GWAS and eQTL assays differently, mostly preventing the discovery of eQTLs at important genes. This shrinks the overlap in the discovery regions of the two assays compared to the scenario without selection (Figure 6A,B).

Principles of gene regulation make eQTLs more proximal. We next considered how features of gene regulation may affect GWAS-eQTL overlap. Previous eQTL studies have demonstrated that variants close to the target gene's TSS tend to have larger effect sizes than more distal variants [61,62]. Similarly, promoter-enhancer contact frequencies decay with genomic distance between the enhancer and promoter elements [64,65]. Thus, in terms of our model, average β^2 should decay with distance from the TSS.

Under this scenario, eQTLs will be skewed towards regulatory elements near the TSS (with large β^2), and depleted from distal elements (with small β^2) regardless of the phenotypic importance of the target genes (γ^2) (Figure 6A). GWAS hits, however, will be skewed towards phenotypically important genes (with large γ^2) in a distance-dependent manner: distal regulatory elements are more likely to include a GWAS hit if acting on genes with larger phenotypic effects (Figure 6A). These patterns are predicted to be even more pronounced with selection, as large-effect TSS-proximal eQTLs should be most depleted from highly important genes, thereby further reducing eQTL–GWAS overlap (Figure 6B).

In the Supplementary Note we further discuss the role of gene regulatory architecture.

Multi cell-type model. So far we have described the simplest case where there is only a single relevant cell type. But in practice, most current eQTL studies sample across mixtures of cell types (e.g., whole tissues), and/or environmental contexts. More fundamentally, genetic effects on a phenotype may be mediated by gene expression in multiple regulatory contexts (cell types, developmental stages, environmental stimuli, etc.).

To explore these scenarios, Figure 7 outlines a simple model of GWAS and eQTL mapping in a bulk tissue context. We derive a one-dimensional representation of this scenario in order to use the insights gained from our single-context model discussed earlier. Here, β_{agg} denotes an aggregate effect of a genetic variant on expression across contexts, and γ_{agg} denotes an aggregate effect of gene expression levels across contexts on the phenotype (Figure 7A; Supplementary Note).

13

A One-dimensional model of effects across multiple contexts

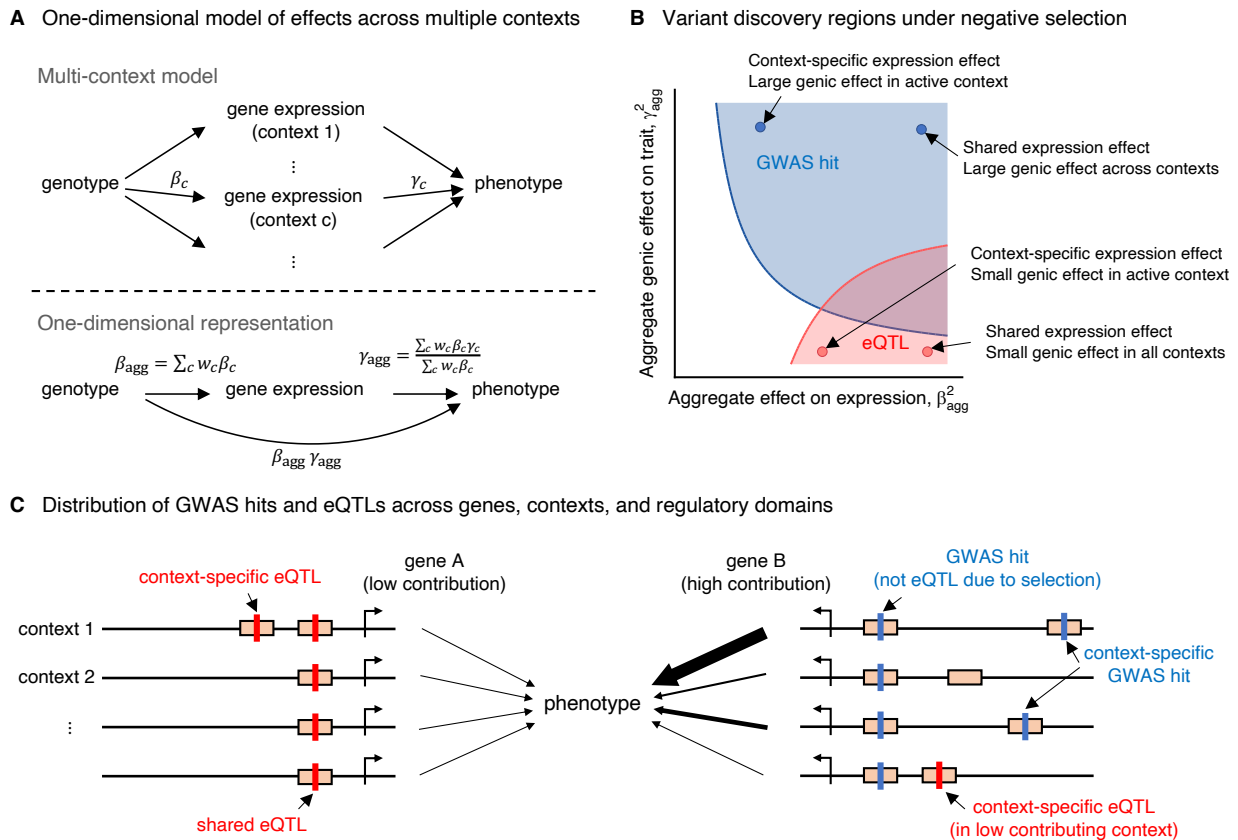**B** Variant discovery regions under negative selection

Figure 7: **A model for variant discovery across multiple contexts.** *A) One-dimensional representation of the pathway from genotype to phenotype mediated by gene expression across multiple contexts (e.g., cell types in a tissue), by defining parameters β_{agg} and γ_{agg}. **B**) Variant discovery in the space defined by β_{agg} and γ_{agg} for a phenotype under selection; same as in Figure 6B. Shading colors represent parameter space for the discovery of GWAS hit only (blue), eQTL only (red), and both types (purple). **C**) Schematic of variant discovery across multiple contexts mapped to regulatory domains (represented by orange rectangles) for two genes at different ends of the phenotypic importance axis.*

eQTLs are biased toward less-important genes and less-important contexts. Similar to the previous analysis, we can gain insight into the GWAS and eQTL discovery process by considering variants in terms of β_{agg} and γ_{agg} (Figure 7B). GWAS will tend to detect variants if $\beta_{\mathrm{agg}}\gamma_{\mathrm{agg}}$ is large enough, and this can occur through a combination of large expression effects and/or large phenotypic effects. But eQTL mapping will be most powerful for detecting large – shared – expression effects, and thus the two types of assays may have limited overlap. Figure 7C illustrates how this may play out for different variants: GWAS hits will be skewed towards functionally important genes and highly contributing contexts. Crucially, eQTLs will be skewed towards unimportant genes; meanwhile, eQTLs discovered at important genes will be skewed toward low contributing contexts. Thus, we may even find context-specific eQTLs for the "right" genes but at the "wrong" variants in the "wrong" contexts.

4   Discussion

Most GWAS hits are in the noncoding portion of the genome, and they are highly enriched within active chromatin. It is generally believed that the large majority of these act via effects on cis gene

regulation. However, it has long been recognized that a large fraction of noncoding GWAS hits do not colocalize with known eQTLs [15, 16, 66]. This raises the question: Where are the missing eQTLs?

Certainly, part of the colocalization gap is due to the fact that we do not yet have complete measurement of all cell types, and some is due to other regulatory mechanisms such as SNP effects on splicing. But here we argue that a more fundamental issue is that GWAS and eQTL mapping are powered to identify different types of variants.

We argue that any explanation for limited colocalization must account for the fact that GWAS variants differ from eQTL variants in many important axes. Using carefully matched analyses, we have shown here that GWAS hits are biased toward more constrained genes, toward genes with functions in many GO categories, toward transcription factors, and toward genes with complex regulatory landscapes; eQTLs are biased away from all these categories. In contrast, eQTLs show a strong promoter bias that is largely absent from GWAS hits. These systematic differences cannot easily be explained by the fact that we have not yet studied all cell types (Figure S6).

Instead, to understand these observations we note that, like any mapping method, GWAS and eQTL mapping have incomplete power. In the case of eQTL mapping, a variant must explain at least a few percent of the variance to be discovered in a typical study (Supplementary Note). What types of variants, and what types of genes, are likely to cross this threshold? They tend to be variants with relatively large effects on gene regulation – large β in our model – especially in promoters. Moreover, they tend to be at genes where selective constraint is low, thereby enabling variants to drift to high frequencies.

In contrast, GWAS hits are, by design, variants that have measurable effects on an organismal trait or disease – large $\beta\gamma$ – and hence these are biased toward functionally important genes. While selection does play an important role in reducing allele frequencies for these variants, it has a "flattening" effect, and does not systematically bias *against* discovery at important genes [40, 44].

Given these conclusions, what are the most promising routes forward for linking GWAS hits to their cognate genes and relevant cell types?

