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

In the format provided by the authors and unedited

Contents

List of Figures

1 Robustness of the data analyses

In this section we test the robustness of our results and conclusions to many of the choices made for our main analyses, showing that our results replicate for different choices of traits and tissues, sources of GWAS and eQTL data sets, and strategies to link variants to their target genes.

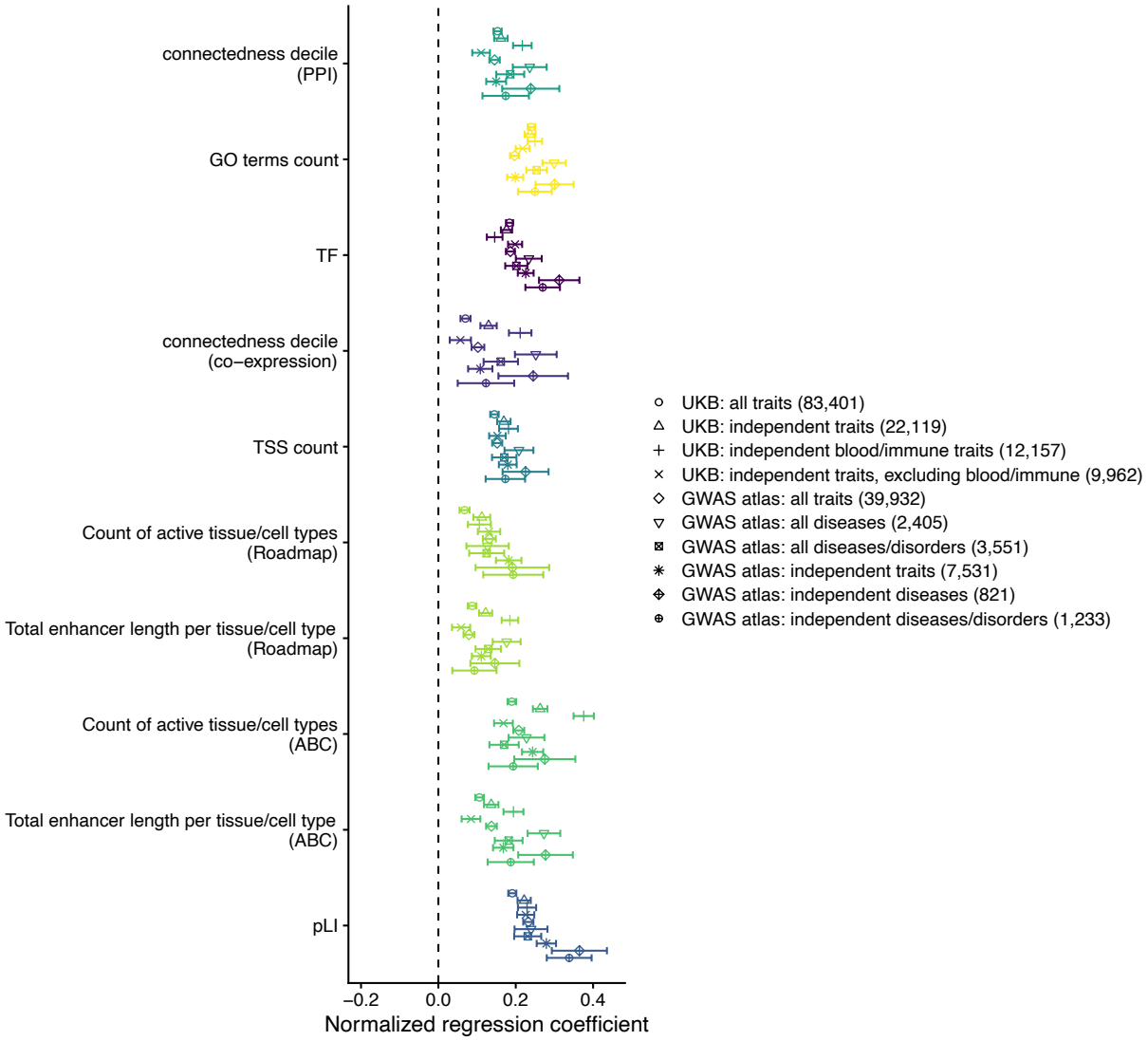
1.1 GWAS analyses

Choice of traits. In the main analyses presented, we focused on 44 complex traits from the UK Biobank (UKB), chosen such that no pairs of traits are highly correlated (see Online Methods). 14 of the resulting traits are blood cell and immune related traits, contributing about half of the final GWAS SNPs (12,157 out of 22,119). Furthermore, UKB is a prospective study, and UKB-based GWAS for disease traits may not be as powered for variant discovery as case-control studies.

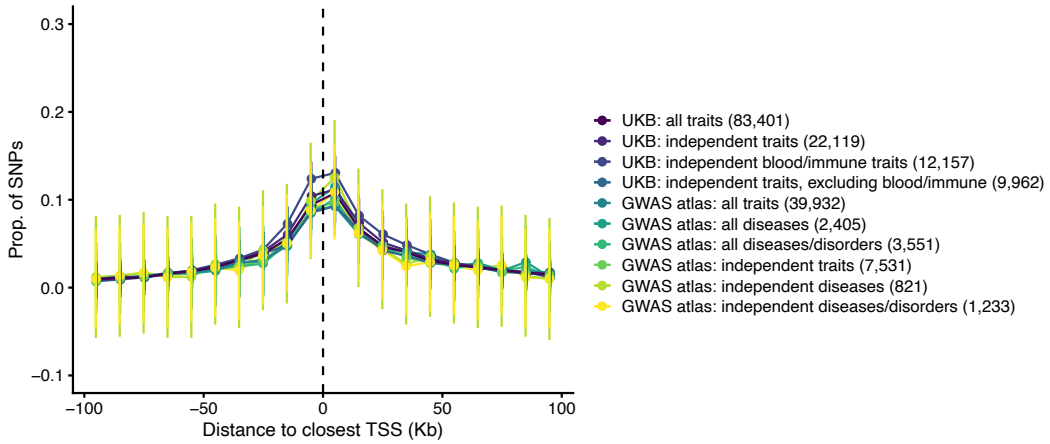
To test the robustness of our results to the choice of traits, we first extended our analysis of UKB to construct three additional groups of traits: (Group 1) 1,083 traits analyzed by the Neale lab [1] (83,401 GWAS hits ascertained following the same procedures described in the Online Methods). We also split the 44 traits used for our main analyses into 14 blood/immune related traits (Group 2, 12,157 GWAS hits) and 30 non-blood/immune related traits (Group 3, 9,962 GWAS hits). We selected blood/immune related traits based on GWAS variants enrichment in myeloid/erythroid or lymphoid specific open chromatin regions [2]. See Supplementary Methods for details.

Second, we analyzed 39,932 lead GWAS SNPs for 1,488 traits curated by the GWAS ATLAS [3]. We note that there is a substantial overlap between the traits in GWAS ATLAS and UKB. Yet, the GWAS ATLAS includes GWAS data for tens of complex diseases and disorders that are not well-represented in UKB. We constructed six groups of traits. Group 1: 1,488 traits in GWAS ATLAS (39,932 GWAS hits); Group 2: 154 traits labeled with the term "disease" or "disorder" (3,551 GWAS hits); Group 3: 92 traits labeled with the term "disease" (2,405 GWAS hits); Group 4: a pruned set of 173 traits from set (1) (7,531 GWAS hits); Group 5: a pruned set of 40 traits from group 2 (1,233 GWAS hits); Group 6: a pruned set of 23 traits from group 3 (821 GWAS hits). For the last three groups pruning was performed as described in Online Methods to exclude highly correlated trait pairs (genetic correlation > 0.5). See Supplementary Methods for details.

We used our logistic regression framework (as used for the analysis presented in Fig. 3B) to evaluate the genic features differentiating the sets of GWAS SNPs detailed above and random SNPs after adjusting for potential confounders (see Online Methods). For all genic features we studied, the regression coefficients are similar (in both magnitude and direction of effect) across all sets of traits and are consistent with the trends reported in the main text: GWAS hits are more likely to be near genes that are under strong selective constraint, have complex regulatory landscapes, and are linked with functional annotations (Supplementary Fig. 1). We also show that enrichment of GWAS variants near TSSs is similar across all sets of traits (Supplementary Fig. 2). Note that the trait groups considered here are potentially overlapping.

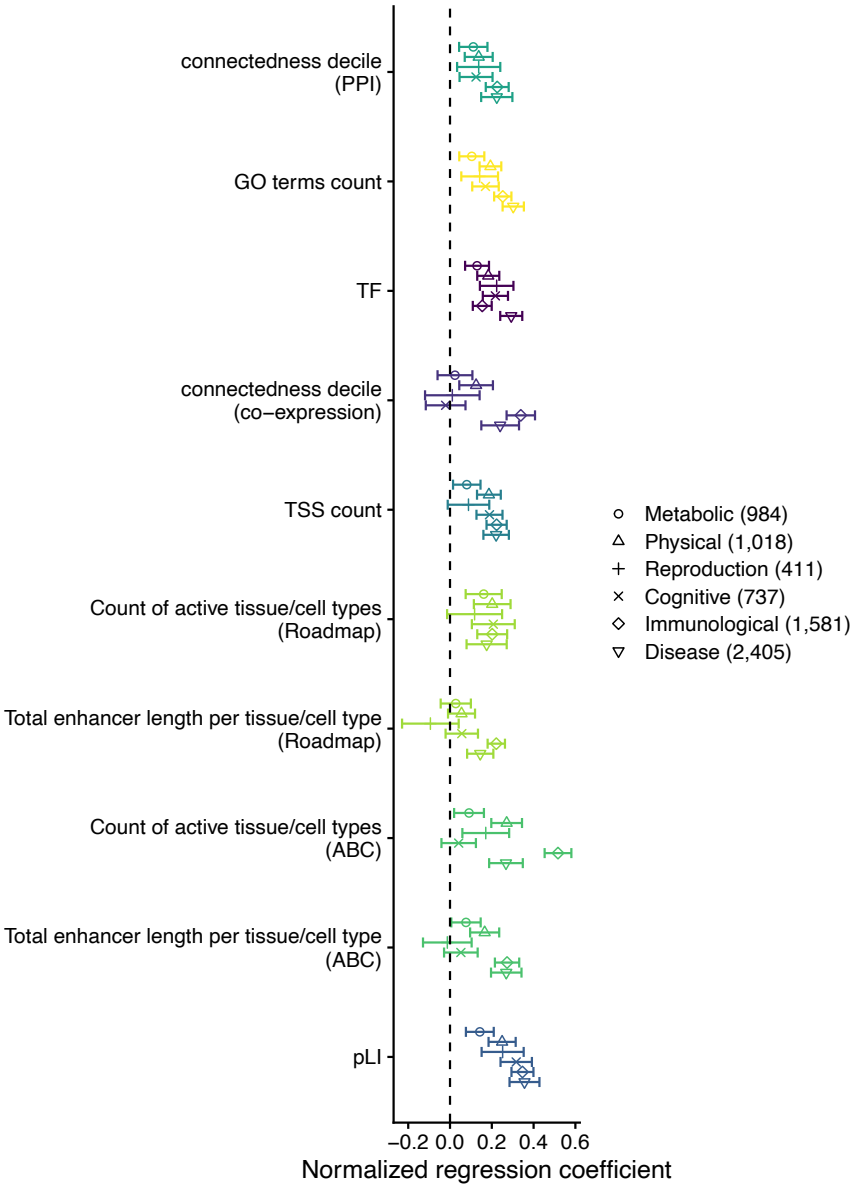


Supplementary Fig. 1: **Robustness of GWAS gene properties to trait choice.** *Properties of GWAS hits from the UK Biobank [1] and GWAS ATLAS [3] for different strategies of choosing traits. Points show logistic regression coefficients corresponding with different genic features for predicting GWAS hits versus random SNPs after adjusting for confounders. Results are plotted as regression coefficients* ± 2 *standard errors. Colors demonstrate regression models: features are tested one at a time, with the exception of the two enhancer features that are tested in a joint model. Shapes correspond to different strategies for grouping traits. The group "UKB: independent traits" includes all 44 complex traits studied in the main text. The numbers in the legend represent the count of GWAS hits in each trait set. See Supplementary Methods for details.*



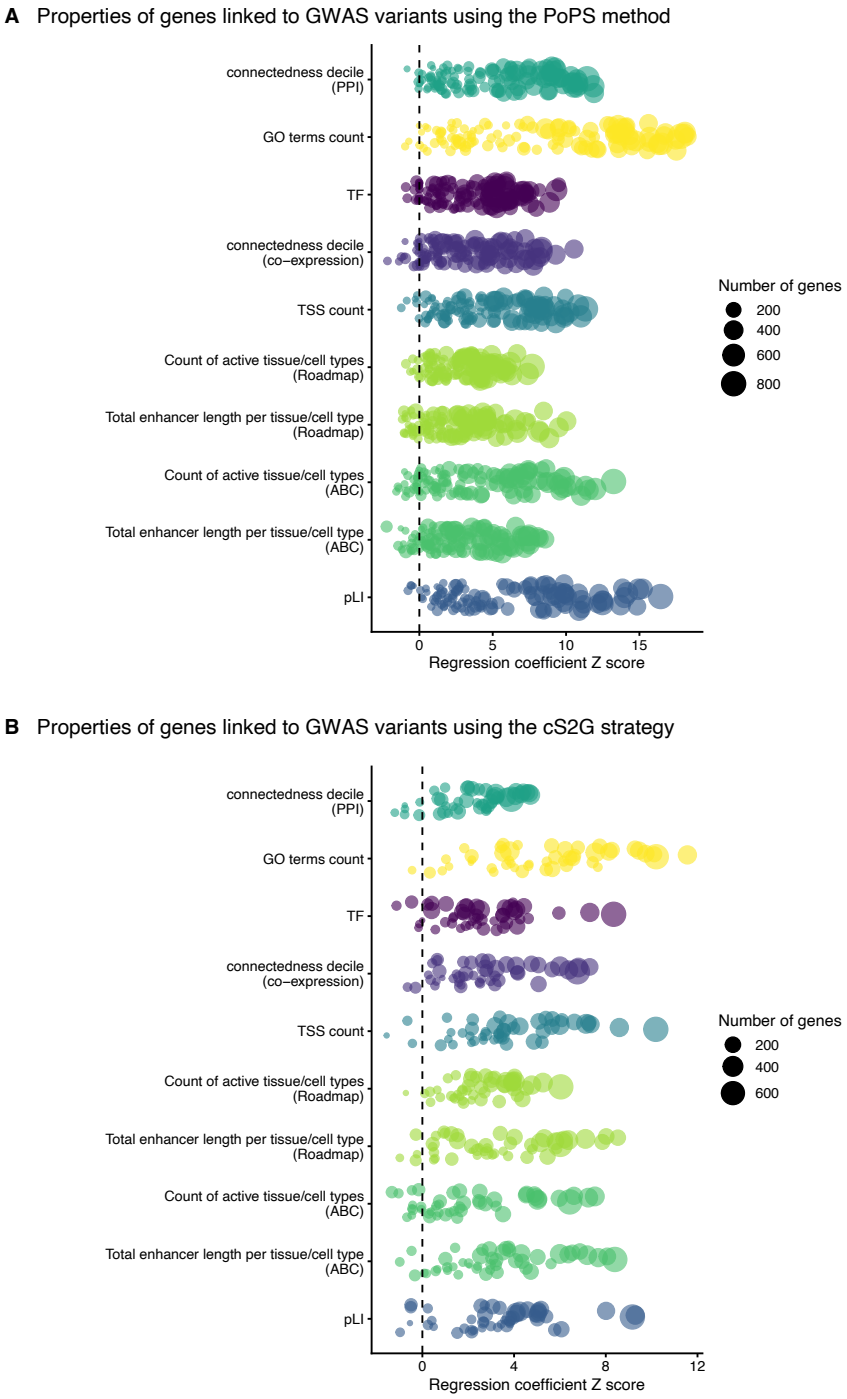
Supplementary Fig. 2: **Robustness of GWAS SNPs enrichment near TSSs to trait choice.** *Distance of GWAS hits to the nearest TSS. Points show fraction of GWAS hit SNPs in 10Kb bins. Results are plotted as fractions* ± 2 *standard errors. Standard errors are computed as* $\sqrt{2f(1-f)/M}$*, where* f *is the estimated fraction, and* M *is the number of hits per group. SNPs more than 100Kb away from their closest TSS are not shown for clarity. Colors correspond to different strategies for grouping traits. The group "UKB: independent traits" includes all 44 complex traits studied in the main text. The numbers in the legend represent the count of GWAS hits in each trait set. See Supplementary Methods for details.*

We further sliced the GWAS ATLAS traits in group 4 described above into two "disease" and "non-disease" categories, and then divided the "non-disease" category into 5 non-overlapping domains: cognitive, reproduction, metabolic, physical, and immunological traits (Supplementary Fig. 3). We note that enrichment of different gene features seems to be more pronounced for disease and immunological traits, and less so for metabolic and physical traits. For example, high pLI genes are most enriched near GWAS genes for disease traits, which appears to be consistent with some previous studies inferring stronger selection on variants underlying complex diseases compared to other categories [4]. That said, we caution against over-interpretation of these trends, because: one, there are relatively few (around 900) GWAS hits per category and so trends for individuals trait categories are noisy. Two, there is a large variation in the number of GWAS hits for different traits, and trait categories. For example, we have around 6 traits and around 170 GWAS hits per trait in the physical domain, compared to 23 disease traits and around 35 GWAS hits per disease. It is plausible that for a given trait, the properties of top GWAS hits are different from less significant hits. Three, the properties of variants discovered in GWAS depend on features such as heritability and the degree of polygenicity, which are highly variable across traits [5]. Four, many variants or genes underlying complex traits contribute to multiple traits, i.e., are pleiotropic, and thus some of their properties, such as the strength of natural selection acting on them, would depend on their effect in the multi-phenotype space. Therefore, prior expectations in the trait space, e.g., stronger selection on disease traits, may not translate to similar trends on the underlying variants and genes.



Supplementary Fig. 3: **GWAS gene properties by trait category.** *Properties of GWAS hits from the GWAS ATLAS [3] grouped by trait category. Points show logistic regression coefficients corresponding with different genic features for predicting GWAS hits versus random SNPs after adjusting for confounders. Results are plotted as regression coefficients ±2 standard errors. Colors demonstrate regression models: features are tested one at a time, with the exception of the two enhancer features that are tested in a joint model. Shapes correspond to different strategies for grouping traits. The numbers in the legend represent the count of GWAS hits in each trait set. See Supplementary Methods for details.*

Choice of SNP-to-gene linking method. For our main analyses, we linked GWAS SNPs to genes with the closest TSS as a proxy for their target genes. The closest gene approach is applicable to all SNPs, which is not the case for other more sophisticated approaches to nominate GWAS genes, while it yields a comparable (and in many cases higher) accuracy [6, 7]. Nevertheless, we show that our main results are robust to this choice.



Supplementary Fig. 4: **Properties of GWAS genes prioritized by different SNP-to-gene linking strategies.** *Properties of GWAS genes from the PoPS method by Weeks et al. [8] (A) and the cS2G approach by Gazal et al. [7] (B). Each point corresponds to a trait, showing logistic regression Z-score (regression coefficient divided by the standard error) corresponding with different genic features for predicting GWAS genes versus non-GWAS genes. The size of the points show the number of genes assigned per trait. Colors demonstrate regression models: features are tested one at a time, with the exception of the two enhancer features that are tested in a joint model. See Supplementary Methods for details, and Supplementary Data for Z-scores for individual traits.*

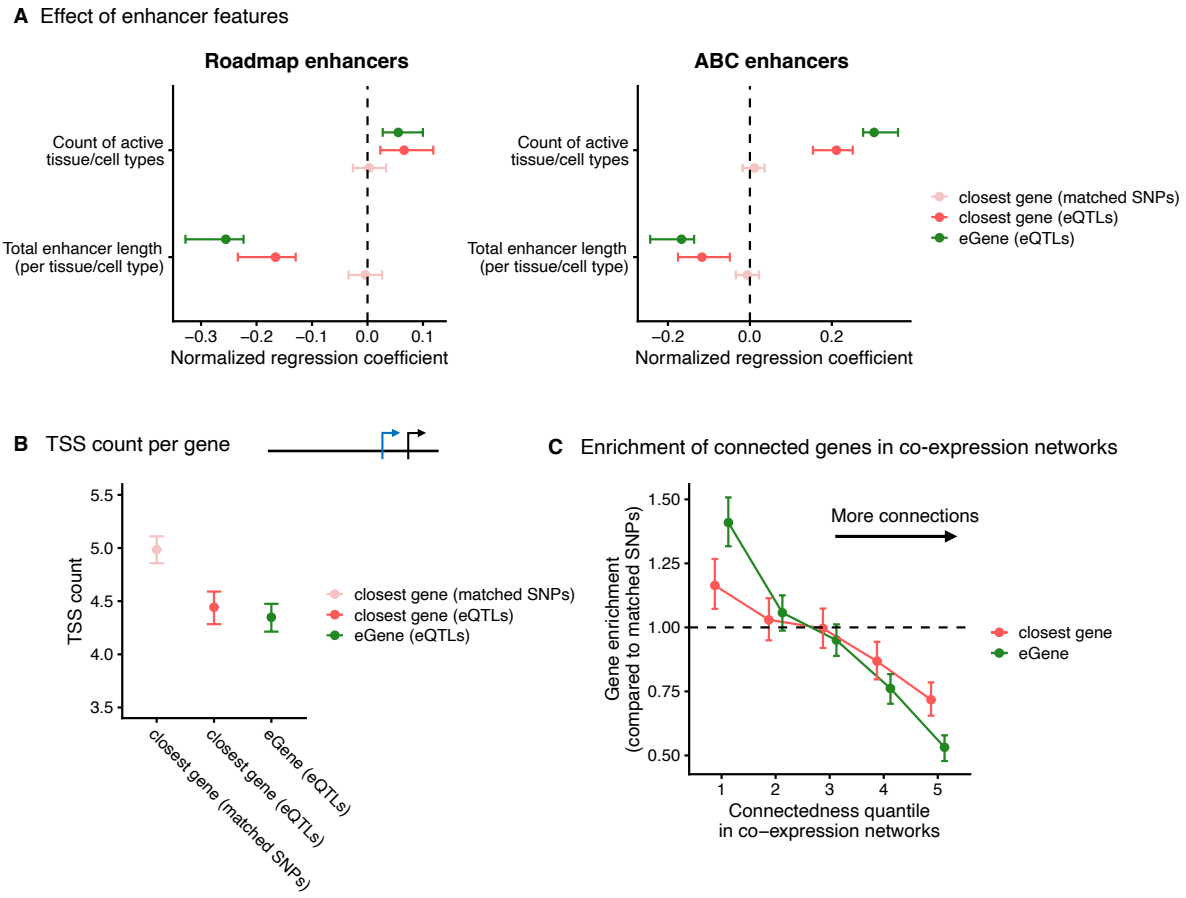
To this end, we analyzed the genes nominated by two independent methods: (i) 25,252 SNP-gene links across 113 traits provided by Weeks et al., using a machine learning approach, named PoPS, which integrates GWAS data with external functional annotations [8]; (ii) 6,655 SNP-gene links across 47 traits provided by Gazal et al., by integrating earlier SNP-to-gene linking strategies into a combined score (cS2G) for gene prioritization [7]. For each trait studied by these two approaches, we used a logistic regression model to evaluate genic features that differentiate GWAS genes versus other protein-coding genes, showing results consistent with the trends we show for the closest genes (Supplementary Fig. 4; each point represents a trait). See Supplementary Methods for details.

