nature genetics

Article

https://doi.org/10.1038/s41588-023-01529-1

Systematic differences in discovery of genetic effects on gene expression and complex traits

Received: 15 May 2022

Accepted: 14 September 2023

Published online: 19 October 2023

Check for updates

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

Most signals in genome-wide association studies (GWAS) of complex traits implicate noncoding genetic variants with putative gene regulatory effects. However, currently identified regulatory variants, notably expression quantitative trait loci (eQTLs), explain only a small fraction of GWAS signals. Here, we show that GWAS and *cis*-eQTL hits are systematically different: eQTLs cluster strongly near transcription start sites, whereas GWAS hits do not. Genes near GWAS hits are enriched in key functional annotations, are under strong selective constraint and have complex regulatory landscapes across different tissue/cell types, whereas 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 variant, and support the use of complementary functional approaches alongside the next generation of eQTL 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 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 with eQTL mapping to gain a 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 Genotype-Tissue Expression (GTEx) Consortium[13,14], most GWAS hits are not explained by currently known eQTLs[15–17]. For example, one analysis by the GTEx Consortium found that only 43% of GWAS hits (median 21% of hits per trait) were colocalized with eQTLs[14]. Similarly, averaged across traits, only 11% of heritability is estimated to be 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; for example, 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 postmortem adult whole tissues. Nevertheless, the contribution of context-specific eQTLs in explaining trait-associated variants has thus far been modest[17,31]. For example, a study of eQTLs during differentiation of induced pluripotent stem (iPS) cells toward neural fate added ~10% more colocalizations with neurological trait loci beyond GTEx eQTLs[22]. Also in GTEx, cell type-specific eQTLs colocalized with only ~8% of GWAS hits[25]. Although 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.

[1]Department of Genetics, Stanford University, Stanford, CA, USA. [2]Department of Chemical and Systems Biology, Stanford University, Stanford, CA, USA. [3]Department of Biology, Stanford University, Stanford, CA, USA. e-mail: hmostafavi@stanford.edu; pritch@stanford.edu

A different type of explanation is that perhaps many trait-relevant eQTLs have not yet been discovered because of incomplete statistical power or the challenges of colocalization analysis, even if mapping was performed in the correct contexts[18,32]. However, others have proposed that trait-eQTL discovery 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), whereas 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 detectable as *cis*-eQTLs[16,38]. Together, these observations suggest that most GWAS hits are indeed *cis*-eQTLs, but 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 (UKB), and eQTL data for 38 tissues in the GTEx dataset. We show that in fact GWAS hits and eQTLs differ systematically: GWAS hits lie at greater distances from transcription start sites (TSSs); they are enriched near genes associated with key 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. By contrast, eQTLs are tightly clustered near the TSSs of genes that are typically depleted of most functional annotations, show reduced 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 effects on GWAS discovery compared with eQTL discovery.

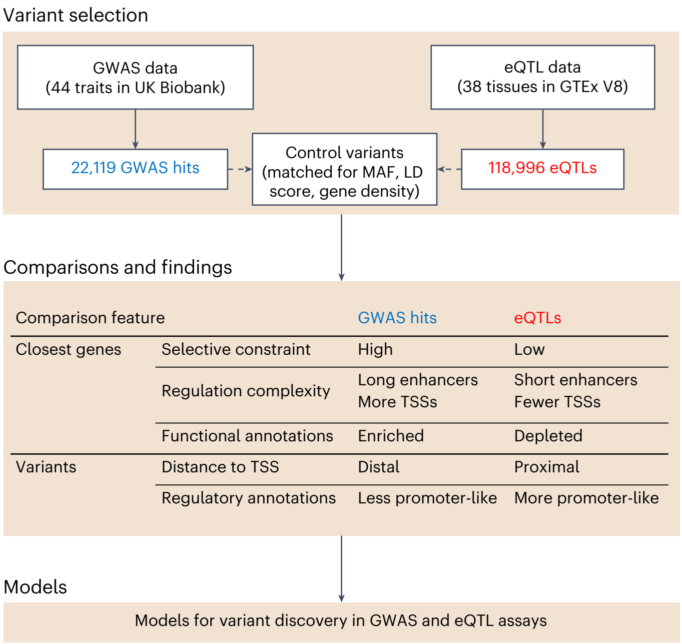
In summary, GWAS and eQTL mapping tend to maximize power for different types of variant, and current eQTL mapping has limited discovery power at the most trait-relevant genes. Although further context-specific eQTL studies will help somewhat in explaining GWAS hits, we argue here that these efforts should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

Results

Analysis overview

For GWAS analyses, we used publicly available summary statistics for 44 traits in the UKB (Supplementary Table 1). For eQTL analyses, we used the GTEx V8 data for 38 tissues[14] (Supplementary Table 1), focusing on *cis*-eQTLs associated with 18,332 protein-coding genes (Supplementary Table 2). To make the GWAS hits and eQTLs more comparable, we used identical quality control and SNP selection procedures for both datasets. We removed lead SNPs in strong linkage disequilibrium (LD) with protein-altering variants, 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 (Supplementary Tables 3 and 4). See Methods for all details.

For each GWAS hit and 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 nearest TSS among the same 18,332 protein-coding genes. Although the true causal genes for GWAS hits are often unknown, the nearest gene serves as an extremely useful proxy[39]. For eQTLs, the relevant SNP–gene pairs are known. However, in most analyses, we masked the true eGenes and instead linked the eQTL SNP to the nearest gene. This strategy maximizes the similarity between the GWAS and eQTL pipelines, although it introduces some mis-assignments of genes, particularly for weaker eQTLs (Extended Data Fig. 1a). Nevertheless, we demonstrate that this strategy is conservative, and all genic properties we report for the nearest genes replicate even more strongly for the eGenes (Supplementary Figs. 5 and 6). We refer to the genes assigned this way as 'GWAS genes' and 'eQTL genes' in the rest of this paper.



Variant selection
GWAS data (44 traits in UK Biobank)
eQTL data (38 tissues in GTEx V8)
22,119 GWAS hits
Control variants (matched for MAF, LD score, gene density)
118,996 eQTLs
Comparisons and findings
Models
Models for variant discovery in GWAS and eQTL assays

Comparison feature		GWAS hits	eQTLs
Closest genes	Selective constraint	High	Low
	Regulation complexity	Long enhancers More TSSs	Short enhancers Fewer TSSs
	Functional annotations	Enriched	Depleted
Variants	Distance to TSS	Distal	Proximal
	Regulatory annotations	Less promoter-like	More promoter-like

Fig. 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 Supplementary Table 1. We compare GWAS hits and eQTLs with respect to a number of genic and SNP features. A summary of our main results is presented. We describe models for variant discovery in GWAS and eQTL assays to conceptualize these results. See Methods for details.

We observed several systematic differences between eQTL and GWAS lead SNPs (Extended Data Fig. 2). Specifically, eQTL SNPs have a higher minor allele frequency (MAF) (median 0.24 compared with 0.2) and are located in more gene-dense regions (median 11 genes per megabase compared with 8) than GWAS lead SNPs. To account for these differences and minimize potential confounding effects, 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. Our key findings are robust to various data and processing choices (Supplementary Note).

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]. Despite this, SNP heritability for complex traits is enriched near selectively constrained genes[45–47], as measured by the pLI (probability of being loss-of-function intolerant) score[45]. Although pLI does not directly quantify selection acting on a gene[48], high-pLI genes are reasonable candidates for genes under strong selection.

Consistent with previous findings, GWAS genes are enriched for high-pLI genes (pLI > 0.9) (Fig. 2a): 26% compared with 21% for genes linked to control SNPs ($P = 2 \times 10^{-4}$; see Methods for statistical details regarding the computation of all reported *P* values in the text). By

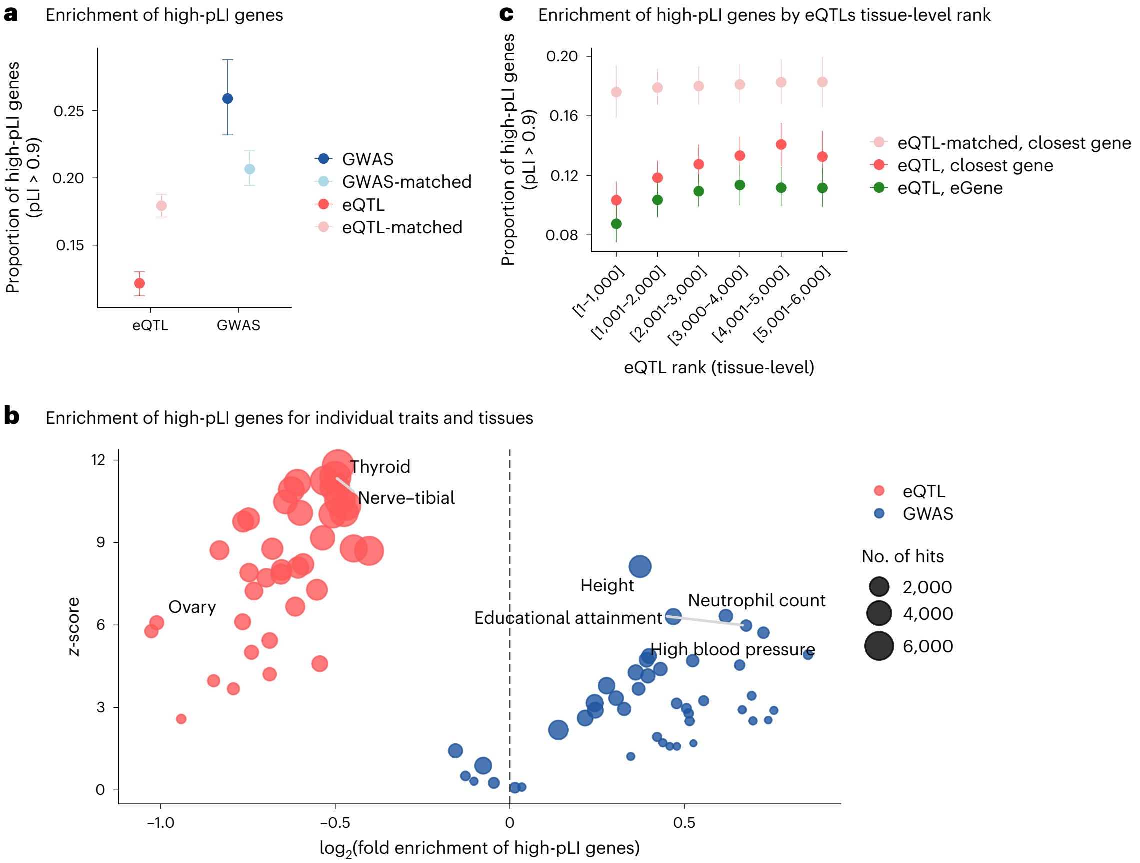


Fig. 2 | GWAS and eQTL genes are under different selective constraints. **a**, Fraction of genes with high pLI (pLI > 0.9, a measure of selective constraint) among genes nearest to 22,119 GWAS hits (blue), 118,996 eQTLs (red) and control SNPs matched for MAF, LD score and gene density. **b**, Enrichment of high-pLI genes in GWAS genes for individual traits and eQTL genes for individual tissues. Enrichment values (*x* axis) and *z*-scores (*y* axis) were derived from 1,000 control SNP sampling iterations. The *z*-scores were calculated as the difference between the mean of control samples and the values for GWAS hits or eQTLs, divided by the standard deviation of control samples. The Bonferroni correction threshold for absolute *z*-score values is 3.43 (Methods). **c**, Fraction of high-pLI genes 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 1,000 eQTLs, and then pooled eQTLs across tissues by the ranked bins. This procedure resulted in 33,123, 25,643, 18,256, 13,839, 10,875 and 8,064 eQTLs in the rank bins [1–1,000], [1,001–2,000], [2,001–3,000], [3,001–4,000], [4,001–5,000] and [5,001–6,000], respectively. In panels **a** and **c**, error bars show 95% confidence intervals as determined by quantile bootstrapping over 1,000 sampling iterations. For matched SNPs, points show mean values in sets of matched SNPs corresponding to bootstrapped samples (Methods).

contrast, eQTL genes are depleted of high-pLI genes: 12% compared with 18% for control SNPs ($P = 10^{-21}$). This depletion is more pronounced for eGenes (true target genes), indicating that our nearest gene assignment approach merely makes the underlying trends noisier (Extended Data Fig. 1b). These results align with previous reports that genes without detectable eQTLs have relatively higher pLI[37,45]. We observed similar results with selection-related measures other than pLI (Extended Data Fig. 3).