Of course, larger sample sizes in eQTL mapping will help. But for the reasons we have out-lined above, the discovery regions for eQTL and GWAS mapping are systematically distinct, and extremely large eQTL samples will be needed to colocalize all desired GWAS hits (Figure S7). For less-accessible cell types this may simply not be practical. Moreover, even with extraordinar-ily large samples, as for blood where there is now a cis-eQTL for most expressed genes [37], our model predicts that the detected eQTLs will still be biased toward enhancers that are active in less-constrained cell types, or environmental contexts that are less relevant for interpreting GWAS.

Rather, it seems most likely that the colocalization gap will be solved with a multipronged approach, as no single method can be expected to resolve all GWAS hits. Certainly, it will be helpful to collect more cell types, more developmental stages, and larger samples. Some eQTLs may be active only in specific contexts that are underrepresented in conventional eQTL samples from bulk adult tissues such as GTEx. Likewise, some eQTLs may be represented only in rare cell types that are difficult to detect in bulk data.

Meanwhile, other types of molecular QTL assays – such as ATAC-QTLs (e.g., [67]) – may also help to plug some of the gaps. Our modeling arguments about eQTLs also hold for chromatin QTLs, but each type of assay discovers distinct variant sets, each only partially overlapping with GWAS hits and with each other (as in Figure 6B). Then, if one can identify the likely causal enhancers, these may be provisionally linked to genes using computational or functional approaches [54, 68–70].

Lastly, there are various orthogonal methods that do not suffer from quite the same discovery biases as eQTL mapping. Examples include models that predict regulatory activity of variants from DNA sequence, without the need to measure effects on gene expression (e.g., [71, 72]). Also, deciphering regulatory networks (e.g., [73–75]) can help prioritize variants that may not necessarily have large cis-effects but strong effects on the expression of tentatively functional genes. Lastly, emerging techniques using MPRA reporter assays, or CRISPR-based variant-editing or enhancer-silencing, offer powerful experimental alternatives to QTL-based techniques [76–78]. We anticipate that combinations of genome-scale techniques will ultimately help close the colocalization gap.

In summary, we have shown here that eQTLs and GWAS hits differ dramatically in several important ways. We have argued that this likely reflects essential differences in what these assays detect, shaped in large part by selection.

References

[1] Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577(7789):179-89.

[2] Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012;337(6099):1190-5.

[3] Gusev A, Lee S Trynka G, Finucane H, Vilhjálmsson B, Xu H, et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. The American Journal of Human Genetics. 2014;95(5):535-52.

[4] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.

[5] Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature. 2020:1-8.

[6] Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genetics. 2010;6(4):e1000888.

[7] Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics. 2015;47(11):1228-35.

[8] Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genetics. 2014;10(5):e1004383.

[9] Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature genetics. 2015;47(9):1091-8.

[10] Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics. 2016;48(3):245-52.

[11] Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics. 2016;48(5):481-7.

[12] Hormozdiari F, van de Bunt M, Segrè A, Li X, Joo J, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. The American Journal of Human Genetics. 2016;99(6):1245-60.

[13] Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204-13.

[14] Consortium G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318-30.

[15] Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, Jager PLD, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nature Genetics. 2017;49(4):600-5.

[16] Umans BD, Battle A, Gilad Y. Where Are the Disease-Associated eQTLs? Trends in Genetics. 2020;37(2):109-24.

[17] Connally N, Nazeen S, Lee D, Shi H, Stamatoyannopoulos J, Chun S, et al. The missing link between genetic association and regulatory function. medRxiv. 2021:2021.06.08.21258515.

[18] Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nature genetics. 2020;52(6):626-33.

[19] Strober BJ, Elorbany R, Rhodes K, Krishnan N, Tayeb K, Battle A, et al. Dynamic genetic regulation of gene expression during cellular differentiation. Science. 2019;364(6447):1287-90.

[20] D'Antonio-Chronowska A, Donovan MKR, Fujita K, Salgado BM, Matsui H, Arthur TD, et al. iPSC-derived pancreatic progenitors are an optimal model system to study T2D regulatory variants active during fetal development of the pancreas. bioRxiv. 2021:2021.03.17.435846.

[21] Walker RL, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, Torre-Ubieta Ldl, et al. Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. Cell. 2019;179(3):750-71.e22.

[22] Jerber J, Seaton DD, Cuomo AS, Kumasaka N, Haldane J, Steer J, et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. Nature genetics. 2021;53(3):304-12.

[23] Zhernakova DV, Deelen P, Vermaat M, Iterson Mv, Galen Mv, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nature Genetics. 2017;49(1):139-45.

[24] Young AMH, Kumasaka N, Calvert F, Hammond TR, Knights A, Panousis N, et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. Nature Genetics. 2021;53(6):861-8.

[25] Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, et al. Cell type–specific genetic regulation of gene expression across human tissues. Science. 2020;369(6509):eaaz8528.

[26] Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, Andersen S, et al. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. Science. 2022;376(6589):eabf3041.

[27] Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science. 2014;343(6175):1246949.

[28] Calderon D, Nguyen MLT, Mezger A, Kathiria A, Müller F, Nguyen V, et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nature Genetics. 2019;51(10):1494-505.

[29] Gutierrez-Arcelus M, Baglaenko Y, Arora J, Hannes S, Luo Y, Amariuta T, et al. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. Nature genetics. 2020;52(3):247-53.

[30] Ota M, Nagafuchi Y, Hatano H, Ishigaki K, Terao C, Takeshima Y, et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. Cell. 2021;184(11):3006-21.e17.

[31] Mu Z, Wei W, Fair B, Miao J, Zhu P, Li YI. The impact of cell type and context-dependent regulatory variants on human immune traits. Genome Biology. 2021;22(1):122.

[32] Hukku A, Pividori M, Luca F, Pique-Regi R, Im HK, Wen X. Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. The American Journal of Human Genetics. 2021;108(1).

[33] Li YI, Van De Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016;352(6285):600-4.

[34] Li L, Huang KL, Gao Y, Cui Y, Wang G, Elrod ND, et al. An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. Nature Genetics. 2021;53(7):994-1005.

[35] Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169(7):1177-86.

[36] Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell. 2019;177(4):1022-34.e6.

[37] Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nature Genetics. 2021;53(9):1300-10.

[38] Pierce BL, Tong L, Chen LS, Rahaman R, Argos M, Jasmine F, et al. Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians. PLoS Genetics. 2014;10(12):e1004818.

[39] Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nature Genetics. 2021:1-7.

[40] O'Connor LJ, Schoech AP, Hormozdiari F, Gazal S, Patterson N, Price AL. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. The American Journal of Human Genetics. 2019;105(3):456-76.

[41] Gazal S, Finucane HK, Furlotte NA, Loh PR, Palamara PF, Liu X, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature Genetics. 2017;49(10):1421-7.

[42] Zeng J, Vlaming Rd, Wu Y, Robinson MR, Lloyd-Jones LR, Yengo L, et al. Signatures of negative selection in the genetic architecture of human complex traits. Nature Genetics. 2018;50(5):746-53.

[43] Koch EM, Sunyaev SR. Maintenance of Complex Trait Variation: Classic Theory and Modern Data. Frontiers in genetics. 2021:2198.

[44] Simons YB, Bullaughey K, Hudson RR, Sella G. A population genetic interpretation of GWAS findings for human quantitative traits. PLoS biology. 2018;16(3):e2002985.

[45] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.

[46] Siewert-Rocks KM, Kim SS, Yao DW, Shi H, Price AL. Leveraging gene co-regulation to identify gene sets enriched for disease heritability. The American Journal of Human Genetics. 2022;109(3):393-404.

[47] Weiner DJ, Gazal S, Robinson EB, O'Connor LJ. Partitioning gene-mediated disease heritability without eQTLs. The American Journal of Human Genetics. 2022;109(3):405-16.

[48] Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. Nature Genetics. 2019;51(5):772-6.

[49] Wang X, Goldstein DB. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. The American Journal of Human Genetics. 2020;106(2):215-33.

[50] Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nature reviews Molecular cell biology. 2015;16(3):144-54.

[51] Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved Role of Intragenic DNA Methylation in Regulating Alternative Promoters. Nature. 2010;466(7303):253-7.

[52] Landry JR, Mager DL, Wilhelm BT. Complex controls: the role of alternative promoters in mammalian genomes. Trends in Genetics. 2003;19(11):640-8.

[53] Liu Y, Sarkar A, Kheradpour P, Ernst J, Kellis M. Evidence of reduced recombination rate in human regulatory domains. Genome Biology. 2017;18(1):193.

[54] Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021;593(7858):238-43.

[55] Forrest ARR, Kawaji H, Rehli M, Baillie JK, Hoon MJLd, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462-70.

[56] Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, McDowell IC, et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Research. 2017;27(11):1843-58.

[57] Kim SS, Dai C, Hormozdiari F, Geijn Bvd, Gazal S, Park Y, et al. Genes with High Network Connectivity Are Enriched for Disease Heritability. The American Journal of Human Genetics. 2019;104(5):896-913.

[58] Dey KK, Gazal S, Geijn Bvd, Kim SS, Nasser J, Engreitz JM, et al. Contribution of enhancer-driven and master-regulator genes to autoimmune disease revealed using functionally informed SNP-to-gene linking strategies. bioRxiv. 2021:2020.09.02.279059.

