CRISPRi perturbation screens and eQTLs provide complementary and distinct insights into GWAS target genes

Samuel Ghatan[1], Jasper Panten[2], Winona Oliveros[1,3], Neville E Sanjana[1,4], John A Morris[1,5], Tuuli Lappalainen[1,2]

1. New York Genome Center, New York, USA
2. Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Solna, Sweden
3. Department of Life Sciences, Barcelona Supercomputing Center (BCN-CNS), Barcelona, Catalonia, 08034, Spain
4. Department of Biology, New York University, New York, NY, USA
5. Donnelly Centre for Cellular and Biomolecular Research, Department of Molecular Genetics, University of Toronto, Toronto, Canada

Abstract

Most genetic variants associated with human traits and diseases lie in noncoding regions of the genome [1], and a key challenge is determining which genes they affect [2,3]. A common approach has been to leverage associations between natural genetic variation and gene expression to identify expression quantitative trait loci (eQTLs) in the population [4,5]. At the same time, a newer method uses pooled CRISPR interference (CRISPRi) perturbations of noncoding loci with single-cell transcriptome sequencing [6,7]. Here, we systematically compared the results from these approaches across hundreds of genomic regions associated with blood cell traits. We find that while the two approaches often identify the same target genes, they also capture distinct features of gene regulation. CRISPRi tends to detect genes that are physically closer to regulatory variants and more constrained, whereas eQTL studies are sensitive to detecting multiple, often distal, genes. By comparing these discoveries to a gold-standard set of genes linked to blood traits, we demonstrate that the two approaches provide highly complementary insights with distinct strengths and caveats. Our results offer guidance for improved design of CRISPRi and eQTL studies and highlight their potential as a powerful toolkit for interpreting disease-associated loci.

Introduction

Genome-wide association studies (GWAS) have discovered hundreds of thousands of loci that associate with complex traits and diseases, but elucidating the functional mechanism of these variants remains challenging [2,3]. Notably, ~90% of GWAS-implicated variants lie in noncoding regions [1], making identifying causal genes difficult. Expression quantitative trait loci (eQTL) studies were the first approach for associating variants with gene regulation genome-wide. Nevertheless, despite their popularity, an emerging body of work highlights the limitations of relying exclusively on eQTL data to link GWAS variants to their putative target genes [8,9]. We still lack eQTL associations for many GWAS variants [4,5]. When an eQTL does exist for a GWAS variant, it exhibits a high false positive rate [4,8], and its reliability in identifying the true causal gene is relatively poor compared to simple metrics such as assigning the closest gene as the causal gene [10,11]. Recently, single-cell pooled CRISPR interference (CRISPRi) screens have emerged as a potentially powerful strategy for pinpointing causal genes at hundreds of GWAS loci [6,7]. However, CRISPRi inhibits entire enhancers [12], producing effects that may not be analogous to single-nucleotide variation, and the studies are typically conducted in non-primary human cell lines, raising questions about the transferability of biological effects into human cells and disease associations. The extent to which eQTL and CRISPRi approaches converge—or diverge—in identifying GWAS gene targets remains poorly understood.

GWAS target gene discovery

We analyzed and fine-mapped GWAS loci from 44 blood trait studies from the UK Biobank and the Blood Cell Consortium (BCX) (Supplementary Table 1). To identify *cis*-eQTL target genes (eGenes) for GWAS variants, colocalization was performed with fine-mapped eQTL variants from 50 studies based on RNA-sequencing of whole blood, bulk RNA-seq of isolated blood cell types, and single cell RNA-sequencing (scRNA-seq) of peripheral blood mononuclear cells (PBMCs) (Supplementary Table 2) [5,13–15]. To identify CRISPRi cis target genes (cGenes), we re-analyzed data from 14 enhancer-targeting CRISPRi studies in K562 erythroleukemia cells (Supplementary Table 3) [6,7,16] and identified fine-mapped GWAS variants intersecting gRNAs (Fig. 1a, mean PIP = 0.20).

Overall, we identified 882 GWAS cis-regulatory elements (CREs) that were targeted by a CRISPRi gRNA. Of these, 145 CREs had a cGene (FDR < 0.10, Extended Data Fig. 1a), and 515 had a colocalized eGene (PP.H4 > 0.5, Extended Data Fig. 1b) within 1 Mb (Supplementary Tables 4 & 5). In 97 CREs, both eGenes (n = 266) and cGenes (n = 122) were

identified, with 69 overlapping CRE–gene target pairs shared between the two approaches (Fig. 1b).

CRISPRi and eQTL power shape the discovery of GWAS target genes

We first investigated how these discoveries were affected by various gene, CRE, and study properties, statistical power, and context-specific gene expression by analyzing patterns in the empirical data, complemented by simulations of both eQTL and CRISPRi discovery in each locus (Methods, Supplementary Tables 6 & 7). The largest numbers of eGenes were derived from bulk RNA-sequencing of isolated cell types, with only 3 of the overlapping genes uniquely identified in scRNA-seq eQTLs (Fig. 1c, Extended Data Fig. 1c). This points to limited power in current single-cell eQTL data in identifying GWAS target genes, particularly for less abundant cell types. This suggests that when cell type isolation is technically feasible, bulk cell type eQTL studies hold advantages over sc-eQTL and bulk tissue studies in providing both good statistical power and cell type specificity. GWAS variants with eGenes that intersected cGenes had a higher colocalization probability (PP.H4) ($P = 1.4 \times 10^{-6}$, Fig. 1d) than eGenes that did not, among colocalized eQTLs (PP.H4 > 0.5), even after controlling for eQTL and GWAS p-values (Extended Data Fig. 2a).

cGenes were identified for GWAS variants at lower minor allele frequencies (MAFs) than eGenes ($P = 2.9 \times 10^{-2}$, Fig. 1e), a pattern recapitulated in simulations (Extended Data Fig. 2b). This is expected, as eQTL discovery power, similarly to other variant association studies, declines for low-frequency variants. In contrast, CRISPRi-based mapping is unaffected by allele frequency. Conversely, eGenes have significantly lower expression than cGenes in K562 cells ($P = 2.2 \times 10^{-16}$, Fig. 1f) and GTEx whole blood ($P = 1.5 \times 10^{-3}$, Fig. 1g), with simulations further indicating that CRISPRi discovery power is more greatly affected by lower expression than eQTLs, likely deriving from sparsity of scRNA-seq data (Extended Data Fig. 2c). We then compared the simulated power of eGene and cGene discovery for intersecting and non-intersecting gene targets. eGene discovery power is similar for genes that intersect cGenes and for those that do not ($P = 0.70$, Fig. 1h), while cGene discovery power is significantly higher for intersecting target genes compared to non-intersecting ones ($P = 7.4 \times 10^{-4}$, Fig. 1i). This suggests that low eQTL discovery power is not driving the missing overlaps, while better-powered CRISPRi targets align better with eQTL results. Some of the differences are due to gene expression: in the 97 CREs with both cGenes and eGenes, 98% of cGenes were

expressed in eQTL datasets, whereas only 76% of eGenes were expressed in K562 cells, underscoring the limitations of inferring causal genes from a single cell line. Altogether, these analyses highlight both fundamental and study design features that affect the power of eQTL and CRISPRi studies and their impact on GWAS gene discovery.