These trends are not driven by a few outlier traits/tissues (Fig. 2b). Notably, the depletion of high-pLI genes near eQTLs is replicated in all tissues, although to lesser degrees for tissues with a larger number of eQTLs (Fig. 2b). This is likely due to the inclusion of weaker eQTLs that reach significance in tissues with larger sample sizes, because the depletion of high-pLI genes is increasingly pronounced among top ranked eQTLs (Fig. 2c).

Together, these results show that although a substantial fraction of GWAS hits are located near selectively constrained genes, most identified 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 reports that the fraction of trait heritability estimated to be mediated via gene expression is dominated by genes with low *cis*-heritability for expression levels[18,46].

GWAS genes have more complex regulatory landscapes than eQTL genes

Wang and Goldstein previously demonstrated that genes near GWAS hits and eGenes 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 TSS usage across diverse tissues and cell types.

We computed two enhancer features per gene using the enhancer–gene links inferred by Liu et al. based on the Roadmap Epigenomics Consortium dataset[50]: (1) the number of biosamples in which the gene is linked to at least one enhancer element, and (2) the total length of linked enhancers per biosample (Fig. 3a and Methods). In a logistic regression framework, compared with genes linked with random SNPs, GWAS genes have longer enhancer regions per biosample ($P = 5 \times 10^{-3}$), whereas eQTLs have shorter enhancers ($P = 5 \times 10^{-9}$) (Fig. 3b), consistent with Wang and Goldstein[49]. By contrast, both types of gene have enhancer activity across more biosamples relative to genes linked with random SNPs ($P = 5 \times 10^{-3}$ for GWAS, and 10^{-2} for eQTLs) (Fig. 3b). These results are replicated when using enhancer–gene links from the activity-by-contact model[51] (Extended Data Fig. 4).

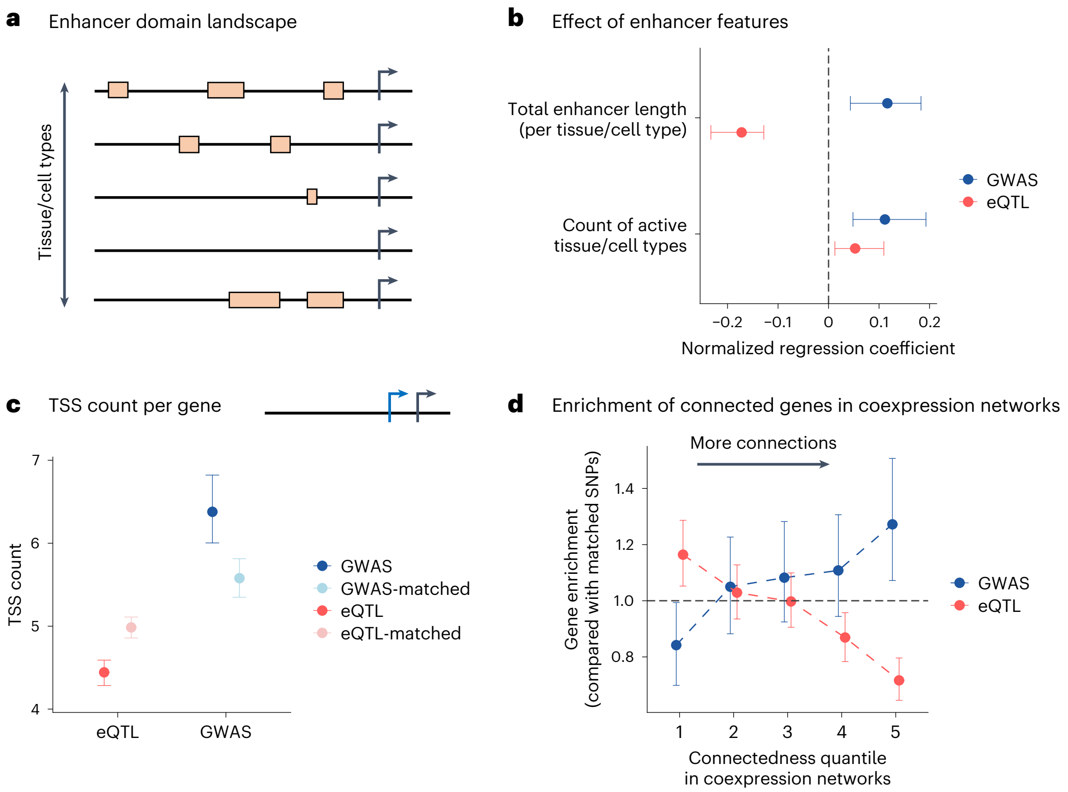


Fig. 3 | GWAS and eQTL genes have different transcriptional regulatory landscapes. a, Schematic of enhancer activity across a number of tissue/cell types. Based on enhancer–gene links inferred from the Roadmap dataset[50], for a given gene, we computed: (1) the number of tissue/cell types in which a gene has an enhancer, and (2) 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 ($N = 100{,}000$) after adjusting for confounders (Methods). **c**, Mean count of TSSs per gene across cell types in the FANTOM project[52] for GWAS and eQTL genes. **d**, Enrichment of GWAS and eQTL genes relative to genes linked to matched SNPs (y axis) in gene bins ranked by connectedness values computed based on coexpression networks from Saha et al.[53] (x axis). In **b**–**d**, error bars show 95% confidence intervals as determined by quantile bootstrapping over 1,000 sampling iterations. For matched SNPs, points show mean values in sets of matched SNPs corresponding to bootstrapped samples. All statistics were computed for 118,996 eQTLs and 22,119 GWAS hits. See Supplementary Table 5 for the counts of genes in each bin shown in panel **d**.

We then considered another regulatory feature: how many different TSSs are used for a given gene across a diverse set of cell types in the FANTOM project[52] (Methods). On average, GWAS genes have more, whereas eQTL genes have fewer TSSs than control SNPs (Fig. 3c): 6.4 versus 5.6 for GWAS ($P = 4 \times 10^{-5}$) and 4.4 versus 5 for eQTLs ($P = 8 \times 10^{-11}$).

We hypothesized that the regulatory landscape of genes, in part, corresponds with their regulatory function in gene regulatory networks. 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 coexpression networks inferred by Saha et al. for GTEx tissues[53]. For each gene, we constructed a connectedness measure based on the gene's number of neighbors in tissue-specific networks (Methods). We found that genes with more connections are progressively enriched near GWAS hits ($P = 5 \times 10^{-3}$ for comparing enrichment in top versus bottom connectedness quantiles) (Fig. 3d), consistent with previous reports that trait heritability is enriched near such genes[54,55]. By contrast, genes with more connections are progressively depleted for eQTLs ($P = 6 \times 10^{-9}$) (Fig. 3d).

Taken together, these observations suggest that targets of GWAS hits are often genes with complex regulatory architecture, in the 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 types of gene are depleted of eQTLs.

Key Gene Ontology terms are enriched in GWAS genes but depleted in eQTL genes

Differences in the selective constraint and regulatory landscapes of genes likely reflect the functional roles of different types of gene. To explore functional disparities between GWAS and eQTL genes, we analyzed 577 Gene Ontology (GO) biological process terms. For each term, we evaluated the enrichment of its associated genes in GWAS and eQTL genes relative to control SNPs across all traits and tissues (Methods). For data visualization, we focused on 41 terms that are broadly unrelated, prioritizing terms that are informative of GWAS/eQTL SNPs (Supplementary Table 6 and Methods).

Notably, we found that GO terms are pervasively enriched among GWAS genes for many traits (Fig. 4a). For some terms, the enriched term highlight evidently relevant traits: for example, 'skeletal system development' for height, 'lipid localization' for high cholesterol and 'adaptive immune response' for a number of blood-related traits. That said, many terms show enrichment across multiple traits.

By contrast, many GO terms show clear depletion of eQTLs in all tissues (Fig. 4a). Consistent with our earlier results, depletion is most prominent for gene sets with high average pLI, suggesting that selection purges eQTLs for genes with evolutionarily important functions. By contrast, certain gene sets, such as 'DNA repair', are enriched for eQTL genes, suggesting that they may be more tolerant of expression changes.

We wondered whether transcription factors exhibit a similar pattern, given their essential role in development and cellular functions.

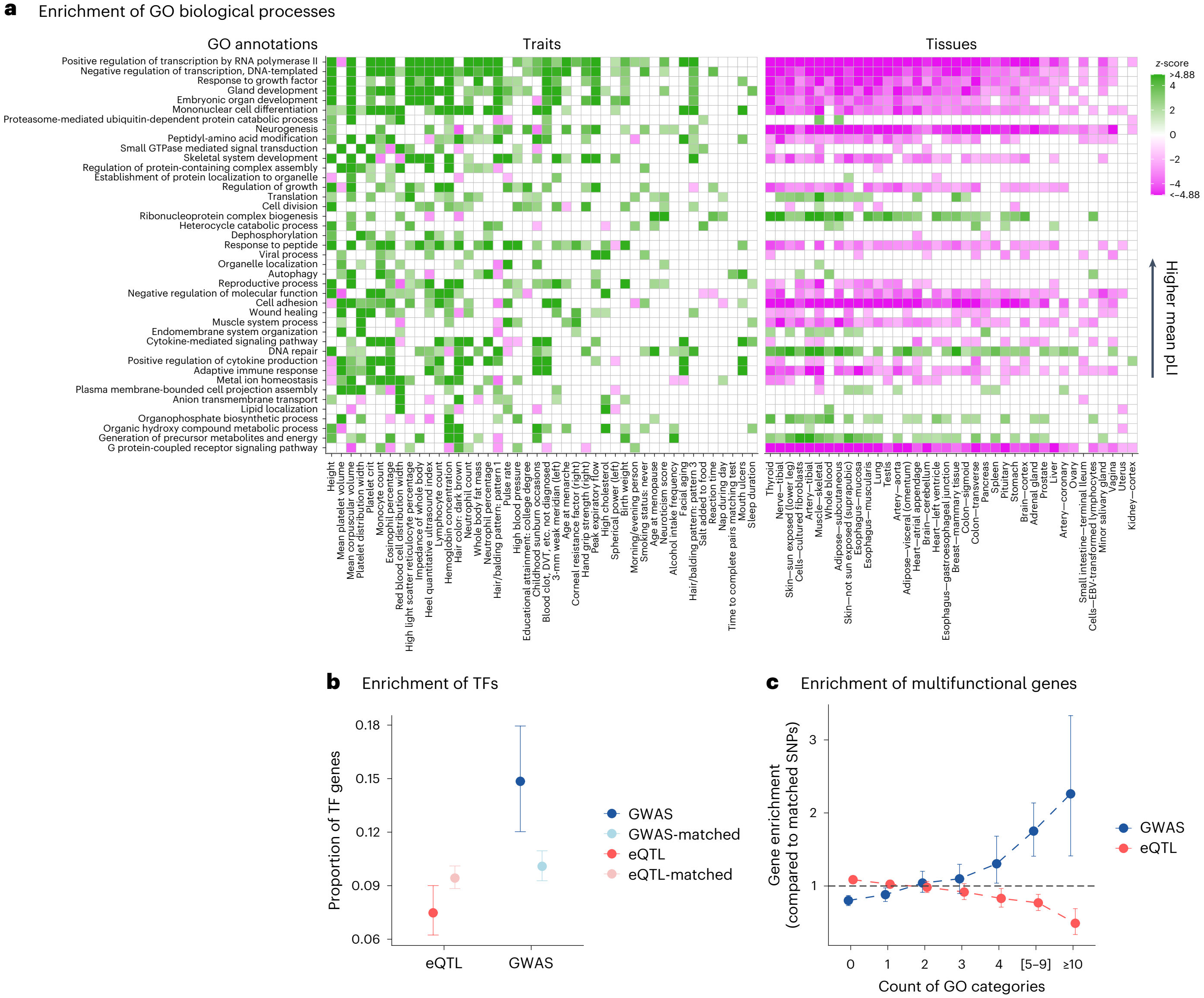


Fig. 4 | Diverse categories of functional genes are enriched among GWAS genes but not among eQTL genes. a, Enrichment of genes associated with 41 GO terms among GWAS and eQTL genes (Methods). Traits and tissues (*x* axis) are sorted by hit count (decreasing from left to right), and GO terms (*y* axis) are sorted by the mean pLI value of associated genes. For each trait- or tissue–GO term pair we computed enrichment *z*-scores based on 1,000 sampling iterations of SNPs matched for MAF, LD score and gene density (Methods). The color map represents enrichment (green) or depletion (magenta) of a given gene set among GWAS or eQTL genes. Absolute *z*-scores less than 1.86, which is the threshold value corresponding to a 5% false discovery rate, are colored white. The maximum color intensity is capped at absolute *z*-score of 4.88 corresponding to the Bonferroni correction threshold (Methods). See Supplementary Table 7 for enrichment and *z*-score values. DVT, deep vein thrombosis; EBV, Epstein-Barr virus. **b**, Fraction of GWAS genes (of 22,119) and eQTL genes (of 118,996) that are transcription factor (TFs). **c**, Enrichment of GWAS and eQTL genes relative to genes linked to matched SNPs (*y* axis) in different bins of genes ranked by the counts of GO terms they belong to (*x* axis). See Supplementary Table 5 for the counts of genes in each bin. Error bars show 95% confidence intervals as determined by quantile bootstrapping over 1,000 sampling iterations. For matched SNPs in **b**, points show mean values in sets of matched SNPs corresponding to bootstrapped samples.