[59] Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research. 2014;24(1):14-24.

[60] Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, et al. High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genetics. 2008 10;4(10):e1000214.

[61] Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, et al. Common Regulatory Variation Impacts Gene Expression in a Cell Type–Dependent Manner. Science. 2009;325(5945):1246-50.

[62] Brown CD, Mangravite LM, Engelhardt BE. Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. PLoS Genetics. 2013;9(8):e1003649.

[63] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

[64] Zuin J, Roth G, Zhan Y, Cramard J, Redolfi J, Piskadlo E, et al. Nonlinear control of transcription through enhancer–promoter interactions. Nature. 2022:1-7.

[65] Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature genetics. 2019;51(12):1664-9.

[66] Connally N, Nazeen S, Lee D, Shi H, Stamatoyannopoulos J, Chun S, et al. The missing link between genetic association and regulatory function. medRxiv. 2021.

[67] Liang D, Elwell AL, Aygün N, Krupa O, Wolter JM, Kyere FA, et al. Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation. Nature neuroscience. 2021;24(7):941-53.

[68] Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature Genetics. 2019;51(12):1664-9.

[69] Boix CA, James BT, Park YP, Meuleman W, Kellis M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature. 2021;590(7845):300-7.

[70] Grubert F, Srivas R, Spacek DV, Kasowski M, Ruiz-Velasco M, Sinnott-Armstrong N, et al. Landscape of cohesin-mediated chromatin loops in the human genome. Nature. 2020;583(7818):737-43.

[71] Nair S, Kim DS, Perricone J, Kundaje A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. Bioinformatics. 2019;35(14):i108-16.

[72] Avsec Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods. 2021;18(10):1196-203.

[73] Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nature Methods. 2016;13(4):366-70.

[74] Sonawane AR, Platig J, Fagny M, Chen CY, Paulson JN, Lopes-Ramos CM, et al. Understanding Tissue-Specific Gene Regulation. Cell Reports. 2017;21(4):1077-88.

[75] Freimer JW, Shaked O, Naqvi S, Sinnott-Armstrong N, Kathiria A, Chen AF, et al. Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. bioRxiv. 2021:2021.04.18.440363.

[76] Abell NS, DeGorter MK, Gloudemans MJ, Greenwald E, Smith KS, He Z, et al. Multiple causal variants underlie genetic associations in humans. Science. 2022;375(6586):1247-54.

[77] Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell. 2019;176(1-2):377-90.

[78] Morris JA, Daniloski Z, Domingo J, Barry T, Ziosi M, Glinos DA, et al. Discovery of target genes and pathways of blood trait loci using pooled crispr screens and single cell rna sequencing. bioRxiv. 2021.

[79] Consortium G, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648-60.

[80] Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC genome browser database: update 2006. Nucleic acids research. 2006;34(suppl_1):D590-8.

[81] Aygün N, Elwell AL, Liang D, Lafferty MJ, Cheek KE, Courtney KP, et al. Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. The American Journal of Human Genetics. 2021.

[82] Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695. Available from: https://igraph.org.

[83] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2020. Available from: https://www.R-project.org/.

[84] Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols. 2009;4(8):1184-91.

[85] Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology; 2021. R package version 2.44.0.

[86] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell. 2018;172(4):650-65.

[87] Pintacuda G, Lassen FH, Hsu YHH, Kim A, Martín JM, Malolepsza E, et al. Genoppi is an open-source software for robust and standardized integration of proteomic and genetic data. Nature Communications. 2021;12(1):2580.

[88] Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, et al. A scored human protein–protein interaction network to catalyze genomic interpretation. Nature Methods. 2017;14(1):61-4.

[89] Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics. 2015;47(3):291-5.

[90] Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics. 2016;32(2):283-5.

[91] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1-22. Available from: `https://www.jstatsoft.org/v33/i01/`.

[92] Schoech AP, Jordan DM, Loh PR, Gazal S, O'Connor LJ, Balick DJ, et al. Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. Nature communications. 2019;10(1):1-10.

[93] Sinnott-Armstrong N, Naqvi S, Rivas M, Pritchard JK. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. Elife. 2021;10:e58615.

[94] Lloyd-Jones LR, Holloway A, McRae A, Yang J, Small K, Zhao J, et al. The genetic architecture of gene expression in peripheral blood. The American Journal of Human Genetics. 2017;100(2):228-37.

[95] Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. Nature Reviews Genetics. 2019;20(8):437-55.

5 Methods

Datasets.

GWAS data. We used publicly available GWAS summary statistics for traits in the UK Biobank provided by Ben Neale's lab (see URLs). We focused on traits with the lower bound of confidence interval on SNP heritability estimates exceeding 0.05, and with at least 50 hits (with association p-value $< 5 \times 10^{-8}$) that passed our SNP selection criteria (see section *SNP selection*). For binary traits we used heritability estimates on the liability scale, and further conditioned on traits with prevalence > 0.05. We pruned the traits list such that genetic correlation, ρ_g, was < 0.5 for all trait pairs in the final list. To this end, we first sorted traits by hit count, and then starting from the trait with most hits, iterated through the list removing traits with $\rho_g > 0.5$ with the focal trait. This procedure resulted in 44 traits listed in Table S1. Genetic correlation and SNP heritability estimates used for this procedure were downloaded from the Neale lab website.

eQTL data. We used eQTLs from the GTEx v8 data that were based on analyzing the subset of individuals with European ancestry in the dataset (`*EUR.signif_pairs.txt.gz` files from `GTEx_Analysis_v8_eQTL_EUR.tar` data, see URLs) to match the GWAS data. To avoid over-representation of brain tissues, for brain-related eQTLs we retained those identified in "Brain - Cerebellum" or "Brain - Cortex", which have relatively distinct expression profiles among the brain regions [79]. In total 38 tissue were included listed in Table S1. To match the GWAS data, all genomic coordinates were mapped to the hg19 assembly using `LiftOver` [80]. In one analysis (Figure S6) we also used eQTLs detected (i) in fetal brain samples by Aygün et al. [81], (ii) at multiple stages of iPSCs differentiation towards neuronal fate by Jerber et al. [22], and (iii) in single-cell analyses of blood cell types by Yazar et al. [26]. In all analyses we conditioned on eQTLs whose target eGenes were among 18,332 protein-coding genes (see section *Gene selection*)

UK Biobank (UKB). We used the UKB resource, specifically for the computation of allele frequencies, and as a linkage disequilibrium (LD) reference panel. For these purposes, we used quality control (QC) measures provided by UKB to select participants who: their reported gender, "Submitted.Gender", matched their "Inferred.Gender" from genotypes; were not identified as heterozygosity outliers ("het.missing.outliers"== 0); did not have excessive number of relatives in the data ("excess.relatives"== 0), and were not predicted to carry sex chromosome aneuploidies ("putative.sex.chromosome.aneuploidy"==0). We further restricted our analysis to individuals identified by UKB to be of "White British" (WB) ancestry ("in.white.British.ancestry.subset"==1), and to be unrelated ("used.in.pca.calculation"==1). A total of 337,123 individuals passed these filters and were used for MAF computation. We randomly selected 10,000 of these participants as a LD reference panel.

Gene selection.

We selected 18,332 genes (Table S2) that (i) were annotated as protein-coding, and (ii) were linked with a HUGO Gene Nomenclature Committee (HGNC)-approved gene nomenclature (i.e., linked with a HGNC ID) in the GENCODE Basic gene annotations (release 39, see URLs). We used the HGNC IDs to link genic features from multiple resources to avoid issues with regard to gene names mismatching.

24

SNP selection.

For our SNP selection process we started with the list of 13.7 million variants that passed quality control measures for the UK Biobank analyses released by the Neale lab (labeled as "imputed-v3 Variant QC" in the Neale lab pipeline). We further applied the following filters: biallelic autosomal SNP; MAF > 0.01 among the unrelated WB individuals in the UKB (see section *Datasets*); polymorphic in the 1000 Genomes Project phase 3 data (used by Alkes Price's lab for LD score regression, see URLs). This yielded 8,136,100 filtered SNPs.

In both GWAS and eQTL data, we first extracted SNPs that were among the list of filtered SNPs above. We then performed LD-based clumping separately for each trait in the GWAS data, and for each gene-tissue pair in the eQTL data. To this end, we used `plink`'s (version 1.90b6.12) --clump flag using the same parameters in both data types: p-value threshold of 5×10^{-8}, LD threshold of $r^2 = 0.1$ and physical distance threshold of 1Mb. The UKB resource was used as the LD reference panel (see section *Datasets*). We refer to the resulting clumped SNPs as "lead SNPs". To make the comparisons between GWAS hits and eQTLs consistent, for both sets of lead SNPs, we removed (i) SNPs in LD ($r^2 > 0.8$) with predicted protein-truncating or missense mutations (annotated as "ptv" or "missense" using the Variant Effect Predictor, VEP, in the Neale lab data), to condition on SNPs putatively acting through gene regulation, and (ii) SNPs > 1 Mb away from the transcription start site (TSS) of any of 18,332 protein-coding genes (see section *Gene selection*), which are not tested for eQTLs in GTEx. We further removed SNPs in the MHC region (chr6:28477797-33448354). This resulted in 22,119 GWAS hits across traits, and 118,996 eQTLs across all gene-tissue pairs (Tables S3,S4).