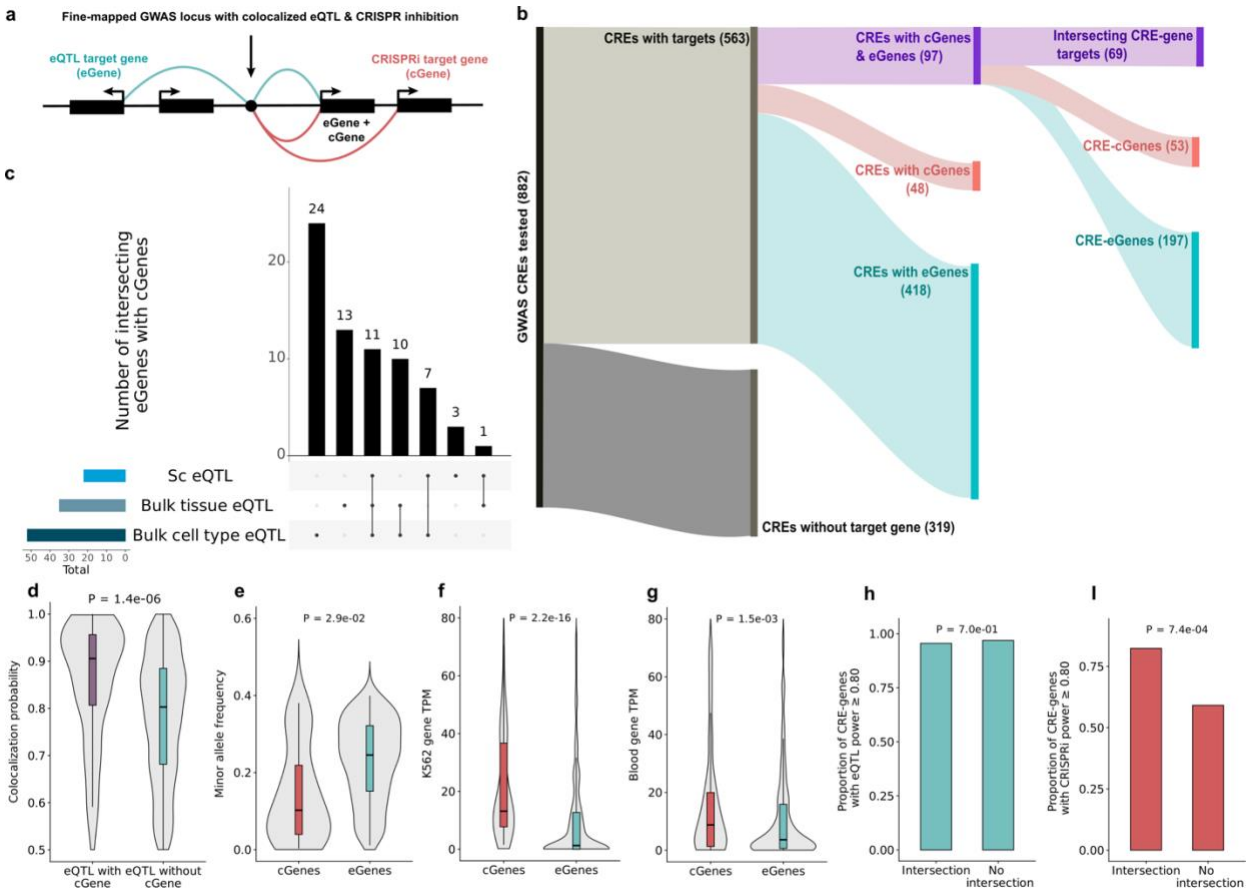

Figure 1. Discovery of eQTL and CRISPRi target genes for GWAS loci. a*, Illustration of the comparison of gene targets for GWAS variants identified through eQTL and CRISPRi approaches.* ***b****, The number of CREs and genes discovered by eQTL and CRISPRi.* ***c****, The number of eGenes that intersect with cGenes across different eQTL study types.* ***d****, Comparison of GWAS–eQTL colocalization probabilities between eQTLs whose eGenes intersect a cGene and those that do not.* ***e****, Minor allele frequencies (MAFs) for GWAS variants uniquely mapped to eGenes or cGenes.* ***f-g****, Gene expression (transcripts per million; TPM) for cGenes and eGenes in K562 erythroleukemia cells (f) and in GTEx whole blood (g).* ***h-i****, Simulated eGene* ***(h)*** *and cGene* ***(i)*** *discovery power between CRE-gene targets that intersect between CRISPRi and eQTL methods versus those that do not, at GWAS-associated CREs with both target types.*