1.2 eQTL analyses

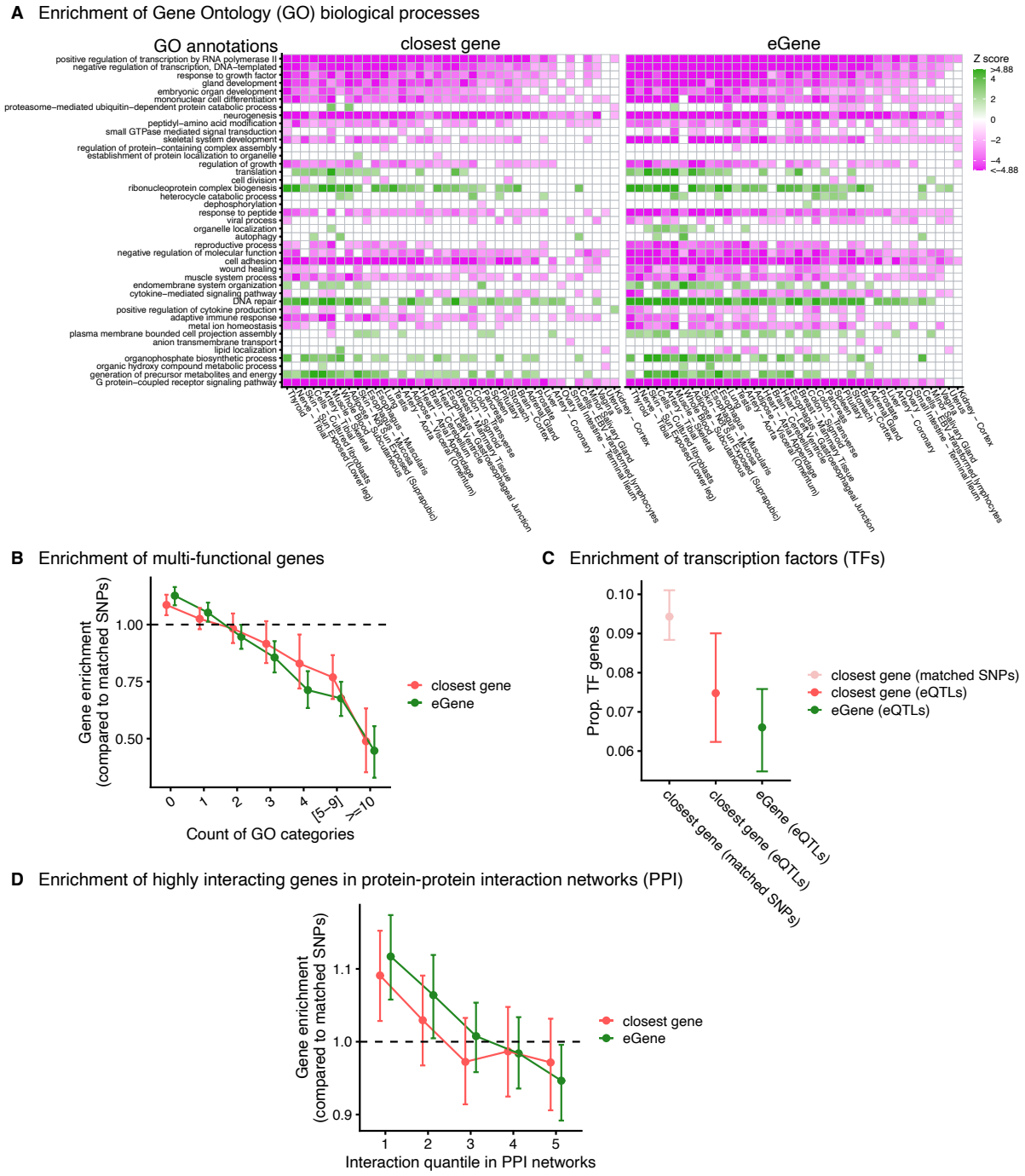
eGenes versus closest genes. In our main analysis we linked eQTLs to their closest genes despite knowing the true target genes, i.e., eGenes. The main rational for this approach is consistency with our GWAS analysis (as the target genes for GWAS SNPs are unknown and we link those to their closest genes), although it introduces noise to our eQTL analysis by mis-assigning eQTL targets. In fact, for 48% of GTEx eQTLs that we analyzed, the eGene was not the gene with the closest TSS. The closest gene approach is more accurate for eQTLs with stronger association signal (Extended Data Fig. 1A), presumably because stronger eQTLs lie at closer distances to their target gene's TSS.

We re-performed all of our gene-level analyses for eQTLs using eGenes instead of the closest genes. For all genic features, the difference between eGenes and control genes (i.e., genes closest to control SNPs matched to eQTLs for MAF, LD score and gene density) are more pronounced than that for closest genes, demonstrating that gene mis-assignments merely dilute the trends towards random SNPs (Extended Data Fig. 1B, Supplementary Fig. 5, Supplementary Fig. 6). It follows that the systematic differences between true GWAS and eQTL targets are likely larger than what is observed for the closest genes.

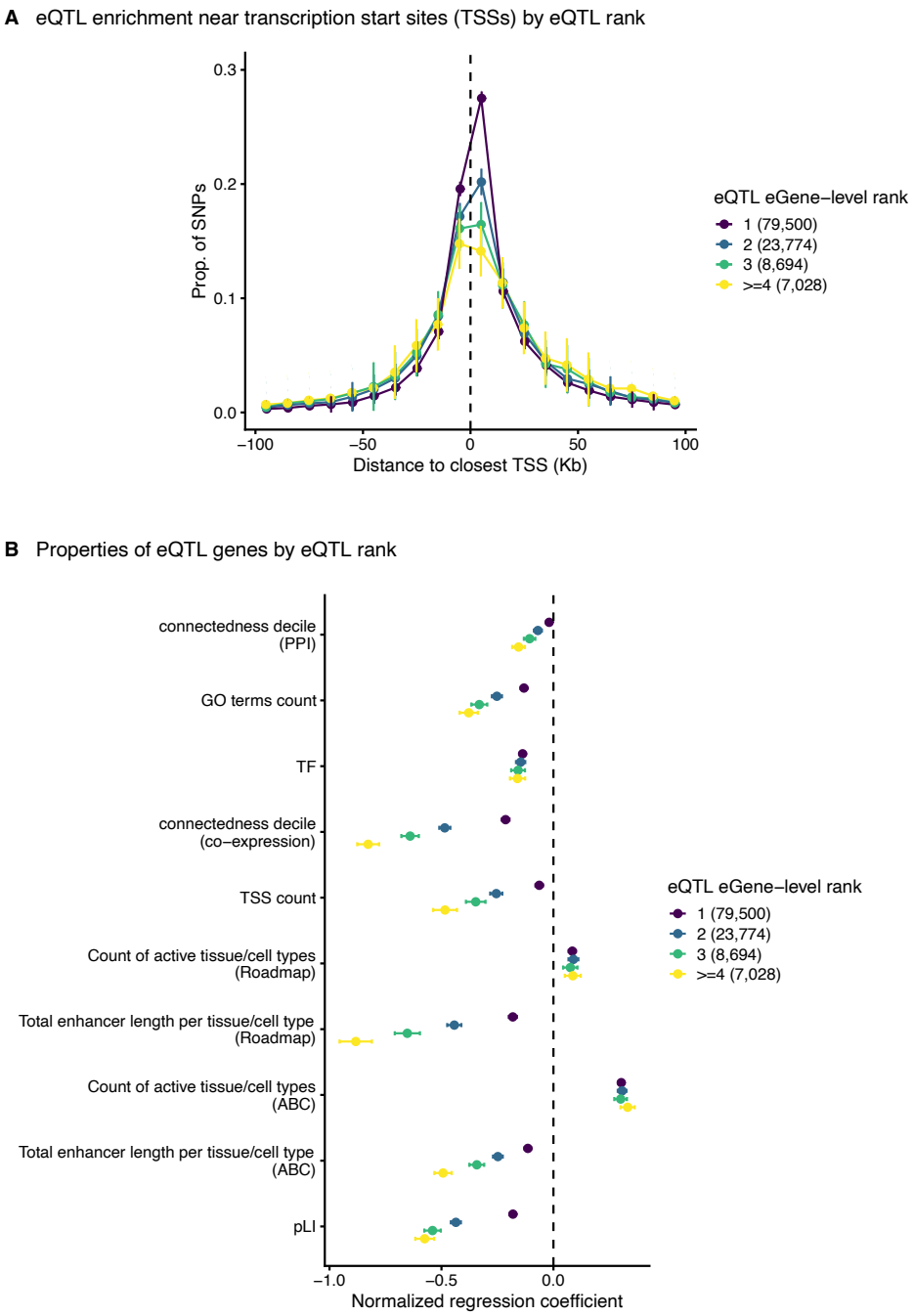
Primary versus secondary eQTLs. As we show in the main text, GWAS SNPs lie at longer distances to TSSs compared to eQTLs (Fig. 5A), and under our model include weak eQTLs acting on phenotypically important genes. This raises the question: are weaker eQTLs discovered in eQTL studies more similar to GWAS SNPs? To get at this question, we studied eQTL properties by the ranking of the lead eQTLs at eGenes (after LD clumping), i.e., first most significant eQTLs, second most significant eQTLs, and so on. Weaker discovered eQTLs lie at longer distances to TSS (Supplementary Fig. 7A), consistent with previous findings [9], and on par with GWAS SNPs. However, these weaker eQTLs are discovered at genes that are more dissimilar to GWAS genes (Supplementary Fig. 7B), that is are under weaker selection, have simpler regulatory landscapes, and are more depleted of functional annotations. Note that this does not mean that GWAS hits are not secondary or tertiary eQTLs. Rather, genes for which we have the power to detect multiple eQTLs are even more biased toward unimportant genes. In our model terms, secondary eQTLs on average have smaller β^2 than primary eQTLs and so are more similar to GWAS variants in that regard. That said, detection power for these eQTLs is higher at genes with smaller γ^2, and so are overall less GWAS-like than primary eQTLs. These trends would not hold at the limit where most/all eQTLs are discovered.



Supplementary Fig. 5: **eGenes have simpler regulatory landscape than eQTL closest genes.** *Replication of analyses shown in the main Fig. 3.* ***A)*** *Logistic regression coefficients corresponding with the two enhancer features for predicting 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) versus closest genes to random SNPs after adjusting for confounders (see Online Methods). The enhancer features are computed as (i) the number of tissue/cell types in which a gene has an enhancer, and (ii) the average total enhancer length (in base pairs) across tissue/cell types with enhancer activity, based on enhancer-gene links inferred from the Roadmap dataset [10] or by the activity-by-contact model [11].* ***B)*** *Mean count of TSSs per gene across cell types in the FANTOM project [12].* ***C)*** *Enrichment of eGenes and eQTL closest genes in gene bins ranked by connectedness values computed based on co-expression networks from Saha et al. [13], relative to control SNPs. Error bars show 95% confidence intervals as determined by quantile bootstrapping over 1000 sampling iterations. For control SNPs (light red), points show mean values in sets of matched SNPs corresponding to bootstrapped samples. All statistics were computed for 118,996 eQTLs. See Supplementary Table 5 for the counts of genes in each bin shown in panel (C).*

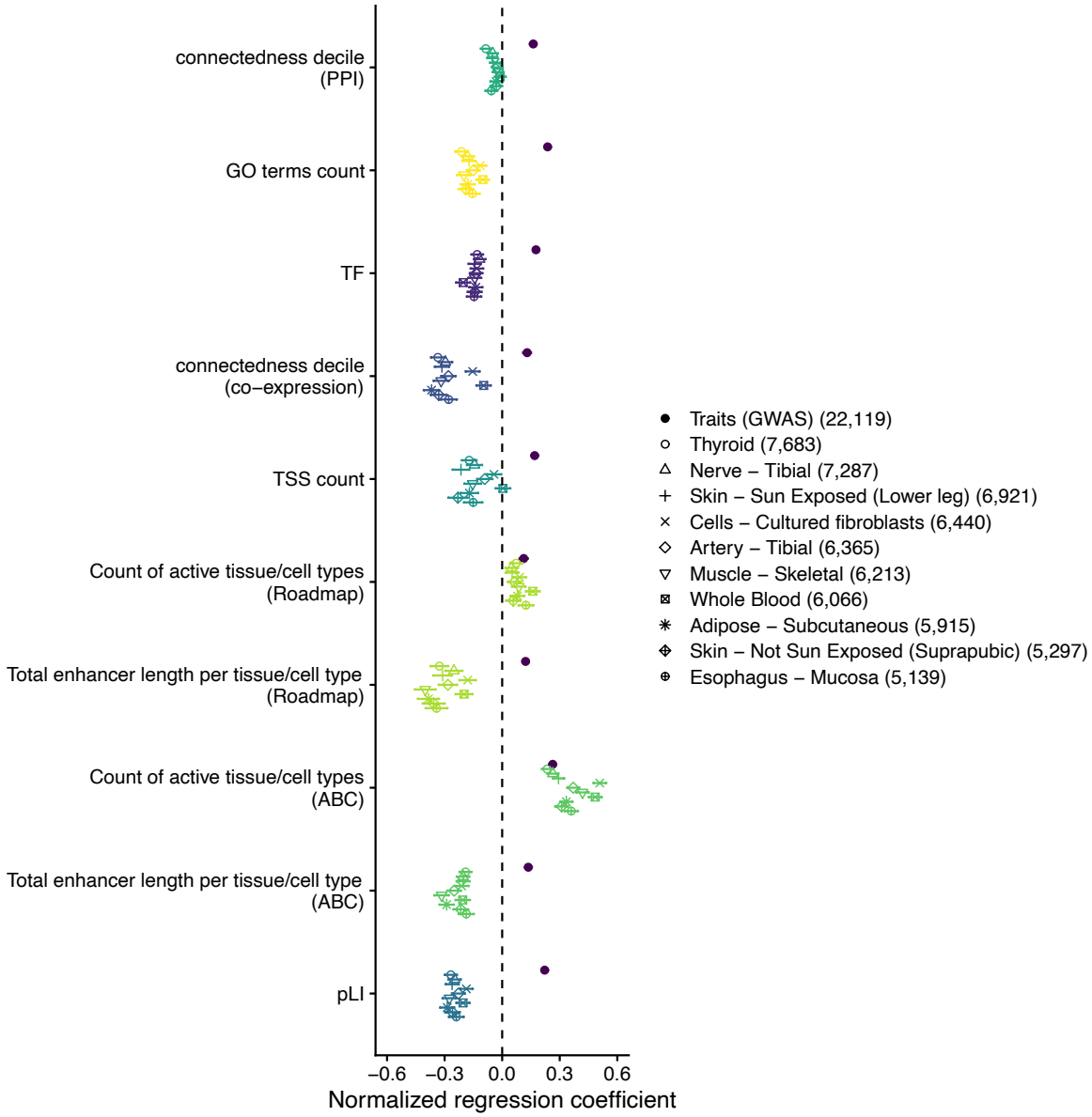


Supplementary Fig. 6: **eGenes are more depleted of functional annotations than eQTL closest genes.** ***A)*** *Enrichment of genes in 41 GO terms shown in Fig. 4A relative to control SNPs, for eQTL genes (left) and eGenes (right). Color map indicates enrichment (green) or depletion (magenta) of gene sets. See Fig. 4A for more details.* ***B)*** *Replication of Fig. 4C. Gene enrichments relative to control SNPs (y-axis) across gene bins ranked by the counts of GO terms they belong to (x-axis).* ***C)*** *Replication of main Fig. 4B. Fraction of transcription factors among genes.* ***D)*** *Replication of Extended Data Fig. 6B. Gene enrichments relative to control SNPs (y-axis) in gene bins ranked by the number of interactions in the InWeb PPI network [14] (x-axis). In panels (B-D), error bars show 95% confidence intervals based on quantile bootstrapping over 1000 sampling iterations. For control SNPs (light red), the point shows mean value in sets of matched SNPs corresponding to bootstrapped samples. All statistics are based on 118,996 eQTLs. See Supplementary Table 5 for gene counts in each bin shown in panels (B) and (D).*



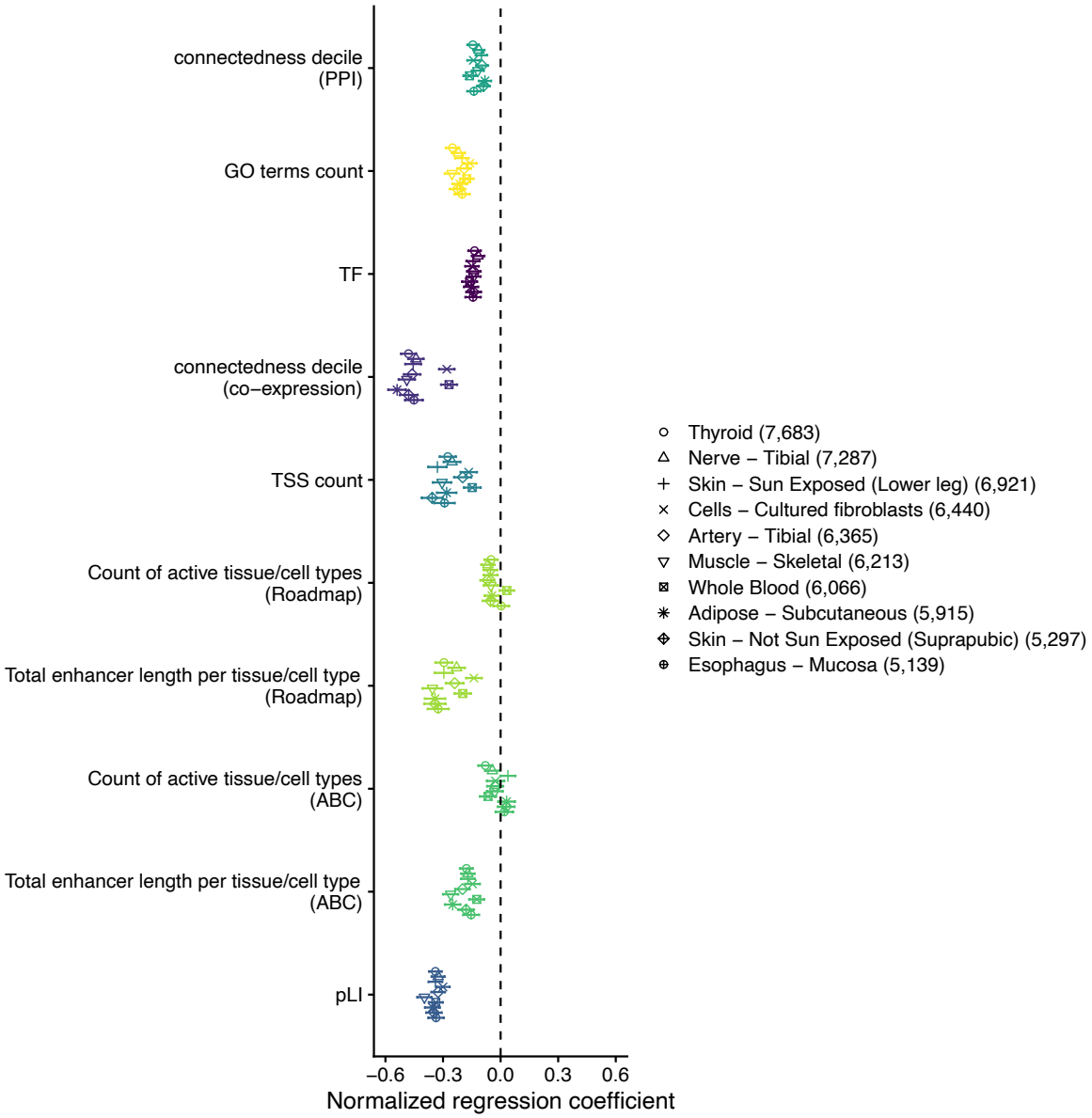
Supplementary Fig. 7: **Properties of primary versus secondary eQTLs.** ***A)*** *Distance of eQTLs to the nearest TSS for different eQTL groups. Points show fraction of eQTLs in 10Kb bins. Results are plotted as fractions* ± 2 *standard errors. Standard errors are computed as* $\sqrt{2f(1-f)/M}$*, where* f *is the estimated fraction, and* M *is the number of eQTLs per group. SNPs more than 100Kb away from their closest TSS are not shown for clarity.* ***B)*** *Properties of eGenes linked with different eQTL groups. Points show logistic regression coefficients corresponding with different genic features for predicting eGenes linked with eQTLs versus genes linked with random SNPs (closest genes) after adjusting for confounders. Results are plotted as regression coefficients* ± 2 *standard errors. In both panels, colors correspond to different eQTL groups based on the ranking of lead variants (after LD clumping) with respect to eGenes. Specifically, for each eGene we labeled the first most significant eQTL as rank 1, second most significant eQTL as rank 2, and so on. We then grouped eQTLs across all eGenes by these rank labels. The numbers in the legend represent the count of eQTLs in each set.*

eQTL data set. Our main analyses focused on eQTLs identified by the GTEx project [15], pooling eQTLs across 38 tissues. We first show that the eQTL properties we studied are robust to the choice of GTEx tissues. Focusing on top 10 tissues with most eQTLs we obtained similar regression coefficients across tissues for eGene features in our logistic regression model classifying eQTLs from random SNPs (Supplementary Fig. 8).



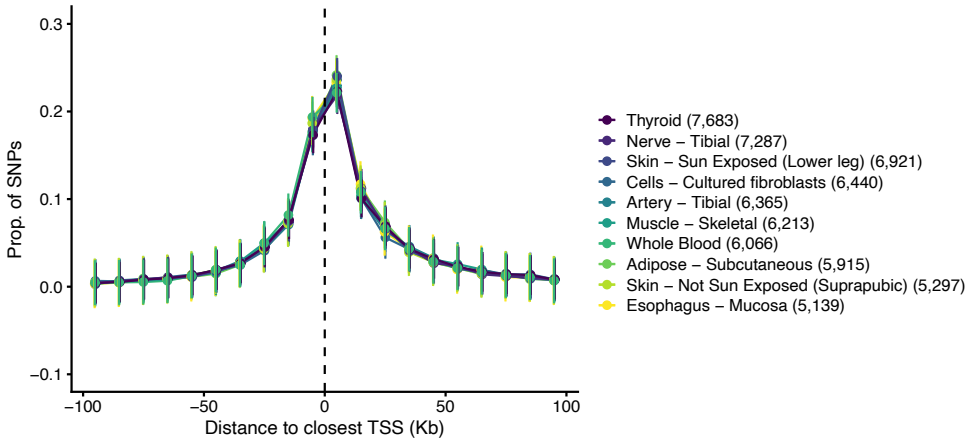
Supplementary Fig. 8: **Robustness of eQTL properties across GTEx tissues.** *Points show logistic regression coefficients corresponding with different genic features for predicting eGenes linked with eQTLs versus genes linked with random SNPs (closest genes) after adjusting for confounders. Results are plotted as regression coefficients ±2 standard errors. Colors demonstrate regression models: features are tested one at a time, with the exception of the two enhancer features that are tested in a joint model. Shapes correspond to the top 10 tissues in GTEx with the highest number of detected eQTLs. For comparison, regression coefficients computed with the GWAS data for the 44 complex traits from the UKB are also shown (solid circles), as previously presented in Supplementary Fig. 1. The numbers in the legend represent the count of variants in each set.*

We note that eQTL discovery power is higher for genes with higher expression levels. Given that gene expression patterns vary across different tissues, we extended our regression model to adjust for tissue-specific expression levels as provided by GTEx (Supplementary Methods). The similarity across tissues shown in Supplementary Fig. 8 is robust to this adjustment (Supplementary Fig. 9). Notably, the distinction between eQTLs and random SNPs with respect to most genic features is more pronounced after adjusting for tissue-specific expression levels. That said, the enrichment of eQTLs at genes with active regulation across multiple tissue/cell types is explained by increased expression levels of these genes (Supplementary Fig. 9).