Indeed, GO terms related to transcriptional regulation, along with several developmental terms to which TFs substantially contribute (Extended Data Fig. 5a), are particularly enriched in GWAS genes and depleted in eQTL genes (Fig. 4a). Similarly, we found that transcription factors are enriched in GWAS genes (15% compared with 10% for matched SNPs, $P = 2 \times 10^{-3}$) and depleted in eQTL genes (7% compared with 9% for matched SNPs, $P = 5 \times 10^{-3}$) (Fig. 4b), in line with previous reports that transcription factors are underrepresented among eGenes[56].

Lastly, noting that most GO terms show enrichment for complex traits (Fig. 4a), we reasoned that genes associated with multiple functional categories, multifunctional genes, would be particularly enriched among GWAS genes. Indeed, genes belonging to more GO terms are progressively enriched in GWAS genes ($P = 2 \times 10^{-3}$ for comparing enrichment among genes linked with ≥10 biological processes versus genes with no linked functional annotation), while being modestly depleted for eQTL genes ($P = 8 \times 10^{-10}$) (Fig. 4c). Multifunctional genes tend to be highly connected in protein–protein interaction networks (Extended Data Fig. 6a). Similarly, highly interacting genes are enriched for GWAS hits (Extended Data Fig. 6b).

In summary, our GO analysis shows that a diverse range of biological processes are enriched near GWAS hits; moreover, multifunctional genes are especially enriched for GWAS hits. In contrast, most trait-related terms are underrepresented near eQTLs.

https://doi.org/10.1038/s41588-023-01529-1

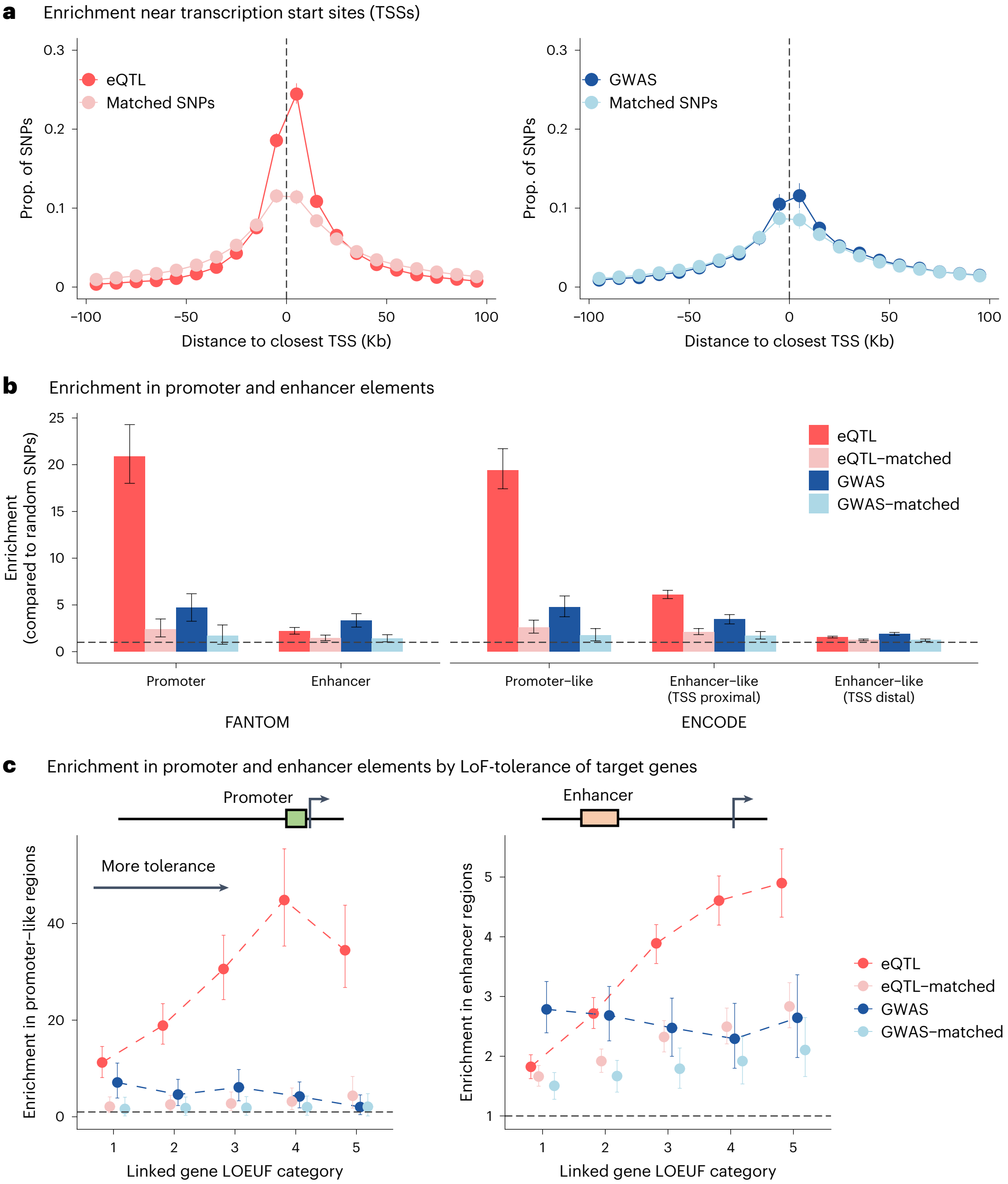


Fig. 5 | GWAS hits are less enriched at TSSs than are eQTLs. a, Distance of eQTLs (left) and GWAS hits (right) to the nearest TSS. Points show fraction of SNPs in 10-kb bins. SNPs more than 100 kb away from their closest TSS are not shown for clarity. **b**, Enrichment of eQTLs and GWAS hits in promoter 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. See Supplementary Table 8 for the counts of SNPs in each annotation. **c**, Enrichment of eQTLs and GWAS hits relative to random SNPs ($N = 6{,}971{,}256$) in ENCODE promoter elements (left) and Roadmap enhancer elements (right) as a function of the quintile of their target gene LOEUF score (a measure of selective constraint[67]). For promoter elements, linking to genes was done taking the closest TSS within 1 kb. For enhancers, enhancer–gene links from Liu et al. were used[50]. See Supplementary Table 8 for the counts of SNPs linked with each LOEUF bin. In **b** and **c**, the black dashed lines mark the value of 1 on the *y* axis. In all panels, error bars show 95% confidence intervals as determined by quantile bootstrapping over 1,000 sampling iterations. For matched SNPs, points show mean values in sets of matched SNPs corresponding to bootstrapped samples. All statistics were computed for 118,996 eQTLs and 22,119 GWAS hits. LoF, loss-of-function.

GWAS hits are further from 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 hits and eQTLs.

It is well known that eQTLs tend to be tightly clustered near TSSs[57–59]. We observed similar results (Fig. 5a): 43% of eQTLs lie within 10 kb of the nearest TSS, compared with 23% for control SNPs (enrichment of 1.88, $P = 7 \times 10^{-168}$). However, GWAS hits are only modestly enriched near TSSs (Fig. 5a): 22% lie within 10 kb of the nearest TSS, compared with 17% for control SNPs (enrichment of 1.28, $P = 2 \times 10^{-9}$). GWAS hits typically lie at greater distances from the nearest TSS (median 36 kb) compared with eQTLs (median 13 kb).

Consistent with these trends, although both GWAS hits and eQTLs are enriched in annotated promoter and enhancer domains (Fig. 5b), the relative enrichment in promoter versus enhancer domains is much stronger for eQTLs than GWAS hits. Also, eQTL enrichment in promoter and enhancer elements sharply increases with decreasing target gene constraint, surpassing GWAS enrichments for the least constrained genes (Fig. 5c). Therefore, differential localization of GWAS hits and eQTLs within regulatory regions cannot be solely attributed to the promoter versus enhancer distinction (eQTLs being promoter variants and GWAS hits being enhancer variants).

A model for variant discovery in GWAS and eQTL assays

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

- At the gene level, GWAS hits are enriched near selectively constrained genes, at genes with complex regulatory landscapes across different cell types and more coexpression partners, and across a wide variety of GO terms and among transcription factors; whereas eQTLs are relatively depleted in all these features.
- At the variant level, GWAS hits show only slight promoter enrichment and tend to be located far from TSSs; by contrast, eQTLs are strongly enriched at promoters and clustered near TSSs.

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. We now describe a model that predicts these qualitative patterns.

We start by examining the case in which 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 in which specific cell types can be studied by cell-sorting or single-cell sequencing. We discuss an extended model for tissue-level eQTLs in the Supplementary Note.

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 discovered variants satisfy the conditions:

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

where c^* is a study-dependent discovery threshold. Specifically, $c^* \propto \chi^2_c/n$, where χ^2_c is the relevant chi-squared critical value for significance and n is the sample size (Methods). Equivalently, in expectation, a variant is discovered if the fraction-of-variance it explains exceeds χ^2_c/n. The fraction-of-variance explained by trait-associated variants is usually much smaller than for *cis*-eQTLs (~10^3-fold smaller), but this is roughly balanced by the much-larger GWAS sample sizes (~10^3-fold larger). We estimate that at current typical sample sizes, both assays are generally low-powered and far from saturation, likely discovering around 10–20% of all causal variants (Supplementary Note). The key question is: should we expect GWAS and eQTL mapping to find the same variants?

We can address the question of GWAS–eQTL overlap in terms of what parts of the parameter space have appreciable power for each assay (Fig. 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 the 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 (Supplementary Fig. 22).

So far, we have described a neutral model assuming independence of effect sizes and p. 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 and Fig. 6b).

Intuitively, selection strength is inversely related to the magnitude of the phenotypic effect, such that the reduction in $E[2p(1-p)]$ at top variants is compensated by their larger $\beta^2\gamma^2$. Consequently, selection does not systematically alter the expected rankings of variants discovered in GWAS compared with the neutral scenario (Fig. 6b). However, it leads to a more uniform distribution of heritability across variants with intermediate and large effects, a phenomenon referred to as 'flattening' (Extended Data Fig. 7a)[40,44]. In the case of eQTLs, selection affects them depending on their phenotypic impact: for eQTLs with a given magnitude of regulatory effect (β^2), selection is stronger on variants acting on genes with larger γ^2 values (Extended Data Fig. 7b). This disproportionately reduces the discovery power for the most trait-relevant eQTLs.

These predicted differences between GWAS and eQTL genes in terms of γ^2 align with the gene-level differences observed in data. Specifically, most complex traits are subject to natural selection, either directly or indirectly. Therefore, a higher γ^2 value, on average, corresponds to stronger selection (Fig. 2). In addition, the phenotypic impact of a gene (γ^2 in our model) likely depends on its involvement in specific biological processes and its connectivity within regulatory networks (Figs. 3 and 4).