Biological properties of eQTL and CRISPRi discoveries

Next, we examined the biological properties of the discovered eGenes and cGenes. While most cGenes (81%) and eGenes (68%) lie within 100 kb of the GWAS CRE (Fig. 2a), cGenes are more proximal: 41% are within 10 kb, compared to only 19% of eGenes. As a baseline, 21% of CREs without a cGene or eGene have the closest gene within 10 kb (Fig. 2b). The proximity of CRE–cGene pairs is notable given that CRISPRi screens specifically targeted regions more than 1 kb from transcription start sites, and repression by KRAB/MeCP2 is generally limited to local chromatin. Thus, we do not expect CRISPRi to cause gene repression at long distances. eQTL mapping more often identifies genes beyond the closest gene (97% vs. 37% for cGene-linked CREs, $P = 4.6 \times 10^{-08}$, Fig. 2c). Additionally, CRISPRi identifies a smaller number of genes, with 82% of cGene-linked CREs and 38% of eGene-linked CREs being associated with a single gene ($P = 5.9 \times 10^{-21}$, Fig. 2d). This difference partly reflects the broader range of eQTL studies and cell types analyzed, with a median of two distinct eQTL studies identifying unique genes per CRE (Extended Data Fig. 2d). In contrast, the CRISPRi data comes from a single cell line. These results indicate that eQTL mapping captures weaker regulatory effects that span larger genomic distances, including multiple targets of the same gene. This is likely at least partially due to the higher power and diversity of eQTL studies, as more distal CRE-promoter interactions are, on average, weaker [17] and thus harder to detect, especially with CRISPRi. This reasoning is supported by the fact that genes closer to CREs have a lower log-fold change ($P = 7.5 \times 10^{-05}$, Extended Data Fig. 2e). Larger future CRISPRi studies with more diverse cell types might bring the two approaches closer to alignment. Ultimately, it depends on the study questions whether detecting even small effects on distal target genes is necessary for biological insights and downstream applications.

It has been proposed that, due to natural selection, eQTLs that rely on natural genetic variation are biased against selectively constrained effects[18]. Indeed, we observed that cGenes were more haplosensitive, measured by Episcore predictions ($P = 5.7 \times 10^{-3}$, Fig. 2e) and pHaplo scores ($P = 0.03$), as well as a suggestive signal of higher loss-of-function intolerance, with 29% of cGenes and 21% of eGenes having pLI ≥ 0.9 (P = 0.06, Fig. 2f). This suggests that CRISPRi is better at identifying genes that do not tolerate regulatory variation. Furthermore, cGenes had a higher number of associated enhancers ($P = 2.0 \times 10^{-4}$, Fig. 2g) and super-enhancers (P =

0.02), suggesting a more complex regulatory architecture and tighter regulatory control than eGenes. These results highlight the potential of experimental perturbations such as CRISPRi to discover genes under selective and regulatory control, which are more likely to be relevant for diseases and traits [19]. However, it remains unclear how often highly constrained genes undergo more buffering [20] to the extent that their expression changes are harder to detect even after CRISPR perturbations.

From ATAC-seq data spanning 18 blood cell types (Supplementary Methods), we observed that CREs with both eQTL and CRISPRi had more ubiquitous open chromatin, and cGenes tend to reside in more broadly shared open chromatin than eGenes (Fig. 2h; Kruskal-Wallis rank sum test $P = 0.03$). This suggests that cell type differences between primary cell types in eQTL studies and the K562 line contribute to the sharing of detected regulatory effects, and that experiments in K562s tend to pick up regulation that is shared between many primary cell types, potentially due to K562 being a progenitor cell type. eQTL studies identified target genes for 12% of GWAS CREs that were not accessible in any cell type tested (excluding K562). In comparison, only 5% of CREs with CRISPRi target genes were not annotated as open chromatin other than K562 ($P = 0.02$), supporting the prior notion that CRISPRi-based mapping is particularly sensitive to classical biochemical enhancer signatures [6,7]. We then examined the relationship between CRE-gene targets and chromatin interaction data from Hi-C in K562 cells to assess how physical chromatin contacts align with each approach. Hi-C data in K562 cells revealed that enhancer-gene targets were more likely to overlap with cGenes than eGenes ($P = 1.8 \times 10^{-5}$, Extended Data Fig. 2f), although 53% of the GWAS CREs tested had no annotated Hi-C contacts.

Shared cis-gene anchors reveal variable *trans*-regulatory landscapes across methods

We next examined how CRISPRi-identified *trans*-regulatory genes (*trans*-cGenes) compared with eQTL-based *trans* genes (*trans*-eGenes) for the same GWAS CREs acting through a shared *cis*-gene. We detected significant *trans*-cGene effects for 16 GWAS CREs (FDR < 0.10; 72–3,870 trans-cGenes per CRE; Extended Data Fig. 3a, Supplementary Tables 8 & 9). At two of these loci (*GFI1B* and *HHEX*), trans-effects were detected by both Morris et al. and Gasperini et al.[6,7], with high correlation in effect sizes (*GFI1B*: $r = 0.88$, $P < 2 \times 10^{-16}$, $n = 921$; *HHEX*: $r = 0.92$, $P = 9.6 \times 10^{-8}$, $n = 18$, Extended Data Figs. 3b & c). To identify *trans*-eGenes for these loci, we analyzed *trans*-eQTLs (FDR < 0.05) from single-cell PBMC data from OneK1K [15] and a large meta-analysis of lymphoblastoid cell lines (LCLs) (n = 3,892) [21]. We discovered *trans*-

eGenes for 11 out of 16 CREs (Supplementary Table 10), of which 6 had overlapping *trans*-genes (Extended Data Fig. 3d). The correlation between trans-eGene and trans-cGene effects was modest for most loci and often variable between cell types (r = -0.28–0.41, Fig. 2i), and significant only for the CRE regulating *MRPL4* in LCLs (r = 0.32, FDR = 9.9×10^{-3}, Supplementary Table 11). Additionally, the overlap of trans-cGenes and trans-eGenes was not significant for any CREs or cell types (Extended Data Fig. 3e, Supplementary Table 11). These data indicate the potential for robust detection of trans-regulatory networks with both methods, but also the need for further work to fully understand the drivers of shared and distinct patterns across a larger number of loci.