For both the GWAS hits and the eQTLs selected above, we included control SNPs in most of our analyses. To this end, similar to our ascertainment procedure for GWAS hits and eQTLs, we extracted 6,971,257 SNPs (from the 8,136,100 filtered SNPs) that were not in LD ($r^2 > 0.8$) with predicted protein-truncating or missense mutations, were not > 1 Mb away from the TSSs of the 18,332 protein-coding genes, and were not in the MHC region. From this set, we randomly sampled 1000 SNPs for each GWAS hit or eQTL matching for MAF, LD score, and gene density (see section *SNP annotations* for the definitions and matching scheme).

Gene annotations.

We compiled a number of genic features from various resources:

- *Basic annotations.* We computed total gene length using genomic locations for transcription start and end sites extracted from GENCODE Basic annotations (see URLs). We also retrieved total coding sequence length for the longest transcript of genes from Ensembl's BioMart tool (see URLs). These annotations were used as covariates in our logistic regression models (see section *Statistical methods*).

- *Selective constraint.* We used pLI and LOEUF measures of intolerance to LoF mutations extracted from gnomAD.v2.1's pLoF metrics by gene data ([45, 63], see URLs).

- *Enhancer features.* We considered enhancer-gene links based on two different approaches: (i) links inferred by Liu et al. based on correlation of chromatin marks with gene expression [53] (see URLs), and (ii) links predicted based on the activity-by-contact (ABC) model from Nasser et al. [54] (see URLs). In both data, we first complied the union of all enhancer intervals per gene per biosample. We then computed two features for a given gene: (i) the number of

biosamples in which the gene is linked with at least one enhancer interval (i.e., count of active biosamples), and (ii) the total length of intervals averaged across active biosamples. Genes not present in the data were assigned the value 0 for both features.

- *TSS count.* We analyzed promoter regions identified by the FANTOM consortium using Cap Analysis of Gene Expression (CAGE) [55]. We downloaded combined hg19 CAGE peaks from FANTOM5 phase 1 and phase 2 data (see URLs). For a given gene, we computed the number of peaks linked with the gene. Genes not present in the data were assigned the value 0.

- *Connectedness in co-expression networks.* We analyzed the co-expression networks from Saha et al. [56] for 16 tissues in GTEx v6p data (available from GTEx portal, see URLs). We first analyzed each tissue separately, focusing on total expression connections ("TE-TE" edges in the data) between protein-coding genes in the Transcriptome-Wide Networks (TWNs). We used the `igraph` package [82] in `R` [83] to rank genes by the number of neighbors they have in the tissue-specific networks. In the case of genes having the same number of neighbors, we used the sum of absolute weights for all edges linked with individual genes to break ties. Genes not present in a network were assigned the rank of the last gene in the network plus one. We then computed the rank-product of genes (i.e., product of the ranks of a given gene) across the 16 tissues, to construct a conntectedness measure (genes with lower rank-product have higher conntectedness).

- *Gene Ontology (GO) annotations.* We focused on 577 GO biological process (BP) terms with at least 400 genes. To this end, we first obtained specific GO annotations linked with genes using the biomaRt package [84] from the Bioconductor project (version 3.13) (attribute `go_id`). We then used these gene-GO links in the topGO package (version 2.44, [85]) (as the `gene2GO` parameter) to extract all genes associated with GO terms (using the `genesInTerm` function).

- *Transcription factors (TFs).* We downloaded a list of 1639 putative human TFs from Lambert et al. [86] (see URLs).

- *Connectedness in protein-protein interaction (PPI) networks.* We used the `Genoppi` package [87] in `R` to retrieve scored InWeb PPI data (loading the `inweb_table` data) [88]. We then used `igraph` to compute the number of interactions per protein weighted by the interaction confidence scores.

To aggregate data across all resources, we first converted gene identifiers (gene symbols or Ensembl gene IDs) to HGNC IDs, and then linked all features to the selected 18,332 protein-coding genes. We used NCBI's Gene resources (see URLs) to link HGNC IDs to gene symbols, including the official/recommended symbol as well as other used symbols labeled as "synonyms". We used the Bioconductor's biomaRt to link HGNC IDs to Ensembl gene IDs in the most recent Ensembl version (version 105) as well as the archived versions.

SNP annotations.

We compiled a number of SNP annotations:

- *Minor allele frequency (MAF).* We used the UKB data (see section *Datasets*) to compute MAFs within unrelated individuals identified as White British.

- *Linkage disequilibrium (LD) score.* We used the `ldsc` software (see URLs, [89]) to compute LD scores using a window size of 1 centiMorgan (cM) (specified with the flag `--ld-wind-cm 1`). To this end, we used genetic distances (in cM) as provided by the Price lab with the "baseline (v1.1)" LD annotations for SNPs in the 1000 Genomes Project phase 3 data (see URLs). For the LD reference panel, we used 10,000 randomly selected unrelated White British individuals in the UKB (see section *Datasets*).

- *Gene density.* For a given SNP, we computed gene density as the number of protein-coding genes with their TSS falling within the 1 Mb window ($\pm$500Kb) around the SNP. The TSS coordinates were extracted from the GENCODE Basic annotations (see section *Gene annotations*).

- *Closest gene assignment.* We linked each SNP to the protein-coding gene with the closest TSS to the SNP. Subsequently, all genic features of the closest genes were assigned to the SNPs. The TSS coordinates were extracted from the GENCODE Basic annotation (see section *Gene annotations*).

- *Overlap with promoter/enhancer elements.* We considered two sets of putative promoter and enhancer elements: (i) mapped by the FANTOM5 consortium using Cap Analysis of Gene Expression (CAGE), and (ii) mapped in the phase 3 of the ENCyclopedia Of DNA Elements (ENCODE) project based on epigenetic signatures. We downloaded "permissive" enhancer regions and combined promoter regions from FANTOM5 phase 1 and phase 2 data (see URLs). In the ENCODE project, epigenetic signatures and proximity to TSSs were integrated to categorize candidate cis-regulatory elements (cCREs) as promoter-like, proximal enhancer-like (within 2Kb of nearest TSS), and distal enhancer-like ($>$ 2Kb away of the nearest TSS) elements. We used ENCODE's Registry V2 of cCREs (see URLs) to download the corresponding regions. Regions reported in GRCh38 were mapped to hg19 using `liftOver`. For our analysis in Figure 5C, we further linked the promoter-like regions to nearest TSS conditional on the distance of the regions midpoints to TSSs were $<$ 1Kb. We constructed indicator variables for belonging to all of the above regulatory elements.

Statistical methods.

Estimation of mean SNP features. For a given set of SNPs (e.g., pooled set of all GWAS hits) and for a given feature (e.g., binary indicator of falling in enhancer elements) we computed mean values across all SNPs in the set. For genic features (e.g., binary indicator of belonging to a given GO annotation) we first linked SNPs to genes with the closest TSSs, and then computed mean feature values corresponding to the linked genes over all SNPs.

Bootstrap confidence intervals. We computed bootstrap confidence intervals for mean SNP features estimated for pooled set of GWAS hits across all traits and pooled set of eQTLs across all tissues. To construct bootstrapped samples, we first sampled traits (for GWAS hits) and tissues (for eQTLs) at random with replacement, and concatenated the sets of SNPs corresponding with sampled traits and tissues. We then sampled with replacement from the set of independent LD blocks (inferred by Berisa et al., [90]) that contain GWAS hits or eQTLs, and then concatenated the sets of SNPs (resulted from the previous bootstrapping step) belonging to the sampled LD blocks. We performed this procedure 1000 times to construct 1000 bootstrapped samples. We computed mean SNP features across all 1000 samples, and determined confidence intervals as the the range between 2.5^{th} and 97.5^{th} percentiles.

Control SNPs. For all GWAS hits and eQTLs, we selected control SNPs matched for MAF, LD score, and gene density. For a given SNP, we extracted SNPs (among a total of 6,971,257 SNPs and excluding the focal SNP; see section *SNP selection*) (i) with the same gene density as the focal SNP, (ii) with MAF within 0.02 of the focal SNP's MAF, and (iii) with LD score within 0.1 standard deviation (estimated across all SNPs) of the focal SNP's LD score. We then sampled 1000 times from this set at random with replacement to construct 1000 instances of control SNPs per SNP of interest. These were then combined across all SNPs in a set of GWAS hits or eQTLs to give 1000 sets of matched SNPs. For all gene or SNP features studied in this paper, we computed mean values across matched SNPs for all 1000 matched samples, and used the distributions to compute and report (i) mean values and confidence intervals (as the the range between 2.5^{th} and 97.5^{th} percentiles), or (ii) Z scores (see section *Analyses of individual traits and tissues*).

Joint analysis of features. In Figure 3B, we jointly considered the effect of multiple genic features in classifying GWAS hits or eQTLs from random SNPs. To this end, we constructed an indicator variable for GWAS hits or eQTLs (labeled 1s) versus 100,000 SNPs chosen at random from the full set of 6,971,257 SNPs (labeled 0s). We then used a logistic regression framework to predict this indicator variable using the genic features of interest. Genic feature values were normalized. We included the following covariates in the regression: MAF, LD score, gene density, absolute distance to nearest TSS, total gene length, total length of gene coding sequence, as well as dummy variables for 20 quantiles of MAF, LD score, gene density, and absolute distance to nearest TSS.