Supplementary Fig. 9: **Properties of eQTLs adjusting for tissue-specific eGene expression levels in GTEx.** *Same as Supplementary Fig. 8, but adjusting for tissue-specific eGene expression levels in the logistic regression model in addition to other confounders. Results are plotted as regression coefficients* ± 2 *standard errors. See Supplementary Methods for details.*

We further demonstrate similar clustering of eQTLs near TSS for different tissues (Supplementary Fig. 10). These results recapitulate the trends reported in the main text for the pooled set of eQTLs.

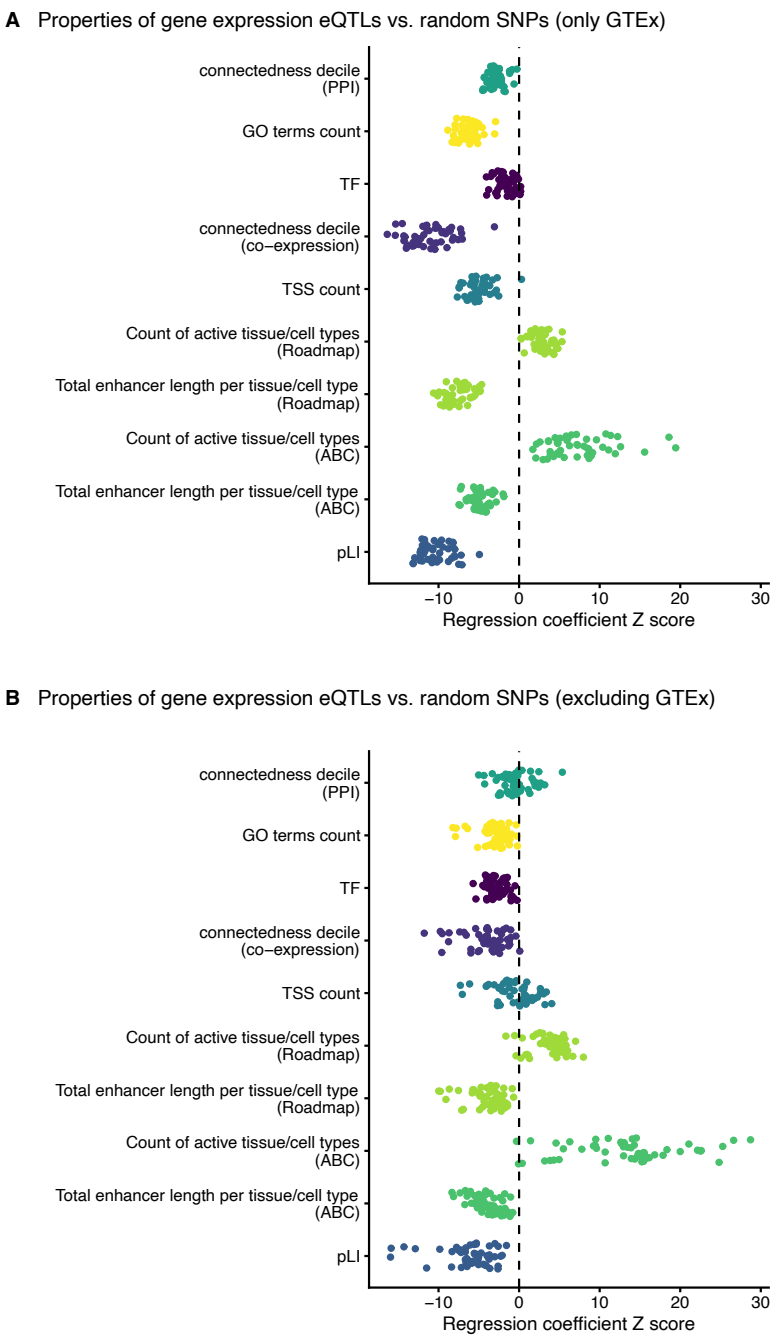


Supplementary Fig. 10: **Robustness of eQTLs enrichment near TSSs to tissue choice in GTEx.** *Distance of eQTLs to the nearest TSS. Points show fraction of SNPs in 10Kb bins ±2 standard errors. Standard errors are computed as* $\sqrt{2f(1-f)/M}$*, where* f *is the estimated fraction, and* M *is the number of eQTLs per group. SNPs more than 100Kb away from their closest TSS are not shown for clarity. Colors correspond to the top 10 tissues in GTEx with the highest number of detected eQTLs. The numbers in the legend represent the count of eQTLs in each set.*

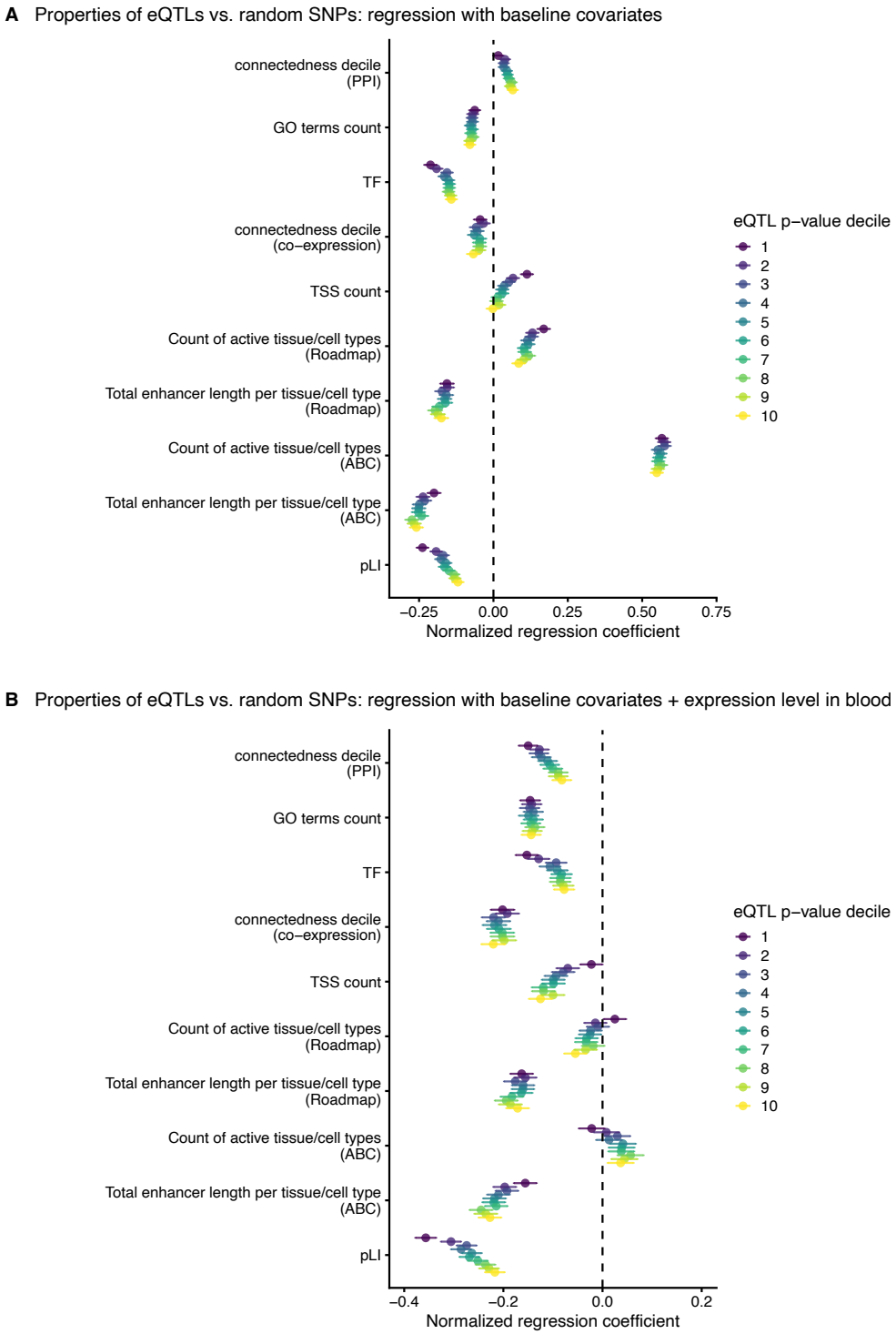
Second, we extended our analysis to include eQTLs and eGenes from two resources other than GTEx: (i) the eQTL catalogue [16], and (ii) the eQTLGen consortium [17].

The eQTL catalogue provides uniformly processed eQTLs curated from several public data sets including GTEx [16]. Notably, it includes cell type specific eQTLs from a variety of cell types. Focusing on gene expression eQTLs in the eQTL catalogue, we observed similar eGene features for GTEx (68,009 eQTLs) and non-GTEx eQTLs (74,366 eQTLs), and consistent with previous results (Supplementary Fig. 11). See Supplementary Methods for details.

The eQTLGen consortium is a large-scale meta-analysis of several blood eQTL studies (net sample size of around 32K) [17], and is a suitable data set to study discovery trends as eQTL sample sizes grow. Most expressed genes in blood (88%) are discovered as eGenes in the eQTLGen data [17]. We processed eQTLGen eQTLs similarly to GTEx eQTLs, resulting in 230,032 cis-eQTLs (Supplementary Methods). We sliced these eQTLs into 10 groups based on the deciles of association p-values, mimicking the progressive discovery of eQTLs with sample size, and to avoid pooling a large number of eQTLs which could complicate the interpretations. Most properties of eQTLGen eGenes are consistent with previous results (Supplementary Fig. 12A), with the exception of TSS count and connectedness in PPI networks. Compared to other genic features, these features also show weaker consistency between GTEx and non-GTEx eQTLs in the eQTL catalogue (Supplementary Fig. 11), and across tissues in GTEx (Supplementary Fig. 8). That said, we find that after adjusting for gene expression levels in blood, all eGene properties in eQTLGen are consistent with GTEx (Supplementary Fig. 12B and Supplementary Fig. 9). These observations suggest that differences between eQTLGen and GTEx are potentially due to covariance between gene expression levels and other gene properties of interest, considering that eQTLGen study is powered to detect eQTLs for many low-expressed genes that are not detected as eGenes in GTEx.



Supplementary Fig. 11: **Genic features of eQTLs from the eQTL catalogue [16].** *Properties of eGenes linked with eQTLs in the GTEx study (A), and in other eQTL studies processed by the eQTL catalogue, i.e., excluding GTEx (B). Each point corresponds to an eQTL study, showing logistic regression Z-score (regression coefficients divided by the standard errors) corresponding with different genic features for predicting eGenes linked with eQTLs versus genes linked with random SNPs (closest gene). Colors demonstrate regression models: features are tested one at a time, with the exception of the two enhancer features that are tested in a joint model. See Supplementary Data for Z-scores for individual studies.*



Supplementary Fig. 12: **Genic features of eQTLs from the eQTLGen consortium [17].** ***A)*** *Points show logistic regression coefficients corresponding with different genic features for predicting eGenes linked with eQTLs versus genes linked with random SNPs (closest genes) after adjusting for our baseline confounders. 230,032 eQTLs are evenly split into the 10 p-value bins shown. See Supplementary Methods for details.* ***B)*** *Same as (A), but adjusting for blood-specific gene expression levels in addition to other confounders. In both panels, results are plotted as regression coefficients ±2 standard errors. Colors correspond to different eQTL groups based on the deciles of association p-values.*

In summary, taken together with our analyses of brain eGenes from fetal samples and iPSC differentiations, and cell type specific eGenes from blood presented in Extended Data Fig. 8, these results suggest that the eQTL properties we find are general to all types of eQTL studies, e.g., bulk assays or assays based on single purified cell types, studies of eQTLs in adult samples or developmental stages, etc. and not specific to GTEx post-mortem whole tissues. That said, some properties are robustly observed across all studies (e.g., depletion of genes under strong selection from eQTL genes), while trends related to gene regulatory features and networks (e.g., TSS count and connectedness in PPI networks) seem to be more variable and possibly tissue or cell type dependent.

1.3 Other QTLs

Our QTL analyses in this paper focused on eQTLs, i.e., genetic effects on gene expression levels. Specifically, we show that discovered eQTLs are systematically different from GWAS variants, and present a model to explain these differences, highlighting the role of natural selection. We reason that natural selection likely hampers the discovery of trait-relevant variants in QTL assays for other molecular intermediates. Comprehensive analyses of other QTLs is beyond the scope of this paper, nevertheless, we show a few examples that are broadly consistent with this argument. Specifically, we studied four QTL classes (see Supplementary Methods for details):

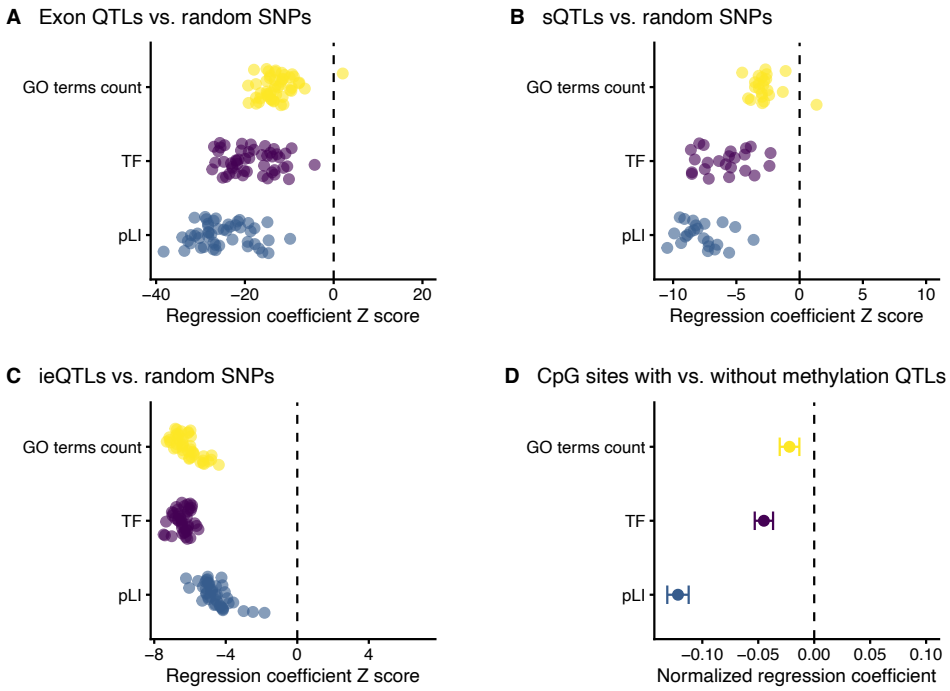
Exon QTLs. We analyzed 1,088,880 QTLs linked with exon level expression in 49 GTEx tissues ascertained by the eQTL catalogue [16], comparing the properties of genes with exon QTLs and genes nearest to random SNPs after adjusting for potential confounders.

Splicing QTLs. We processed 67,250 splicing QTLs (or sQTLs) across 23 tissues in GTEx [15], comparing the properties of genes with sQTLs (or sGenes) and genes nearest to random SNPs after adjusting for potential confounders.

Interaction eQTLs. We analyzed 530,819 eQTLs in GTEx that show cell type specific effects (cell type-interaction eQTLs or ieQTLs [18]), comparing the properties of genes with ieQTLs (or ieGenes) and genes nearest to random SNPs after adjusting for potential confounders.

Methylation QTLs. We analyzed 336,732 CpG sites that were tested for methylation QTLs by Hawe et al., [19]. We linked each CpG site to its nearest gene, and compared genes with and without detected methylation QTLs.

We used our logistic regression framework to compare features of gene sets described above. We focused on gene annotations related to broad biological functions and selective constraint, as we expect these features to be less sensitive to the particular biology of each molecular phenotype. Genes linked with all QTL classes are depleted of these features (Supplementary Fig. 13) similar to eQTLs.



Supplementary Fig. 13: **Properties of other QTLs.** ***A)*** *Properties of genes linked with exon QTLs from the eQTL catalogue [16] versus genes linked with random SNPs (closest gene). Each point corresponds to a GTEx tissue.* ***B)*** *Properties of genes linked with splicing QTLs in GTEx [15] versus genes linked with random SNPs (closest gene). Each point corresponds to a GTEx tissue.* ***C)*** *Properties of genes linked with cell type interaction QTLs (ieQTLs) in GTEx [18] versus genes linked with random SNPs (closest gene). Each point corresponds to a tissue-cell type pair tested by the GTEx team.* ***D)*** *Properties of genes linked with CpG sites for which methylation QTLs were detected by Hawe et al. [19] (N=64,478) versus CpG sites without any detected methylation QTL (N=272,254). Results are plotted as logistic regression coefficients ±2 standard errors. In panels (A-C), points show logistic regression Z-scores (regression coefficients divided by the standard errors) corresponding with different genic features (colors) to predict the QTLs of interest versus control SNPs described for each panel. See Supplementary Data for values for individual data points.*

2 Power considerations in GWAS and eQTL mapping

In this section we present power considerations for variant discovery in GWAS and eQTL assays to show that: (i) per-SNP contribution of variants to complex traits and gene expression phenotypes, and thus sample size requirements for discovery in the two mapping strategies, vary by orders of magnitude. (ii) GWAS and eQTL assays are generally low-powered, likely ranging from 10–20% at typical sample sizes that are currently available for association studies.

Per-SNP heritability, and sample size requirements. Complex traits are usually extremely polygenic, with estimates for the number of causal variants on the order of 10K-100K [7, 20, 21]. Therefore, a typical per-SNP contribution to trait heritability, h^2_{SNP}/h^2, is on the order of 10^{-4}–10^{-5}, where h^2 is the total trait heritability and h^2_{SNP} is the fraction of trait variance explained by a given SNP. Gene expression is expected to be much less polygenic than organismal phenotypes which result from aggregate effects of hundreds to thousands of genes [7, 22]. For example, we estimate a median value of 15 roughly independent cis-eQTLs across eGenes in the eQTLGen study [17] (Supplementary Methods), on the same order of magnitude as previous analyses [23]. Considering tens of causal variants with local effect on gene expression (i.e., in cis), $h^2_{SNP}/\text{cis-}h^2$ for cis-eQTLs is on the order of 10^{-1}. Thus, causal variants typically have much larger (10^3–10^4 times larger) contributions to cis-expression heritability than to complex trait heritability.

As a result, with respect to variant discovery, much larger sample sizes are required in GWAS than eQTL assays. As detailed in the Online Methods section, in expectation, the discovered variants in GWAS and eQTL assays satisfy: $h^2_{SNP} > \chi^2_c/n$, where n is the sample size, and χ^2_c is the study-dependent discovery threshold. The conventional GWAS threshold of p-value$= 5 \times 10^{-8}$ corresponds to $\chi^2_c = 29.7$. In an eQTL study, if we assume a p-value threshold of $\sim 2 \times 10^{-4}$ (on par with nominal p-value threshold values computed in GTEx at a gene-level false discovery rate threshold of 0.05 [15]), the corresponding $\chi^2_c \approx 14$. With typical sample sizes of around 500K for GWAS (e.g., the UK Biobank) and 500 for eQTL assays (e.g., the GTEx consortium), the discovered GWAS variants and eQTLs will have $h^2_{SNP} > 6 \times 10^{-5}$ and $h^2_{SNP} > 0.028$, respectively. Now, consider a SNP which contributes 10% to cis-expression heritability of its target gene, and 0.01% to the heritability of the downstream complex trait. Considering a heritability of 0.2 for a typical complex trait, a GWAS sample size of around 1.5 million is required for the SNP's discovery. On the other hand, considering a typical cis-expression heritability of 0.05 [23, 24], sample sizes on the order of 3K are required for its discovery as an eQTL.

Variant discovery power at current sample sizes. Sample sizes that are currently available for association studies are typically on the order of 10^5–10^6 for GWAS (e.g., the UK Biobank) and 10^2–10^3 for eQTL assays (e.g., the GTEx consortium). A quantitative analysis of the variant discovery power with such sample sizes requires knowledge about the genetic architecture of complex trait and gene expression phenotypes, specifically the joint distribution of the effect sizes and allele frequencies of the causal SNPs. This is a challenging task due to the extreme polygenicity of most traits and confounding by LD (linkage disequilibrium) which complicates identifying the causal variants within a given genomic locus with evidence for trait association.

Nevertheless, many lines of evidence suggest that current association studies are far from saturation in terms of discovering the causal loci:

(i) For most traits, discovered variants at current sample sizes explain a small fraction of heritability. For example, a GWAS for systolic blood pressure using over 1 million samples discovered 901 loci explaining only about 27% of heritability [25]. Similarly, in an eQTL analysis for blood from 2,765 individuals, averaged across genes, discovered eQTLs explained about 31% of heritabil-

ity for gene expression, most of which (87%) could be attributed to the single top eQTL of the genes [23]. These suggest that variants with smaller effect sizes than the currently discovered ones likely constitute a larger fraction of heritability and thus many more are yet to be discovered. In line with these observations, using a statistical method, O'Connor recently estimated that on the order of 1-100 millions of samples are needed to discover variants that explain up to 90% of variance for most studied complex traits [21].

(ii) Association studies with growing sample sizes over time keep discovering new variants. This is mostly notable for complex traits that are easily measured such as height and educational attainment, and for tissue and cell types which are easier to sample, particularly blood. The latest GWAS for height by Yengo et al., reached a sample size of 5.4 million, discovering 12,111 independent loci at the genome-wide significance level [26], which is much more than discoveries with smaller sample sizes (Table 2 in Yengo et al., [26]). For comparison, with sample sizes of around 240K and 700K (which is typical for most GWAS), around 600 and 2,800 loci were discovered, corresponding to power estimates less than 5% and 23%, respectively, based on the number of loci discovered by Yengo et al. [26]. Similarly for eQTLs, a recent study of blood eQTLs by the eQTLGen consortium reached a sample size of around 32K [17]. Our analysis of this data set yields around 250K roughly independent cis-eQTLs for 12,659 protein-coding autosomal eGenes (Supplementary Methods). The same ascertainment procedure applied to whole blood in GTEx for 670 individuals (which is on the same scale as sample sizes available for most eQTL studies) yields 28,645 cis-eQTLs for 7,953 eGenes, corresponding to a power estimate of less than 11% based on the number of discoveries in eQTLGen.