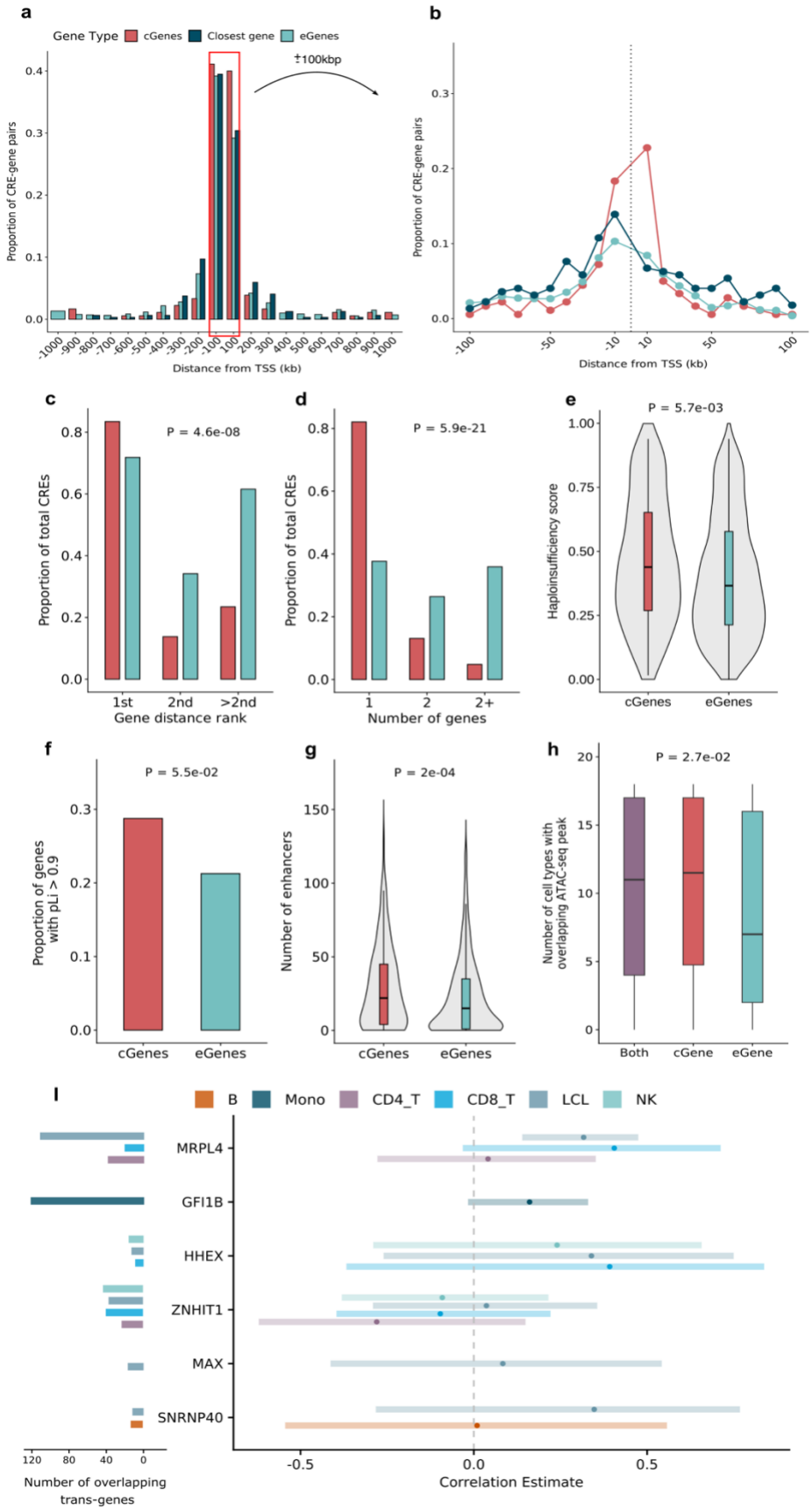
a
Gene Type cGenes Closest gene eGenes
±100kbp
Proportion of CRE-gene pairs
Distance from TSS (kb)
b
Proportion of CRE-gene pairs
Distance from TSS (kb)
c
P = 4.6e-08
Proportion of total CREs
1st
2nd
>2nd
Gene distance rank
d
P = 5.9e-21
Proportion of total CREs
1
2
2+
Number of genes
e
P = 5.7e-03
Haploinsufficiency score
cGenes
eGenes
f
P = 5.5e-02
Proportion of genes with pLi > 0.9
cGenes
eGenes
g
P = 2e-04
Number of enhancers
cGenes
eGenes
h
P = 2.7e-02
Number of cell types with overlapping ATAC-seq peak
Both
cGene
eGene
I
B Mono CD4_T CD8_T LCL NK
MRPL4
GFI1B
HHEX
ZNHIT1
MAX
SNRNP40
Number of overlapping trans-genes
Correlation Estimate

***Figure 2. Biological properties of eQTL and CRISPRi discoveries**, **a,** The distance between GWAS-associated CREs and target gene TSS, comparing cGenes, eGenes, and the closest gene to GWAS CREs where no significant association was identified, with **b**, showing the region within 100 kb of TSS. **c,** The proportion of GWAS-associated CREs linked to their closest, second closest, or further away genes. CREs may be counted in multiple categories if they are associated with multiple genes. **d**, The proportion of GWAS-associated CREs linked to one, two, or more genes. **e**, Haploinsufficiency (predicted by Episcore) for cGenes and eGenes. **f**, The proportion of cGenes and eGenes with loss-of-function intolerance metric pLI > 0.9. **g**, The number of enhancers for cGenes and eGenes. **h**, Number of blood-cell types in which a CRE has open chromatin for CREs with both eGenes and cGenes, only cGenes, and only eGenes. **i**, Trans-gene analysis, with the y-axis indicating the cis-genes through which trans-effects occur. The left-hand side of the plot shows the number of overlapping trans eGenes and cGenes, and the correlation and its confidence interval of the effect sizes are shown on the right-hand side.*

Detection of gold-standard GWAS genes by eQTL and CRISPRi

To evaluate the ability of each method to identify genes that are likely to be true causal genes for blood trait phenotypes, we constructed a “gold-standard' gene list of 704 genes. This included genes identified through loss-of-function (LoF) burden tests for 29 blood cell traits in the UK Biobank[22] and a curated set of genes linked to Mendelian blood disorders (Methods). We then filtered this list to include only protein-coding genes located within 1 Mb of our 882 GWAS-associated CREs, resulting in 421 genes (Supplementary Table 12). eQTL mapping identified 53 of these, and 23 were discovered as cGenes, corresponding to 11% (53/471) of all eGenes and 22% (23/107) of all cGenes in these loci (Fig. 3a & b). Only 9 genes were detected by both approaches. While the gold-standard gene discovery is significantly enriched over the null of a random set of protein-coding genes for both cGenes (OR: 14.2, $P = 3.2 \times 10^{-6}$) and eGenes (OR: 5.3, $P = 1.9 \times 10^{-8}$), the enrichment is higher for cGenes (OR: 2.2, $P = 6.9 \times 10^{-3}$). The enrichment for Hi-C gene targets was not significantly different from that of eGenes (Fig. 3b, $P = 0.57$). The higher enrichment of cGenes over eGenes is largely due to the high rate of eGenes overall: at 72% and 33% of the gold-standard gene loci, other eGenes or cGenes were detected, respectively, providing potentially misleading gene targets. When a cGene overlapped a gold-standard gene, it was the closest gene to the GWAS CRE in 74% of cases, compared to 58% for eGenes, indicating the further potential of eQTLs to identify more distal target genes that are more difficult to identify as putative targets of GWAS loci. The patterns of gene target discovery are highly variable between the loci (Fig. 3d, Extended Data Figs. 4-10).