Analyses of individual traits and tissues. In Figures 2B and 4A we studied gene features separately for individual traits and tissues. For a given feature (e.g., proportion of SNPs that are near high pLI genes), we computed mean values for the sets of GWAS hits for individual traits and eQTLs for individual tissues. We also computed mean feature values in sets of matched SNPs (1000 sets of SNPs for each set of GWAS hits and eQTLs; see section *Control SNPs*), and used the distributions to compute (i) enrichment values as the estimated values for GWAS hits or eQTLs divided by the matched samples mean, and (ii) Z scores as the matched samples mean subtracted from the estimated values for GWAS hits or eQTLs divided by the matched samples standard deviation.

Analyses of eQTLs by rank. For the analysis in Figure 2C we grouped eQTLs by their rank based on association strengths (i.e., p-values) in individual tissues. To this end, in all 38 tissues, we first bin ranked eQTLs by association p-values in groups of 1000 eQTLs: first top 1000 eQTLs as group 1, second top 1000 eQTLs as group 2, and so on. We then pooled eQTLs across tissues by the ranked bins.

Selection of broadly unrelated GO annotations. In Figure 4A we show enrichment Z scores for GWAS hits and eQTLs across 41 GO annotations. GO annotations are hierarchical and thus interdependent. We selected these 41 gene categories by pruning 577 GO biological process annotations (see section *Gene annotations*) to give a set of broadly unrelated categories while retaining those relevant to the traits and tissues studied here. To this end, we first determined enrichment Z scores for all 577 categories (Table S6). Then, we selected the top category (i.e., most enriched or depleted) for each tissue and trait, conditioning on GO annotations with < 3000 associated genes. This gives 52 unique categories. We then pruned this set as follows: we sorted categories by the count of associated genes in an ascending order, and iterated over categories starting with the category with the least number of genes. At each iteration, we retained the focal category if gene associations with that category could not be well-predicted from the previously included categories, defined as the AUC value of < 0.75 estimated using penalized logistic regression (as implemented in the `glmnet` package in R using the `cv.glmnet` function for cross-validation, [91]) over all protein-coding genes. This gives 27 broadly unrelated categories. Using the same procedure, we built upon this set by

iterating over the rest of the categories that were not among the top categories with respect to GWAS hit or eQTL enrichments, resulting in 41 categories (Table S5).

Analysis of non-GTEx eGenes. In Figure S6, we analyzed eGenes identified (i) in fetal brain samples by Aygün et al. [81], (ii) at multiple stages of iPSCs differentiation towards neuronal fate by Jerber et al. [22], and (iii) in single-cell analyses of blood cell types by Yazar et al. [26]. For each sample, we computed the proportion of high pLI genes among the eGenes. We then sampled the same count of genes as eGenes, 10,000 times at random from the set of all protein-coding genes. We computed the proportion of high pLI genes in each set of random genes and used the distributions to compute (i) enrichment values as the estimated values for eGenes divided by the random samples mean, and (ii) Z scores as the random samples mean subtracted from the estimated values for eGenes divided by the random samples standard deviation. For comparison, we performed the same procedure for GTEx eGenes in brain and whole blood tissues.

Modeling variant discovery.

Here we describe how we derived the discovery regions for GWAS and eQTL assays (shown in Figures 6A,B):

Effect of selection. A quantitative treatment of the role of selection is beyond the scope of this paper: it requires knowledge of the joint distributions of variant effects on gene expression (β), genic effects on phenotypes (γ), and allele frequency (p) as a function of the selection strength on the phenotype. We make simplistic assumptions to gain a *qualitative* understanding of how selection affects variant discovery in GWAS and eQTL assays.

Under a neutral model, the effect sizes and allele frequencies are independent. Thus for a given pair of (β, γ) values, the expected contribution of variants to phenotypic variance $E[2p(1-p)\beta^2\gamma^2|\beta,\gamma] = E[V_p|\beta,\gamma]\beta^2\gamma^2 \propto \beta^2\gamma^2$ (Figure S5A), where we defined $V_p = 2p(1-p)$ as the variance in allele frequency. Under selection, $E[V_p|\beta,\gamma]$ is negatively correlated with $\beta^2\gamma^2$. This is supported by empirical evidence [43, 92], as well as modeling work on the effect of selection on genetic architecture of complex traits [44]. To mimic the qualitative effect of selection, we used an asymptotic exponential form to describe the relationship $E[V_p\beta^2\gamma^2|\beta,\gamma] \propto \kappa(1 - e^{-\beta^2\gamma^2/\kappa})$ (Figure S5A). In the main text we set $\kappa = 2.986$; under this model such a κ reduces $E[V_p]$ compared to the neutral case by $\sim 10\%$ assuming that β and γ are drawn form independent standard Normal distributions. We show in Figure S8 that as long as $E[V_p|\beta,\gamma]$ decreases with increasing phenotypic effect size, i.e., $\beta^2\gamma^2$, our qualitative conclusions are robust to (i) the mathematical form describing the relationship between expected phenotypic variance and effect size, and (ii) the magnitude of the relative reduction in phenotypic variance due to selection compared to neutrality.

Estimation of discovery regions. Given a GWAS or eQTL assay discovery threshold, c^* as defined in the main text, the power to discover a given variant depends on its β, γ, and $E[V_p|\beta,\gamma]$. The previous section describes $E[V_p|\beta,\gamma]$ for a given (β,γ) pair (under selection or neutrality). Therefore, in principle, discovery regions can be solved for sets of β and γ values that satisfy $E[V_p|\beta,\gamma]\beta^2 > c^*_{\text{eQTL}}$ for eQTLs, and $E[V_p|\beta,\gamma]\beta^2\gamma^2 > c^*_{\text{GWAS}}$ for GWAS hits.

For discovery regions in Figures 6,7, we set c^*_{eQTL} and c^*_{GWAS} thresholds such that under the assumption that β and γ are drawn form independent standard Normal distributions, the power in GWAS and eQTL assays are matched (i.e., same fraction of causal SNPs is discovered in both assays). We set the power threshold at 15% for the visibility of discovery trends, such that there is incomplete overlap between GWAS and eQTL regions. [In Figure S7 we vary the discovery threshold to mimic the effect of sample size.] To this end, we first sampled 10 million pairs of (β,γ) $\sim N(\mathbf{0}, \mathbf{I_2})$.

For each pair we computed (i) the expected contribution to variance in phenotype, $E[V_p|\beta,\gamma]\beta^2\gamma^2$, and (ii) the expected contribution to variance in expression, $E[V_p|\beta,\gamma]\beta^2$, taking $E[V_p|\beta,\gamma]$ values based on the exponential equation described in the previous section for the selection scenario, and equal to 1 for the neutral scenario (the ranking of variants is invariant to the scaling of $E[V_p]$). For each of the four distributions (i.e., variance in phenotype and expression, in the presence or absence of selection) we computed the discovery threshold, c^*, as the 85th percentile of the distribution. To plot the regions delineated by these c^* values, we specified a grid of β and γ values from 0 to 3.84 (corresponding to the 95th percentile of γ^2 and $\beta^2 \sim \chi^2(1)$), with 0.0025 increments. For each pair of (β,γ) we computed expected contributions to variance in phenotype and expression as described above for the 10 million random pairs. We then identified the regions on the grid with values greater than c^*. The borders of the regions were smoothed using the `loess` function in `R`, applied to points on the grid between the 84.9^{th} and 85.1^{th} percentiles of the distributions of variance in phenotype or expression used to determine each c^*.

Dependence of discovery thresholds on sample size. In this section, we derive the dependence of discovery thresholds on assay sample size, n, and other basic parameters. We consider a simple linear regression model, estimating the effect of a single SNP, β, on a quantitative phenotype Y (a gene's expression level or a complex trait):

$$Y = G\beta + \epsilon,$$

where G is the genotype, and $\epsilon \sim N(0, \sigma_\epsilon^2)$ is the noise term capturing the effect of environment as well as other causal SNPs (i.e., genetic background). In OLS regression, the effect estimate $\hat{\beta}$, is Normally distributed with expectation $E[\hat{\beta}] = \beta$. The variance of $\hat{\beta}$ is:

$$Var[\hat{\beta}] = \frac{Var[\epsilon]}{nVar[G]} \approx \frac{Var[Y]}{nVar[G]},$$

where we made the assumption that the contribution of individual SNPs to phenotypic variance is small such that $Var[\epsilon] \approx Var[Y]$. Now the effect is deemed significant if the Z score is large enough:

$$\chi^2 = \frac{n\hat{\beta}^2 Var[G]}{Var[Y]} > \chi_c^2,$$

where χ_c^2 is the significance threshold. (The conventional GWAS threshold of p-value$= 5 \times 10^{-8}$ corresponds to $\chi_c^2 = 29.7$). By definition, c^* is the critical value such that for discovered variants $Var[G]\beta^2 = 2p(1-p)\beta^2 > c^*$. Now, this condition is in expectation satisfied if $c^* := \chi_c^2 Var(Y)/n$.