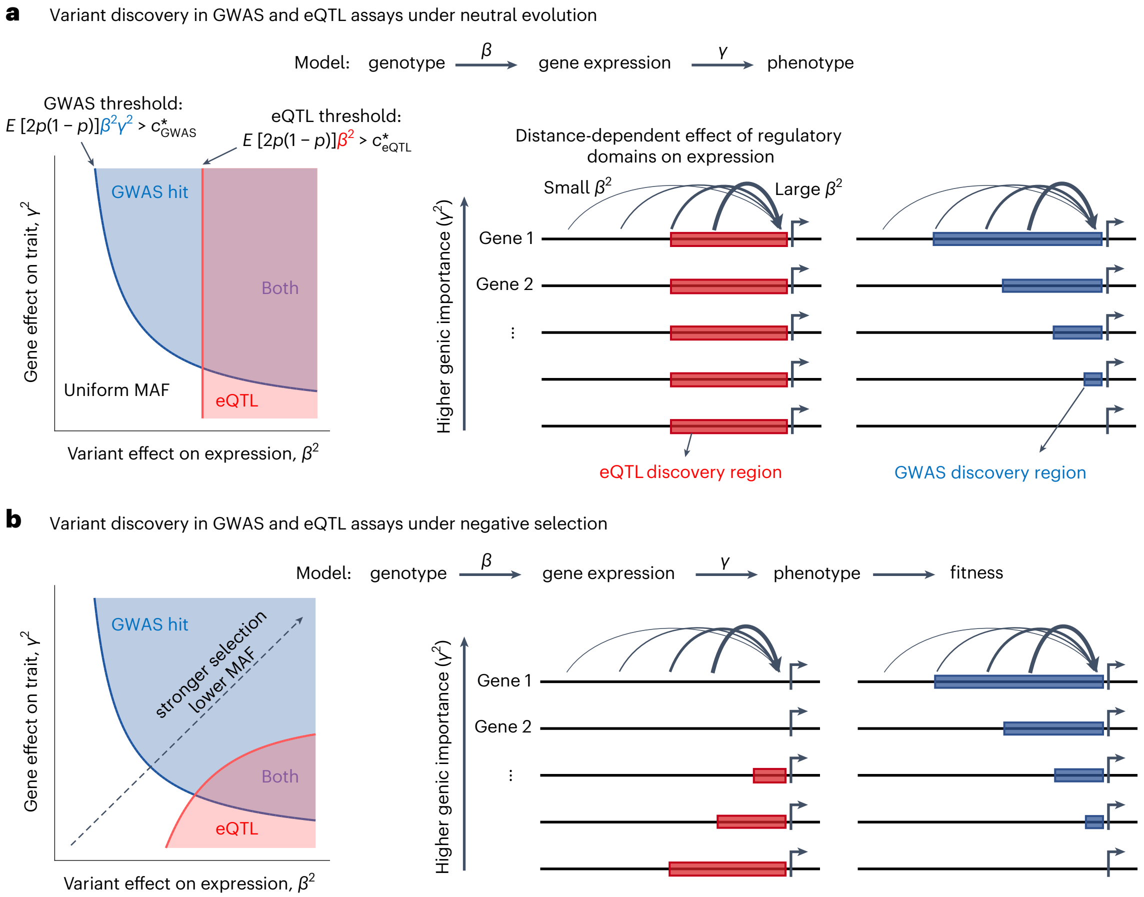
We next considered how features of gene regulation relate to our model. Previous eQTL studies have demonstrated that variants close to the target gene's TSS tend to have larger effect sizes than more distal variants[58,59]. Similarly, promoter–enhancer contact frequencies decay with genomic distance between the enhancer and promoter elements[60,61]. Thus, in terms of our model, average β^2 should decay with distance from the TSS.

Therefore, in the neutral scenario, eQTLs will be skewed towards TSS-proximal regulatory elements (with large β^2), and depleted from distal elements (with small β^2) regardless of the phenotypic importance of the target genes (γ^2) (Fig. 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 (Fig. 6a). Under selection, large-effect TSS-proximal eQTLs should be most depleted from highly important genes, thereby further reducing eQTL–GWAS overlap (Fig. 6b). These predictions are in line with the findings depicted in Fig. 5.

In summary, GWAS and eQTL mapping have power in different areas of the parameter space. Additionally, selection affects the GWAS and eQTL assays differently, primarily hindering the discovery of eQTLs at important genes. In the Supplementary Note, we illustrate that these conclusions are robust to various modeling assumptions (Supplementary Figs. 14–20). We further discuss the limitations of our single-cell type model, and explore more complex scenarios, uncovering additional systematic differences between GWAS and eQTL mapping. Notably, eQTL mapping in bulk assays is skewed towards less trait-relevant cell types (Supplementary Figs. 23 and 24).

Discussion

Most GWAS hits are in the noncoding portion of the genome, and they are highly enriched within active chromatin. It is generally believed


a Variant discovery in GWAS and eQTL assays under neutral evolution

b Variant discovery in GWAS and eQTL assays under negative selection

a
Model: genotype $\xrightarrow{\beta}$ gene expression $\xrightarrow{\gamma}$ phenotype
GWAS threshold: $E\,[2p(1-p)]\beta^2\gamma^2 > c^*_{\mathrm{GWAS}}$
eQTL threshold: $E\,[2p(1-p)]\beta^2 > c^*_{\mathrm{eQTL}}$
GWAS hit
Both
Uniform MAF
eQTL
Gene effect on trait, γ^2
Variant effect on expression, β^2
Distance-dependent effect of regulatory domains on expression
Small β^2
Large β^2
Higher genic importance (γ^2)
Gene 1
Gene 2
⋮
eQTL discovery region
GWAS discovery region
b
Model: genotype $\xrightarrow{\beta}$ gene expression $\xrightarrow{\gamma}$ phenotype ⟶ fitness
GWAS hit
stronger selection
lower MAF
Both
eQTL
Gene effect on trait, γ^2
Variant effect on expression, β^2
Higher genic importance (γ^2)
Gene 1
Gene 2
⋮

Fig. 6 | A model for variant discovery in GWAS and eQTL assays. a, Case of a neutrally evolving phenotype. (Left) 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) 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 discovery regions in the left panel are determined based on how $E[2p(1-p)]$ covaries with effect sizes under a model of selection. Note that the discovery lines shown represent qualitative (and not quantitative) trends, derived under simplifying assumptions for illustrative purposes (Methods). In the Supplementary Note we demonstrate the robustness of discovery trends to various modeling assumptions and choices of parameters.

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[15–17]. 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 fundamental issue is that GWAS and eQTL mapping are powered to identify different types of variant.

We argue that any explanation for limited colocalization must account for the fact that GWAS variants differ from eQTL variants along many important axes. Using carefully matched analyses, we have shown 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. Meanwhile, 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 (Extended Data Fig. 8).

Instead, to understand these observations we note that, like any mapping method, GWAS and eQTL mapping have incomplete power (Supplementary Note). 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. What types of variant, and what types of gene, 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.

By 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. Although 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 outlined 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 (Extended Data Fig. 9). For less-accessible cell types this may simply not be practical. Moreover, even with extraordinarily 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.

Indeed, we observe that these limitations are present even for blood eQTLs from the eQTLGen Consortium, which has a sample size of ~32,000[37]. Of particular note, many critical cell types, such as regulatory T cells, are at low frequency in whole blood, and hence these studies may be underpowered for constrained genes even at vast sample sizes (Supplementary Figs. 12 and 28).

Rather, it seems most likely that the colocalization gap will be solved using a multipronged approach, because 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 for which cell-sorting or single-cell approaches may be more informative[26,30].

Moreover, other types of molecular QTL assay, including chromatin QTLs and splicing QTLs, may help link additional variants to functional effects; although we note that similar discovery biases are liable to act on any such trait (Supplementary Fig. 13). Alternatively, various orthogonal methods, including models that predict the regulatory activity of variants from the DNA sequence[62,63], and emerging functional assays including massively parallel reporter assays, or CRISPR-based variant-editing or enhancer-silencing should not be biased by selection in the same way, although every method has its own limitations[64–66]. 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.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01529-1.

References

1. Claussnitzer, M. et al. A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
2. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
3. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
4. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
5. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
6. Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
7. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
8. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
9. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
10. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
11. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
12. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
13. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
14. GTEx Consortium The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
15. Chun, S. et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
16. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
17. Connally, N. J. et al. The missing link between genetic association and regulatory function. *eLife* **11**, e74970 (2022).
18. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
19. Strober, B. J. et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).
20. D'Antonio-Chronowska, A. et al. iPSC-derived pancreatic progenitors are an optimal model system to study T2D regulatory variants active during fetal development of the pancreas. Preprint at *bioRxiv* https://doi.org/10.1101/2021.03.17.435846 (2021).
21. Walker, R. L. et al. Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell* **179**, 750–771.e22 (2019).
22. Jerber, J. et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53**, 304–312 (2021).
23. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
24. Young, A. M. H. et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nat. Genet.* **53**, 861–868 (2021).
25. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
26. Yazar, S. et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
27. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
28. Calderon, D. et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).
29. Gutierrez-Arcelus, M. et al. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat. Genet.* **52**, 247–253 (2020).
30. Ota, M. et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006–3021.e17 (2021).
31. Mu, Z. et al. The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biol.* **22**, 122 (2021).
32. Hukku, A. et al. Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am. J. Hum. Genet.* **108**, 25–35 (2021).

33. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
34. Li, L. et al. An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat. Genet.* **53**, 994–1005 (2021).
35. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
36. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034.e6 (2019).
37. Võsa, U. et al. Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
38. Pierce, B. L. et al. Mediation analysis demonstrates that *trans*-eQTLs are often explained by *cis*-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS Genet.* **10**, e1004818 (2014).
39. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
40. O'Connor, L. J. et al. Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).
41. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
42. Zeng, J. et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
43. Koch, E. M. & Sunyaev, S. R. Maintenance of complex trait variation: classic theory and modern data. *Front. Genet.* **12**, 763363 (2021).
44. Simons, Y. B., Bullaughey, K., Hudson, R. R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.* **16**, e2002985 (2018).
45. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
46. Siewert-Rocks, K. M., Kim, S. S., Yao, D. W., Shi, H. & Price, A. L. Leveraging gene co-regulation to identify gene sets enriched for disease heritability. *Am. J. Hum. Genet.* **109**, 393–404 (2022).
47. Weiner, D. J., Gazal, S., Robinson, E. B. & O'Connor, L. J. Partitioning gene-mediated disease heritability without eQTLs. *Am. J. Hum. Genet.* **109**, 405–416 (2022).
48. Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
49. Wang, X. & Goldstein, D. B. Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).
50. Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced recombination rate in human regulatory domains. *Genome Biol.* **18**, 193 (2017).
51. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
52. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
53. Saha, A. et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27**, 1843–1858 (2017).
54. Kim, S. S. et al. Genes with high network connectivity are enriched for disease heritability. *Am. J. Hum. Genet.* **104**, 896–913 (2019).
55. Dey, K. K. et al. SNP-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. *Cell Genom.* **2**, 100145 (2022).
56. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
57. Veyrieras, J. B. et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
58. Dimas, A. S. et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246–1250 (2009).
59. Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* **9**, e1003649 (2013).
60. Zuin, J. et al. Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, 571–577 (2022).
61. Fulco, C. P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
62. Nair, S., Kim, D. S., Perricone, J. & Kundaje, A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics* **35**, i108–i116 (2019).
63. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
64. Abell, N. S. et al. Multiple causal variants underlie genetic associations in humans. *Science* **375**, 1247–1254 (2022).
65. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390 (2019).
66. Morris, J. A. et al. Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science* **380**, eadh7699 (2023).
67. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Inclusion and ethics

The study solely relied on genetic associations extracted from publicly available summary statistic data. Individual-level data was not utilized for any association analyses. Specifically, genetic data from individuals labeled as 'white British' within the UKB dataset was exclusively used to compute allele frequencies and serve as a reference panel for assessing LD. The inclusion of these data was granted under application no. 24983 and did not necessitate Institutional Review Board approval. Importantly, no phenotype data from participants within the UKB were involved in this study. All other data used are publicly available (see 'Data availability' section).