Next, we analyzed loci at which the gold-standard gene was not identified to understand why it was missed. In eQTL analysis, lack of power is an infrequent explanation, as only 8% of the loci had simulated statistical power <0.8. Again, other eGenes were detected in 70% of the loci, which may lead to a misleading inference. Conversely, in CRISPRi analysis, as much as 56% of gold-standard genes were either underpowered in CRISPRi study simulations (27%) or not expressed in K562 cells (29%), and in 33% of the loci, a different cGene was identified (Fig. 3c). We note that both eQTL and CRISPRi power calculations assume uniform effect sizes across CRE–gene pairs; however, distal or more constrained genes likely exhibit weaker regulatory effects, leading to overestimation of power for these targets.

These results indicate substantial and distinct shortcomings in both approaches in the identification of causal genes for complex traits. Assuming gold-standard genes represent true causal targets within a locus, eQTL-based approaches demonstrated higher sensitivity (0.08 vs. 0.03) but lower specificity (0.95 vs. 0.99) compared to CRISPRi (Supplementary Methods). The respective importance of specificity or sensitivity depends on the downstream applications, and these metrics may change with more powerful future CRISPRi studies. However, the largely non-overlapping subsets of gold-standard genes identified by the two methods point to their orthogonal nature in discovering candidate causal genes for GWAS loci. Lastly, we assessed how many *trans*-genes identified by each method overlapped our genome-wide catalogue of 704 gold-standard genes. For every CRE and across all cell types examined, neither the *trans*-cGene set nor the *trans*-eGene set showed a significant enrichment for gold-standard genes (Supplementary Table 13). These findings suggest that *trans*-regulatory targets identified by both CRISPRi and eQTL approaches differ from genes identified by rare variant association studies and diseases, underscoring the complexity of interpreting *trans* associations in variant-to-function studies.

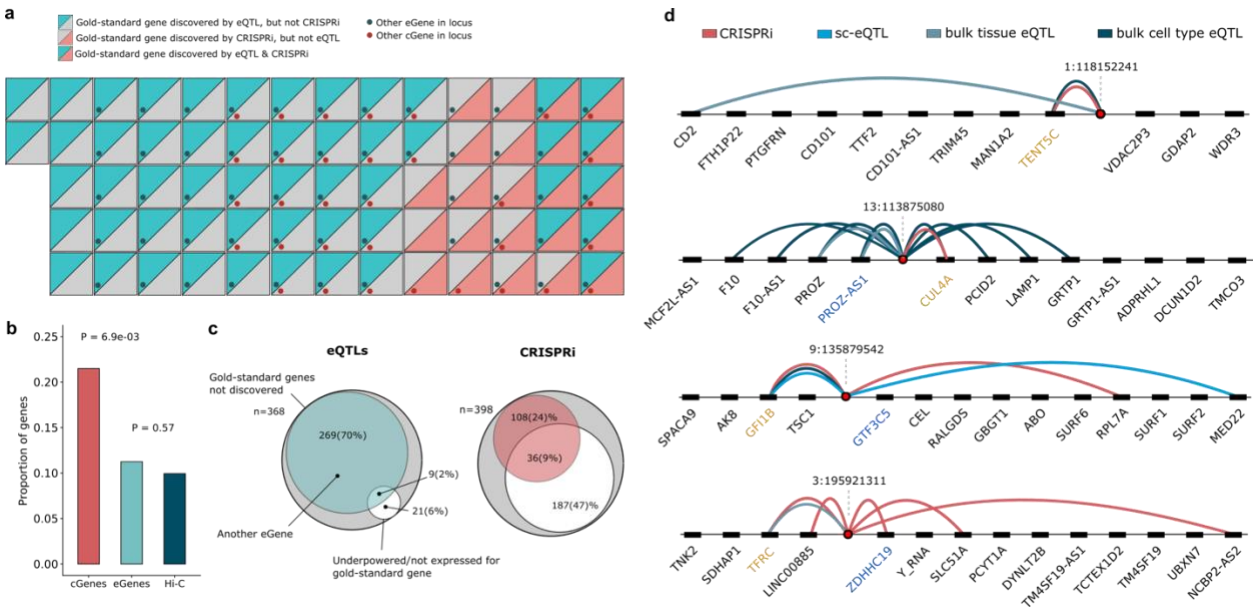

Figure 3 Discovery of gold-standard genes for blood traits. a, *Gold-standard genes identified as cGenes and/or eGenes, with each tile representing a single gold-standard gene.* ***b****, The proportion of cGenes, eGenes, and Hi-C interaction targets overlapping gold-standard genes.* ***c****, The gold-standard genes that were not identified as either cGenes or eGenes, with the proportion of undiscovered gold-standard genes for which eQTL and CRISPRi studies suggest a different causal gene at the same locus, as well as the proportion that remain undetected due to power limitations or lack of gene expression.* ***d****, Example loci showing the position of a GWAS variant and its gene associations as identified by different study types. Gold labels indicate the gold-standard gene in each locus, while blue labels mark the closest gene to the variant. If no blue label is present, the gold-standard gene is also the closest gene.*

Discussion

We present a systematic comparison of eQTL- and CRISPRi-derived GWAS *cis* target genes for blood cell traits, revealing that the gene targets captured by each approach are often different. Some discordance is expected due to the differences in the designs of the analyzed studies, which rely on CRISPRi silencing in a cancer cell line versus genetic variation in primary cells. However, as both approaches are popular for identifying *cis* target genes for GWAS loci, understanding the yield from each approach and the reasons for any differences will improve downstream biological inferences, practical applications, and optimal study designs.

Our results show that blood eQTL studies with bulk RNA-sequencing are highly sensitive to detect proximal and distal cis-target genes of regulatory variants. While sc-eQTLs have the potential to provide crucial insights into cellular contexts, further investments to increase power are needed for robust target gene identification. Conversely, CRISPRi studies detect a smaller

number of mostly proximal target genes. Our results indicate that statistical power limits the discovery of weaker (often distal[17]) effects, and better power from higher cell numbers and/or deeper sequencing would improve the alignment with eQTL results. The ability to discover distal genes is important for novel discoveries, as genes within tens of kilobases of a GWAS locus would be the first candidates based on gene annotation alone.