Taken together, these considerations suggest that GWAS and eQTL assays are generally low-powered at currently available sample sizes.

3 Robustness of the model

In this section we further examine our model for variant discovery in GWAS and eQTL assays, and perform an extensive survey of our modeling assumptions and choices. We focus on our single cell type (or 1-D) model in this part, and will explore more complicated scenarios in a later section (*Model extensions*). We conclude that: (i) the *qualitative* predictions of our model about what types of genes and variants are discovered in GWAS and eQTL assays are robust, and (ii) making *quantitative* predictions requires several additional assumptions, with many parameters that are unknown and not easily estimated with current data.

3.1 Key qualitative predictions

Our first key consideration is that in eQTL assays variants are more likely to be discovered if they have large effects on gene expression levels, β^2 in our model. On the other hand, in GWAS, the likelihood of discovery is not only dependent on a variant's effect on expression, but also how relevant the target gene is to the phenotype, γ^2 in our model. We model the net effect of the variant as $\beta\gamma$. These arguments do not rely on strong assumptions.

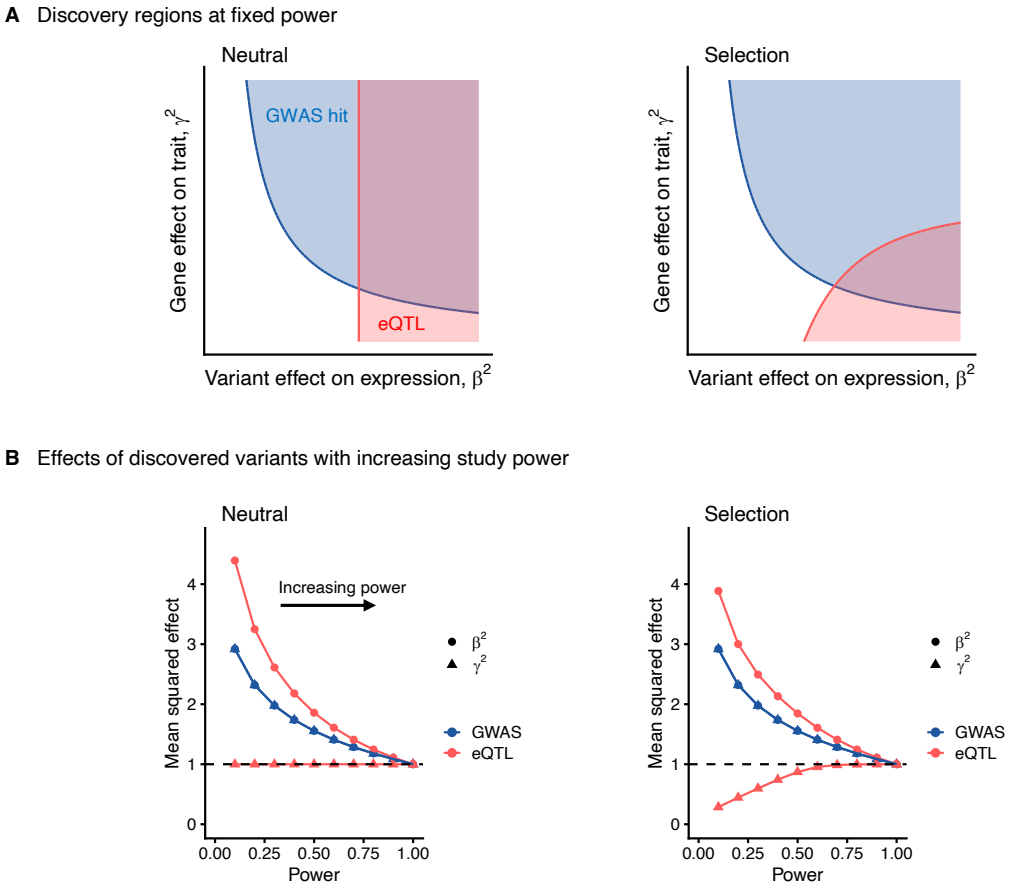
Our second key consideration is that discovery of a given variant in both GWAS and eQTL assays also depends on its minor allele frequency, p; discovery power is higher for more common variants. Under neutral evolution, p and effect sizes are independent, and thus variation in p, on average, does not modify the trends described above. When a phenotype is under natural selection however, variants with large phenotypic effects (i.e., large $\beta^2\gamma^2$) are kept at lower frequencies.

Based on previous evolutionary analyses of traits under stabilizing selection [27, 28], we considered selection to have a "flattening" effect: for variants with small effect sizes selection is weak and thus contribution to variance, $E[2p(1-p)\beta^2\gamma^2]$, scales roughly linearly with effect size. For variants with very large effect sizes selection is strong, lowering p such that $E[2p(1-p)\beta^2\gamma^2]$ plateaus. Under this model of selection, selection more strongly constrains regulatory variants acting on phenotypically important (high γ^2) genes, disproportionately degrading eQTL signal, $2p(1-p)\beta^2$, at such genes. On the other hand, considering $2p(1-p)\beta^2\gamma^2$ for GWAS discovery, lowering of p at these genes is compensated by high γ^2 values, and thus selection on average does not change the ranking of variants in GWAS.

Determining exactly how discovery trends are affected by selection requires a model for the joint distribution of p and the effect sizes, β and γ. Importantly however, our qualitative predictions on the effect of selection hold as long as p decreases monotonically with increasing phenotypic effect $\beta^2\gamma^2$, and do not rely on how exactly the flattening effect is formulated. Nevertheless, for visualization purposes we made some modeling choices: we used an asymptotic exponential form to describe the relationship $E[2p(1-p)\beta^2\gamma^2|\beta,\gamma] \propto \kappa(1-e^{-\beta^2\gamma^2/\kappa})$. We also assumed β and γ values are drawn from independent standard Normal distributions. We made an arbitrary choice of $\kappa = 2.986$ such that $E[2p(1-p)]$ is reduced by $\sim$10% compared to the neutral scenario.

We used simulations under these assumptions to show discovery regions of GWAS and eQTL assays with and without natural selection (Supplementary Fig. 14A, same as trends shown in main Fig. 6). We also show the average properties of variants and genes that fall in the discovery regions with increasing study power, by progressively including variants based on $2p(1-p)\beta^2$ values as eQTLs, and based on $2p(1-p)\beta^2\gamma^2$ values as GWAS hits (Supplementary Fig. 14B). These trends are consistent with the intuitions provided above, and with the data we present in the main text. Notably, under selection, eQTL assays prioritizes high β^2 variants but at low γ^2 genes.

In the remaining we systematically modify the assumptions listed above showing that these qualitative trends are robust to the specific modeling choices made.



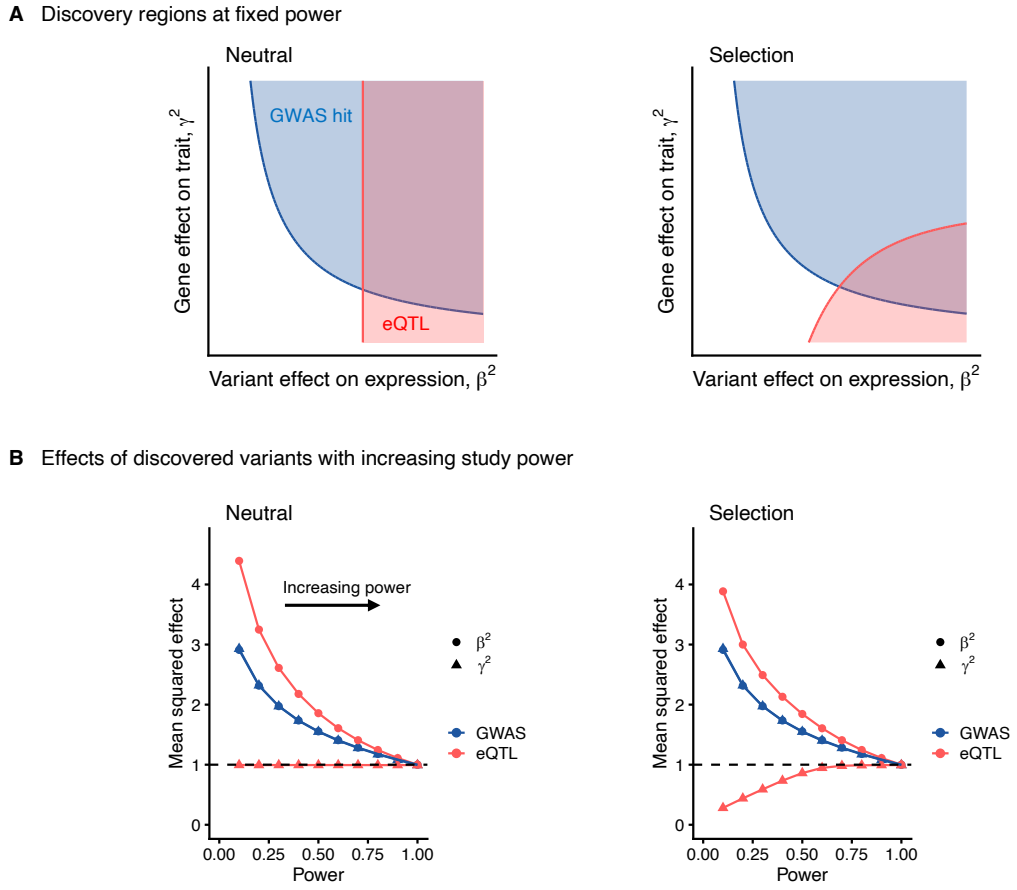
Supplementary Fig. 14: **Main qualitative model predictions.** *Model predictions under our baseline assumptions and choices of parameters: (i) a single cell type affecting a single phenotype, (ii) the SNP effect on gene expression, β, and the gene effect on phenotype, γ, drawn from independent standard Normal distributions, (iii) modeling selection to have a flattening effect formulated as $E[2p(1-p)\beta^2\gamma^2|\beta,\gamma] \propto \kappa(1-e^{-\beta^2\gamma^2/\kappa})$, where p is the allele frequency, and κ determines the strength of selection tuned to give $\sim$10% reduction in the average contribution to phenotypic variance across all causal SNPs relative to the neutral scenario. All results are based on 10 million simulated causal SNPs. See Supplementary Methods for details.* ***A)*** *Shading colors represent parameter space for the discovery of GWAS hit only (blue), eQTL only (red), and both types (purple) for the case of a neutrally evolving phenotype (left panel) and the case of a phenotype under selection (right panel). The discovery lines are derived assuming 15% power in both assays (see Online Methods).* ***B)*** *The mean properties of variants discovered as GWAS hits (blue) or eQTLs (red) with progressively increasing the discovery power in either assay. For a given study power X, points show the mean expression effect (mean β^2 values, circle) and the mean gene effect (mean γ^2 values, triangle) of the top X% of variants, ranked based on their strength of association signal in either assay, that is $2p(1-p)\beta^2\gamma^2$ for GWAS and $2p(1-p)\beta^2$ for eQTL mapping. The dashed lines show the mean values for all simulated causal SNPs.*

3.2 Joint distribution of β and γ

In this part, we derive results under different scenarios for the distributions of β and γ, while keeping other modeling choices (e.g., selection model parameters) fixed unless explicitly mentioned. Our

reference for comparison is Supplementary Fig. 14 which is based on β and γ values drawn from independent standard Normal distributions.

We first show that a different strategy for drawing gene effect sizes yields similar results. Specifically, in our main figures, we simulate 10 million SNPs with β and γ values drawn from independent standard Normal distributions. We varied this step by first drawing gene effects for 20K genes, and then sampling from these genes to assign to each SNP. Discovery trends under this sampling scheme (Supplementary Fig. 15) are almost identical to Supplementary Fig. 14.

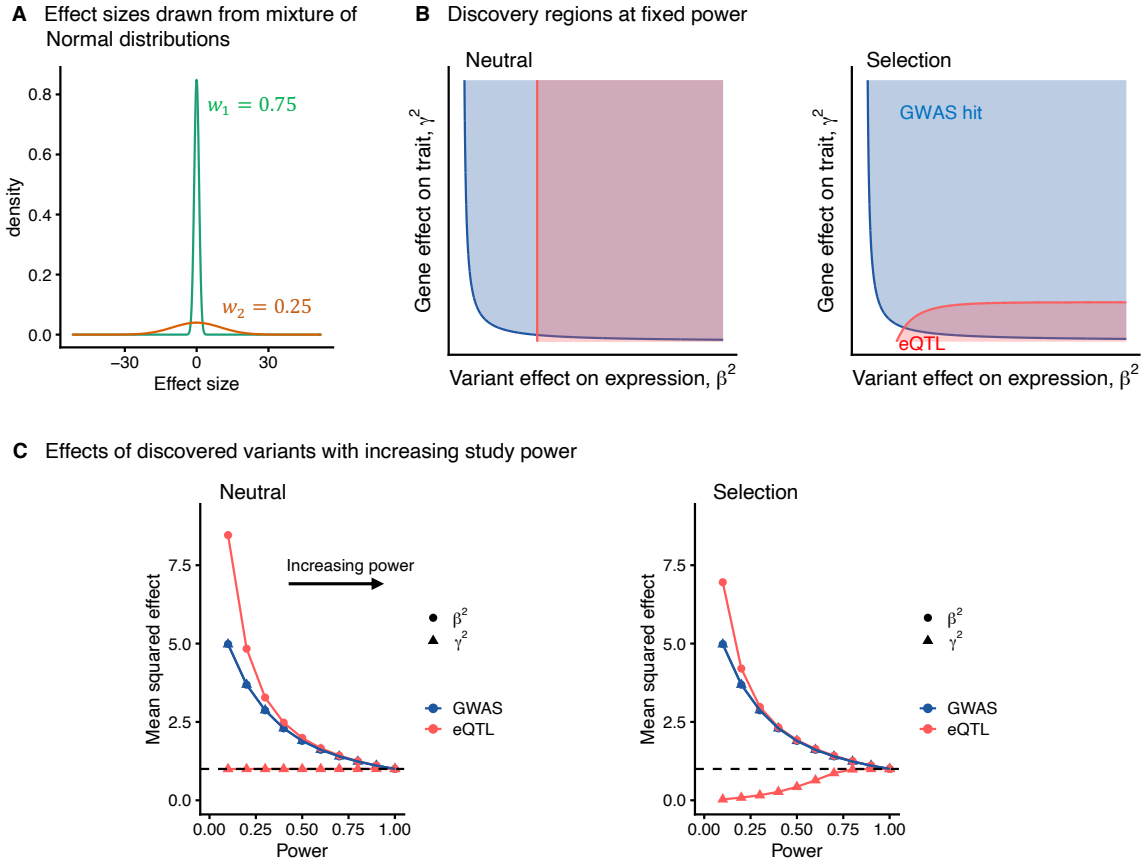


Supplementary Fig. 15: **Model results with modified gene effect sampling.** *Modeling and simulation details are similar to Supplementary Fig. 14, but using a two step sampling scheme for gene effects, γ^2: we first sampled 20K γ values from $N(0,1)$ corresponding to 20K genes, and then randomly sampled 10 million values with replacement from this set corresponding to 10 million linked SNPs.*

We next show that the discovery regions are qualitatively similar if we consider effects drawn from a mixture of two Normal distributions with different widths (Supplementary Fig. 16A), representing two sets of regulatory variants (e.g., promoter and enhancer variants) and two sets of genes. Also, the trends in the mean effects of variants prioritized by GWAS and eQTL designs in this scenario (Supplementary Fig. 16B) mirror trends in Supplementary Fig. 14B, though the absolute differences between GWAS and eQTL effects are larger.

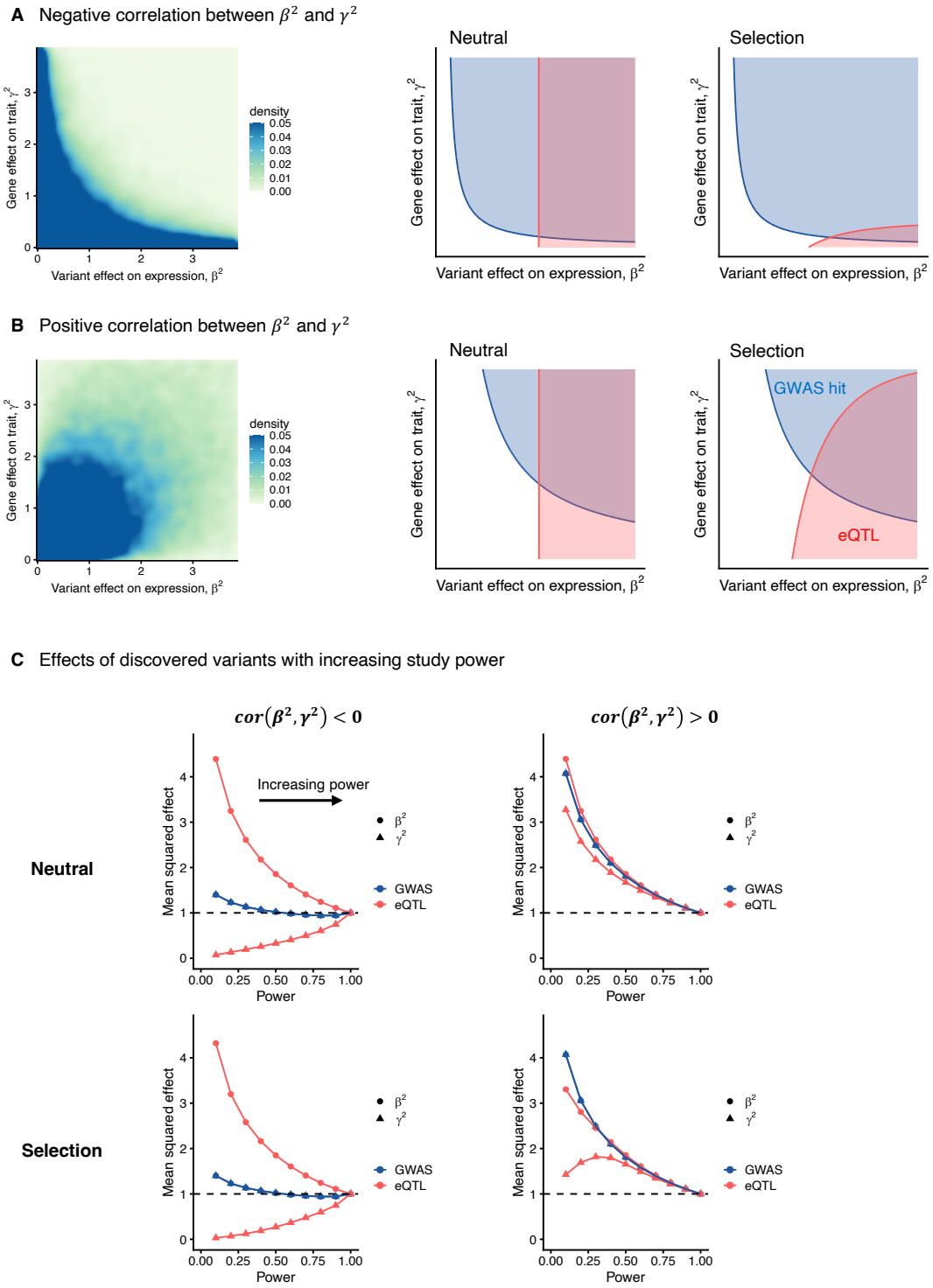
Next we varied the correlation between β^2 and γ^2, ρ_{β^2,γ^2} (Supplementary Fig. 17, see Supplementary Methods for details.) The discovery curves are qualitatively similar, whether ρ_{β^2,γ^2} is positive or negative (Supplementary Fig. 17A,B). That said, the exact shape of the discovery curves, the mean effect sizes of the variants that fall in the discovery regions, and how much these properties

are affected by selection are sensitive to ρ_{β^2,γ^2}, particularly for eQTL discovery: when $\rho_{\beta^2,\gamma^2} < 0$, detected eQTLs with high β^2 are disproportionately located at low γ^2 genes (Supplementary Fig. 17C). Furthermore, although selection has a strong effect on variants in the top-right corner of the parameter space, i.e., large β^2 and large γ^2, few such variants exist in this scenario, and thus the mean effect of selection across variants is small (Supplementary Fig. 17C). When $\rho_{\beta^2,\gamma^2} > 0$, detected eQTLs with high β^2 are located at high γ^2 genes, albeit with average genic effects that are smaller than GWAS variants (Supplementary Fig. 17C). Also, under this scenario, there are many more variants in the high selection region of the parameter space, and the net impact of selection is higher.



Supplementary Fig. 16: **Model results with effects drawn from a mixture of two Normal distributions.** *Modeling and simulation details are similar to Supplementary Fig. 14, but sampling β and γ values from a mixture of two Normal distributions as shown in panel (A). Discovery regions and trends shown in panels (B) and (C) are derived and presented similar to panels (A) and (B) in Supplementary Fig. 14, respectively.*

What biological factors determine the joint distribution of β^2 and γ^2? A key factor is the variability of regulatory architecture across genes: SNPs regulating genes with more/longer enhancers likely, on average, have smaller β^2s, due to the dispersion of cis-heritability for gene expression across more regulatory variants. On the other hand, it is conceivable that the complexity of the regulatory architecture of a gene, in part, reflects its functional importance, e.g., developmental genes are typically regulated by several enhancers elements [29]. Under these assumptions, it follows that genes with more/longer enhancers, on average, have higher γ^2s, and thus ρ_{β^2,γ^2} is plausibly negative. Similar points are made by Wang and Goldstein [30], though our argument does not rely on the assumption that multiple enhancer elements regulating the same gene are functionally redundant.



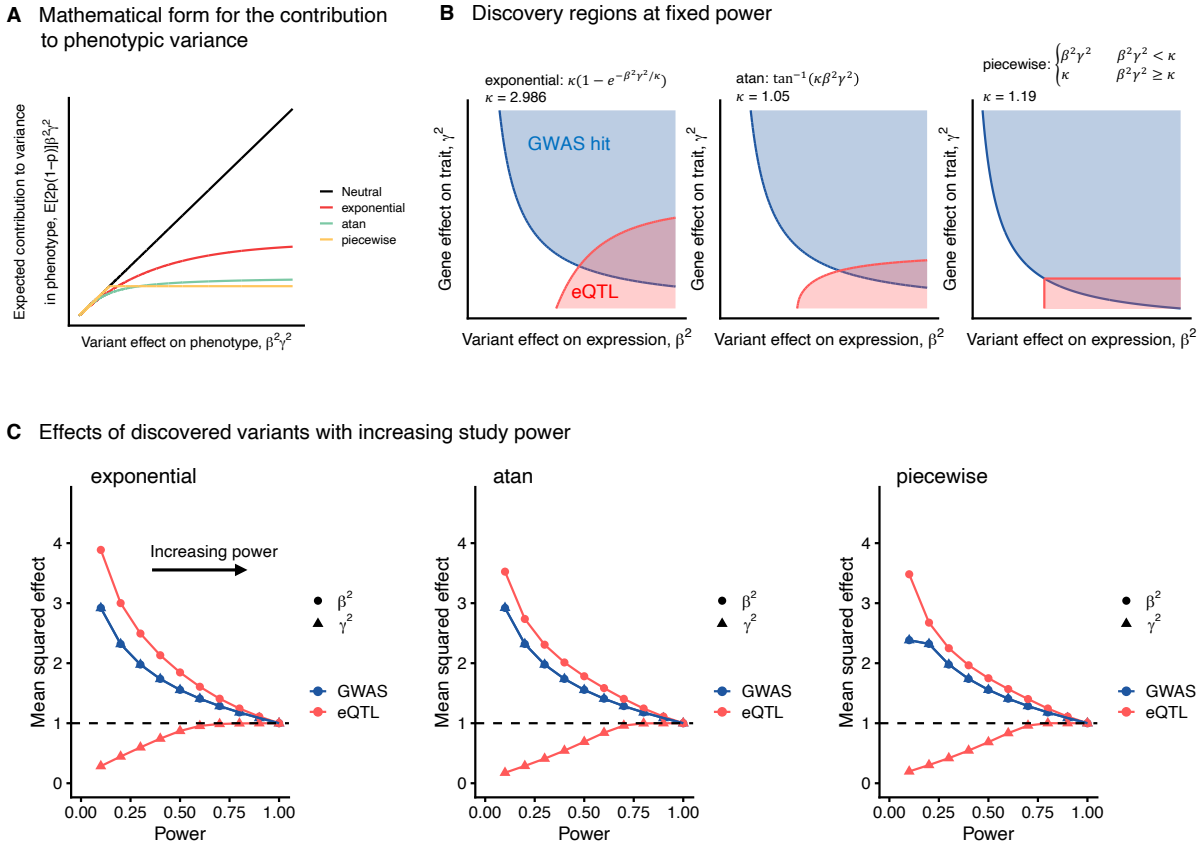
Supplementary Fig. 17: **Model results with varying correlation between effect sizes.** *Other than the correlation between β and γ, all modeling and simulation details are the same as in Supplementary Fig. 14.* ***A)*** *Joint distribution of negatively correlated β^2 and γ^2 values (left panel), simulated setting the parameter $\rho = -0.75$ (see Supplementary Methods) and the corresponding discovery regions (right panel) derived and presented similar to Supplementary Fig. 14A.* ***B)*** *Same as (A) but with positively correlated β^2 and γ^2 values, simulated setting the parameter $\rho = 0.75$.* ***C)*** *Discovery trends with increasing power under different distribution of effect sizes and selection scenarios, derived and presented similar to Supplementary Fig. 14B.*

In summary, the trends observed in data for GWAS and eQTL findings are consistent with our model with selection (which has been extensively documented for complex traits) and/or with a negative ρ_{β^2,γ^2} (which is biologically plausible as discussed above). Both likely contribute in reality.

3.3 Natural selection

In this part we evaluate two aspects of our model for the role of selection: (i) the mathematical model for selection, and (ii) the strength of selection. Other key model parameters are kept fixed, and identical to parameters used in Supplementary Fig. 14.

As we detailed earlier, motivated by previous evolutionary analyses [27,28], we modeled selection to have a flattening effect on phenotypic variance with effect size (Extended Data Fig. 7), and we used an asymptotic exponential form to describe the relationship $E[2p(1-p)\beta^2\gamma^2|\beta,\gamma] \propto \kappa(1-e^{-\beta^2\gamma^2/\kappa})$. We show that our qualitative results are similar when we use other mathematical forms with similar asymptotic behavior (Supplementary Fig. 18).

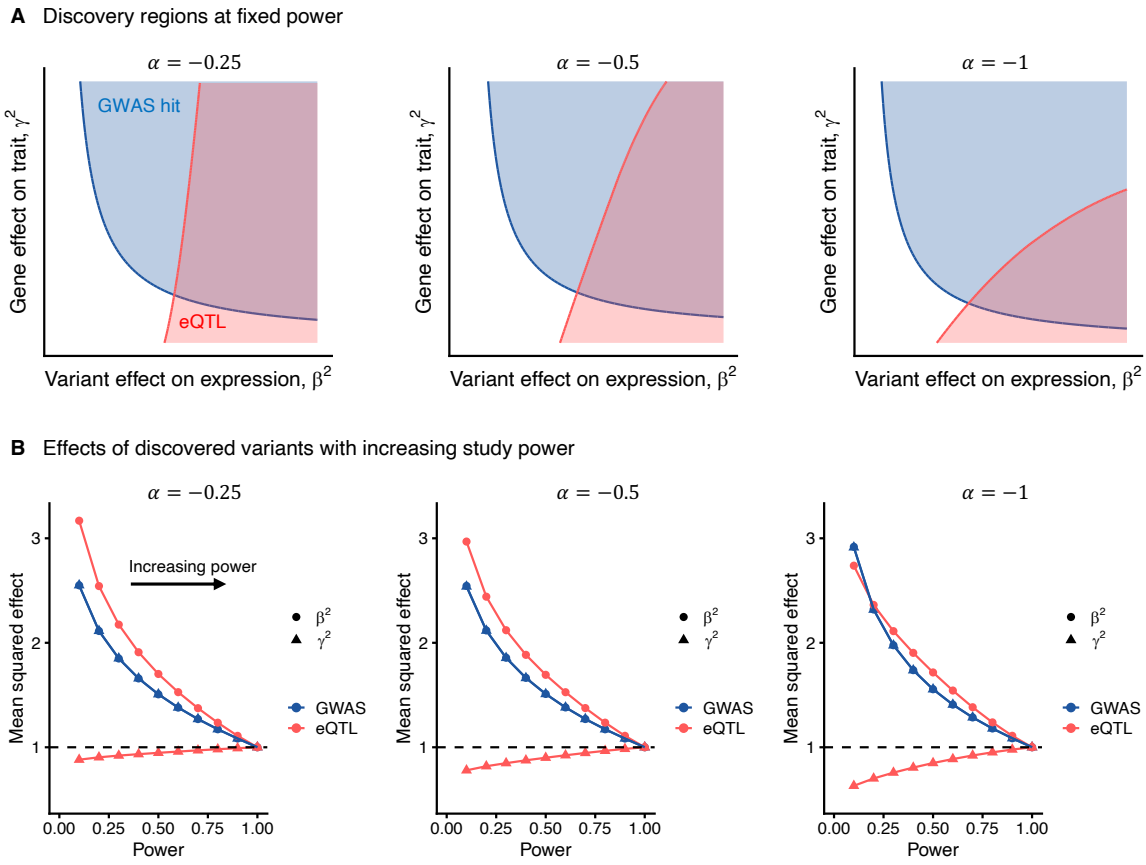


Supplementary Fig. 18: **Model results with varying mathematical form of selection's flattening effect.** *Modeling and simulation details are similar to Supplementary Fig. 14 for the case of a phenotype under natural selection, but using different mathematical forms to describe the flattening effect of selection, as illustrated in panel (A). The exact formulations are shown in panel (B). The κ values are tuned to give an average reduction in contribution to phenotypic variance of ~10% for all three formulations. Discovery regions and trends shown in panels (B) and (C) are derived and presented similar to panels (A) and (B) in Supplementary Fig. 14, respectively.*

Several studies have used a model previously referred to as the α model, to describe the relationship between allele frequency and effect size [4, 31, 32]. Under this model $E[\beta^2\gamma^2|p] \propto [2p(1-p)]^\alpha$. This model is not based on a particular evolutionary model; rather it is based on mathematical convenience as previously acknowledged [32]. Previous estimates gave $\hat{\alpha} < 0$, indicating a negative correlation between p and effect size as is consistent with the effect of selection [4, 31, 32].

We investigated how our results change under the α model. The α model describes $E[\beta^2\gamma^2|p]$ whereas our model is based on the reverse expectation $E[p|\beta^2\gamma^2]$. Therefore, to incorporate the α model, following a previous study [32], we approximated $E[p|\beta^2\gamma^2]$ as follows: we first sampled β and γ values from standard Normal distributions, and then multiplied those by a factor of $[2p(1-p)]^{\alpha/4}$, where p values are drawn from an exponential distribution. We then numerically solved for $E[p|\beta^2\gamma^2]$. See Supplementary Methods for details.

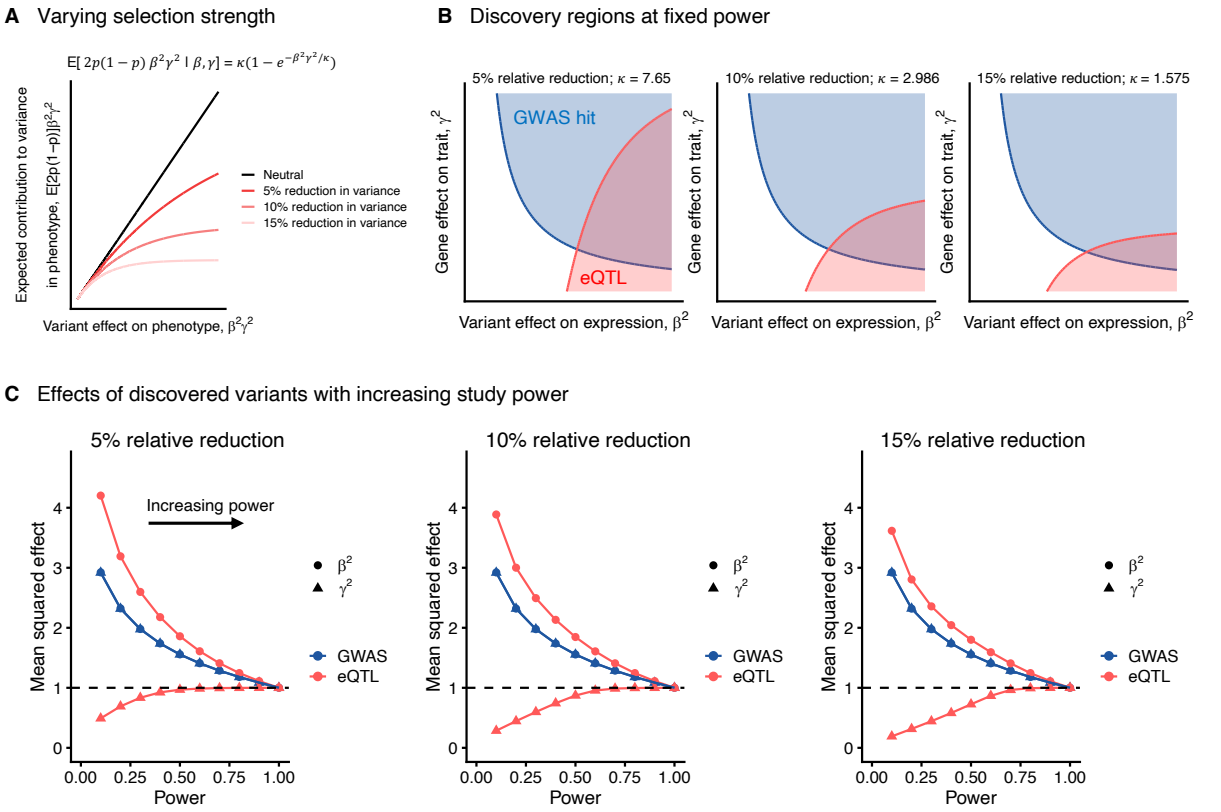
We experimented with negative α values on par with previous estimates and consistent with the effect of negative selection [4, 31, 32]. These produced results similar to our flattening model (Supplementary Fig. 19).



Supplementary Fig. 19: **Model results under the α model of selection.** *Modeling and simulation details are similar to Supplementary Fig. 14 for the case of a phenotype under natural selection, but using the α model to describe the relationship between phenotypic effect of a SNP and allele frequency as $E[\beta^2\gamma^2|p] \propto [2p(1-p)]^\alpha$. We considered negative α values, on par with previous estimates, and consistent with the effect of negative selection; a more negative α value corresponds to a stronger selection [4, 31, 32]. See the text and Supplementary Methods for details on the implementation of the α model in our simulations. Discovery regions and trends shown in panels (A) and (B) are derived and presented similar to panels (A) and (B) in Supplementary Fig. 14, respectively.*

Next, we explored the effect of varying selection strength, which in our flattening model of selection is achieved by varying κ. We tuned κ values based on the average per-SNP reduction in contribution to phenotypic variance compared to the neutral scenario (i.e., no correlation between p and effect sizes), i.e., $E[\frac{2p_{\text{selection}}(1-p_{\text{selection}})}{2p_{\text{neutral}}(1-p_{\text{neutral}})}]$. As expected, the gap between GWAS and eQTL variants grows with increasing selection strength, κ, (Supplementary Fig. 20), also consistent with trends observed with decreasing α in the α model described above (Supplementary Fig. 19).

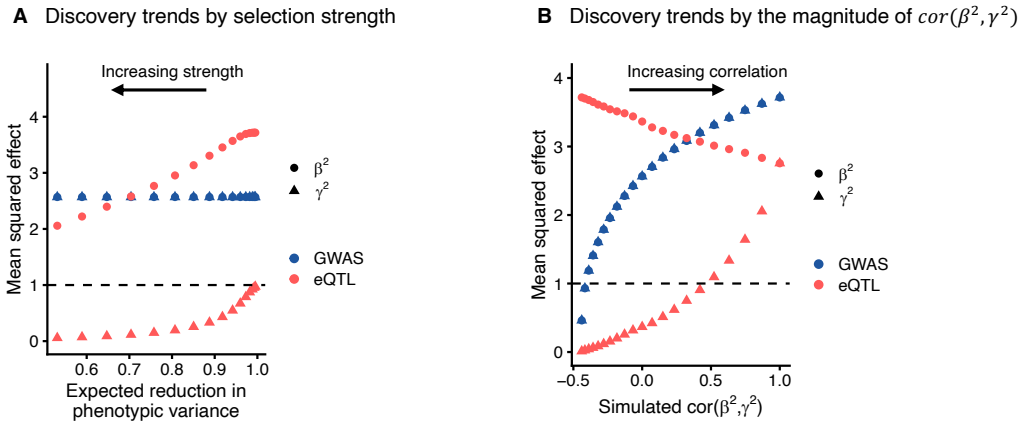
In summary, the qualitative effect of selection is robust to how selection is modeled: selection always reduces the amount of overlap between GWAS and eQTL findings by "bending" the eQTL discovery region down (as shown in Supplementary Fig. 14), skewing the discovered eQTLs away from the functionally important genes. Note that although our selection model was motivated by the effect of stabilizing selection on complex traits (which is considered the default mode of selection on quantitative traits [33]), the consequence for variant discovery would be similar under negative selection (e.g., as shown in Supplementary Fig. 19). That said, the exact shape and reduction in overlap depends on the selection model and joint distribution of β and γ. See below for an additional discussion on the challenges of quantitative analyses.



Supplementary Fig. 20: **Model results with increasing selection strength.** *Modeling and simulation details are similar to Supplementary Fig. 14 for the case of a phenotype under natural selection, but with varying strength of selection. Selection strength is determined by the κ parameter in our flattening model formulated as $E[2p(1-p)\beta^2\gamma^2|\beta,\gamma] \propto \kappa(1-e^{-\beta^2\gamma^2/\kappa})$ and illustrated in panel (A). The κ values are shown in panel (B), tuned to give ∼5%, ∼10%, or ∼15% reduction in the average contribution to phenotypic variance across all causal SNPs relative to the neutral scenario; higher reduction corresponds to stronger selection. Discovery regions and trends shown in panels (B) and (C) are derived and presented similar to panels (A) and (B) in Supplementary Fig. 14, respectively.*

3.4 Limitations of the model

The above analyses show that although our qualitative results are generally robust, making quantitative predictions is sensitive to the modeling and parameter choices. As an example, even under our simple model, the mean effects of discovered variants vary by selection strength and the correlation between β^2 and γ^2 effects (Supplementary Fig. 21). A detailed quantitative analysis requires an estimation of the true joint distribution of selection coefficients, allele frequencies, and effect sizes, β and γ. Selection coefficients and genic effects, γ^2, are especially hard to estimate (e.g., see [5,34]).



Supplementary Fig. 21: **Sensitivity of quantitative results to model parameters.** *The mean effect sizes of discovered variants as GWAS hits (blue) or eQTLs (red), with varying the strength of selection (A) or varying the correlation between the SNP effect on gene expression, β^2, and the gene effect on phenotype, γ^2 (B), keeping the discovery power fixed at 15% in both assays. In each panel, except the changing parameter, i.e., κ or the degree of flattening in panel (A) and ρ or the correlation between effect sizes in panel (B), all other modeling and simulation details are the same as in Supplementary Fig. 14. Shape of the points demonstrate the two model parameters. The dashed lines show the mean of all simulated causal SNPs.*

Furthermore, the picture becomes complicated very quickly as one goes beyond a single gene, single cell type, single phenotype model to a multi-gene, multi-cell type, multi-phenotype model. Specifically, a given SNP can affect multiple genes (G), in multiple cell types or contexts (C), and affect multiple phenotypes (P). Therefore, in reality, the joint distribution of β^2 and γ^2 effects needs to be quantified in a very high dimensional space ($G \times C \times P$). Our model in its simplest form should be viewed as a one dimensional representation of this high dimensional space.

For example, in our analysis of the eQTLGen data (Supplementary Fig. 12), we note that some discovery trends with increasing association p-value (as a proxy for increasing study sample size), notably with respect to the regulatory features of genes, are inconsistent with the predictions of our single cell-type model. Specifically, the depletion of eQTLs at genes with longer enhancers is more pronounced for weaker eQTLs. Explaining this observation in this data requires accounting for the variability of blood-specific enhancer architecture across genes and blood cell types, as well as its correlation with features influencing eQTL discovery, such as expression level, selective constraint, cis-heritability of gene expression, etc. These detailed considerations are challenging, and beyond the scope of this study.

Also note that while our model provides insight on how to conceptualize the effect of natural selection on variant discovery, understanding how selection affects specific GWAS or eQTL assays is complicated. This is mainly because trait-variants are usually pleiotropic [35]. Thus, the selective

constraint on a given SNP or gene is determined by its net effect in a multi-dimensional trait space. Also, considering that many complex traits are under selection [4,27,28,31], the contribution of any single trait to fitness is likely small. As a result, there is likely no one-to-one correspondence between a SNP's or gene's effect on a single trait and selection. Consistent with this picture, recent work suggests that the distribution of selection coefficients for GWAS variants, as well as the contribution of high pLI genes to trait heritability from common variants, is similar across a diverse range of traits [5,36].

Regardless, our model provides a useful framework to conceptualize current findings (and lack of findings) in GWAS and eQTL assays, and to guide future quantitative efforts in characterizing and estimating key biological parameters.

4 Model extensions

In this section we explore additional qualitative properties of our model for variant discovery and provide insight about more complicated scenarios than the single cell type, single phenotype model presented in the main text.

4.1 Dependency on sample size

The discovery regions in main Fig. 6 contain variants that explain variance in expression levels (for eQTLs) or phenotype (for GWAS variants) more than the study-dependent discovery thresholds, c^*:

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

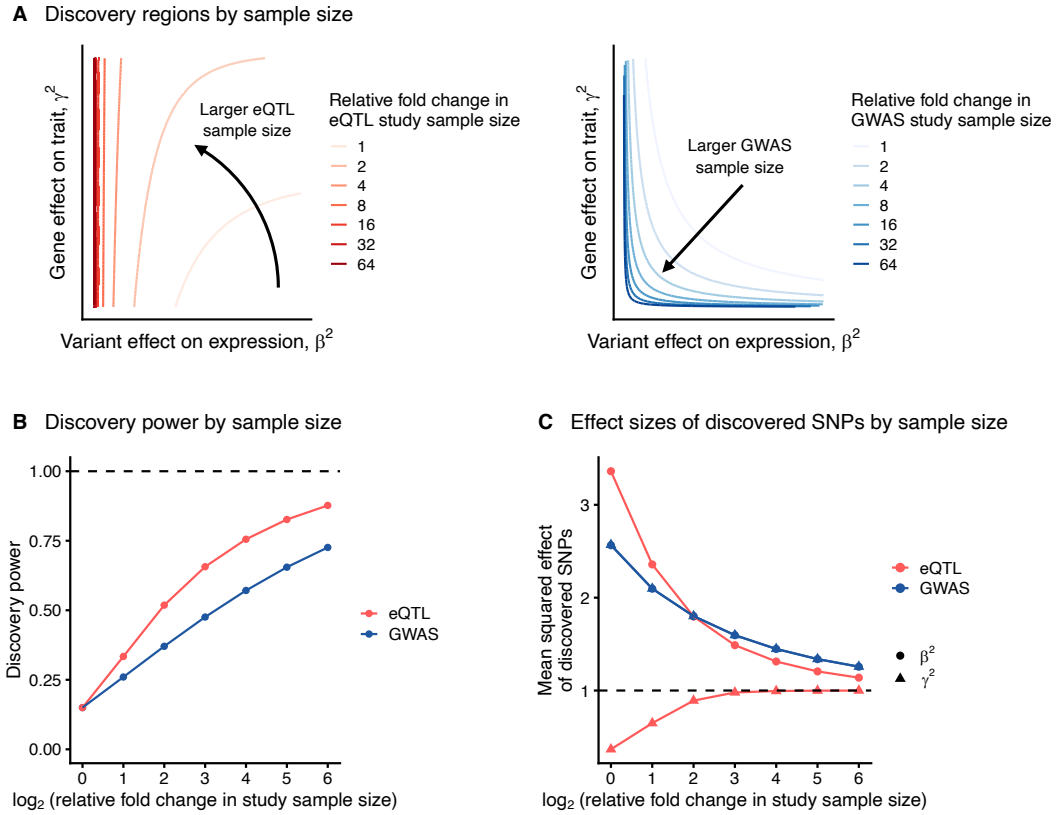
As detailed in the Online Methods section, $c^* \propto 1/n$, where n is the study sample size. For visualization purposes we set c^*_{GWAS} and c^*_{eQTL} values such that the discovery power in both assays is 15% under our modeling assumptions, on par with rough estimates of discovery power at current samples sizes (see section *Power considerations in GWAS and eQTL mapping*). Taking this point as reference, c^*_{ref}, we studied how discovery regions change as sample sizes increase by adjusting c^*: a k-fold increase in study sample size corresponds to setting $c^* = c^*_{\text{ref}}/k$.

For both GWAS and eQTL assays, the discovery regions expand with increasing sample size towards covering the whole parameter space (Supplementary Fig. 22A). Also in each assay, discovery lines remain qualitatively similar in shape across discovery thresholds (Supplementary Fig. 22A). However, the rate of increasing power with sample size is slower for GWAS (Supplementary Fig. 22B), considering that compared to eQTL effect sizes, β^2, GWAS effect sizes, $\beta^2\gamma^2$, depend on both β^2 and γ^2, and thus are relatively more variable across SNPs. Importantly, the discovered SNPs in the two assays are systematically different at any sample size, and the bias against functional genes in eQTL assays remains, though becomes smaller with sample size (Supplementary Fig. 22C).

We emphasize that the main point of this analysis is to demonstrate qualitative (and not quantitative) trends, for two main reasons: (i) The shifts in the position of discovery lines depend on c^*_{ref}. In other words, predicting the number of new discoveries with increasing sample size in practice depends on the discovery power at current sample sizes. Precise estimation of power at current sample sizes for each assay is not straightforward and is beyond the scope of this study. That said, we reason that, at current sample sizes, the discovery power in both GWAS and eQTL studies is on the low end (see section *Power considerations in GWAS and eQTL mapping*). (ii) Quantitating the discovery gains with samples size is sensitive to modeling choices and parameters, and we refer to our discussion in the previous section on the challenges of deriving quantitative results.