Also, defining $h_{SNP}^2 := \frac{\beta^2 Var[G]}{Var[Y]}$ as the fraction of trait variance explained by the SNP effect, discovered variants in expectation satisfy:

$$h_{SNP}^2 > \chi_c^2/n.$$

URLs.

Neale lab UKB data: `http://www.nealelab.is/uk-biobank`

GTEx data: `https://gtexportal.org/home/datasets`

NCBI's gene_info file: `ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz`

GENCODE Basic annotations: `https://www.gencodegenes.org/human/release_39lift37.html`

Ensembl's BioMart: `http://uswest.ensembl.org/biomart/martview`

gnomAD: `https://gnomad.broadinstitute.org/downloads`

ABC enhancer-gene links: `https://www.engreitzlab.org/resources`

Liu et al.'s enhancer-gene links: `https://ernstlab.biolchem.ucla.edu/roadmaplinking`

FANTOM5 promoters: `https://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks`

FANTOM5 enhancers: `https://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers`

Transcription factors: `http://humantfs.ccbr.utoronto.ca`

ldsc software: `https://github.com/bulik/ldsc`

LD annotations: `https://alkesgroup.broadinstitute.org/LDSCORE`

ENCODE cCREs: `https://screen-v2.wenglab.org`

Acknowledgments

Supplementary Material

Supplementary Note.

In this note we elaborate on a few arguments relevant to our model for variant discovery in GWAS and eQTL assays described in the main text:

Power considerations in GWAS and eQTL mapping. For both GWAS and eQTL mapping we consider a simple regression model, estimating the effect of a single SNP on a complex phenotype or a gene's expression level, respectively. As detailed in the Methods section, in expectation, the discovered variants satisfy: $h_{SNP}^2 > \chi_c^2/n$, where h_{SNP}^2 is the fraction of trait variance explained by the SNP effect, n is the sample size, and χ_c^2 is the study-dependent discovery threshold.

The conventional GWAS threshold of p-value$= 5 \times 10^{-8}$ corresponds to $\chi_c^2 = 29.7$. Considering a typical GWAS sample size of around 500K, discovered variants will have $h_{SNP}^2 > 6 \times 10^{-5}$. For comparison, considering a trait with a total SNP-heritability of 0.5 and 100K causal variants (which is typical of complex traits, [93]), trait variants on average have $h_{SNP}^2 \approx 5 \times 10^{-6}$.

In an eQTL study of a given gene, if we assume a p-value threshold of $\sim 2 \times 10^{-4}$ (on par with nominal p-value threshold values computed in GTEx at a gene-level false discovery rate threshold of 0.05), the corresponding $\chi_c^2 \approx 14$. Thus, considering sample sizes on the order of 500, a variant is discovered if $h_{SNP}^2 > 0.028$. For comparison, if we assume a typical total gene expression heritability of 0.15, 30% of which is from the cis component [94] (i.e., cis-heritability of about 0.05), and 10 independent cis-regulatory variants per gene, regulatory effects on expression on average have $h_{SNP}^2 \approx 0.005$. Also, if we assume the top regulatory variants per gene typically explain 25% of total heritability [94] ($\sim$83% of cis-heritability), eGenes discovered, in expectation, satisfy cis-heritability > 0.033.

Covariance of gene function and regulatory architecture. Figure 6 illustrates discovery regions conditional on the β and γ values. Where actual genetic variants lie in this parameter space depends on the distributions of β and γ across variants and genes. A determinant of the distribution of β is the variability of regulatory architecture across genes: SNPs regulating genes with more/longer enhancers likely, on average, have smaller βs, due to the dispersion of cis-heritability for gene expression across more regulatory variants.

On the other hand, it's conceivable that complexity of the regulatory architecture of a gene, in part, reflects its functional importance, e.g., developmental genes are typically regulated by several enhancers elements [95]. Under this assumption, genes with more/longer enhancers, on average, have higher γs. Therefore, such genes are expected to lie in the top-left region of the parameter space (i.e., small β, large γ), and to be enriched in GWAS and depleted in eQTL assays.

These trends are similar to those predicted by Wang and Goldstein [49], though our argument does not rely on the assumption that multiple enhancer elements regulating the same gene are functionally redundant.

Multi cell-type model. In Figure 7 we provide a one-dimensional representation of a multi-context scenario. As described in the main text, we assign parameters β_{agg} denoting an aggregate effect of a genetic variant on expression across contexts, and γ_{agg} denoting an aggregate effect of gene expression levels across contexts on the phenotype (Figure 7A). We define $\beta_{\text{agg}} = \sum_c w_c \beta_c$, as the weighted sum of effects over all causal contexts, where β_c denotes the genetic effect on expression in context c with the weighting w_c. As an example, considering a situation where a gene's expression in a tissue influences a phenotype, β_c represents the effect in cell type c, w the relative abundance

of different cell types, and β_{agg} the effect estimate in a bulk assay of the causal tissue.

We define $\gamma_{\mathrm{agg}} = \frac{\sum_c w_c \beta_c \gamma_c}{\sum_c w_c \beta_c}$, where γ_c denotes the effect of a unit change in target gene's expression in context c on the phenotype. The rationale for this definition is such that, similar to our single-context model, the quantity $\beta_{\mathrm{agg}}\gamma_{\mathrm{agg}}$ gives the net effect of the genetic variant on phenotype. For a given variant, the interpretation of γ_{agg} is the mean genic effect across contexts weighted by the effect of the variant on gene expression. For example, if a variant is active in a single-context c, γ_{agg} is γ_c, and if the variant has the same effect on expression across all contexts γ_{agg} is $\sum_c w_c \gamma_c$.

In this multi-context model, for a given variant, the role of the degree of sharing of the expression effects across multiple contexts is analogous to the role of distance to TSS in the single-context model. In this view, using the result that discovered eQTLs are expected to be more TSS-proximal than GWAS hits, eQTLs are expected to be more shared across contexts, or on the flip side, GWAS hits are expected to be more context-specific than eQTLs.
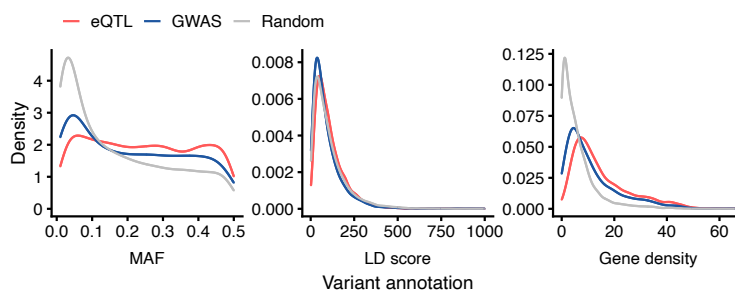
Supplementary Figures.

Figure S1: **Basic variant-level differences between GWAS hits and eQTLs.** *Distribution of minor allele frequency (MAF), linkage disequilibrium (LD) score and gene density for eQTLs (red), GWAS hits (blue), and 100,000 randomly chosen variants. LD score values are cut at 1000 for clarity.*
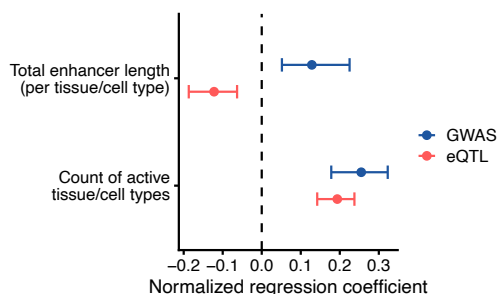

Figure S2: **GWAS and eQTL genes have different enhancer architectures.** *Same as Figure 3B, but using enhancer-gene links predicted by the activity-by-contact (ABC) model from Nasser et al. [54] (Methods). For a given gene, we computed (i) the number of biosamples in which a gene has an enhancer, and (ii) the average total enhancer length (in base pairs) across active biosamples. Shown are logistic regression coefficients corresponding with the two enhancer features for predicting GWAS hits (blue) and eQTLs (red) versus random variants after adjusting for confounders (Methods). Error bars show 95% confidence intervals as determined by quantile bootstrapping.*

A Proportion of transcription factors (TFs) in Gene Ontology (GO) biological processes
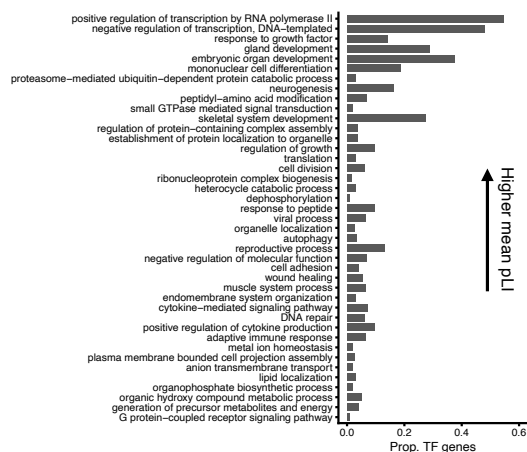
B Enrichment of Gene Ontology (GO) biological processes excluding transcription factors (TFs)
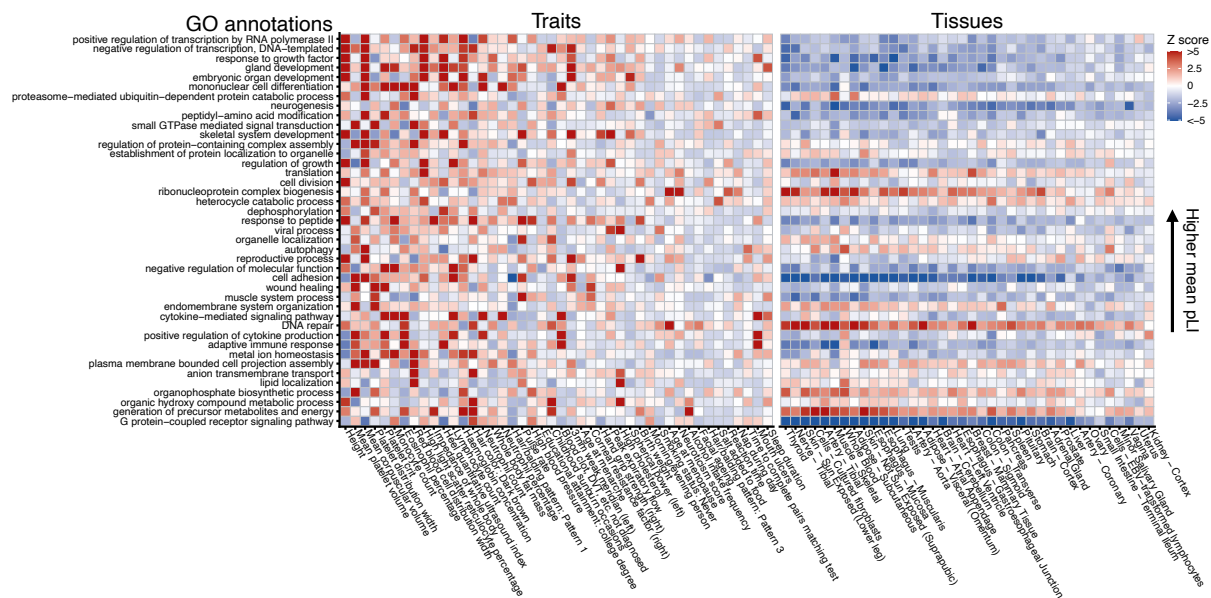

Figure S3: **Contribution of transcription factors (TFs) in Gene Ontology (GO) annotations and their enrichment in GWAS and eQTL genes.** *A) Proportion of TFs in 41 GO biological processes shown in Figure 4A. B) Same as Figure 4A, but now excluding TFs from all 41 gene categories before computing enrichment values among GWAS and eQTL genes. Traits and tissues (x-axis) are sorted by hit count (decreasing from left to right), and GO annotations (y-axis) are sorted by the mean pLI value of genes within categories (before removing TFs, replicating the ordering in Figure 4A). For each trait- or tissue-GO category pair we computed enrichment Z scores based on 1000 sampling iterations of variants matched for MAF, LD score, and gene density (see Methods). The color map represents enrichment (red) or depletion (blue) of a given gene set among GWAS or eQTL genes.*
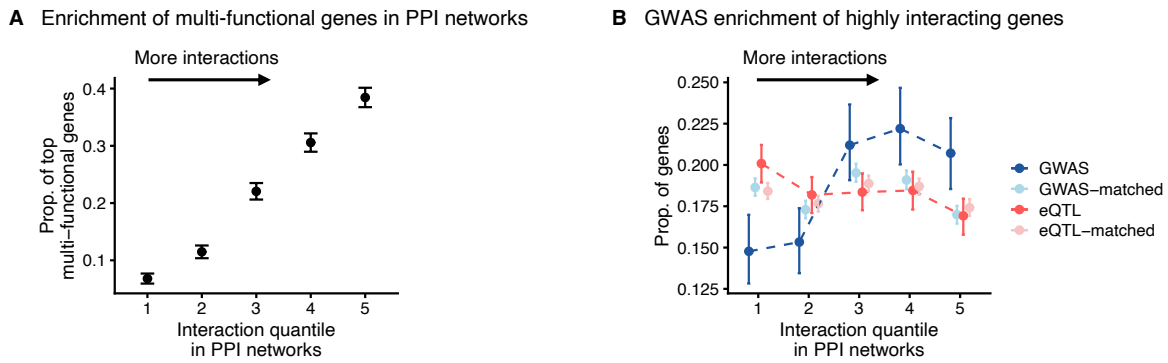
A Enrichment of multi-functional genes in PPI networks

B GWAS enrichment of highly interacting genes

Figure S4: **Multi-functionality of highly interacting genes in protein-protein interaction (PPI) networks and their enrichment in GWAS genes.** *A) Proportion of genes in bins ranked by the number of interactions in the InWeb PPI network [88] that are among the top multi-functional genes (defined as top 20% of genes ranked by the count of Gene Ontology (GO) categories they belong to, see Methods). Error bars show 2 standard errors.* ***B)*** *Fraction of GWAS and eQTL genes in gene bins ranked by the number of interactions in the InWeb PPI network. For GWAS hits and eQTLs, error bars show 95% confidence intervals as determined by quantile bootstrapping. For matched variants (for MAF, LD score and gene density, shown in light blue and red colors), points and error bars show mean values and 95% confidence intervals in 1000 sampling iterations.*
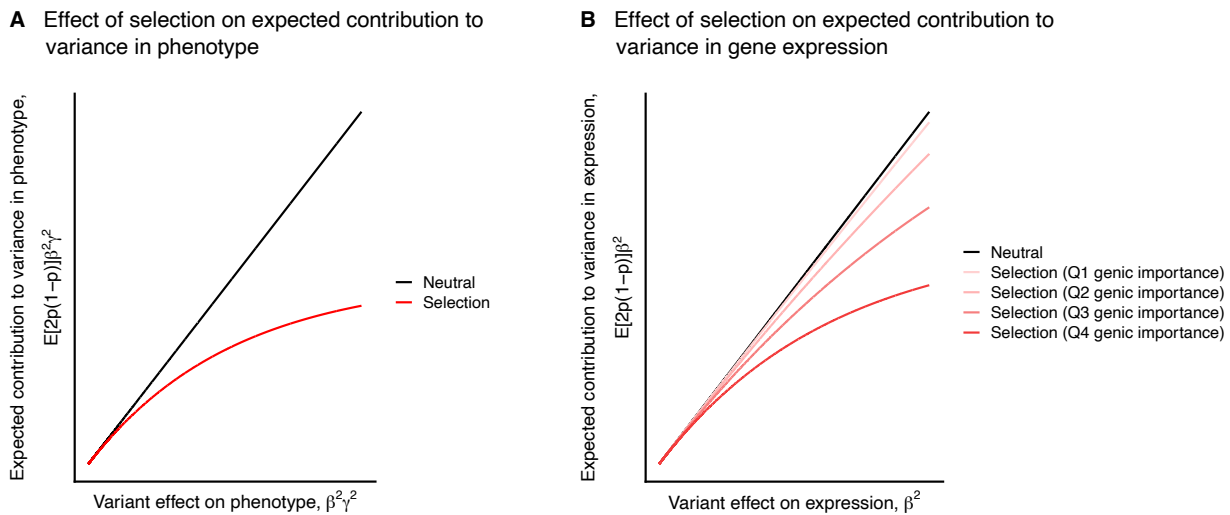