Our results show that cGenes are more selectively constrained and tightly regulated than eGenes, in line with prior work showing that eQTL mapping, due to being based on natural population variation, is biased against molecular effects under purifying selection [9,18]. As these genes are more likely to be relevant to disease, it is an important complication of eQTL approaches: analyses that are sufficiently powered to detect small constrained effects will readily detect more variable molecular features, leading to loss of specificity.

However, CRISPRi and other experimental perturbation approaches suffer from another fundamental limitation. These methods are nearly impossible to apply to the full diversity of primary cell types, as many disease-relevant cell types cannot even be cultured, let alone manipulated at scale. For many complex traits, the relevant cell types are not even identified. Our results indicate that cell type differences between primary cell types and K562 cells of the CRISPRi studies analyzed here partly contribute to the discoveries and discordance with eQTLs. The ongoing expansion of CRISPR studies to accessible primary cell types will be an exciting avenue to expand the catalogs of cGenes, to compare the results to eQTLs from fully matching cell types, and to understand the extent and mechanisms of context-specificity. Furthermore, while accurate variant-level editing at scale [23,24,25] is still technically challenging, it will be essential to understand whether CRE modulation leads to discoveries that may differ from the actual GWAS variant effects, not only quantitatively but even qualitatively.

This work compares the results of the first generation of CRISPRi studies to results from more than a decade of eQTL work. The discoveries of the first eQTL studies were not too dissimilar from the patterns of CRISPRi results here: biased towards strong regulatory effects close to the promoter, detected in a single cell type [26,27]. Analogously, while we anticipate further discoveries as the CRISPR field grows and matures, the improvements in sensitivity, resolution, and diversity are likely to come with some loss of specificity, given that multiple gene targets of CREs across different cell types are a core feature of genome function [28–30].

Our analysis of the detection of gold-standard trait-affecting genes at the studied GWAS loci clearly highlights how both the eQTL and CRISPRi approaches have substantial limitations and orthogonal strengths. However, the gold-standard gene set from rare disease studies and rare coding variants is not without caveats, as it may capture fundamentally different sets of genes than those underlying complex traits [10,31–34]. We also note that while cataloguing disease-causing genes is a major research focus, given its applications in drug target identification [35], it is not the only outcome or objective for eQTL and CRISPRi studies that can also provide invaluable, generalizable information about the biology of gene regulatory mechanisms.

Altogether, our results show the complementary insights obtained from functional genomics approaches that rely on population variation and experimental perturbations. Our results highlight the importance of continued investment in well-powered studies and the opportunities for integrating multiple approaches to create a comprehensive toolkit for GWAS interpretation.

Methods

Data

GWAS data

We utilized the genome-wide association study (GWAS) results for 29 different blood cell traits from 361,194 white British participants in the UK Biobank and trans-ethnic meta-analyses for 15 hematological traits in 746,667 participants from the BCX consortium [36]. These traits included measures such as white blood cell count, red blood cell count, hemoglobin concentration, hematocrit percentage, and various others listed in supplementary table 1.

eQTL data

A total of 50 eQTL studies [5,13–15], focusing on blood and/or blood cell traits were analyzed (Supplementary Table 2). Summary statistics and log Bayes factors for most studies were obtained from the eQTL Catalogue [13]. Summary statistics for GTEx [5], the Multi-Ancestry Analysis of Gene Expression (MAGE) study [14], and Meta LCL [21] were acquired from their respective study data sources listed in supplementary table 2. OneK1K association summary statistics were generated through re-analysis of the individual-level data (see Supplementary Methods).

CRISPRi perturbation data

CRISPRi data were obtained from Morris et al. [6], Gasperini et al. [7], and Gschwind et al [16]. The

Morris and Gasperini studies involved high-MOI CRISPRi screens targeting enhancers in K562 cells and were reanalyzed using SCEPTRE [37] with the default "union" setting, which groups gRNAs targeting the same element for association testing. Gschwind et al. compiled a harmonized gold-standard dataset of 10,411 element–gene pairs in K562 cells by reanalyzing CRISPR-based perturbation data from multiple studies. We used this curated dataset in our analyses.

Determination of CRISPRi target genes

To link CRISPRi perturbations to GWAS variants, we obtained the SNP coordinates of the genomic regions targeted by gRNAs from Morris et al. For Gasperini et al., gRNA targeting coordinates were determined by aligning spacer sequences to the GRCh37 reference genome and identifying their positions within the targeted enhancers. Next, we intersected these CRISPRi-targeted regions with fine-mapped GWAS variants within ±2 kb to determine which credible set variants were perturbed by a particular CRISPRi experiment.

Gold-standard gene list

Summary statistics for pLoF burden tests on 29 blood cell traits were obtained from Genebass [38]. Protein-coding genes with a SKAT p-value $< 5 \times 10^{-7}$ and located within 1 Mb of GWAS CREs targeted by CRISPRi gRNAs were selected for analysis. In addition to these burden test genes, curated gene lists associated with rare inherited Mendelian blood disorders [39,40] that were also in cis (within 1 Mb) of a GWAS CRE were also included. Fisher's exact tests were performed to assess the enrichment of overlap between Mendelian genes, burden test genes, cGenes, eGenes, and Hi-C gene targets. Hi-C, genes were filtered to protein-coding genes that were associated with a CRE within 1 Mb of a gold-standard gene.

Statistical analysis

Statistical fine-mapping of UK Biobank blood cell traits

For each of the 29 UK Biobank blood trait GWASs, statistical fine-mapping was conducted using Sum of Single Effects regression (SuSiE), as implemented in the susieR R package [41]. We defined 899 disjoint fine-mapping regions (Supplementary Methods) containing genome-wide significant variants ($P < 6.6 \times 10^{-9}$). We generated linkage disequilibrium (LD) matrices from a subset of 50,000 UK Biobank white British participants (UK Biobank accession code: 47976) using PLINK. We extracted all variants in the 95% credible set with a posterior inclusion probability (PIP) > 0.01.