Statistics and reproducibility

No preliminary statistical analyses were conducted to determine sample sizes. Publicly available GWAS and eQTL data were downloaded. The inclusion of individuals in these association analyses was based on criteria established by the original research groups, in alignment with study-specific quality control measures. The choice of traits and tissues followed an unbiased procedure, as outlined in the section 'Datasets'. The list of SNPs underwent selection procedures detailed in the section 'SNP selection'. All statistical analyses on the chosen set of SNPs were performed using R (v.3.5.1) unless specified otherwise. The codes to replicate the analyses are publicly available (see 'Code availability' section).

Datasets

GWAS data. We used publicly available GWAS summary statistics for traits in the UKB provided by Ben Neale's lab (see 'Data availability' section). 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 $P < 5 \times 10^{-8}$) that passed our SNP selection criteria ('SNP selection'). For binary traits we used heritability estimates on the liability scale, and further filtered 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 the 44 traits listed in Supplementary Table 1. 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 'Data availability' section) 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[68]. In total, 38 tissues were included and are listed in Supplementary Table 1. To match the GWAS data, all genomic coordinates were mapped to the hg19 assembly using LiftOver[69]. In one analysis (Extended Data Fig. 8) we also used eQTLs detected: (1) in fetal brain samples by Aygün et al.[70], (2) at multiple stages of iPS cells differentiation towards neuronal fate by Jerber et al.[22], and (3) in single-cell analyses of blood cell types by Yazar et al.[26]. In all analyses we selected eQTLs whose target eGenes were among 18,332 protein-coding genes ('Gene selection').

UK Biobank. We used the UKB resource, specifically for the computation of allele frequencies, and as an LD reference panel. For these purposes, we used quality control (QC) measures provided by UKB to select participants for whom their reported gender, 'Submitted.Gender', matched their 'Inferred.Gender' from genotypes; who were not identified as heterozygosity outliers ('het.missing.outliers'== 0); did not have an 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 the UKB to be of 'white British' 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 an LD reference panel.

Gene selection

We selected 18,332 genes (Supplementary Table 2) that (1) were annotated as protein-coding and (2) were linked with a HUGO Gene Nomenclature Committee (HGNC)-approved gene nomenclature (linked with a HGNC ID) in the GENCODE Basic gene annotations (release 39; see 'Data availability' section). We used the HGNC IDs to link genic features from multiple resources to avoid issues with regard to gene names mismatching.

SNP selection

For our SNP selection process we started with the list of 13.7 million variants that passed quality control measures for the UKB 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 white British individuals in the UKB ('Datasets'); polymorphic in the 1000 Genomes Project phase 3 data (used by A. Price's lab for LD score regression; see 'Data availability' section). 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 (v.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 1 Mb. The UKB resource was used as the LD reference panel ('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 (1) SNPs in LD ($r^2 > 0.8$) with predicted protein-truncating or missense mutations (annotated as 'ptv' or 'missense' using the Variant Effect Predictor, in the Neale lab data), to condition on SNPs putatively acting through gene regulation, and (2) SNPs >1 Mb away from the TSS of any of 18,332 protein-coding genes ('Gene selection'), which are not tested for eQTLs in GTEx. We further removed SNPs in the major histocompatibility complex region (chr6:28477797-33448354). This resulted in 22,119 GWAS hits across traits, and 118,996 eQTLs across all gene–tissue pairs (Supplementary Tables 3 and 4).

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 6,971,256 SNPs (from the 8,136,100 filtered SNPs) that were not among variants in LD with predicted protein-truncating or missense mutations ($r^2 > 0.8$), were within 1 Mb of the TSSs of the 18,332 protein-coding genes, and were not in the major histocompatibility complex region. From this set, we randomly sampled 1,000 SNPs for each GWAS hit or eQTL matching for MAF, LD score and gene density (see '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 'Data availability' section). We also retrieved total coding sequence length for the longest transcript of genes from Ensembl's BioMart tool (see 'Data availability' section). These annotations were used as covariates in our logistic regression models ('Statistical methods').

Selective constraint. We used pLI and LOEUF (loss-of-function observed/expected upper bound fraction) measures of intolerance to loss-of-function mutations extracted from gnomAD.v2.1's pLoF metrics by gene data[45,67] (see 'Data availability' section). Measures such as pLI and LOEUF are proxies for selection acting on a gene, and do not provide an interpretable measure of selection[48]. We therefore also considered a direct measure of selection, the *hs* parameter estimated by Agarwal et al.[71], which quantifies the fitness cost of losing one copy of a gene.

Enhancer features. We considered enhancer–gene links based on two different approaches: (1) links inferred by Liu et al. based on correlation of chromatin marks with gene expression[50] (see 'Data availability' section), and (2) links predicted based on the activity-by-contact model from Nasser et al.[51] (see 'Data availability' section). In both data, we first compiled the union of all enhancer intervals per gene per biosample. We then computed two features for a given gene: (1) the number of biosamples in which the gene is linked with at least one enhancer interval (that is, count of active biosamples); and (2) 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)[52]. We downloaded combined hg19 CAGE peaks from FANTOM5 phase 1 and phase 2 data (see 'Data availability' section). 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 coexpression networks. We analyzed the coexpression networks from Saha et al.[53] for 16 tissues in GTEx v6p data (available from the GTEx portal; see 'Data availability' section). 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. We used the igraph package (v.1.3.5)[72] in R[73] (v.4.1) 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 (product of the ranks of a given gene) across the 16 tissues, to construct a connectedness measure (genes with lower rank-product have higher connectedness).

GO annotations. We focused on 577 GO biological process terms with at least 400 genes. To this end, we first obtained specific GO terms linked with genes using the biomaRt package (v.2.48.3)[74] from the Bioconductor project (v.3.13) (attribute 'go_id'). We then used these gene–GO links in the topGO package (v.2.44)[75] (as the 'gene2GO' parameter) to extract all genes associated with GO terms (using the 'genesInTerm' function).

Transcription factors. We downloaded a list of 1,639 putative human transcription factors from Lambert et al.[76] (see 'Data availability' section).

Connectedness in protein–protein interaction networks. We used the Genoppi package (v.1.0.13)[77] in R (v.4.1) to retrieve scored InWeb protein–protein interaction data (loading the 'inweb_table' data)[78]. We then used igraph (v.1.3.5) 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 'Data availability' section) 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 (v.105) as well as the archived versions.

SNP annotations

We compiled a number of SNP annotations.

MAF. We used the UKB data ('Datasets') to compute MAFs within unrelated individuals identified as white British.

LD score. We used the ldsc software (v.1.0.1; see 'Data availability' section)[79] to compute LD scores using a window size of 1 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 (v.1.1)' LD annotations for SNPs in the 1000 Genomes Project phase 3 data (see 'Data availability' section). For the LD reference panel, we used 10,000 randomly selected unrelated white British individuals in the UKB ('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 (±500 kb) around the SNP. The TSS coordinates were extracted from the GENCODE Basic annotations ('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 ('Gene annotations').

Overlap with promoter/enhancer elements. We considered two sets of putative promoter and enhancer elements: (1) mapped by the FANTOM5 consortium using CAGE, and (2) 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 'Data availability' section). 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 2 kb of nearest TSS), and distal enhancer-like (>2 kb away from the nearest TSS) elements. We used ENCODE's Registry V2 of cCREs (see 'Data availability' section) to download the corresponding regions. Regions reported in GRCh38 were mapped to hg19 using liftOver. For our analysis in Fig. 5c, we further linked the promoter-like regions to nearest TSS conditional on the distance of the regions midpoints to TSSs were <1 kb. 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 (for example, pooled set of all GWAS hits) and for a given feature (for example, binary indicator of falling in enhancer elements) we computed mean values across all SNPs in the set. For genic features (for example, binary indicator of belonging to a given GO annotation) we first linked SNPs to genes, and then computed mean feature values corresponding to the linked genes over all SNPs. In all analyses we linked SNPs to the genes with the closest TSSs, with the exception of Fig. 2c and Extended Data Fig. 1 where we also linked eQTLs to their target eGenes.

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.[80]) 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 1,000 times to construct 1,000 bootstrapped samples, and computed confidence intervals as the range between 2.5th and 97.5th percentiles across all samples.

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 6,971,256 SNPs and excluding the focal SNP; 'SNP selection') (1) with the same gene density as the focal SNP, (2) with MAF within 0.02 of the focal SNP's MAF and (3) with LD score within 0.1 s.d. (estimated across all SNPs) of the focal SNP's LD score. We then sampled 1,000 times from this set at random with replacement to construct 1,000 instances of control SNPs per SNP of interest. For our analyses of individual traits and tissues (for example, shown in Fig. 2b), we linked these control SNPs to all SNPs in a set of GWAS hits or eQTLs to give 1,000 sets of matched SNPs per trait or tissue. We then used the distribution of genic features across these matched sets to compute *z*-scores ('Analyses of individual traits and tissues'). For our analyses of the pooled set of GWAS hits (across all traits) and the pooled set of eQTLs (across all tissues), we paired each instance of sampled control SNPs to each bootstrapped sample described above, forming 1,000 sets of bootstrapped matched SNPs. For all gene or SNP features studied (for example, Figs. 2a and 5a), we computed and report mean values and confidence intervals (as the range between 2.5th and 97.5th percentiles) across these bootstrapped samples.

Computing *P* values. For a number of analyses we report *P* values comparing genic or SNP features of GWAS hits (pooled across traits) or eQTLs (pooled across tissues) and their corresponding control SNPs. Specifically, for each bootstrapped sample described above, we computed the difference in features between GWAS hits or eQTLs and matched SNPs. Assuming that the difference values are normally distributed, we utilized the distributions across 1,000 bootstrapped samples to compute *z*-scores as the mean values divided by the standard deviations. Subsequently, we calculated the *P* values using a two-tailed test. No adjustments were made for multiple comparisons. For comparing the enrichment of GWAS hits or eQTLs in different gene sets, specifically the top versus the bottom quantile of connectedness in Fig. 3d and genes with high versus low number of linked GO terms in Fig. 4c, we computed *P* values based on the distribution of difference values as: (enrichment in gene set with high feature value relative to matched SNPs) − (enrichment in gene set with low feature value relative to matched SNPs).

Gene comparison analysis using logistic regression. In Fig. 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 6,971,256 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 the nearest TSS. We used the same logistic regression framework in Extended Data Fig. 3 but with different gene-level measures relating to selective constraint as predictors (one at a time).

Analyses of individual traits and tissues. In Figs. 2b and 4a we studied gene features separately for individual traits and tissues. For a given feature (for example, 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 (1,000 sets of SNPs for each set of GWAS hits and eQTLs; 'Control SNPs'), and used the distributions to compute (1) enrichment values as the estimated values for GWAS hits or eQTLs divided by the matched samples mean, and (2) *z*-scores as the matched samples mean subtracted from the estimated values for GWAS hits or eQTLs divided by the matched samples standard deviation. For these analyses we report *z*-score thresholds corresponding to the Bonferroni correction for multiple testing. In Fig. 2b we analyzed 82 traits and tissues corresponding to a multiple testing correction *P* value threshold of $0.05/82 = 6.1 \times 10^{-4}$, and a *z*-score threshold of 3.43. The analysis in Fig. 4a is based on 577 GO terms (see below) for 82 traits and tissues, and thus the conservative *P* value threshold for multiple testing correction is $0.05/(577 \times 82) = 1.1 \times 10^{-6}$, corresponding to a *z*-score threshold of 4.88. We also report the *z*-score threshold corresponding to a 5% false discovery rate computed using the qvalue package (v.2.24) in R (v.4.1)[81].

Analyses of eQTLs by rank. For the analysis in Fig. 2c we grouped eQTLs by their rank based on association strengths (*P* values) in individual tissues. To this end, in all 38 tissues, we first bin ranked eQTLs by association *P* values in groups of 1,000 eQTLs: first top 1,000 eQTLs as group 1, second top 1,000 eQTLs as group 2 and so on. We then pooled eQTLs across tissues by the ranked bins.