A Effect of selection on expected contribution to variance in phenotype

B Effect of selection on expected contribution to variance in gene expression

Figure S5: **Effect of selection on variants contribution to variance in phenotype and gene expression.** *A,B) As described in the main text, we consider a model of phenotypic effects mediated by effects on gene expression intermediates: a genetic variant affects the expression of the target gene with effect β, and the gene expression intermediate affects the downstream phenotype with effect size γ.* ***A)*** *Contribution to phenotypic variance. Under a neutral model, contribution to phenotypic variance, $E[2p(1-p)]\beta^2\gamma^2$, is proportional to phenotypic effect, $\beta^2\gamma^2$, as effect size and allele frequency are uncoupled. Selection keeps higher effect variants at lower frequencies (i.e., lowering $E[p(1-p)]$) and thus "flattens" the expected contribution to variance. The red line shows a flattened curve taking $E[2p(1-p)\beta^2\gamma^2|\beta,\gamma] \sim \kappa(1-e^{-\beta^2\gamma^2/\kappa})$, with $\kappa = 2.986$ (Methods).* ***B)*** *Contribution to variance in gene expression. Similar to the argument in (A), under neutrality, contribution to variance in gene expression, $E[2p(1-p)]\beta^2$, is proportional to the effect on expression, β^2. Under selection, flattening (i.e., lowering of $E[p(1-p)]$) is more pronounced for variants regulating high-effect (i.e., high γ^2) genes. Red lines show trends for four quantiles of γ^2, where $\gamma \sim N(0,1)$; darker colors show higher γ^2 values. See Methods for modeling details.*
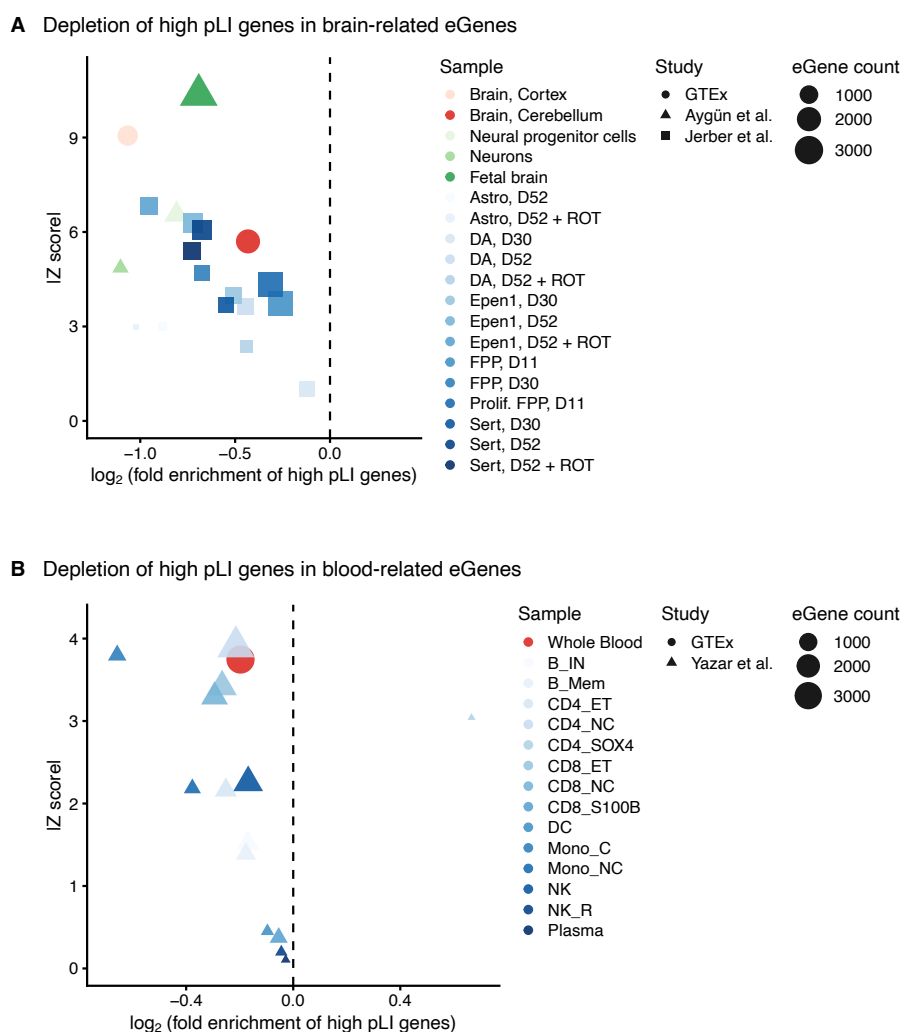
36

Figure S6: **Depletion of selectively constrained genes among non-GTEx eGenes.** *The factors we described against the discovery of trait-eQTLs likely bias eQTL assays in any context. As proof of concept, we show that similar to GTEx eGenes, eGenes identified in non-conventional eQTL assays are also depleted of strongly selected genes. **A)** Enrichment of high pLI genes in eGenes identified (i) in fetal brain samples by Aygün et al. [81], (ii) at multiple stages of iPSCs differentiation towards neuronal fate by Jerber et al. [22], and (iii) in GTEx brain tissues. Sample labels for Jerber et al. refer to different ascertained cell types, at different days of differentiation, and in the presence or absence of stimulation by rotenone (ROT) (see Jerber et al. for details). **B)** Enrichment of high pLI genes in eGenes identified in (i) single-cell analyses of blood cell types by Yazar et al. [26], and (ii) GTEx whole blood. Sample labels for Yazar et al. refer to different blood cell types (see Yazar et al. for details). Enrichment values (on the x-axis) and Z scores (on the y-axis) were computed based on values observed in 10,000 sampling iterations of random genes (Methods).*
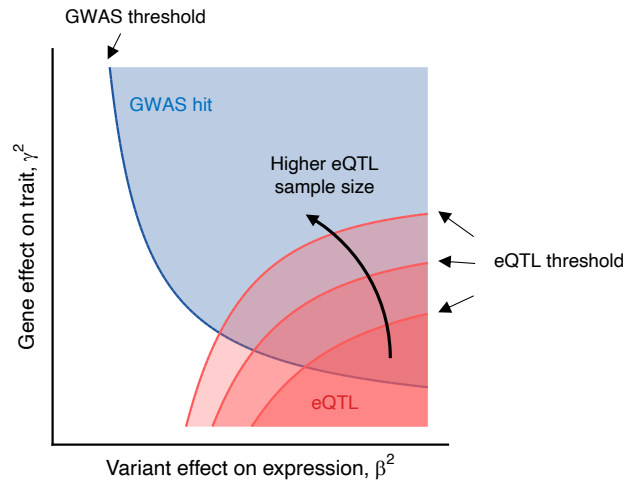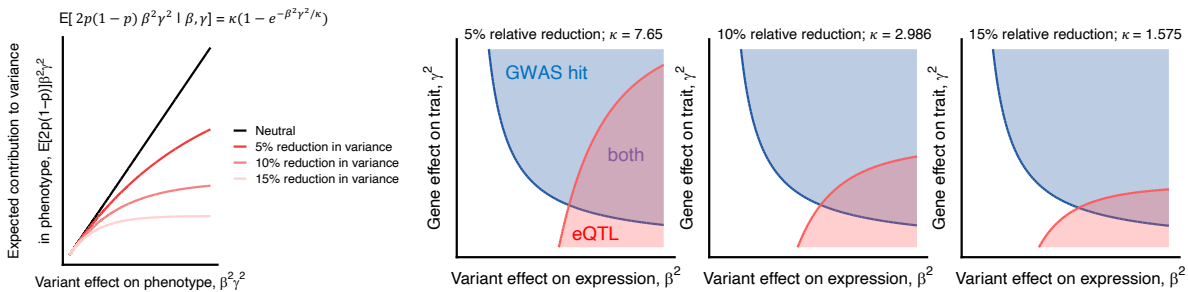
Figure S7: **Effect of eQTL assay sample size on discovery.** *Same as Figure 6B, but with three eQTL discovery thresholds corresponding to different sample sizes. The discovery thresholds are derived by setting the power rate to 15% for GWAS under the assumptions detailed in the Methods section, and to 10%, 15% and 20% for eQTLs.*
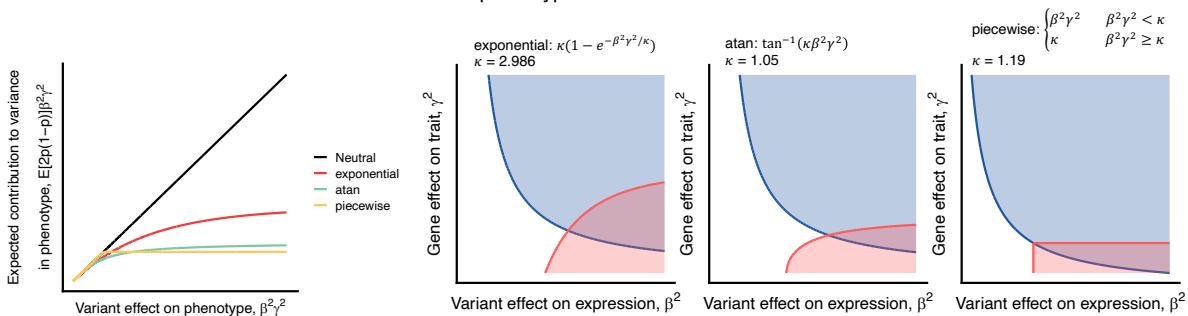

Figure S8: **Robustness of qualitative results to modeling choices.** *In Figure 6B we show GWAS and eQTL discovery regions under the scenario that the phenotype is under selection. We incorporate selection by modeling a reduction in contribution to phenotypic variance (compared to the neutral scenario) as a function of the phenotypic effect (see Methods for details). For plots shown in the the main text we chose the exponential form $E[2p(1-p)\beta^2\gamma^2|\beta,\gamma] \sim \kappa(1-e^{-\beta^2\gamma^2/\kappa})$, with $\kappa = 2.986$ to obtain an average reduction of $\sim 10\%$ across randomly sampled β and γ values. We show here that our qualitative results are robust to the choice of (A) average reduction in contribution to phenotypic variance and (B) the functional form of the variance under selection. In panel (B), κ values are tuned to give an average reduction in contribution to phenotypic variance of $\sim 10\%$ for all three models.*

Supplementary Tables.

Table S1: **List of traits and tissues.**

Table S2: **List of autosomal protein-coding genes.**

Table S3: **List of GWAS hits.**

Table S4: **List of eQTLs.**

Table S5: **List of broadly unrelated Gene Ontology (GO) biological process terms.**

Table S6: **Enrichment of Gene Ontology (GO) biological processes in GWAS and eQTL genes for individual traits and tissues.**