It is noteworthy that typical GWAS sample sizes (∼500K) are orders of magnitude larger than typical sample sizes of eQTL studies (∼500), and thus the range of fold increase in sample sizes considered here are perhaps more practical for eQTL studies, e.g., the eQTLGen study of blood eQTLs has sample size of ∼32K [17], about 64 times larger than GTEx whole blood sample size [15]. That said, translating sample size increases in multi-cell type, whole-tissue assays to the sample size parameter in our single cell type model is not straightforward. This is for two reasons: one, different cell types have different abundance in tissues; more abundant cell types contribute more to the total number of new eQTL discoveries with sample size. Two, cell types likely vary by their contribution to phenotypes, i.e., the distribution of genic effects, γ^2, varies by cell type. Therefore, different cell

types fall on different discovery trajectories as shown in Supplementary Fig. 22. In the next section we further explore discovery trends in a multi-cell type scenario.



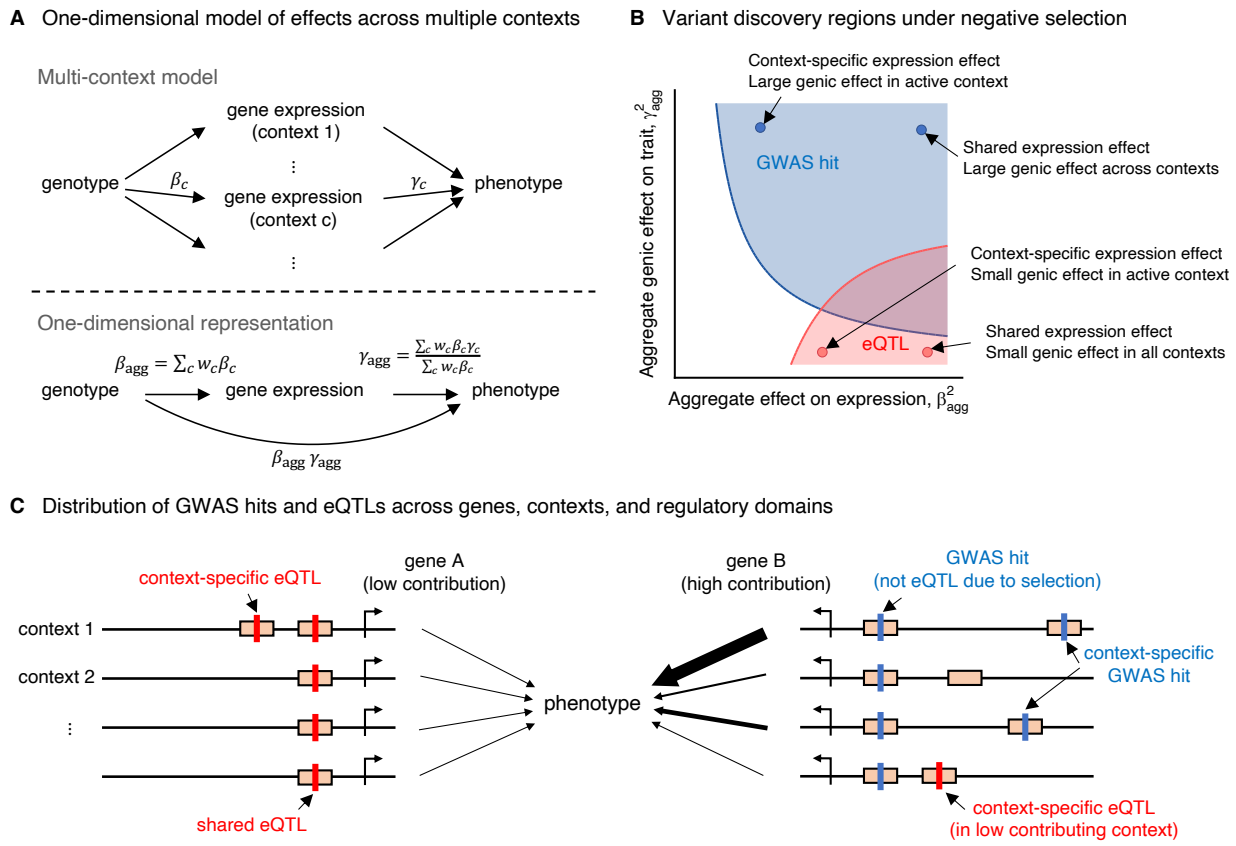
Supplementary Fig. 22: **Model results with increasing sample size.** *All modeling and simulation details are the same as the case of the phenotype under selection in Supplementary Fig. 14.* ***A)*** *Discovery lines for eQTLs (red, left panel) and GWAS hits (blue, right panel). The discovery lines corresponding to our reference point (i.e., fold change of 1 as indicated in the panel legends) are derived by setting the discovery thresholds c^*_{ref} such that 15% of the simulated causal SNPs are discovered in either assay, as described in Supplementary Fig. 14A (see Online Methods for details). Discovery lines corresponding to k fold larger study samples are achieved by setting $c^* = c^*_{\text{ref}}/k$.* ***B)*** *The fraction of all causal SNPs discovered as GWAS hits or eQTLs with fold increase in study sample size relative to the reference sample.* ***C)*** *Points show the mean expression effect (mean β^2 values, circle) and the mean gene effect (mean γ^2 values, triangle) of variants discovered as GWAS hits or eQTLs with fold increase in study sample size relative to the reference sample. The dashed line shows the mean values for all simulated causal SNPs.*

4.2 Multi-cell type model

In the main text, we have described the simplest case where there is only a single relevant cell type. But in practice, most current eQTL studies sample across mixtures of cell types (e.g., whole tissues), and/or environmental contexts. More fundamentally, genetic effects on a phenotype may be mediated by gene expression in multiple regulatory contexts (cell types, developmental stages, environmental stimuli, etc.). To explore these scenarios, Supplementary Fig. 23 outlines a simple model of GWAS and eQTL mapping in a bulk tissue context. We derive a one-dimensional representation of this scenario in order to use the insights gained from our single-context model discussed in the main text (Fig. 6).

In this representation, β_{agg} denotes an aggregate effect of a genetic variant on expression across contexts, and γ_{agg} denotes an aggregate effect of gene expression levels across contexts on the phenotype (Supplementary Fig. 23A). We define $\beta_{\text{agg}} = \sum_c w_c \beta_c$, as the weighted sum of effects over all causal contexts, where β_c denotes the genetic effect on expression in context c with the weighting w_c. As an example, considering a situation where a gene's expression in a tissue influences a phenotype, β_c represents the effect in cell type c, w the relative contribution of different cell types to the overall tissue-level mRNA levels, and β_{agg} the effect estimate in a bulk assay of the causal tissue.

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



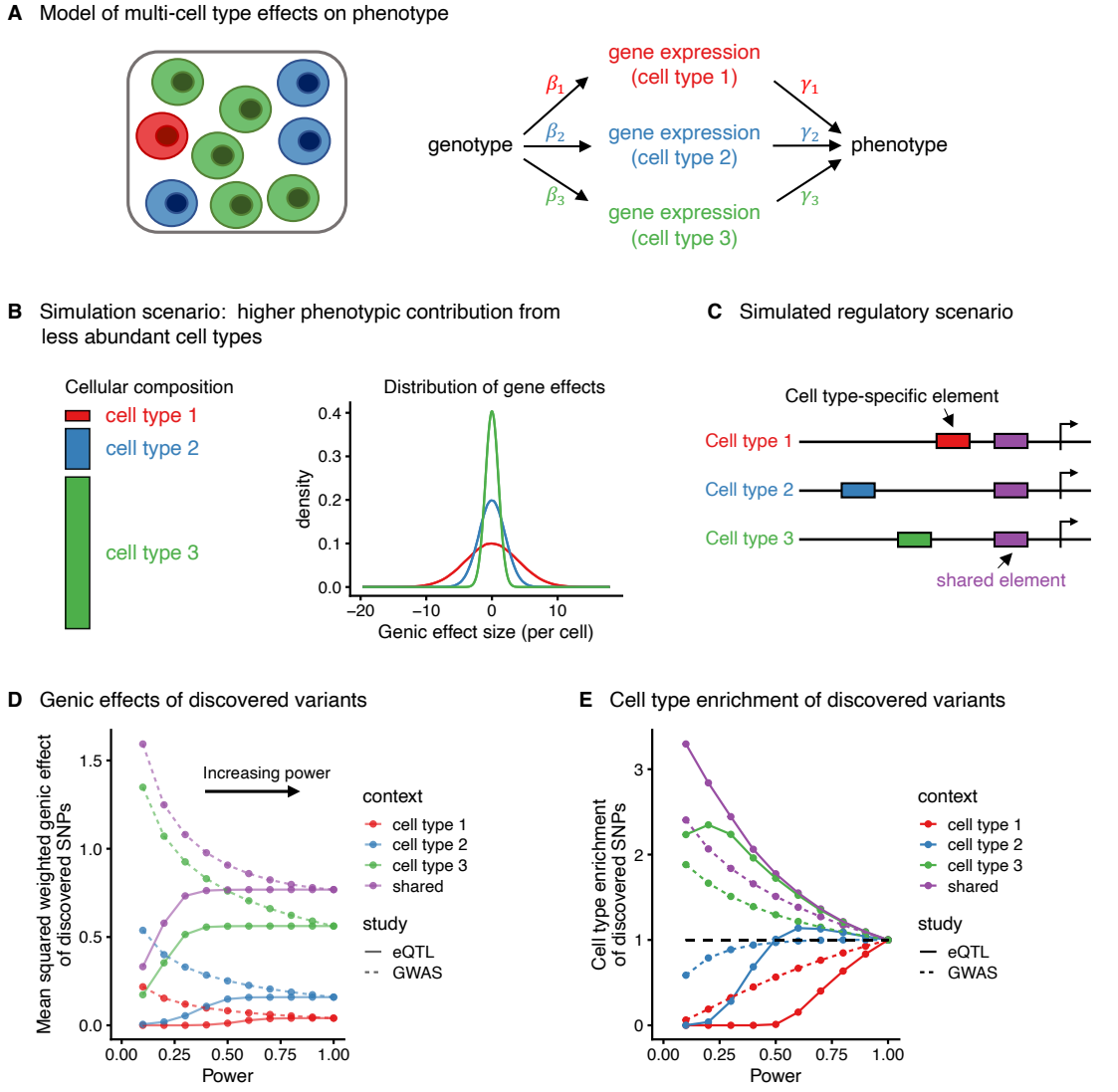
Supplementary Fig. 23: **A model for variant discovery across multiple contexts.** ***A)*** *One-dimensional representation of the pathway from genotype to phenotype mediated by gene expression across multiple contexts (e.g., cell types in a tissue), by defining parameters* β_{agg} *and* γ_{agg}*.* ***B)*** *Variant discovery in the space defined by* β_{agg} *and* γ_{agg} *for a phenotype under selection; same as in main Fig. 6B. Shading colors represent parameter space for the discovery of GWAS hit only (blue), eQTL only (red), and both types (purple).* ***C)*** *Schematic of variant discovery across multiple contexts mapped to regulatory domains (represented by orange rectangles) for two genes at different ends of the phenotypic importance axis.*

Similar to the our analysis in main Fig. 6, we can gain insight into the GWAS and eQTL discovery process by considering variants in terms of β_{agg} and γ_{agg} (Supplementary Fig. 23B). GWAS will tend to detect variants if $\beta_{\text{agg}}\gamma_{\text{agg}}$ is large enough, and this can occur through a combination of large expression effects and/or large phenotypic effects. But eQTL mapping will be most powerful for detecting large – shared – expression effects, and thus the two types of assays may have limited overlap. Supplementary Fig. 23C illustrates how this may play out for different variants: GWAS hits will be skewed towards functionally important genes and highly contributing contexts. Crucially, eQTLs will be skewed towards unimportant genes; meanwhile, eQTLs discovered at important genes will be skewed toward low contributing contexts. Thus, we may even find context-specific eQTLs for the "right" genes but at the "wrong" variants in the "wrong" contexts.

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

We expand on these intuitions using simulations of a multi-cell type scenario, where the causal context is a tissue composed of three different cell types (Supplementary Fig. 24A). We considered a cellular composition of $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3) = (0.05, 0.2, 0.75)$, and modeled genic effect sizes across cell types with independent Normal distributions $\gamma_c \sim N(0, \sigma_c)$, setting $(\sigma_1, \sigma_2, \sigma_3) = (4, 2, 1)$ exploring a scenario where less abundant cell types are more trait-relevant (Supplementary Fig. 24B). We set 10% of regulatory effects to be shared across cell types (Supplementary Fig. 24C), and the expression effects of these shared variants to be correlated across cell types: $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, setting $\Sigma_{i,j} = 0.75$ for non-diagonal and 1 for diagonal elements. The remaining 90% of variants are set to be cell type-specific, equally distributed across cell types (Supplementary Fig. 24C), with expression effect $\beta_c \sim N(0, 1)$ in the relevant cell type, and 0 in other cell types.

We investigated what types of variants are discovered in GWAS and eQTL assays in this scenario. To this end, in simulations we computed β_{agg} and γ_{agg} as described above, and used these in our single cell type framework to model the effect of natural selection. For each variant we defined a weighted genic effect size $\bar{\gamma} = \sum_c w_c I_c \gamma_c$, where I_c is an indicator for the activity of SNP in cell type c. The interpretation of $\bar{\gamma}$ is the net change in phenotype with one unit change in the expression levels of a variant's target gene across all cells in which the variant is active. [Note that $\bar{\gamma}$ is different from γ_{agg}. The motivation to define a new parameter is that γ_{agg} is different for different variants as it depends on the variant's effects on expression, as such when comparing different genes, one needs a variant-independent measure]. We find that across all types of regulatory variants, genes linked with GWAS hits have higher $\bar{\gamma}$ than eQTLs active in the same regulatory contexts (Supplementary Fig. 24D). Also, compared to GWAS variants, eQTLs are more enriched in shared variants and depleted of cell type-specific variants active in the high contributing cell type (Supplementary Fig. 24E).



Supplementary Fig. 24: **Discovery trends in a multi-cell type simulation scenario.** *A scenario whereby a mixture of cell types, e.g., in a causal tissue, affect a single phenotype.* ***A)*** *We considered three cell types: cell type 1 (red), cell type 2 (blue), and cell type 3 (green). For each SNP we define six parameters:* β_c *as the SNP effect on gene expression in cell type c, and* γ_c *as the change in phenotype with one unit change in the expression of the target gene in cell type c.* ***B)*** *We simulated a scenario where less abundant cell types are more trait-relevant. Bar lengths (left panel) show cell type proportions* $(\omega_1, \omega_2, \omega_3) = (0.05, 0.2, 0.75)$*. Considering 20K genes and three cell types, we sampled 20K* $\times$ *3 cell type-specific genic effects drawn from independent Normal distributions* $\gamma_c \sim N(0, \sigma_c)$*, setting* $(\sigma_1, \sigma_2, \sigma_3) = (4, 2, 1)$ *(right panel).* ***C)*** *We simulated 10 million SNPs, considering four categories of regulatory variants: shared SNPs across cell types (purple, 1 million SNPs), and three categories of cell type-specific effects (3 million SNPs with specific effects for each cell type). See the text for expression effect,* β*, assignments per category. To link genes to SNPs, we sampled 10 million times from the 20K genes described in (B) (similar to our procedure in Supplementary Fig. 15).* ***D,E)*** *For each SNP, we computed* β_{agg} *and* γ_{agg}*, and used these in our single cell type, single phenotype model under natural selection (described for Supplementary Fig. 14) to prioritize SNPs based on their signal strength in GWAS or eQTL assays. We show properties of discovered SNPs with progressively including variants with weaker signal in either assay, i.e., with increasing discovery power: (D) mean squared genic effect across cell types,* $\bar{\gamma}^2$ *(see the text for definition), and (E) enrichment of SNPs with different regulatory activities relative to all SNPs (black dashed line). Colors correspond to regulatory activity of SNPs.*

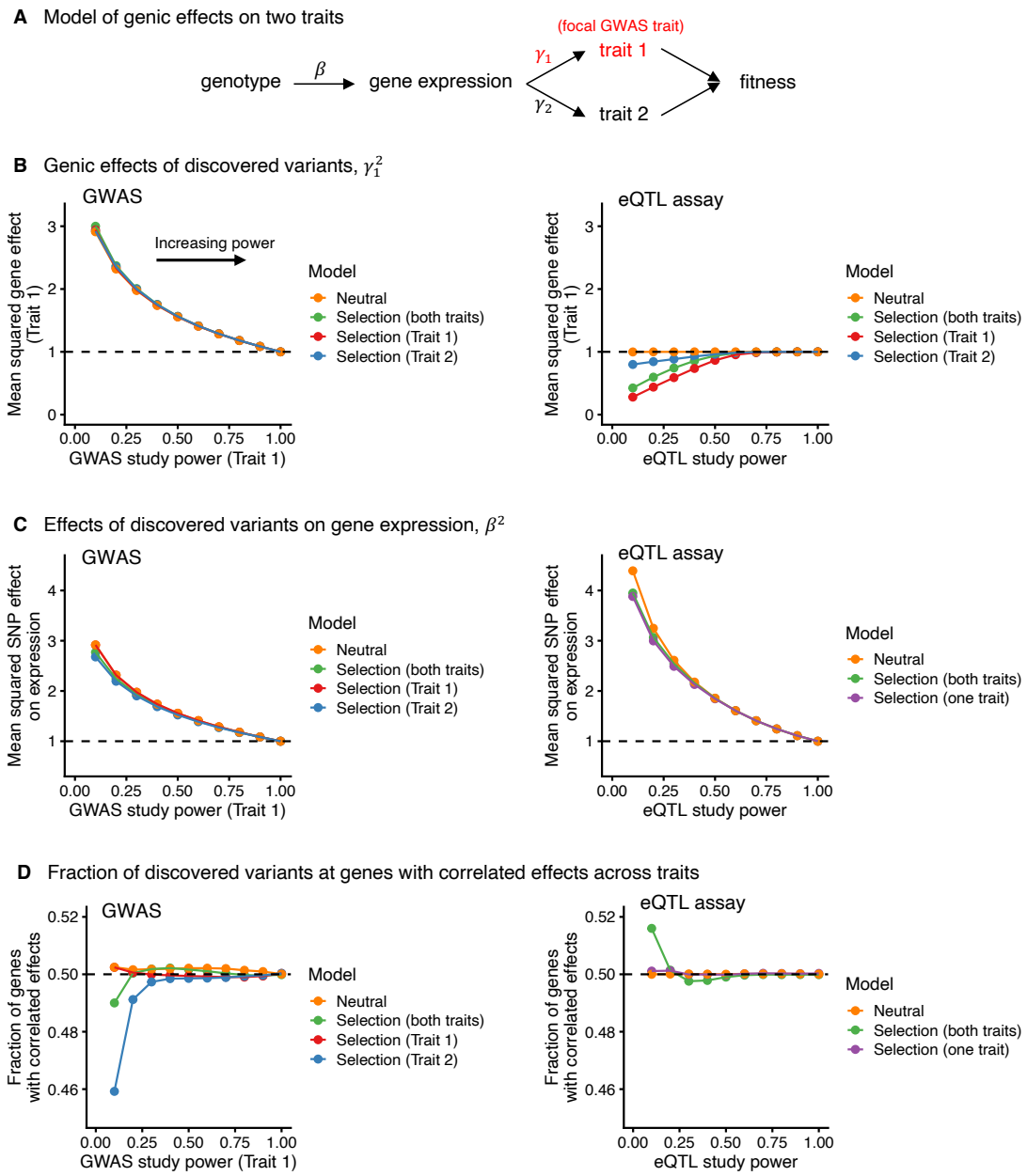
4.3 Multi-phenotype model

We now explore the scenario where a single cell type or context contributes to two traits (Supplementary Fig. 25A), considering that pair-wise genetic correlation between complex traits is very common [35]. We considered 50% of genes to have independent effects on the phenotypes, hereafter referred to as "trait-specific genes", with effects modeled as independent Normal distributions $\gamma_i \sim N(0,1)$, where i denotes the trait index. The remaining 50% of genes we set to have correlated effects across traits, hereafter referred to as "shared genes": $\boldsymbol{\gamma} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, setting $\Sigma_{i,j} = 0.75$ for non-diagonal and 1 for diagonal elements. The gene expression effects were modeled as in the previous simulations: $\beta \sim N(0,1)$.

In reality, it is typically unknown which complex traits are under natural selection directly, or indirectly by being correlated with other traits that are directly under selection. Considering this, we selected one focal trait (trait 1 in our simulation example) for GWAS analysis, and considered three natural selection scenarios, where the effect of selection is mediated by: (i) both traits (assuming equal contribution from each trait), (ii) only the focal GWAS trait (trait 1), or (iii) only the correlated trait not included in the GWAS study (trait 2). We modeled the effect of selection by computing the net effect of variants on fitness as $\beta^2\gamma^2_{\text{net}}$, defining $\gamma^2_{\text{net}} = \sum_i w_i\gamma^2_i$, where w denotes traits relative contributions to fitness.

In all these selection scenarios, the discovery trends in both GWAS and eQTL assays are qualitatively similar to our single trait model, i.e., when selection is directly acting on the GWAS trait (Supplementary Fig. 25B,C). That said, in the eQTL assay, the degree of depletion of high effect genes with respect to trait 1, i.e., γ^2_1, depends on the selection scenario (Supplementary Fig. 25B, right panel). This is because the effect of selection is determined by the net effect in the phenotype space, i.e., γ^2_{net}, but the relative contributions of γ_1 and γ_2 to γ^2_{net} varies by scenario. For example, when trait 1 is directly under selection, all genes contribute to fitness proportional to γ^2_1. But when trait 2 is directly under selection, selection on trait 1-specific effects is decoupled from γ^2_1.

There are also systematic differences between GWAS hits and eQTLs with respect to the contribution of shared genes versus trait-specific genes (Supplementary Fig. 25D, left panel). Compared to trait 1-specific genes, shared genes have higher γ^2_{net} on average when trait 2 is directly under selection. In this case, selection is stronger on shared genes, which lowers the allele frequency of their linked SNPs and disproportionately hampers their discovery in GWAS for trait 1. In contrast, eQTL discovery is skewed towards variants at shared genes, particularly when selection is acting on both traits (Supplementary Fig. 25D, right panel). This is because eQTL discovery is biased towards genes with low γ^2_{net}, i.e., the mean of γ^2_1 and γ^2_2, and the distribution of γ^2_{net} is more dense at low values when the two parameters are correlated. When selection is acting on only one of the traits, correlation with the other effectively neutral trait is irrelevant for gene discovery.



Supplementary Fig. 25: **Discovery trends in a multi-phenotype scenario.** ***A)*** *A scenario in which a single cell type underlies two traits: trait 1, the focus of GWAS, and trait 2, partially correlated with trait 1 but not included in GWAS. For each gene we define two trait-specific genic effects, γ_1 and γ_2. We explore three selection scenarios where the effect on fitness is mediated by: only trait 1, only trait 2, or both traits (equal contribution).* ***B-D)*** *We simulated 10 million SNPs linked with 20K genes with β, γ_1 and γ_2 values sampled as described in the text. To link genes to SNPs, we first sampled 20K × 2 trait-specific genic effects, and then sampled 10 million times from the 20K genes (similar to our procedure in Supplementary Fig. 15). For each SNP and selection scenario, we computed γ^2_{net} (see the text for definition), and used it along with β^2 in our single cell type, single phenotype model (described for Supplementary Fig. 14) to prioritize SNPs based on their signal strength in GWAS or eQTL assays. We show properties of discovered SNPs with progressively including variants with weaker signal in either assay, i.e., with increasing discovery power: (B) mean squared genic effect on trait 1, γ_1^2, (C) mean squared SNP effect on expression, β^2, and (D) fraction of discovered SNPs that are linked with genes with correlated effects on the two traits, i.e., shared genes. Colors correspond to selection scenarios. Dashed lines show average properties of all causal SNPs.*

5 Colocalization of eQTLs and GWAS hits

Our main analyses are focused on the question: what types of variants are prioritized in GWAS and eQTL assays? A different, but related question is: at what types of GWAS loci is there eQTL signal, i.e., GWAS hits and eQTLs colocalize? This is most relevant when eQTL data is used to help identify the target genes and relevant contexts of a given set of GWAS variants. In this section, by combining data analysis and modeling, we show that our main conclusions are applicable to the second question as well.

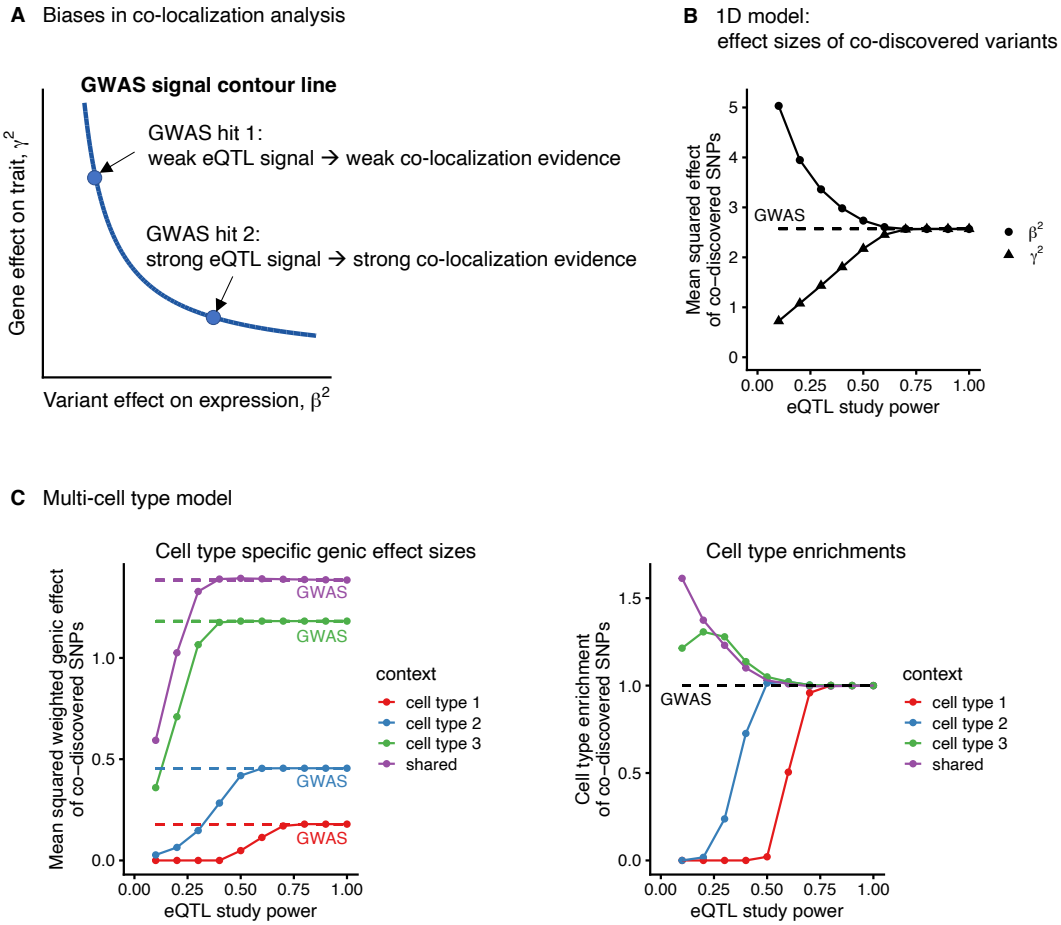
5.1 Insights from our model

We consider two approaches for joint evaluation of GWAS and eQTL signals: (i) for a given set of GWAS hits, investigating which ones are also independently discovered as eQTLs (referred to as "co-discovery" in the following), and (ii) integrating GWAS and eQTL signals across the genome without significance testing in either assay. Transcriptome-wide association studies (TWAS) (e.g., [37–39]), and most statistical approaches for testing for colocalization (e.g., [40]) can be viewed in terms of the second approach. That said, some uses of these statistical methods involve a hybrid of the two approaches, e.g., testing for colocalization only at GWAS loci or loci with some evidence for trait association (e.g., [41]). Without loss of generality, we ignore the problem of LD confounding which complicates distinguishing whether GWAS and eQTL signals are driven by the same causal SNP or two different causal SNPs that are in LD, as it is orthogonal to the factors discussed here.

Intuitively, a colocalization signal is strongest when evidence for both GWAS and eQTL signals is strong. Now, let's consider a set of GWAS hits that explain similar amounts of phenotypic variance, which can be represented by a contour line of $\beta^2\gamma^2 = \text{const.}$ in the parameter space of our 1-D model (Supplementary Fig. 26A). For this set, colocalization evidence becomes stronger with the strength of the eQTL signal, which according to our model is more likely at genes with high β^2 but low γ^2 (Supplementary Fig. 26A). Therefore, in principle, prioritizing GWAS hits based on colocalization evidence, although providing candidates for the downstream target genes, will bias gene nomination towards less phenotypically relevant genes.

We used our simulation framework to demonstrate this intuition: we considered a set of GWAS hits (top 15% of variants based on $2p(1-p)\beta^2\gamma^2$ values) for a phenotype under selection. We then called eQTLs at different discovery power values, progressively including variants based on $2p(1-p)\beta^2$ values, and computed the genic importance (i.e., average γ^2) of variants that were discovered in both assays (co-discovered SNPs). Our point of reference here is the set of all GWAS variants (and not all causal variants as was in previous sections). Consistent with the discussion above, relative to all GWAS hits, co-discovered SNPs are skewed towards high β^2 variants at low γ^2 genes (Supplementary Fig. 26B).

We observed similar biases in our multi-cell type model (Supplementary Fig. 26C), now calling variants as GWAS hits or eQTLs based on β^2_{agg} and γ^2_{agg} values (see section *Multi-cell type model*). For all regulatory contexts, i.e., cell type specific or shared regulatory variants across cell types, SNPs co-discovered as GWAS hits and eQTLs have lower γ^2 than the set of all GWAS hits in each corresponding context (Supplementary Fig. 26C, left panel). Also, co-discovered SNPs are skewed towards variants that are shared across cell types or are active in the least contributing cell types (Supplementary Fig. 26C, right panel). These trends are consistent with eQTL discovery trends shown in Supplementary Fig. 24.

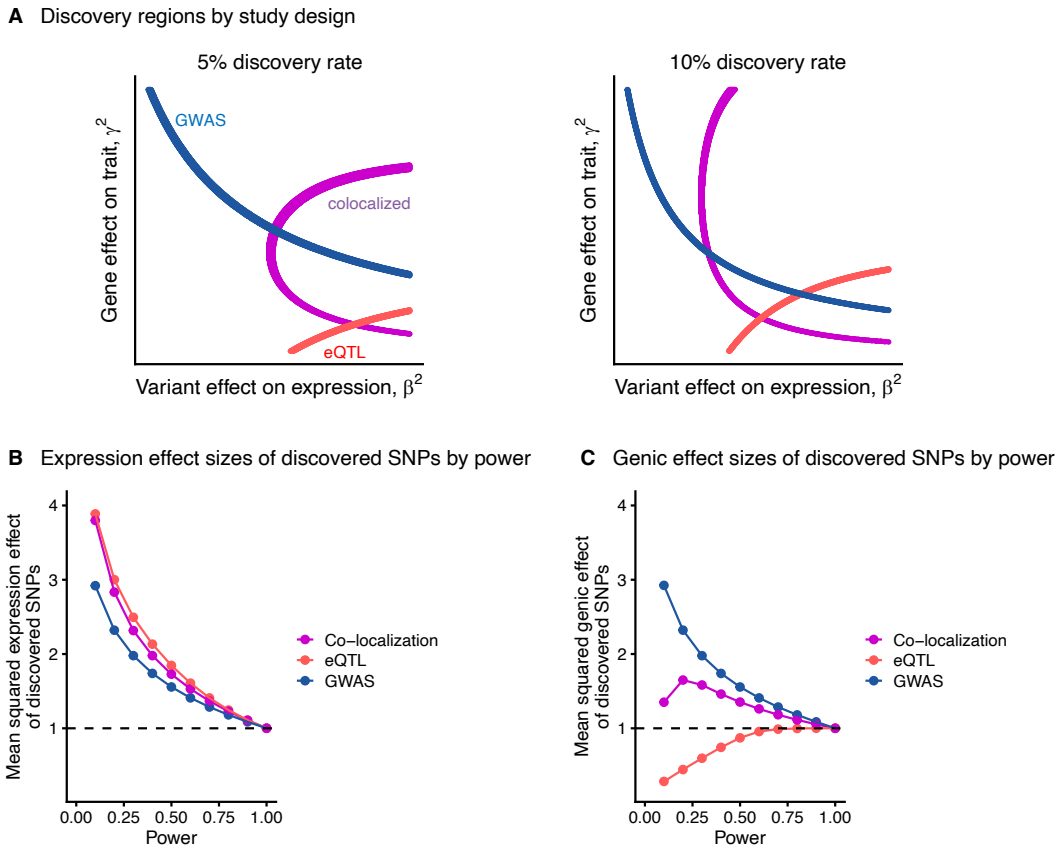


Supplementary Fig. 26: **Model results for co-discovery of GWAS hits and eQTLs.** *GWAS hits with and without eQTL evidence are systematically different.* ***A)*** *Variants with the same strength of GWAS signal are shown by a blue contour line in the parameter space defined by β^2 and γ^2. The eQTL signal and thus colocalization evidence is stronger at GWAS loci with higher β^2 but lower γ^2.* ***B)*** *The mean properties of GWAS hits (at a fixed GWAS discovery power of 15%) that are also discovered as eQTLs in our single cell type, single phenotype model, with progressively increasing the discovery power of the eQTL study. For a given eQTL study power X, points show the mean expression effect (mean β^2 values, circle) and the mean gene effect (mean γ^2 values, triangle) of the GWAS hits that are among the the top X% of variants, ranked based on their strength of association signal in eQTL mapping, that is $2p(1-p)\beta^2$. The dashed line shows the mean effects for all GWAS hits. All modeling and simulations details are similar to Supplementary Fig. 14.* ***C)*** *The properties of GWAS hits (at a fixed GWAS discovery power of 15%) that are also discovered as eQTLs in our multi-cell type model, with progressively increasing the discovery power of the eQTL study: mean squared genic effect across cell types, $\bar{\gamma}^2$ (see section Multi-cell type model for definition) (left panel), and enrichment of SNPs with different regulatory activities relative to all GWAS hits (right panel). All modeling and simulations details are similar to Supplementary Fig. 24. Colors correspond to the regulatory context of GWAS hits. The dashed lines show the average properties of all GWAS hits.*

Statistical approaches such as TWAS or colocalization tests integrate GWAS and eQTL signals without significance testing in either assay as explored above. Based on the study by Hukku et al. [42], on a single-variant level, these approaches can be viewed as testing for $\gamma_j\beta_j\hat{\beta}_i\text{Cov}(g_j, g_i) \neq 0$ for variants i and j, where $\gamma_j\beta_j$ is the effect of variant j on the phenotype, $\hat{\beta}_i$ is the effect estimate of variant i on gene expression in the eQTL assay, and g denotes the genotypes at these loci. Ignoring LD, that is assuming that at a given locus the GWAS and eQTL signals are driven by the same

known variant, the strength of the colocalization signal at the putatively causal variant would be $[\sqrt{2p(1-p)}\beta\gamma][\sqrt{2p(1-p)}\hat{\beta}]$, which in expectation, is equivalent to the geometric mean of GWAS and eQTL signals.

These considerations are inexact and oversimplify the colocalization problem. Nevertheless, they provide the intuition that ranking variants based on evidence for colocalization through integrating GWAS with eQTL data, skews discoveries away from GWAS-like to more eQTL-like variants. We illustrate this intuition using our simulation framework to compare GWAS and eQTL discovery trends, with trends when variants are ranked by the average strength of GWAS and eQTL signals, i.e., $2p(1-p)\beta^2\gamma$, as a proxy for colocalization signal, showing that features of colocalized variants are in-between GWAS hits and eQTLs (Supplementary Fig. 27).



Supplementary Fig. 27: **Model results for integration of GWAS and eQTL signals.** *Features of colocalized variants are in-between GWAS hits and eQTLs. All modeling and simulations details are similar to Supplementary Fig. 14. For a given SNP, the colocalization signal is derived as* $2p(1-p)\beta^2\gamma$ *which is the geometric mean of the signals in GWAS and eQTL assays,* $2p(1-p)\beta^2\gamma^2$ *and* $2p(1-p)\beta^2$*, respectively. See the text for details.* ***A)*** *Discovery lines for GWAS hits (blue), eQTLs (red), and variants prioritized based on colocalization evidence (purple) at* 5% *(left) and* 10% *discovery power (right).* ***B,C)*** *The properties of prioritized variants in GWAS (blue), eQTL assay (red), or based on colocalization signal (purple), with progressively including variants with weaker signal in either approach, i.e., with increasing power: (B) mean squared SNP effects on expression,* β^2*, and (C) mean squared gene effects on phenotype,* γ^2*. The dashed lines show the mean values for all simulated SNPs.*

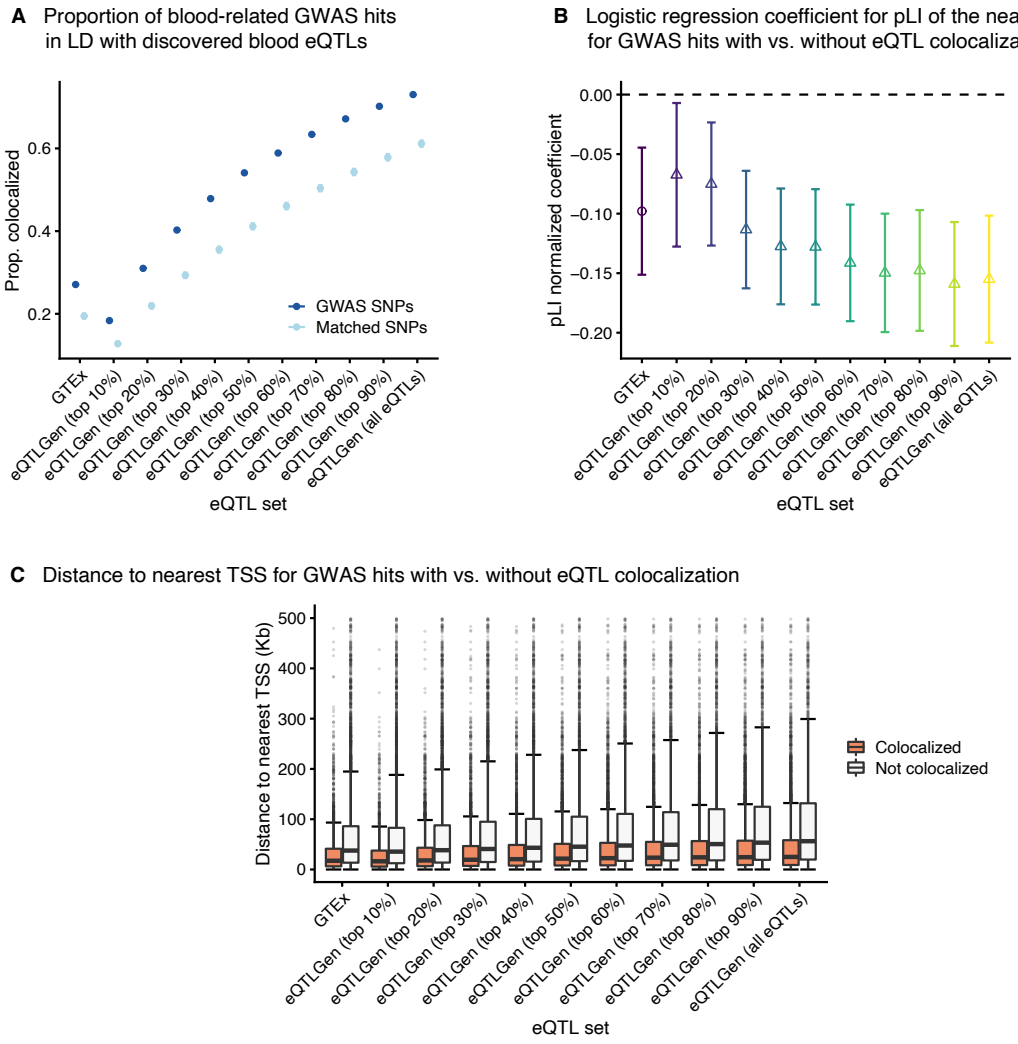
5.2 Colocalization of blood eQTLs and blood-related GWAS hits

As described in the analyses for Supplementary Figs. 1 and 2, out of the 44 complex traits we included in our main analyses, we characterized 14 as blood or immune related based on enrichment in myeloid/erythroid or lymphoid specific open chromatin regions [2] (see Supplementary Methods for details). We focused on 10,980 unique GWAS hits across these 14 traits, and as a proxy for colocalization, we investigated whether they are in LD ($r^2 > 0.8$) with any eQTL discovered by GTEx in whole blood [15], or in eQTLGen data [17]. We note that LD between GWAS and eQTL variants does not necessarily imply that the same causal SNPs underlie GWAS and eQTL signals, and so LD likely leads to some false ascertainment of colocalization events. That said, we find that GWAS hits for these 14 traits are in LD with blood eQTLs more than control SNPs matched for MAF, LD and gene density (Supplementary Fig. 28A). As with previous analyses of eQTLGen data, we sliced these eQTLs into 10 groups based on the deciles of association p-values, mimicking the progressive discovery of eQTLs with sample size, and to avoid pooling a large number of eQTLs which could complicate the interpretations. Furthermore, a lack of LD between GWAS hits and any eQTL in the data is a conservative indication of the absence or weakness of eQTL signal at the GWAS loci. Interestingly, even when using the full eQTLGen data at a sample size of ∼32K, ∼27% of GWAS hits are not in LD with any eQTL (Supplementary Fig. 28A).

We further investigated the properties of GWAS hits with colocalization (i.e., in LD with at least one blood eQTL) versus GWAS hits without colocalization (i.e., not in LD with any blood eQTL). Compared to GWAS hits with colocalization, GWAS hits without colocalization are near genes that are more selectively constrained (Supplementary Fig. 28B), and lie at longer distances to the TSS of their nearest gene (Supplementary Fig. 28C). These trends become more pronounced with increasing power of the eQTL assay indicating that with growing eQTL sample sizes, GWAS hits that remain undiscovered as eQTLs tend to be more TSS-distal and acting on more constrained genes.

These results are consistent with our modeling arguments above, the eQTL discovery trends discussed in the main text, and recent experiments by McAfee et al. [43], who evaluated the regulatory activity of schizophrenia GWAS variants in a massively parallel reporter assay (MPRA) performed in primary human neural progenitors: one, most MPRA-positive variants (i.e., GWAS hits with evidence for regulatory activity) do not overlap eQTLs discovered in adult or developing brain. Two, MPRA-positive variants with and without eQTL support are systematically different with respect to distance to TSSs and genic features of their target genes, in similar ways as described here.

We emphasize however, that although these results can be understood in light of our model, our explanations for lack of colocalizations of GWAS loci with eQTLs are complementary to other hypotheses, such as context-specificity of missing trait-related eQTLs [44]. For example, among the ∼27% GWAS hits for blood-related traits studied here that are not in LD with any yet discovered blood eQTL, some are plausibly eQTLs that are activated upon stimulations that are absent from eQTLGen. That said, the systematic differences between GWAS hits with and without eQTL support, as shown in Supplementary Fig. 28B,C and by McAfee et al., are not readily explained by other hypotheses proposed so far. Thus, as we do not and cannot rule out these alternative explanations, we believe that the key factors we described in this work significantly contribute to the colocalization problem.



Supplementary Fig. 28: **Properties of colocalized GWAS hits for blood or immune related traits with blood eQTLs.** *Properties of 10,980 unique GWAS hits across 14 blood or immune related traits with or without colocalization with eQTLs, defined as being in LD ($r^2 > 0.8$) with any eQTL discovered by GTEx in whole blood [15], or in eQTLGen data [17]. We divided eQTLGen eQTLs into 10 groups based on the deciles of association p-values.* ***A)*** *Fraction of GWAS SNPs (dark blue) that colocalize with eQTLs, compared to control SNPs (light blue) matched for MAF, LD and gene density (see Online Methods).* ***B)*** *Points show logistic regression coefficient for pLI for predicting genes linked with GWAS hits with versus without colocalization with eQTLs after adjusting for confounders. Results are plotted as regression coefficients ±2 standard errors. As proxy for GWAS target genes, we linked GWAS hits to their closest gene that is expressed in blood. See Supplementary Methods for details.* ***C)*** *Box plot of the distance of GWAS variants to the nearest TSS (of genes expressed in blood) by colocalization status. The lower and upper hinges represent the 25th and 75th percentiles, respectively. Whiskers extend up to 1.5 times the interquartile range from the minimum and maximum values. Data points beyond the whiskers are depicted individually, while values greater than 500 Kb are excluded from the plot for clarity. The number of colocalized GWAS hits (out of 10,980) per eQTL group are as follows: 2,974 for GTEx, and 2,019, 3,404, 4,421, 5,254, 5,939, 6,466, 6,961, 7,374, 7,707, and 8,018 for the eQTLGen groups shown.*

6 Supplementary methods

In this section we provide additional details relating to the analyses in this supplementary note, complementing the methods described within the note. All methods relating to the main text are described in the Online Methods.

SNP sets.

As we described in the Online Methods (see section *SNP selection*), 8,136,100 SNPs passed our quality control (QC) measures. For most data sets analyzed in this paper, e.g., GWAS hits, or any type of QTLs, our first QC step is extracting variants overlapping this set. We further processed this SNP set removing variants (i) in LD with putatively protein-altering variants, (ii) more than 1Mb away from autosomal protein-coding genes, (iii) in the MHC region, yielding 6,971,256 SNPs (Online Methods, section *SNP selection*). We focused on variants overlapping this set for most of our analyses of GWAS hits and eQTLs, and further sampled control SNPs per study from this set. In this supplementary note, for our analyses of exon QTLs and splicing QTLs, we removed filters (i) and (iii) above, resulting in 7,776,878 SNPs.

GWAS data.

UK Biobank. In the main text, we analyzed GWAS data for 44 complex traits from the UK Biobank (UKB). These traits were chosen through a trait selection pipeline as described in the Online Methods. We expanded the list of traits for analysis in this note to include 1,083 traits, removing the trait filtering criteria applied previously. This set includes the majority of traits for which GWAS data as well as pairwise genetic correlations were released by the Neale lab [1]. We further removed traits with no obvious biological relevance such as "Day-of-week questionnaire completion requested". We applied the same SNP selection quality control measures that we used for the main 44 traits (see Online Methods) for these 1,083 traits, resulting in 83,401 GWAS hits. We also divided the 44 traits used for our main analyses into 14 "blood or immune related" traits (12,157 GWAS hits), and 30 other "non-blood or immune related" traits (9,962 GWAS hits), based on GWAS variants enrichment in myeloid/erythroid or lymphoid specific open chromatin regions (see *Determination of blood or immune related traits* below).

GWAS ATLAS. We downloaded the list of LD clumped lead GWAS variants provided by the GWAS ATLAS [3] (file `gwasATLAS_v20191115_riskloci.txt.gz`, see URLs below). We overlapped this SNP set with 8,136,100 SNPs that passed our quality control measures (see section *SNP sets*). We focused on GWAS studies performed in European-descent individuals, labeled with "EUR" in the file `gwasATLAS_v20191115.txt`. The data set may include multiple GWAS results for each unique trait; for each trait we selected the GWAS study with the highest number of filtered lead GWAS hits. We then formed three categories from the traits that were retained by these steps: (1) all traits, (2) traits labeled with the term "disease" or "disorder", and (3) traits labeled with the term "disease" in the file `gwasATLAS_v20191115.txt`. For each of these trait groups we preformed two additional filtering steps to form three corresponding filtered sets: one, we conditioned on 558 GWAS studies analyzed by Watanabe et al., [3], and two, we pruned the trait list such that genetic correlation, ρ_g, was < 0.5 for all trait pairs in the final list using the same procedure as described in the Online Methods for the 44 complex traits in UKB, and using ρ_g values released by the GWAS ATLAS (file `gwasATLAS_v20191115_GC.txt`). Following these steps, for each of the 6 trait groups, we conditioned on 6,971,256 set of SNPs that we used in our analysis of GWAS data (see section

SNP sets). The final trait/SNP sets are as follows: (1) "all traits": 1,488 traits, 39,932 GWAS hits, (2) "diseases/disorders": 154 traits, 3,551 GWAS hits, (3) "diseases": 92 traits, 2,405 GWAS hits, (4) "independent traits": 173 traits, 7,531 GWAS hits, (5) "independent diseases/disorders": 40 traits, 1,233 GWAS hits, (6) "independent diseases": 23 traits, 821 GWAS hits.

We further divided the list of 173 "independent traits" into five non-overlapping non-disease categories. To this end, we first excluded all traits labeled with the terms "disease" or "disorder", and then grouped traits based on the "Domain" field provided by GWAS ATLAS in the file `gwasATLAS_v20191115.txt`. We grouped domains "Cognitive", "Neurological", and "Psychiatric" as "cognitive" (17 traits, 737 GWAS hits); domain "Reproduction" as "reproduction" (8 traits, 411 GWAS hits); domain "Immunological" as "immunological" (14 traits, 1,581 GWAS hits); domains "Skeletal" and "Body Structures" as "physical" (6 traits, 1,018 GWAS hits); domain "Metabolic" as "metabolic" (10 traits, 984 GWAS hits). We also relabeled traits "Educational attainment" and "Birth weight" as "cognitive" and "reproduction", respectively.

SNP-gene links. We analyzed GWAS gene assignments from two independent studies: (1) We downloaded SNP-gene links for GWAS hits for 113 complex traits predicted by the PoPS method developed by Weeks et al., [8] (see URLs). We conditioned on 18,332 protein-coding autosomal genes studied in this paper (see Online Methods, section *Gene selection*). For each trait-variant pair, we selected the top gene with the highest rank based on the PoPS score, resulting in 25,252 SNP-gene links. (2) We downloaded gene assignment provided by Gazal et al., [7] for fine-mapped GWAS hits for UKB traits (see URLs), based on their integration of earlier SNP-to-gene linking strategies into a combined score (cS2G). Following Gazal et al., we conditioned on trait-variant pairs with posterior inclusion probability (PIP) > 0.5. Also, similar to our procedure for PoPS genes, we conditioned on the list of 18,332 protein-coding autosomal genes, resulting in 6,655 SNP-gene links across 47 traits.

eQTL data.

eQTL Catalogue. We downloaded fine mapped credible sets for gene expression (ge) eQTLs in 105 studies processed by the eQTL catalogue [16] (`*.purity_filtered.txt.gz` files, see URLs). We mapped the genomic coordinates to the hg19 assembly using `LiftOver` [45], and extracted the overlap between eQTL variants and 8,136,100 SNPs that passed our quality control measures (see section *SNP sets*). For each study, we kept the eQTL SNP with the highest PIP in each credible set. We then overlapped these top eQTLs to the 6,971,256 SNPs that we used in our analysis of GWAS and eQTL data (see section *SNP sets*), as well as eQTLs for eGenes that were among the list of 18,332 protein-coding autosomal genes studied in this paper (see Online Methods, section *Gene selection*).

eQTLGen. We downloaded significant cis-eQTLs ascertained by the eQTLGen consortium [17] (see URLs). We processed these eQTLs similar to the GTEx eQTLs (see Online Methods). Specifically, we focused on eQTLs for eGenes that were among the list of 18,332 protein-coding autosomal genes studied in this paper (see Online Methods, section *Gene selection*). We first extracted eQTLs overlapping 8,136,100 SNPs that passed our quality control measures (see section *SNP sets*). Then, for each eGene, we performed LD-based clumping to ascertain lead eQTLs (as described in the Online Methods, section *SNP selection*), resulting in 249,929 eQTLs for 12,659 eGenes. These numbers were used for our discovery power analysis in section *Power considerations in GWAS and eQTL mapping*. For other analyses of eQTLGen eQTLs, we focused on 230,032 lead eQTLs that were among the 6,971,256 SNPs that we used in our analysis of GWAS and eQTL data (see section

SNP sets). In a subset of analyses, we further divided these 230,032 lead eQTLs into 10 groups based on deciles of association p-values.

GTEx. All methods relating to our analyses of GTEx data are described in the Online Methods. Only for our discovery power analysis in section *Power considerations in GWAS and eQTL mapping*, we reprocessed GTEx eQTLs for whole blood [15], matching our process for eQTLGen data for this purpose. To this end, we focused on eQTLs for eGenes that were among the list of 18,332 protein-coding autosomal genes (see Online Methods, section *Gene selection*). We extracted eQTLs overlapping 8,136,100 SNPs that passed our quality control measures (see section *SNP sets*). Then, for each eGene, we performed LD-based clumping to ascertain lead eQTLs (as described in the Online Methods, section *SNP selection*), resulting in 28,645 eQTLs for 7,953 eGenes.

Other QTL data.

Exon QTLs. We downloaded fine mapped credible sets for exon expression QTLs in 49 GTEx tissues processed by the eQTL catalogue [16] (`*.purity_filtered.txt.gz` files, see URLs). We processed these QTLs using a procedure similar to what we used for gene expression eQTLs from the eQTL catalogue described above. The only difference is that we kept lead QTLs overlapping the 7,776,878 SNPs (described in section *SNP sets*), and not the 6,971,256 SNPs used for gene expression eQTLs. This pipeline resulted in 1,088,880 exon QTLs (median 18,851 QTLs per tissue).

Splicing QTLs. We processed splicing QTLs (sQTLs) for 23 tissues analyzed by GTEx v8 [15], following the same procedures as with our analysis of GTEx eQTLs (see Online Methods, section *eQTL data*). We extracted sQTLs overlapping 8,136,100 SNPs that passed our quality control measures (see section *SNP sets*), focusing on sGenes (genes linked with sQTLs) that were among the list of 18,332 protein-coding autosomal genes studied in this paper (see Online Methods, section *Gene selection*). For each sGene, we performed LD-based clumping to ascertain lead sQTLs (as described in the Online Methods, section *SNP selection*). We kept lead sQTLs that were among the 7,776,878 SNPs described above (see section *SNP sets*). This pipeline resulted in 67,250 sQTLs (median 2,600 sQTLs per tissue).

Interaction eQTLs. We processed cell type interaction eQTL data for 43 tissue-cell type pairs provided by GTEx v8 [18], as described above for GTEx sQTLs, resulting in 530,819 ieQTLs (median 12,492 ieQTLs per tissue-cell type pair).

Methylation QTLs. We analyzed DNA methylation QTLs (meQTLs) in peripheral blood ascertained by Hawe et al., [19] (see URLs). We focused on 336,732 tested CpG sites (file `cpgs2include.RData`), 64,478 of which were linked to at least one meQTL in a conditional analysis to identify independent meQTLs linked with CpG sites (Supplementary Table 8 in Hawe et al.). As a proxy for the CpG sites (and by association meQTLs) target genes, we linked all sites to their closest gene from the 18,332 protein-coding autosomal genes studied in this paper (see Online Methods, section *Gene selection*). We then compared the properties of genes linked with 64,478 CpG sites with meQTLs and 272,254 without meQTLs (see *Gene comparison analysis using logistic regression* below).

Gene expression levels.

For some analyses, we used the data for tissue specific gene expression levels in GTEx v8 tissues. To this end, we extracted median TPM values per gene across GTEx participants as provided by GTEx (file `GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz`).

Determination of blood or immune related traits.

For our analyses in Supplementary Figs. 1 and 28, we divided the 44 complex traits analyzed in the main text into two groups: "blood or immune related" and "not blood or immune related". To this end, we downloaded DNase I Hypersensitive Sites (DHSs) maps created as part of the ENCODE 3 project [2] (see URLs), focusing on regions determined to be specific to myeloid/erythroid or lymphoid components by Meuleman et al., [2]. Similar to other regulatory regions we analyzed from ENCODE 3, we mapped these DHS regions to the hg19 assembly using `LiftOver` [45]. We then computed enrichment of GWAS hits in these regions for the 44 traits relative to their corresponding control SNPs (see Online Methods for the definition of GWAS control SNPs and enrichment computation), and labeled the top traits with significant enrichment after Bonferroni correction (14 traits with enrichment Z score > 3.45, corresponding to p-value $< 0.05/88$) as "blood or immune related".

Gene comparison analysis using logistic regression.

For a number of analyses investigating the genic features linked with a set of SNPs (e.g., GWAS hits, eQTLs, other QTLs, etc.), we used a logistic regression framework similar to what we described in the Online Methods. Each SNP set is linked to genes through different strategies as described for each analysis in the text, e.g., eQTLs to eGenes, or random SNPs to genes with the nearest TSS.

In this regression framework, we first constructed indicator variables for a SNP set of interest (labeled 1s) versus a set of SNPs chosen for comparison or as control (labeled 0s). These comparisons included: GWAS hits vs. random SNPs, eQTLs vs. random SNPs, GWAS hits with vs. without eQTL colocalization, and other QTLs vs. corresponding control sets for each QTL type (see section *Other QTLs*). We chose random/control SNP sets based on the analysis or QTL type as well as computational efficiency. For GWAS hits and eQTLs, the set included 100K SNPs sampled from the set of 6,971,256 SNPs (see section *SNP sets*). For ieQTLs we sampled 20K SNPs at random from the same set. For exon and splicing QTLs we randomly sampled 100K and 20K, respectively, from the set of 7,776,879 SNPs (see *SNP sets* above).

We then used logistic regression models to predict these indicator variables using the genic features of interest (e.g., pLI score, TF status, etc.), one feature at a time (with the exception of the two enhancer features that were used in a joint model). See Online Methods, section *Gene annotations* for details on features. Genic feature values were normalized. For connectedness in gene co-expression or protein-protein interaction (PPI) networks we used as predictors the rank of genes based on the deciles of connectedness scores (described in Online Methods, section *Gene annotations*). We also defined a gene annotation "GO terms count" as the total number of broady unrelated Gene Ontology (GO) terms a given gene contributes to (see Online Methods, section *Selection of broadly unrelated GO annotations*).

We included the following covariates in the regression models: MAF, LD score, gene density, absolute distance to nearest TSS, total gene length, total length of gene coding sequence, as well as dummy variables for 20 quantiles of MAF, LD score, gene density, and absolute distance to nearest TSS. See Online Methods, sections *Gene annotations* and *SNP annotations* for details. In a subset of regression models for GTEx and eQTLGen eQTLs (Supplementary Figs. 9 and 12B), as described in the note, we adjusted for tissue specific expression levels. To this end, we also included TPM values in the corresponding tissues in GTEx (see *Gene expression levels* above), as well as dummy variables for 20 quantiles of the TPM values as covariates. For the comparison of CpG sites with

vs. without meQTLs (Supplementary Fig. 13D), we did not include MAF and LD as covariates (considering that CpG sites may not be polymorphic, and SNP features may not be defined), and also re-computed gene density values around the CpG sites as the number of protein-coding genes within a 1Mb window.

For the analysis of different SNP-to-gene linking strategies in GWAS (Supplementary Fig. 4), we performed the regression on GWAS gene sets vs. other non-GWAS genes, instead of GWAS SNP sets vs. random SNPs, mainly because the considered SNP-to-gene linking methods cannot be applied to random SNPs (unlike the closest gene approach we used in our main analyses). To this end, for each trait, we predicted GWAS genes (labeled 1s) versus other non-GWAS genes among the set of 18,332 protein-coding autosomal genes (labeled 0s), including total gene length, total length of gene coding sequence, and gene density (re-computed as the number of protein-coding genes within 1Mb windows around TSSs) as covariates.

All regression analyses were performed using the `glm` function in `R`. We report the regression coefficients and standard errors, or the corresponding Z values as output by `glm`, unless stated otherwise.

Colocalization analysis.

In our analysis for Supplementary Fig. 28, we investigated the colocalization of GWAS hits for 14 blood or immune-related traits described above, with blood eQTLs. To this end, we focused on 10,980 unique GWAS hits across these 14 traits. As a proxy for colocalization, we investigated whether these GWAS hits, or any of the SNPs in LD with them ($r^2 > 0.8$ among the 13.7 million quality controlled SNPs in UK Biobank, see Online Methods, section *SNP selection*), are among the eQTLs detected in GTEx whole blood, or the eQTLGen study. To be conservative, we included all eQTLs ascertained by these studies, i.e., without any of the processing done for our analyses of these eQTLs described above (e.g., no LD-clumping): 1,323,859 and 7,455,305 total eQTLs in GTEx whole blood and eQTLGen, respectively, were included. As a proxy for target genes, we linked the GWAS hits to their closest gene (with the closest TSS) that is expressed in blood (from genes with median TPM> 0 in GTEx whole blood). We then investigated the genic properties of GWAS hits with and without colocalization with eQTLs using our logistic regression framework described above.

Simulations.

We performed simulations under our model for variant discovery in GWAS and eQTL assays exploring various scenarios and parameter choices (sections *Robustness of the model* and *Model extensions*). Here we provide additional details on a few components of these simulations. All other procedures are described in Online Methods (section *Modeling variant discovery*) and within the supplementary note.

Baseline simulation parameters. In Supplementary Fig. 14 we present our key predictions under our baseline assumptions, serving as a point of comparison for simulations under other scenarios and different modeling parameters. We considered 10 million independent SNPs, and sampled 10 million β and γ values from independent standard Normal distributions. To mimic variant discovery, in the evolutionary neutral scenario, we ranked variants based on β^2 for eQTL mapping, and based on $\beta^2\gamma^2$ values for GWAS mapping. When the phenotype is under natural selection, we ranked variants based on $V_p\beta^2$ for eQTL mapping, and based on $V_p\beta^2\gamma^2$ values, where $V_p = 2p(1-p)$ is the

scaling factor accounting for the reduction in allele frequency and thus phenotypic variance due to selection. Motivated by a flattening model of selection [27, 28], for each SNP with effects (β, γ), we set $V_p = \kappa(1 - e^{-\beta^2\gamma^2/\kappa})\beta^{-2}\gamma^{-2}$. For baseline simulations, we set $\kappa = 2.986$. See Online Methods (section *Modeling variant discovery*) for more details.

Correlation between β^2 and γ^2. In a subset of simulations we explored how our results change with covariance between β^2 and γ^2. To this end, we first generated random variables $(u_\beta, u_\gamma) \sim N(\mathbf{0}, \mathbf{\Sigma})$, setting $\Sigma_{i,j} = \rho$ for non-diagonal and 1 for diagonal elements, where ρ is our tuning parameter to induce correlation between β^2 and γ^2. We then drew β and γ values from independent standard Normal distributions (our baseline parameters), and then rearranged and paired the vectors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to match the ranks of u_β and u_γ, respectively. We then computed the correlation induced between β^2 and γ^2 and reported that on the plots shown.

Strength of selection. As described in the Online Methods, we model selection to have a flattening effect on variants' contribution to phenotypic variance, using an asymptotic exponential form to describe the relationship $E[V_p\beta^2\gamma^2|\beta, \gamma] \propto \kappa(1 - e^{-\beta^2\gamma^2/\kappa})$. See Online Methods (section *Modeling variant discovery*) for more details. In this formulation, the parameter κ determines the strength of flattening or selection. That said, the expected degree of reduction in phenotypic variance across all variants compared to the neutral scenario depends also on the distributions of β and γ. In a subset of simulations we vary distributions of β and γ. With the exception of simulations where we explore the effect of varying the strength of selection (Supplementary Figs. 20 and 21), to keep the net effect of selection fixed, we tuned the κ parameter such that $E[V_p]$ is reduced by $\sim 10\%$ compared to the neutral scenario.

The α model. We also explored our model predictions under the model, termed the α model, which describes the relationship between allele frequency and effect size as $E[\beta^2\gamma^2|p] \propto [2p(1-p)]^\alpha$ [4, 31, 32]. In our simulations, we determine variant discovery based on the reveres expectation $E[p|\beta^2\gamma^2]$. To incorporate the α model in our simulations, we approximated $E[p|\beta^2\gamma^2]$ as follows: we first sampled β and γ values from standard Normal distributions, and then multiplied those by a factor of $[2p(1-p)]^{\alpha/4}$, where p values are drawn from a truncated exponential distribution with mean 0.05 within the range of $[0.001, 0.5]$. We then numerically solved for $E[p|\beta^2\gamma^2]$ using a piecewise linear regression to predict p from $\beta^2\gamma^2$ using the function `segmented` in `R` [46].

URLs.

GWAS ATLAS: `https://atlas.ctglab.nl`

PoPS: `https://www.finucanelab.org/data`

cS2G: `https://alkesgroup.broadinstitute.org/cS2G/finemapping_cS2G_UKBB`

eQTL catalogue: `https://www.ebi.ac.uk/eqtl/`

eQTLGen consortium: `https://www.eqtlgen.org`

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

Methylation QTLs by Hawe et al.,: `https://zenodo.org/record/5196216#.Y6Y6Mi-B28V`

ENCODE DHS regions: `https://www.encodeproject.org/annotations/ENCSR857UZV/`

Supplementary data.

Data generated by or processed for the analyses in this note can be found on Zenodo with the DOI 10.5281/zenodo.6618073. Also see Online Methods, section *Data availability*.

References

[1] Neale lab - UK Biobank; 2018. [Online; accessed 17-June-2020]. http://www.nealelab.is/uk-biobank/.

[2] Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature. 2020;584(7820):244-51.

[3] Watanabe K, Stringer S, Frei O, Mirkov MU, Leeuw Cd, Polderman TJC, et al. A global overview of pleiotropy and genetic architecture in complex traits. Nature Genetics. 2019;51(9):1339-48.

[4] Zeng J, Xue A, Jiang L, Lloyd-Jones LR, Wu Y, Wang H, et al. Widespread signatures of natural selection across human complex traits and functional genomic categories. Nature Communications. 2021;12(1):1164.

[5] Simons YB, Mostafavi H, Smith CJ, Pritchard JK, Sella G. Simple scaling laws control the genetic architectures of human complex traits. bioRxiv. 2022:2022-10.

[6] Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nature genetics. 2021;53(11):1527-33.

[7] Gazal S, Weissbrod O, Hormozdiari F, Dey KK, Nasser J, Jagadeesh KA, et al. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. Nature Genetics. 2022;54(6):827-36.

[8] Weeks EM, Ulirsch JC, Cheng NY, Trippe BL, Fine RS, Miao J, et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Nature Genetics. 2023:1-10.

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

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

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

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

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

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

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

[16] Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nature Genetics. 2021;53(9):1290-9.

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

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

[19] Hawe JS, Wilson R, Schmid KT, Zhou L, Lakshmanan LN, Lehne BC, et al. Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. Nature Genetics. 2022;54(1):18-29.

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

[21] O'Connor LJ. The distribution of common-variant effect sizes. Nature Genetics. 2021;53(8):1243-9.

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

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

[24] Ouwens KG, Jansen R, Nivard MG, Dongen Jv, Frieser MJ, Hottenga JJ, et al. A characterization of cis- and trans-heritability of RNA-Seq-based gene expression. European Journal of Human Genetics. 2020;28(2):253-63.

[25] Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nature Genetics. 2018;50(10):1412-25.

[26] Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022;610(7933):704-12.

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

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

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

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

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

[32] Schoech AP, Jordan DM, Loh PR, Gazal S, O'Connor LJ, Balick DJ, et al. Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. Nature Communications. 2019;10(1):790.

[33] Yair S, Coop G. Population differentiation of polygenic score predictions under stabilizing selection. Philosophical Transactions of the Royal Society B. 2022;377(1852):20200416.

[34] Zhang Y, Quick C, Yu K, Barbeira A, Consortium G, Luca F, et al. PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. Genome biology. 2020;21:1-26.

[35] Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nature Genetics. 2016;48(7):709-17.

[36] Weiner DJ, Nadig A, Jagadeesh KA, Dey KK, Neale BM, Robinson EB, et al. Polygenic architecture of rare coding variation across 394,783 exomes. Nature. 2023;614(7948):492-9.

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

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

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

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

[41] Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature Genetics. 2018;50(11):1505-13.

[42] Hukku A, Sampson MG, Luca F, Pique-Regi R, Wen X. Analyzing and reconciling colocalization and transcriptome-wide association studies from the perspective of inferential reproducibility. The American Journal of Human Genetics. 2022;109(5):825-37.

[43] McAfee JC, Lee S, Lee J, Bell JL, Krupa O, Davis J, et al. Systematic investigation of allelic regulatory activity of schizophrenia-associated common variants. medRxiv. 2022:2022-09.

[44] Umans BD, Battle A, Gilad Y. Where are the disease-associated eQTLs? Trends in Genetics. 2021;37(2):109-24.

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

[46] Muggeo VMR. segmented: an R Package to Fit Regression Models with Broken-Line Relationships. R News. 2008;8(1):20-5. Available from: `https://cran.r-project.org/doc/Rnews/`.