Selection of broadly unrelated GO terms. In Fig. 4a we show enrichment *z*-scores for GWAS hits and eQTLs across gene sets associated with 41 GO terms. GO terms are hierarchical and thus interdependent. We selected these 41 terms by pruning 577 GO biological process terms ('Gene annotations') to give a set of broadly unrelated terms while retaining those relevant to the traits and tissues studied here. To this end, we first determined enrichment *z*-scores for all 577 terms (Supplementary Table 7). Then, we selected the top term (the most enriched or depleted) for each tissue and trait, conditioning on GO terms with <3,000 associated genes. This gives 52 unique terms. We then pruned this set as follows: we sorted terms by the count of associated genes in an ascending order, and iterated over terms starting with the term with the least number of genes. At each iteration, we retained the focal term if gene associations with that term could not be well-predicted from the previously included terms, defined as the area under the receiver operating characteristic curve (AUC) value of <0.75 estimated using penalized logistic regression (as implemented in the glmnet package (v.4.1-3) in R (v.4.0.2) using the 'cv.glmnet' function for cross-validation[82]) over all protein-coding genes. This gives 27 broadly unrelated terms. Using the same procedure, we built upon this set by iterating over the rest of the terms that were not among the top terms with respect to GWAS hit or eQTL enrichments, resulting in 41 terms (Supplementary Table 6).

Analysis of non-GTEx eGenes. In Extended Data Fig. 8, we analyzed eGenes identified (1) in fetal brain samples by Aygün et al.[70], (2) at multiple stages of iPS cell differentiation towards neuronal fate by Jerber et al.[22] and (3) 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 (1) enrichment values as the estimated values for eGenes divided by the random samples mean and (2) *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 provide additional modeling details and describe how we derived the discovery regions for GWAS and eQTL assays (Fig. 6).

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 illustrate the qualitative effect of selection on variant discovery in GWAS and eQTL assays. In the Supplementary Note (Supplementary Figs. 14–21) we show that our key qualitative results are robust to modeling assumptions and choices of parameters, and discuss the challenges of quantitative analyses in more detail.

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 $E[2p(1-p)\beta^2\gamma^2|\beta, \gamma, \text{neutrality}] = E[V_p|\text{neutrality}]\beta^2\gamma^2 \propto \beta^2\gamma^2$ (Extended Data Fig. 7a), where we defined $V_p = 2p(1-p)$ as the variance in allele frequency. Under selection, $E[V_p|\beta, \gamma, \text{selection}]$ is negatively correlated with $\beta^2\gamma^2$, such that $E[V_p|\beta, \gamma, \text{selection}] < E[V_p|\beta, \gamma, \text{neutrality}]$. Therefore, the expected contribution of variants to phenotypic variance increases more slowly with phenotypic effect size under selection compared with the neutral case (Extended Data Fig. 7a); that is, selection has a 'flattening' effect. This is supported by empirical evidence[43,83], as well as modeling work on the effect of selection on genetic architecture of complex traits[44].

Our key insights on the effect of selection on variant discovery in GWAS and eQTL assays are based on the qualitative considerations discussed above, and hold regardless of how the negative correlation between allele frequency and effect size is formulated. Nevertheless, to illustrate the qualitative effect of selection, we used an asymptotic exponential form to describe the relationship $E[V_p\beta^2\gamma^2|\beta, \gamma] \propto \kappa(1 - e^{-\beta^2\gamma^2/\kappa})$ (Extended Data Fig. 7a). In the main text we set $\kappa = 2.986$; under this model, and drawing β and γ effects from independent standard Normal distributions, such a κ reduces $E[V_p]$ compared with the neutral case by ~10%. We show in Supplementary Figs. 14–20, that as long as $E[V_p|\beta, \gamma, \text{selection}]$ decreases with increasing phenotypic effect size, that is, $\beta^2\gamma^2$, our qualitative conclusions are robust to (1) the mathematical form describing the relationship between expected phenotypic variance and effect size, (2) the magnitude of the relative reduction in phenotypic variance due to selection compared with neutrality, and (3) the underlying joint distribution of effects, β and γ.

Estimation of discovery regions using simulations. 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 Fig. 6, to focus on differences between GWAS and eQTL assays that are due to basic features of these approaches rather than power, we set the c^*_{eQTL} and c^*_{GWAS} thresholds such that under our modeling choices the same fraction of causal SNPs is discovered in either assay. For illustrative purposes, we set the power thresholds at 15% on par with rough estimates of discovery power at current samples sizes (Supplementary Note). (In Extended Data Fig. 9 and Supplementary Fig. 22 we vary the discovery threshold to mimic the effect of increasing sample size.) To this end, modeling β and γ effects to be independent and normally distributed, we first sampled 10 million pairs of (β, γ) ~ $N(\mathbf{0}, \mathbf{I}_2)$. For each pair we computed (1) the expected contribution to variance in phenotype, $E[V_p|\beta, \gamma]\beta^2\gamma^2$, and (2) 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 (that is, 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 β^2 ~ $\chi^2(1)$), 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 (v.4.1), applied to points on the grid between the 84.9th and 85.1th percentiles of the distributions of variance in phenotype or expression used to determine each c^*. We show in Supplementary Figs. 14–20 that the discovery lines are qualitatively similar under different distributions of effects, β and γ, and selection models.

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 $\epsilon \sim N(0, \sigma^2_\epsilon)$ is the noise term capturing the effect of environment as well as other causal SNPs (genetic background). In ordinary least squares regression, the effect estimate $\hat{\beta}$, is normally distributed with expectation $E[\hat{\beta}] = \beta$. The variance of $\hat{\beta}$ is:

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

where we made the assumption that the contribution of individual SNPs to phenotypic variance is small such that $\text{Var}[\varepsilon] \approx \text{Var}[Y]$. Now the effect is deemed significant if the squared z-score is large enough:

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

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

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

$$h^2_{\text{SNP}} > \chi^2_{\text{c}}/n.$$

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data generated by or processed for this study can be found in Supplementary Tables, on Zenodo with https://doi.org/10.5281/zenodo.6618073 (ref. 84), and on GitHub (https://github.com/hakha-most/gwas_eqtl) with https://doi.org/10.5281/zenodo.8330029 (ref. 85). Public data used in this study are accessible via URLs cited at appropriate locations in the Methods, as listed: Neale lab UKB data: http://www.nealelab.is/uk-biobank GTEx data: https://gtexportal.org/home/datasets; NCBI's gene_info file: https://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.

Code availability

Codes used to process and analyze GWAS and eQTL data are available on GitHub (https://github.com/hakha-most/gwas_eqtl) with https://doi.org/10.5281/zenodo.8330029 (ref. 85).

References

68. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
69. Hinrichs, A. S. et al. The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
70. Aygün, N. et al. Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. *Am. J. Hum. Genet.* **108**, 1647–1668 (2021).
71. Agarwal, I., Fuller, Z. L., Myers, S. R. & Przeworski, M. Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. *eLife* **12**, e83172 (2023).
72. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).
73. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020); https://www.R-project.org/
74. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
75. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for Gene Ontology. R package version 2.44.0 (2021).
76. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
77. Pintacuda, G. et al. Genoppi is an open-source software for robust and standardized integration of proteomic and genetic data. *Nat. Commun.* **12**, 2580 (2021).
78. Li, T. et al. A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64 (2017).
79. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
80. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).
81. Storey, J. D., Bass, A. J., Dabney, A. & Robinson, D. qvalue: Q-value estimation for false discovery rate control. R package version 2.24.0 http://github.com/jdstorey/qvalue (2021).
82. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
83. Schoech, A. P. et al. Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 790 (2019).
84. Mostafavi, H. Supplementary data for 'Systematic differences in discovery of genetic effects on gene expression and complex traits'. *Zenodo* https://doi.org/10.5281/zenodo.6618073 (2023).
85. Mostafavi, H. Code repository for 'Systematic differences in discovery of genetic effects on gene expression and complex traits'. *Zenodo* https://doi.org/10.5281/zenodo.8330029 (2023).

Acknowledgements

This research has been conducted using the UK Biobank resource under application number 24983. We thank the Rivas lab at Stanford University for assistance with accessing this resource. We are grateful to J. Engreitz, M. Przeworski, G. Sella, A. Kundaje, Y. Simons, I. Agarwal, M. Ota, R. Patel and members of the Pritchard lab for helpful conversations, and to J. Engreitz, B. Pasaniuc, A. Battle, A. Harpak, M. Przeworski, G. Sella and W. Wohns for valuable feedback on an earlier draft of the manuscript. This research was supported by National Institutes of Health grants R01HG008140 and R01HG011432 to J.K.P., and U01HG012069 to A. Kundaje. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

H.M. and J.K.P. conceived and designed the study. H.M. performed all data analyses and developed the model. J.P.S. contributed to the design and interpretation of the statistical analyses and validation of the model. J.P.S. and S.N. provided intellectual contributions to all aspects of the study. H.M. and J.K.P. wrote the paper. J.K.P. supervised the study and acquired funding.

Competing interests

The authors declare no competing interests.

Additional information

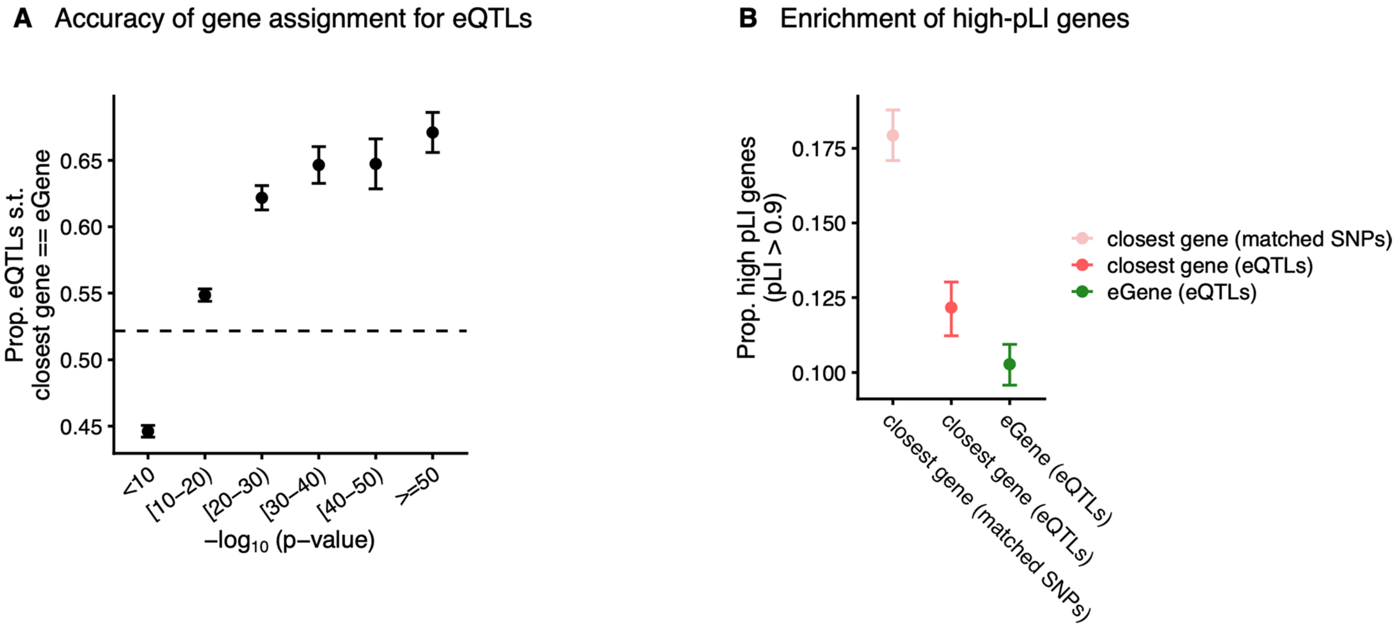
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01529-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01529-1.

Correspondence and requests for materials should be addressed to Hakhamanesh Mostafavi or Jonathan K. Pritchard.