Colocalization with UKBB GWAS variants

We performed colocalization between 29 UK Biobank blood trait GWASs and 49 eQTL studies of European ancestry. Pairwise colocalization was conducted using SuSiE COLOC [42] between each blood trait GWAS and eQTL study. When available, we used fine-mapped eQTL data from the eQTL Catalogue; otherwise, eQTL data were fine-mapped before colocalization using in-sample LD matrices. To ensure consistency in genomic positions, all data were mapped to the GRC37 using bcftools liftover [43]. Colocalization was performed within regions meeting the thresholds of $p < 6.6 \times 10^{-9}$ for GWAS associations and $p < 1 \times 10^{-5}$ for eQTL signals. SuSiE COLOC was conducted with default settings, and if convergence failed at a particular locus, colocalization was performed using COLOC [44] under the assumption of a single causal variant. Additionally, we conducted colocalization between 15 blood cell traits from the BCX GWAS [36] consortium and the multi-ethnic eQTL study MAGE [14]. For this analysis, COLOC was performed under the single causal variant assumption without requiring additional LD information.

Statistical testing

To compare distributions of continuous variables between gene sets, we used the Wilcoxon rank-sum test. This was applied to the number of associated enhancers and super-enhancers,
Episcore, gene expression levels in K562 and GTEx whole blood, minor allele frequency (MAF)
of GWAS variants and colocalization probabilities. For comparisons of proportions, we used Fisher's exact test. This included the number of genes with high loss-of-function intolerance (pLI
> 0.9), the number of gold-standard genes, and the number of genes with estimated CRISPRi or
eQTL power greater than 0.8. Chi-squared tests were used to assess differences in categorical
distributions, including the number of genes per locus and gene–CRE distances. All tests were
Two-sided. To test the association between log-fold change and tss distance, a linear regression
model was used, controlling for the number of cells bearing that particular gRNA.

eQTL power calculation

For genes with a mean count greater than 5, power is estimated analytically using an F-test. This power depends on sample size, the significance threshold α ($P = 1 \times 10^{-3}$), and the proportion of variance explained (R^2) by the variant in the population. Assuming genotypes are in Hardy-Weinberg equilibrium, R^2 is given by:

$$q^2 = 2 \times MAF \times (1 - MAF) \times \beta^2$$ [45]

Where β was set to 0.5. For lowly expressed genes, power is determined through simulations using scPower [46]. In sc-eQTL studies, the overall power is the product of eQTL power and expression probability, which quantifies the likelihood of a gene being detected in a given cell type [46]. Expression probability is determined by sample size, the number of cells per cell type, cell type frequencies, and the proportion of individuals in which a gene is expressed. Read depth is implicitly accounted for through UMI counts. For a detailed description of power calculations, refer to the Supplementary Methods.

CRISPRi power calculation

We estimated the power to detect reductions in gene expression for each of the perturbed CRE–target gene pairs using a modified version of the approach published by Gschwind et al. 2023[16]. For the Morris et al. and Gasperini et al. datasets, we first applied the estimateDispersions function (DESeq2, v1.46.2) to estimate mean and overdispersion parameters across cells for each gene. Using these estimates, we sampled expression values from a negative binomial distribution, introducing a fixed reduction in expression (15%) for each perturbed gene linked to a given CRE. We sampled the same number of cells for each perturbation as the original datasets and aggregated them to ensure that perturbed and non-perturbed cell counts remained consistent. Next, we used SCEPTRE in high-MOI mode (without additional quality filtering) to test each CRE–gene pair for a significant reduction in expression, applying the Benjamini-Hochberg procedure for multiple testing correction. The entire simulation was repeated 10 times, and a CRE–gene pair was considered powered for detection if a significant expression change was observed at FDR < 10% in at least 80% of simulations. For the remaining datasets, power estimates produced by Gschwind et al. 2023 were used [16].

Trans-eQTL calling

To identify GWAS CREs with trans-eQTLs, we selected *cis*-eQTL variants for genes that

displayed significant trans-effects in CRISPRi studies and were cis-targets of GWAS CREs. Trans-eQTL analysis was then performed for these cis-eQTL variants across Azimuth level 1 predicted cell types in the OneK1K dataset using tensorQTL [47]. Trans associations within a ±5 Mb window that met a significance threshold of FDR < 0.05 were retained. Additionally, trans-eQTL results from a meta-analysis of Lymphoblastoid Cell Lines (LCLs) in MetaLCL [21] were incorporated. Using the same approach, we identified cis-eQTL variants for genes with CRISPR-derived trans networks that were also GWAS targets based on MAGE data [14]. We then extracted trans associations for these cis-eQTLs from the MetaLCL results and retained those with FDR < 0.05 to further assess their trans-regulatory effects.

Trans-eQTL comparison

To assess concordance of trans-effects between methods, we first identified significant trans-cGene associations (FDR < 0.10) from Morris et al. and Gasperini et al. using SCEPTRE, filtered to cis-genes linked to GWAS CREs. Trans-eGene effect sizes were then aligned to the allele associated with decreased expression of the cis-gene. We then tested the correlation of effect sizes across overlapping trans-genes using Pearson correlation, adjusting p-values for multiple testing with the Benjamini-Hochberg procedure. Enrichment analyses were conducted using a permutation-based Fisher's exact test to estimate odds ratios and confidence intervals, with multiple testing correction applied similarly.

Gene properties

Enhancer and super-enhancer counts per gene were derived from GeneHancer v5 [48]

[48], with filtering applied to retain enhancers with a score ≥ 0.7 and gene associations

with a "link_score" ≥ 7. Haploinsufficiency (pHaplo) scores and Episcore values were obtained

from Collins et al [49]. pLI scores were retrieved from gnomAD v3, and gene
expression levels (TPM) for whole blood were obtained from GTEx version 8. Hi-C data for enhancer-promoter interactions in K562 was obtained from Zhigulev et al [50].

Data availability

GWAS CRE and eGene and cGene target results are available within the supplementary tables
of this manuscript. The fine-mapped credible sets and log-Bayes factors for the UKBB, GTEx,

and OneK1K are available for download at:
https://drive.google.com/drive/folders/1UdnlUs-xtutWk_hSofk1iX0tnS64P9dO?usp=drive_link.
The summary statistics of the recalled eQTLs of the OneK1K dataset are available at https://drive.google.com/drive/folders/1Q3Zz4EFvvgn4gYlIBJQl-q8605AYOhSO?usp=drive_link.
Reanalyzed CRISPRi data URLs can be found within Supplementary Table 3. We obtained summary statistics and log-Bayes factors for eQTL studies from the eQTL catalog [13].
GWAS summary statistics for 44 blood cell traits were downloaded from the Neale Lab (http://www.nealelab.is/uk-biobank/) and Lettre Lab (http://www.mhi-humangenetics.org/en/resources/). The datasets used in this study are publicly
available from various sources. GeneHancer v5.10 provides enhancer size and the number of enhancers per gene, this dataset is available upon request from GeneCards: https://www.genecards.org/Download/File?file=GeneHancer_v5.10.gff. GTEx v8 transcript-level
TPM data was obtained from GTEx Analysis v8 and can be accessed at:
https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8
_RSEMv1.3.0_transcript_tpm.gct.gz. The pHaplo and pTriplo scores, which estimate the probability of a gene being haploinsufficient or triplosensitive, along with Episcore values, were
obtained from https://ars.els-cdn.com/content/image/1-s2.0-S0092867422007887-mmc7.xlsx.
The pLI and LOEUF scores were retrieved from gnomAD v3 and are available at https://gnomad.broadinstitute.org/. Hi-C data for enhancer-promoter interactions in blood cell lines, including K562, was obtained from:
https://kth-my.sharepoint.com/:f:/g/personal/pelinak_ug_kth_se/Eq0b5SU6LGVFj4My7hiQiwUB5gmQuHEL-unZRrBCRa19ig?e=MinXnm.

Code availability

CRISPRi power calculations are available at:
https://github.com/jasperpanten/sc_crispr_simulations
All other analysis codes are available at: https://github.com/PandaPowell/CRISRPi-eQTLs-

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 47976. The authors would like to thank Angli Xue for valuable discussions and for providing the OneK1K raw VCF file and cell-type classifications.

This work was supported by funding from the National Institutes of Health R01MH106842, 24HG012090, R01HL168247 and R01HG012790, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement no. 101043238), and Knut and Alice Wallenberg Foundation grant KAW 2023.0337. A part of the computations were enabled by resources in project (SNIC 2022-5-560 and NAISS 2023-5-517) provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at UPPMAX, funded by the Swedish Research Council through grant agreement no. 721 2022-06725.

Author contributions

S.G., J.A.M., and T.L. conceived and designed the project. S.G. performed most of the data analysis. J.P. and S.G. performed power analyses and advised on analysis methods. W.O. recalled eQTLs from OneK1K. J.A.M., S.G., and T.L. provided feedback on the study design and data analysis. J.A.M. and T.L. supervised the work. S.G. and T.L. wrote the manuscript with contributions from other authors. All authors read, provided feedback, and approved the manuscript.

Competing interests

T.L. is an advisor with equity at Variant Bio and has received speaker fees from AbbVie. N.E.S. is an advisor to Qiagen and a co-founder and advisor of OverT Bio and TruEdit Bio. The remaining authors declare no competing interests.

Bibliography

1. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science* **373**, 1464–1468 (2021).
3. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating

the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).

4. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).
5. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
6. Morris, J. A. *et al.* Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science* **380**, eadh7699 (2023).
7. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377-390.e19 (2019).
8. Tambets, R., Kolde, A., Kolberg, P., Love, M. I. & Alasoo, K. Extensive co-regulation of neighboring genes complicates the use of eQTLs in target gene prioritization. *HGG Adv.* **5**, 100348 (2024).
9. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
10. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res.* **47**, e3 (2019).
11. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).
12. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8**, 2180–2196 (2013).
13. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
14. Taylor, D. J. *et al.* Sources of gene expression variation in a globally diverse human cohort. *Nature* **632**, 122–130 (2024).
15. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
16. Gschwind, A. R. *et al.* An encyclopedia of enhancer-gene regulatory interactions in the human genome. *BioRxiv* (2023) doi:10.1101/2023.11.09.563812.
17. Zuin, J. *et al.* Nonlinear control of transcription through enhancer-promoter interactions. *Nature* **604**, 571–577 (2022).
18. O'Connor, L. J. *et al.* Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).
19. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in

141,456 humans. *Nature* **581**, 434–443 (2020).

20. Hartman, J. L., Garvik, B. & Hartwell, L. Principles for the buffering of genetic variation. *Science* **291**, 1001–1004 (2001).
21. Freimann, K. *et al. USP18* modulates lupus risk via negative regulation of interferon response. *medRxiv* (2024) doi:10.1101/2024.07.15.24310442.
22. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
23. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
24. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
25. Ren, X. *et al.* High-throughput PRIME-editing screens identify functional DNA variants in the human genome. *Mol. Cell* **83**, 4633-4645.e9 (2023).
26. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
27. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).
28. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
29. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
30. Donohue, L. K. H. *et al.* A cis-regulatory lexicon of DNA motif combinations mediating cell-type-specific gene regulation. *Cell Genomics* **2**, (2022).
31. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2019).
32. Vuckovic, D. *et al.* The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214-1231.e11 (2020).
33. Spence, J. P. *et al.* Specificity, length, and luck: How genes are prioritized by rare and common variant association studies. *BioRxiv* (2024) doi:10.1101/2024.12.12.628073.
34. Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 492–499 (2023).
35. Trajanoska, K. *et al.* From target discovery to clinical drug development with human genetics. *Nature* **620**, 737–745 (2023).
36. Chen, M.-H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667

Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).

37. Barry, T., Wang, X., Morris, J. A., Roeder, K. & Katsevich, E. SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. *Genome Biol.* **22**, 344 (2021).
38. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168 (2022).
39. Westbury, S. K. *et al.* Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Med.* **7**, 36 (2015).
40. Greene, D., NIHR BioResource, Richardson, S. & Turro, E. Phenotype similarity regression for identifying the genetic determinants of rare diseases. *Am. J. Hum. Genet.* **98**, 490–499 (2016).
41. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
42. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* **17**, e1009440 (2021).
43. Genovese, G. *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics* **40**, (2024).
44. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
45. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
46. Schmid, K. T. *et al.* scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat. Commun.* **12**, 6625 (2021).
47. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
48. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**, (2017).
49. Collins, R. L. *et al.* A cross-disorder dosage sensitivity map of the human genome. *Cell* **185**, 3041-3055.e25 (2022).
50. Zhigulev, A. *et al.* Enhancer mutations modulate the severity of chemotherapy-induced myelosuppression. *Life Sci. Alliance* **7**, (2024).