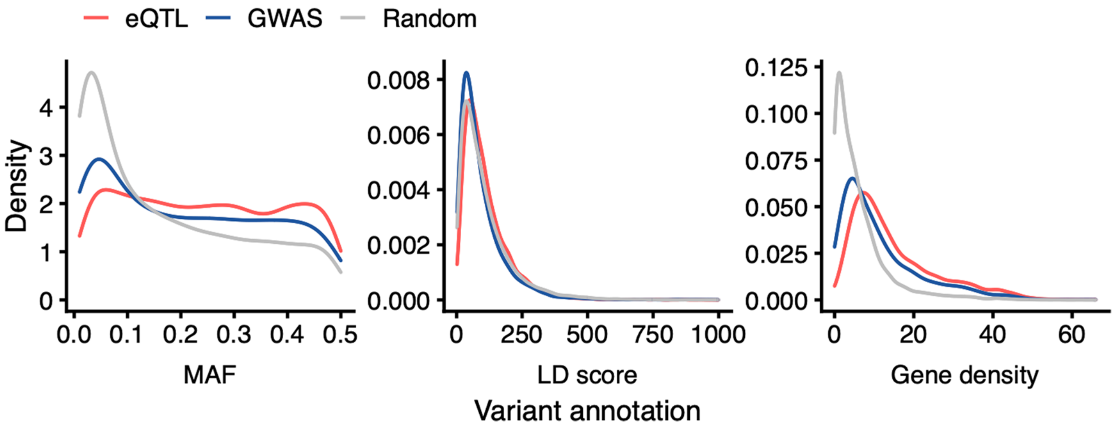
Peer review information *Nature Genetics* thanks Tiffany Amariuta, Andrew Jaffe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

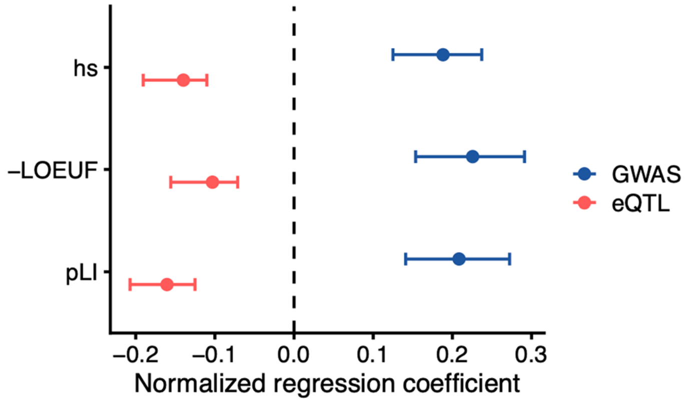


Extended Data Fig. 1 | Genes closest to eQTLs versus eGenes. (A) Fraction of eQTLs for which the target eGene is also the gene with the closest TSS, as a function of eQTL association p-value. Error bars show ± 2 standard errors computed as $\sqrt{2f(1-f)/M}$, where f is the estimated fraction, and M is the number of eQTLs per p-value group. In the p-value groups shown, from left to right, there are 50,859, 45,650, 11,246, 4,781, 2,575, and 3,885 eQTLs, respectively. The dashed line shows the mean value of 0.52 across all eQTLs. (B) Same as Fig. 2a, but with different gene assignments to eQTLs (N=118,996). Fraction of eGenes linked to eQTLs (green), or closest genes to eQTLs (red), or closest genes to control SNPs matched for MAF, LD score and gene density (light red) with high pLI (pLI > 0.9, a measure of selective constraint). Error bars corresponding to eQTL properties (red and green points) show 95% confidence intervals as determined by quantile bootstrapping. For matched SNPs (light red), points and error bars show mean values and 95% confidence intervals in 1000 sampling iterations.

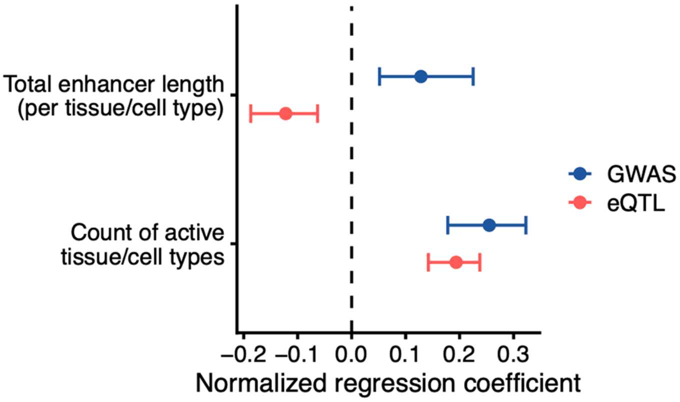
https://doi.org/10.1038/s41588-023-01529-1



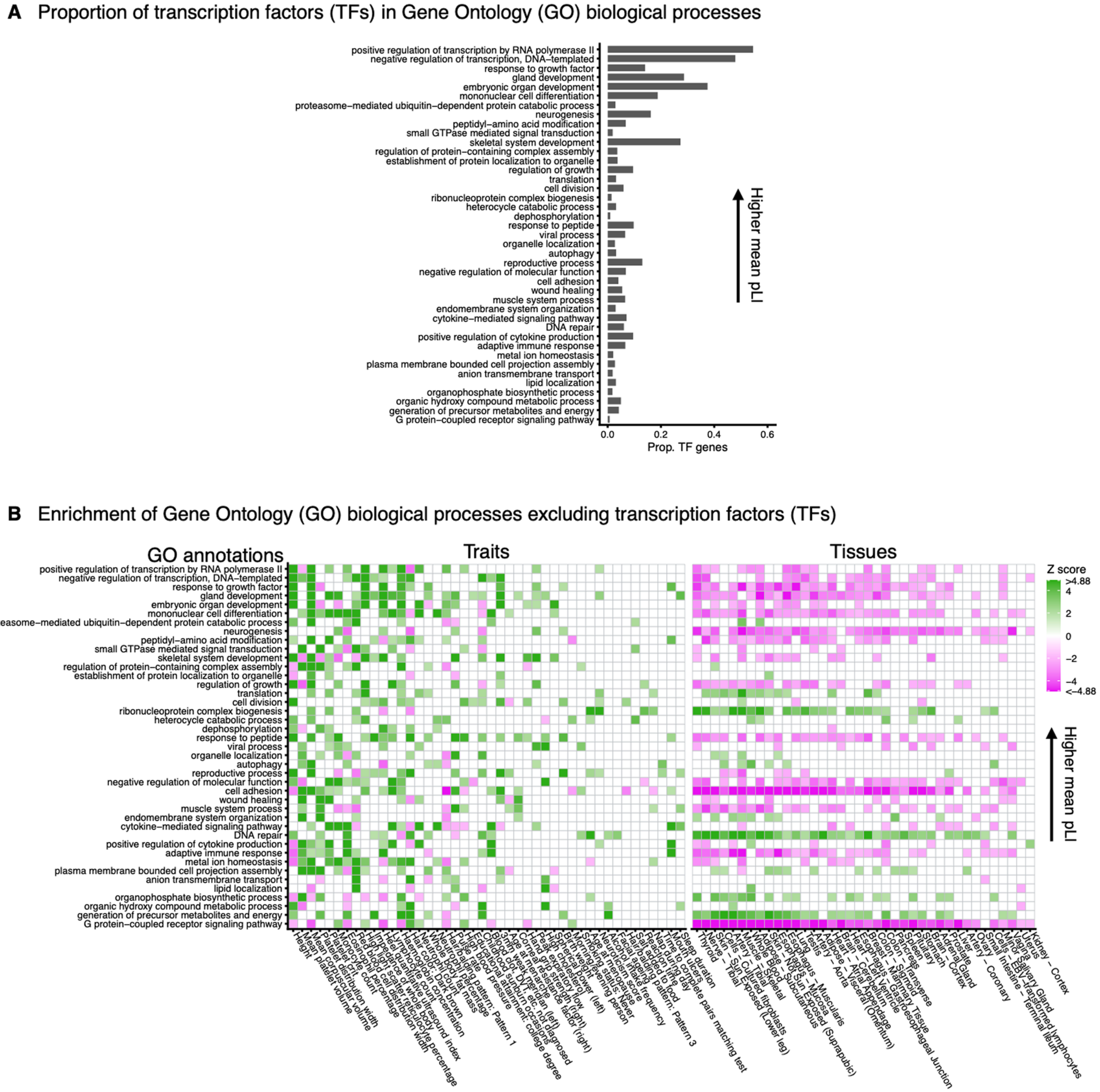
Extended Data Fig. 2 | Basic variant-level differences between GWAS hits and eQTLs. Distribution of minor allele frequency (MAF), linkage disequilibrium (LD) score and gene density for 118,996 eQTLs (red), 22,119 GWAS hits (blue), and 100,000 randomly chosen variants. LD score values are cut at 1000 for clarity.



Extended Data Fig. 3 | GWAS and eQTL genes are under different selective constraints: robustness to gene-level measures of selective constraints. Logistic regression coefficients corresponding with different gene-level measures of selection for predicting GWAS hits (N=22,119) or eQTLs (N=118,996) versus random SNPs (N=100,000) after adjusting for confounders (see Methods). Results are plotted as regression coefficients on the original data with error bars showing the 2.5th and 97.5th percentile over 1000 bootstrap samples. The measures of selection are pLI and LOEUF from the gnomAD study[45,67], and *hs* estimates from Agarwal et al.[71]. Lower LOEUF values correspond to higher selective constraints, therefore we used -LOEUF values to match other measures, such that higher values mean higher constraint levels.

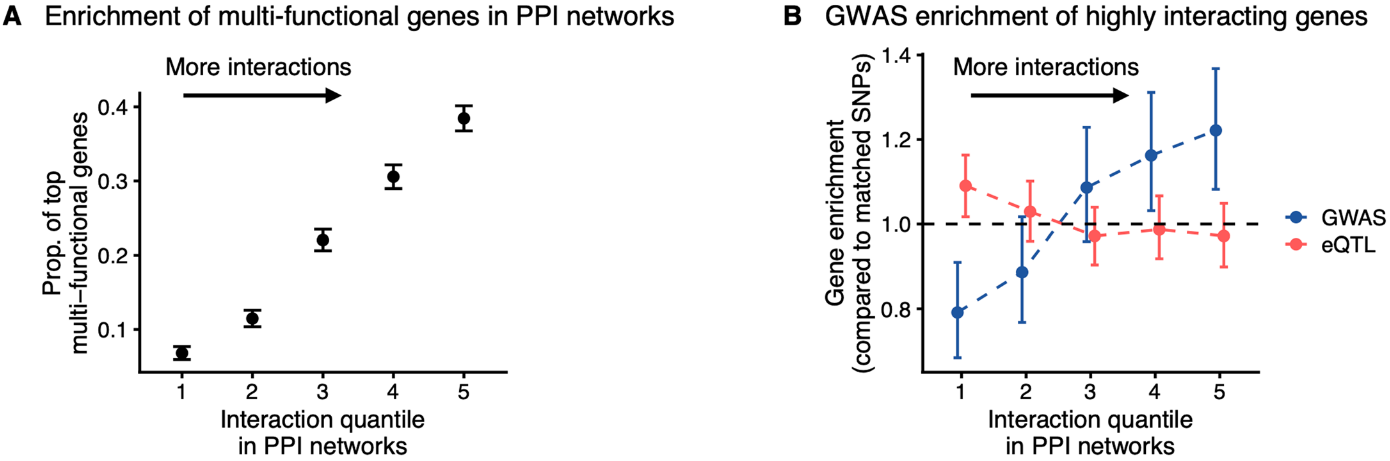


Extended Data Fig. 4 | GWAS and eQTL genes have different enhancer architectures. Same as Fig. 3b, but using enhancer-gene links predicted by the activity-by-contact (ABC) model from Nasser et al.[51] (Methods). For a given gene, we computed (i) the number of biosamples in which a gene has an enhancer, 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 22,119 GWAS hits (blue) and 118,996 eQTLs (red) versus 100,000 random variants after adjusting for confounders (Methods). Results are plotted as regression coefficients on the original data with error bars showing the 2.5th and 97.5th percentile over 1000 bootstrap samples.

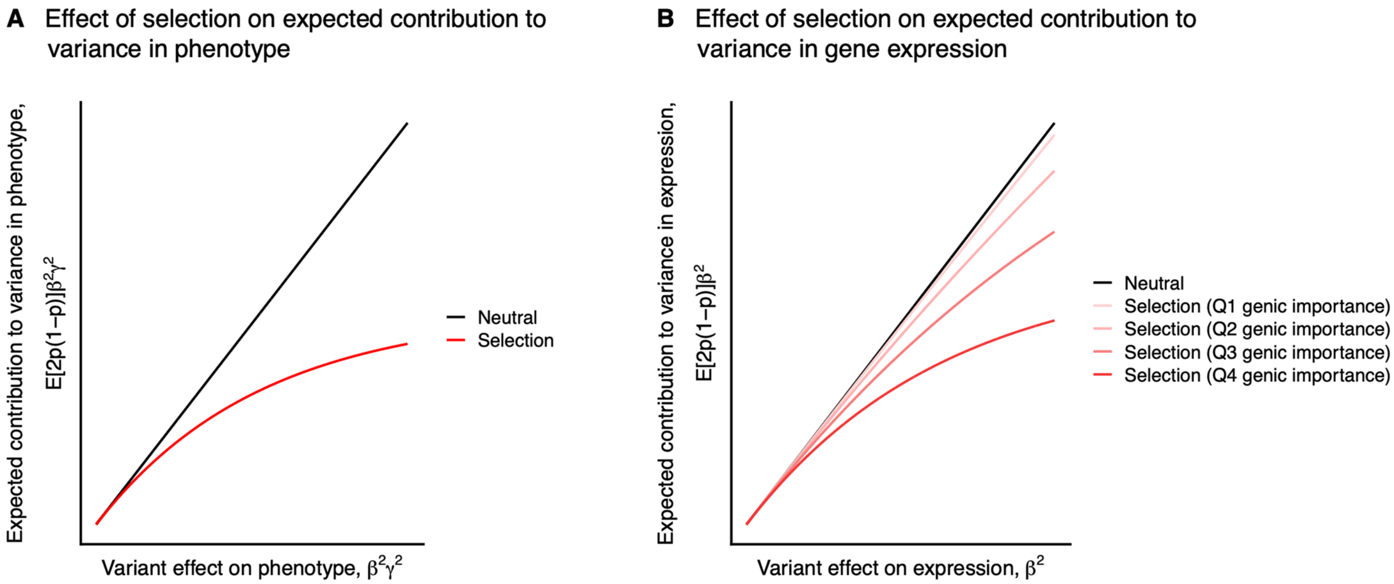


Extended Data Fig. 5 | Contribution of transcription factors (TFs) in Gene Ontology (GO) annotations and their enrichment in GWAS and eQTL genes. (A) Proportion of TFs in 41 GO biological processes shown in Fig. 4a. (B) Same as Fig. 4a, but now excluding TFs from all 41 gene categories before computing enrichment values among GWAS and eQTL genes. Traits and tissues (x-axis) are sorted by hit count (decreasing from left to right), and GO terms (y-axis) are sorted by the mean pLI value of associated genes (before removing TFs, replicating the ordering in Fig. 4a). For each trait- or tissue-GO term 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 (green) or depletion (magenta) of a given gene set among GWAS or eQTL genes. See Fig. 4a for additional details.

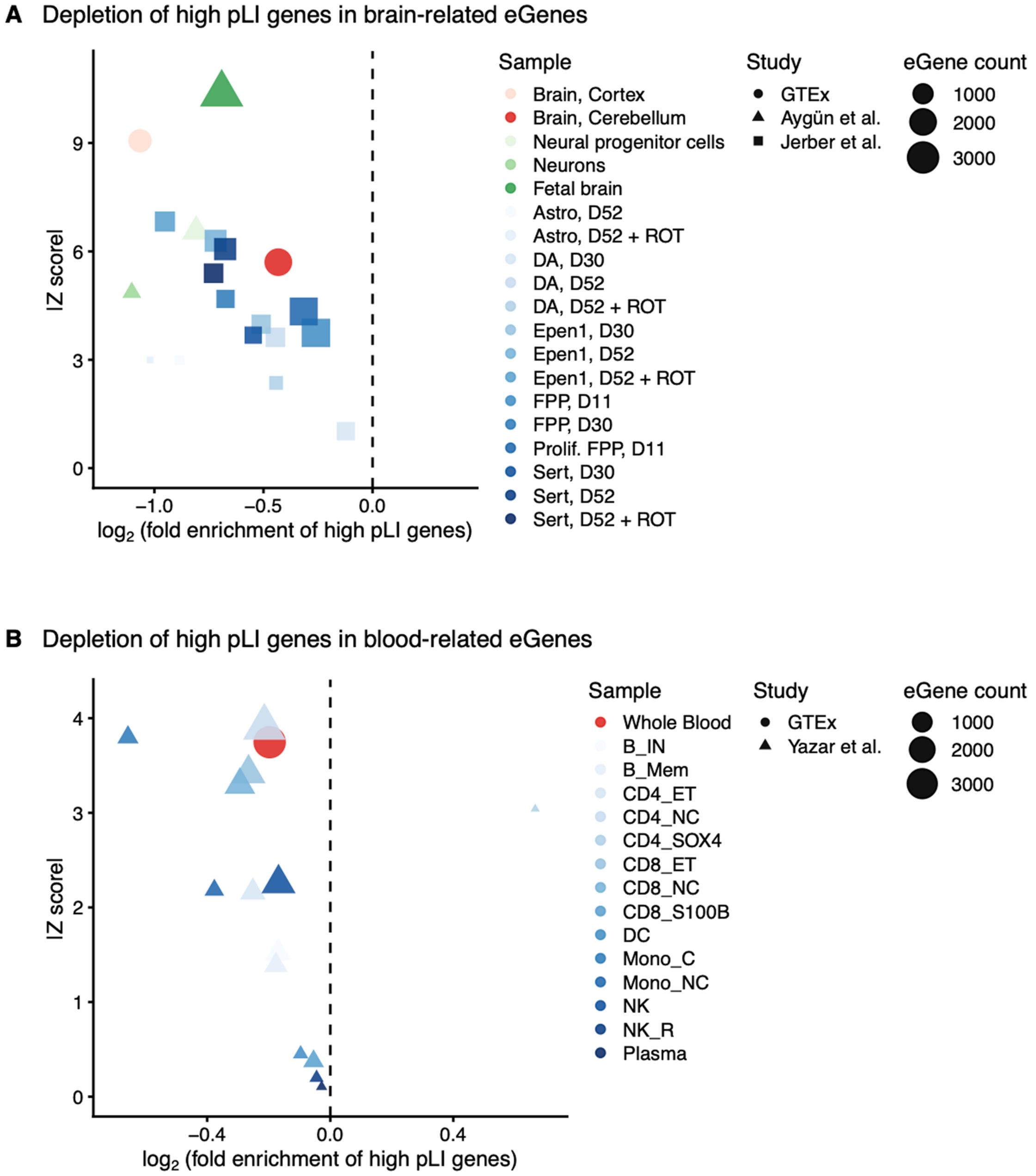
https://doi.org/10.1038/s41588-023-01529-1



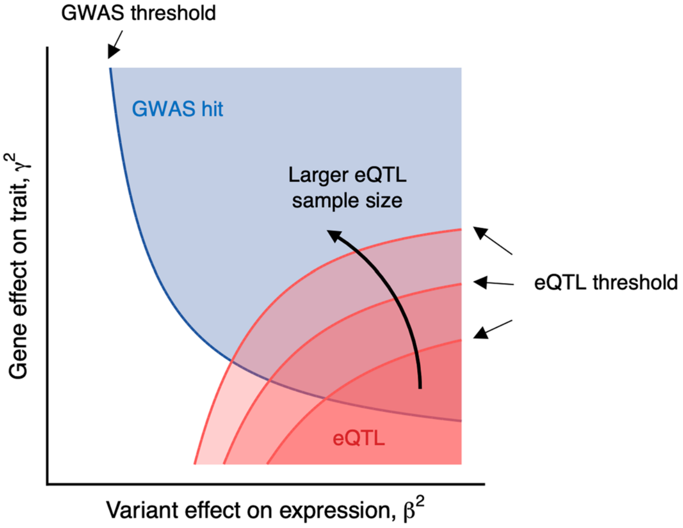
Extended Data Fig. 6 | Multi-functionality of highly interacting genes in protein-protein interaction (PPI) networks and their enrichment in GWAS genes. (A) Proportion of genes in bins ranked by the number of interactions in the InWeb PPI network[77] that are among the top multi-functional genes (defined as top 20% of genes ranked by the count of Gene Ontology (GO) terms they belong to, see Methods). Error bars show 2 standard errors. 16,510 genes with an assigned PPI degree are evenly split into the 5 gene bins shown. (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 over 1000 sampling iterations. 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. See Supplementary Table 5 for the counts of genes in each bin shown.



Extended Data Fig. 7 | 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 (that is, lowering $E[2p(1-p)]$) 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 (that is, lowering of $E[2p(1-p)]$) is more pronounced for variants regulating high-effect (that is, 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.



Extended Data Fig. 8 | 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.[70], (ii) at multiple stages of iPS cells 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). Cell labels for Jerber et al.: Astro, astrocyte-like; DA, dopaminergic neuron; epen1, ependymal-like 1; FPP, floor plate progenitors; prolif. FPP, proliferating floor plate progenitors; sert, serotonergic-like neuron; D11, day 11 of differentiation; D30, day 30; D52, day 52. (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: : B_IN, immature and naive B cell; B_Mem, memory B cell; CD4_ET, CD4+ effector memory and central memory T cell; CD4_NC, CD4+ naive and central memory T cell; CD4_SOX4, CD4+ SOX4 T cell; CD8_ET, CD8+ effector memory T cell; CD8_NC, CD8+ naive and central memory T cell; CD8_S100B, CD8+ S100B T cell; DC, dendritic cell; Mono_C, classical monocyte; Mono_NC, non-classical monocyte; NK, natural killer cell; NK_R, natural killer cell recruiting; Plasma, plasma cell. 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).


GWAS threshold
GWAS hit
Larger eQTL sample size
eQTL threshold
eQTL
Gene effect on trait, γ^2
Variant effect on expression, β^2

Extended Data Fig. 9 | Effect of eQTL assay sample size on discovery. Same as Fig. 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.

nature portfolio

Corresponding author(s): Hakhamanesh Mostafavi
Jonathan Pritchard

Last updated by author(s): Aug 15, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specialized software was used for data collection.
Data analysis	Codes used to process and analyze GWAS and eQTL data is available on GitHub (https://github.com/hakha-most/gwas_eqtl). All statistical analyses were carried out using R v3.5.1, v4.0.2 and v4.1. LD clumping of GWAS and eQTL hits was performed using plink v1.90b6.12. LD scores were computed using the software ldsc v1.0.1. PPI data were retrieved from the package Genoppi v1.0.13. The igraph package v1.3.5 was used to compute connectedness in co-expression and PPI networks. The biomaRt package v2.48.3 from Bioconductor v3.13, along with topGO v2.44 were used to retrieve and process GO terms. The glmnet package v4.1-3 was used to assess the dependency of GO terms. The qvalue package v2.24 was used to compute q-values for GO enrichments across traits and tissues.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data generated by or processed for this study can be found in Supplementary Tables, on Zenodo with the DOI 10.5281/zenodo.6618073 and on GitHub (https://github.com/hakha-most/gwas_eqtl). Public data used in this study are accessible via URLs cited at appropriate locations in the Methods.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For our analyses of GWAS hits and eQTLs we used all variants that passed our filtering procedures. We did not predetermine the sample sizes.
Data exclusions	The list of traits was pruned to exclude highly correlated pairs of GWAS datasets. For brain-related eQTLs in GTEx, to avoid over-representation of brain tissues, we retained data for "Brain - Cerebellum" and "Brain - Cortex" regions and excluded other regions. The list of GWAS and eQTL variants were pruned using LD clumping to retain independent lead variants for subsequent analysis. All analyses conditioned on variants linked with or within 1Mb of 18,332 protein-coding genes. Additional QC filters are described in Methods.
Replication	We constructed bootstrapped samples of GWAS and eQTL variant sets to compute confidence intervals, represented alongside all point estimates, demonstrating that the trends we report are robust (i.e., not highly variable) regarding the choice of tissues/traits and LD blocks containing association loci. Furthermore, in the Supplementary Note (Section 1), we demonstrate that key trends replicate across (i) various choices of traits or tissues, (ii) the sources of GWAS or eQTL data, expanding our analysis to 39,932 lead GWAS SNPs for 1,488 traits curated by the GWAS ATLAS and eQTL data from the eQTL catalogue and the eQTLGen consortium, and (iii) the SNP-to-gene linking strategy. We evaluated replication by considering the consistency of direction of enrichment/depletion patterns.
Randomization	This study utilizes summary statistics from publicly available GWAS data and eQTL data provided by the GTEx and other consortia. We were not engaged in data collection for these GWAS and eQTL analyses, making randomization irrelevant to this study.
Blinding	This study utilizes summary statistics from publicly available GWAS data and eQTL data provided by the GTEx and other consortia. We were not engaged in data collection for these GWAS and eQTL analyses, making blinding irrelevant to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging