Science
AAAS

Supplementary Materials for

Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens

John A. Morris[1,2], Christina Caragine[1,2], Zharko Daniloski[1,2], Júlia Domingo[1], Timothy Barry[3], Lu Lu[1,2], Kyrie Davis[1,2], Marcello Ziosi[1], Dafni A. Glinos[1], Stephanie Hao[4], Eleni P. Mimitou[4], Peter Smibert[4], Kathryn Roeder[3,5], Eugene Katsevich[6], Tuuli Lappalainen[1,7*], Neville E. Sanjana[1,2*]

[1] New York Genome Center, New York, NY, 10013, USA

[2] Department of Biology, New York University, New York, NY, 10003, USA

[3] Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

[4] Technology Innovation Lab, New York Genome Center, New York, NY, 10013, USA

[5] Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

[6] Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA, 19104, USA

[7] Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, 171 65 Solna, Stockholm, Sweden

* Corresponding authors. Emails: tlappalainen@nygenome.org (TL), nsanjana@nygenome.org (NES)

This PDF file includes:

Figs. S1 to S17

Other Supplementary Materials for this manuscript include the following:

Tables S1 to S4

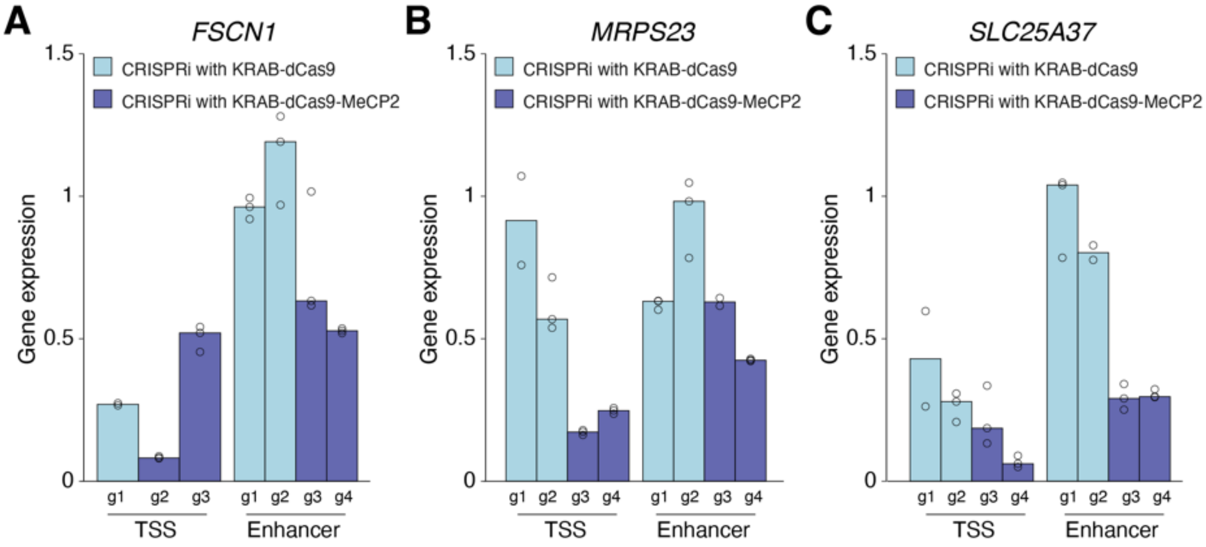


Fig. S1. Digital PCR for CRISPR inhibition (CRISPRi) of genes and guide RNAs. Digital PCR gene expression in K562 cells by targeting the transcription start sites (TSS) and known enhancers of *FSCN1* (**A**), *MRPS23* (**B**), and *SLC25A37* (**C**) with either CRISPRi with KRAB-dCas9 or KRAB-dCas9-MeCP2. Each bar represents one guide RNA (gRNA) (n = 3 biological replicates per gRNA).

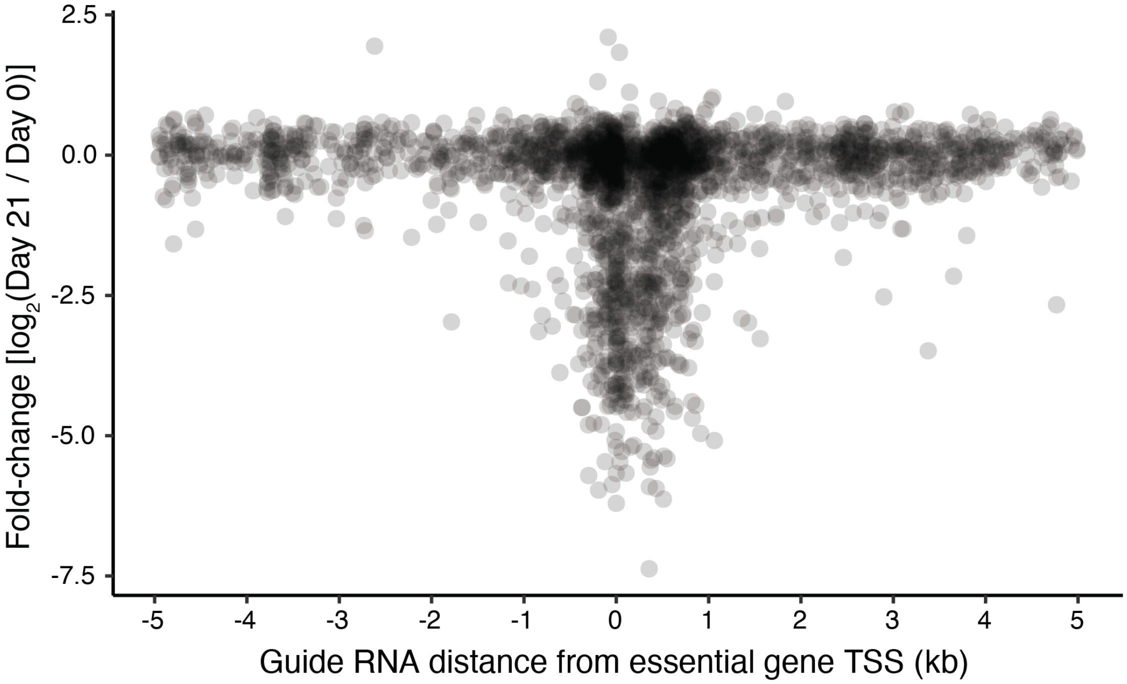


Fig. S2. Genome-scale screen for dual-effector CRISPRi (KRAB-dCas9-MeCP2).
A pooled screen of 1,992 guide RNAs targeting within 5 kb of 263 essential genes (DepMap essentiality score less than -1). We used 100 nt sliding windows to quantify regions where at least 50% of the gRNAs were depleted greater than the median of 1,000 non-targeting gRNA controls. We found that the dual-effector CRISPRi (KRAB-dCas9-MeCP2) is active at a distance of -400 to +850 nt (where +850 indicates 850 nt into the gene body).

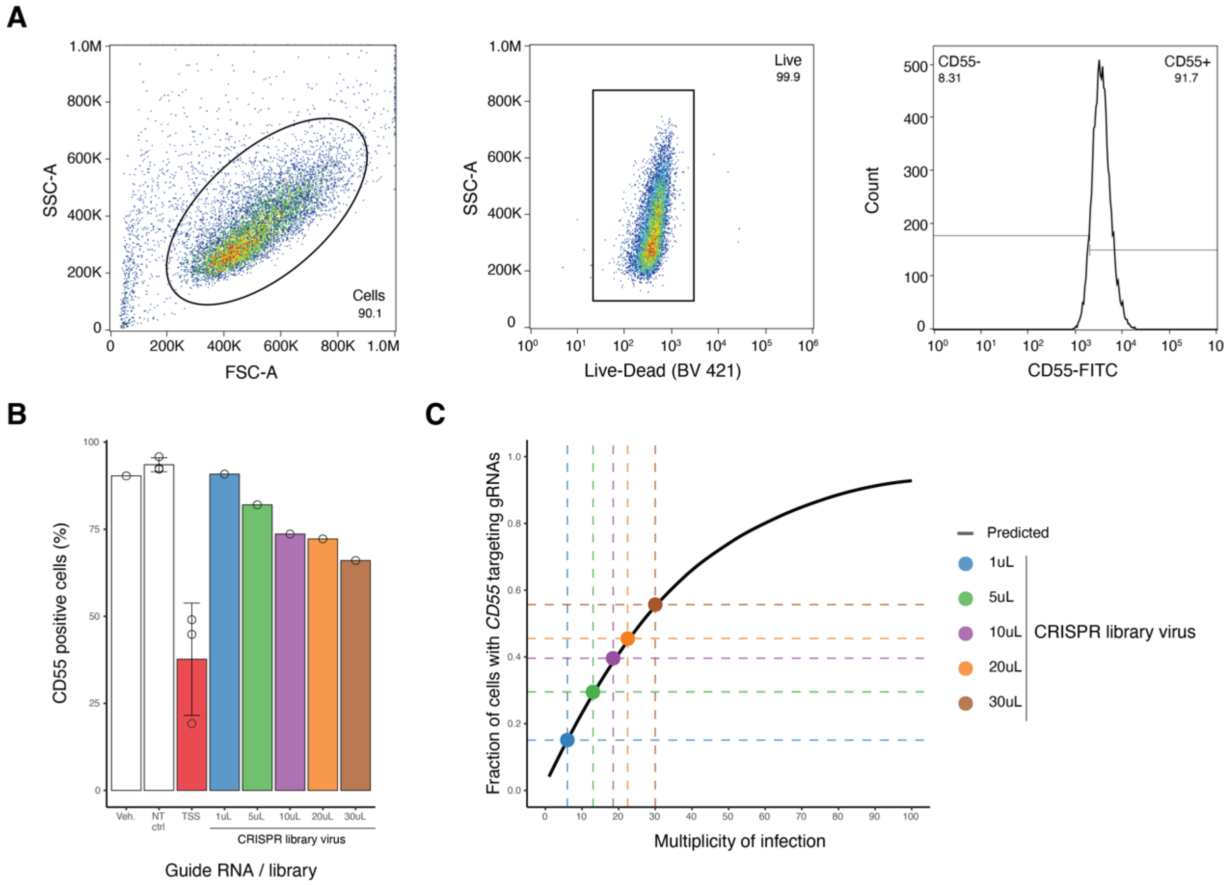


Fig. S3. Flow cytometry for estimation of multiplicity of infection.

(**A**) Flow cytometry gating strategy to quantify cell surface CD55 protein. Live cells were gated by the forward and side scatter area (*left*) then viable cells were selected by gating on side scatter area and LIVE/DEAD Violet (*middle*). Sorting gates were set so that 90% of wild-type K562 cells without any guide RNA (gRNA) are classified as CD55 positive (*right*). (**B**) CD55 positive cells with no transduction (veh.), non-targeting gRNAs, transcription start site (TSS)-targeting gRNAs, and five separate volumes of the STING-seq pooled library. (**C**) Estimation of the multiplicity of infection based on the starting distribution of *CD55* TSS-targeting gRNAs and proportion of cells bearing a gRNA.

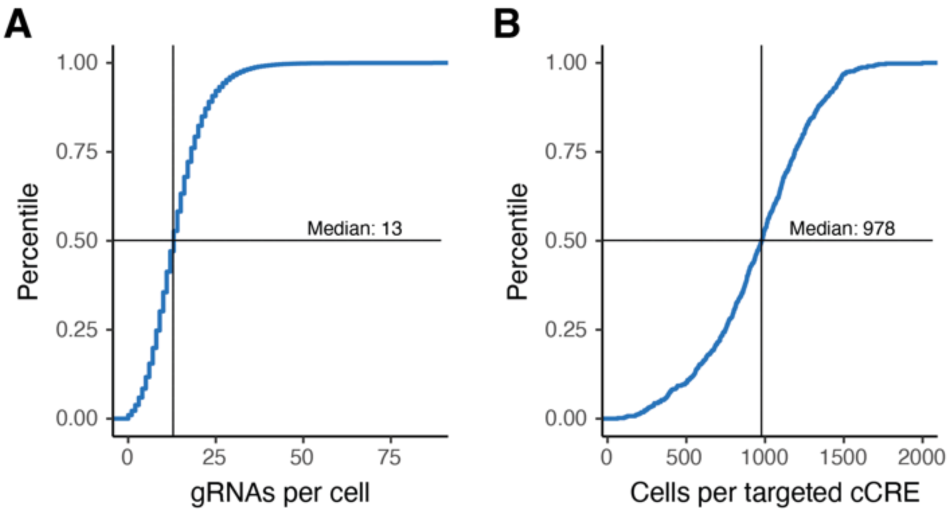


Fig. S4. Guide RNAs per cell and cells per targeted candidate *cis*-regulatory element (cCRE). (**A**) Median number of gRNAs per cell detected with at least three UMIs per gRNA. This number represents the multiplicity of infection. (**B**) Median number of cells per targeted cCRE after gRNA UMI thresholding.

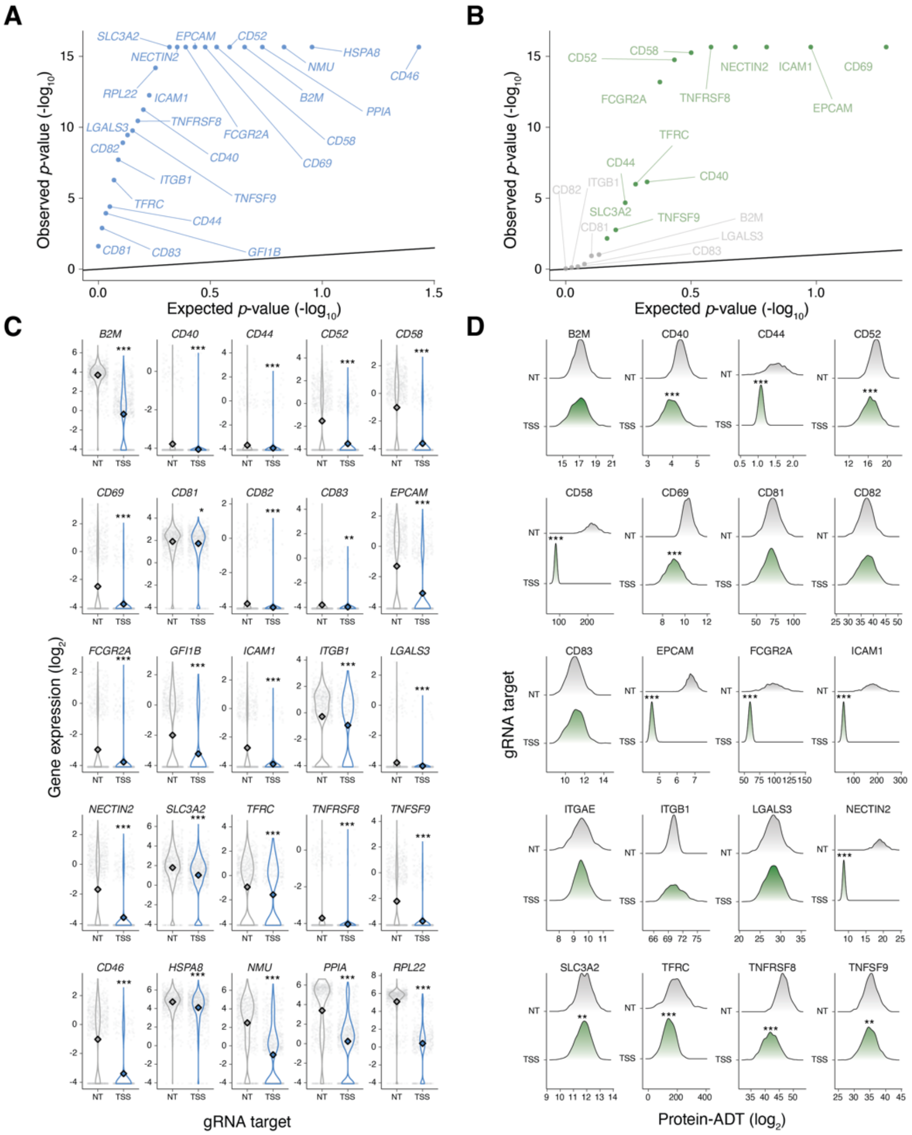


Fig. S5. Gene and protein expression for TSS-targeting positive controls guide RNAs. Quantile-quantile plots for positive control gRNAs and their effects on gene expression (**A**) and protein levels (**B**). Comparison of cells with positive control gRNAs and their effects on target gene expression with non-targeting gRNA controls (**C** for gene expression**, D** for protein). Asterisks denote q-values, Benjamini-Hochberg adjusted SCEPTRE p-values (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$).

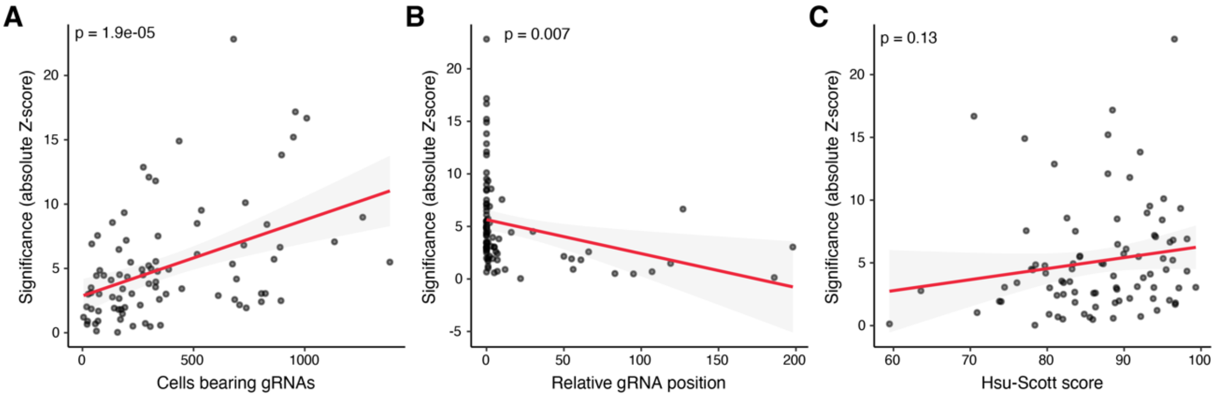


Fig. S6. Individual CRE-targeting gRNA effects on target gene expression.
We explored three facets for possible gRNA discrepancy between gRNAs targeting the same CRE and found that the number of cells bearing each gRNA is the main driver of statistical power (**A**). We observed a weak effect of gRNA position (less significant gRNAs tend to be further from significant gRNAs) (**B**), and no effect of off-target scores (**C**).

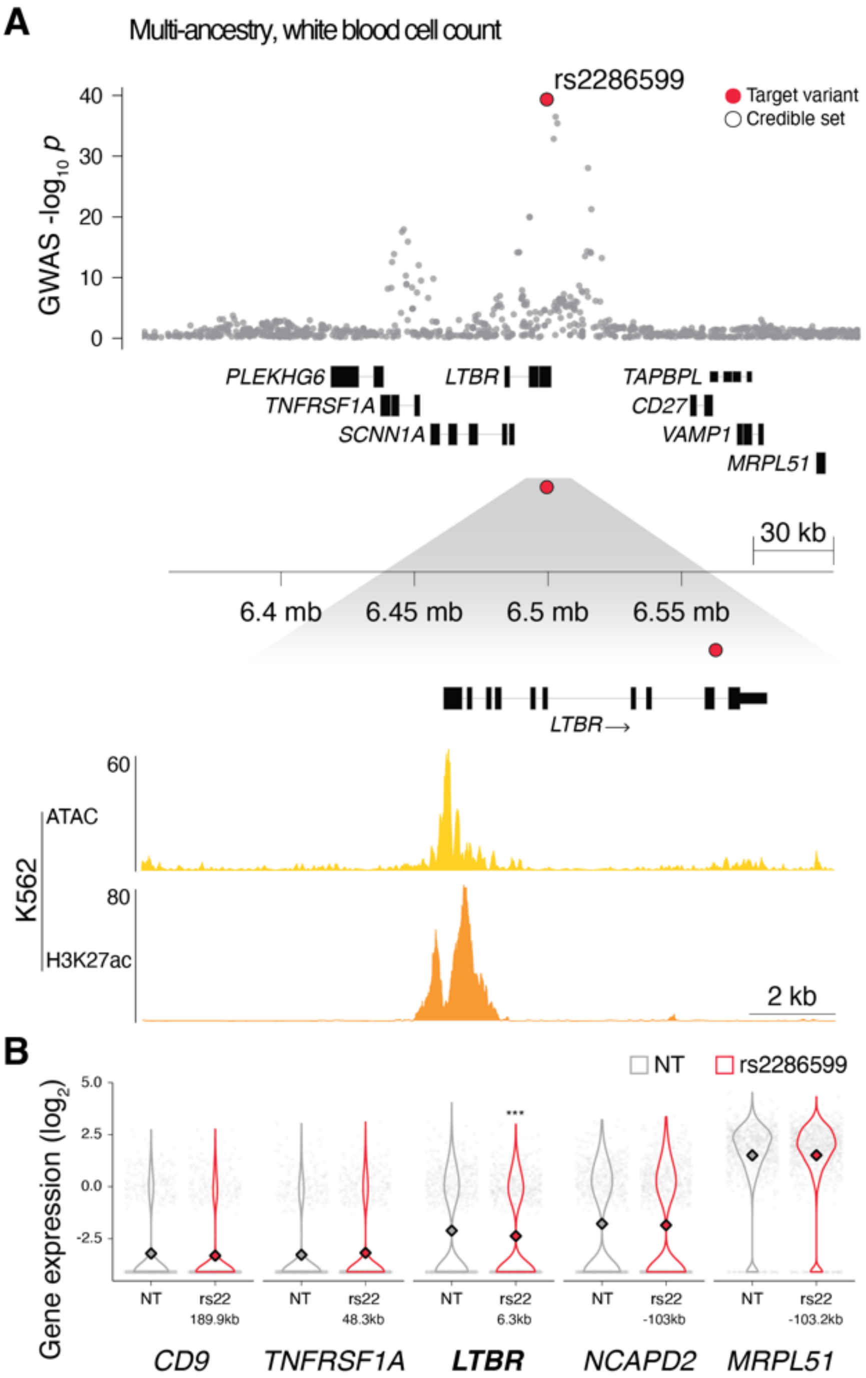


Fig. S7. A multi-ancestry white blood cell count locus with weak enhancer activity.
(**A**) The lead, and only, fine-mapped variant, rs2286599 was targeted as it was highly plausibly causal but did not map to a called peak of biochemical hallmarks of enhancers. However, open chromatin data revealed weak enhancer activity, and we detected in single-cell expression data a *cis*-target gene, *LTBR*, for rs2286599-CRE targeting gRNAs and not for NT gRNAs (**B**). Asterisks denote *q*-values, Benjamini-Hochberg adjusted SCEPTRE *p*-values (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$).

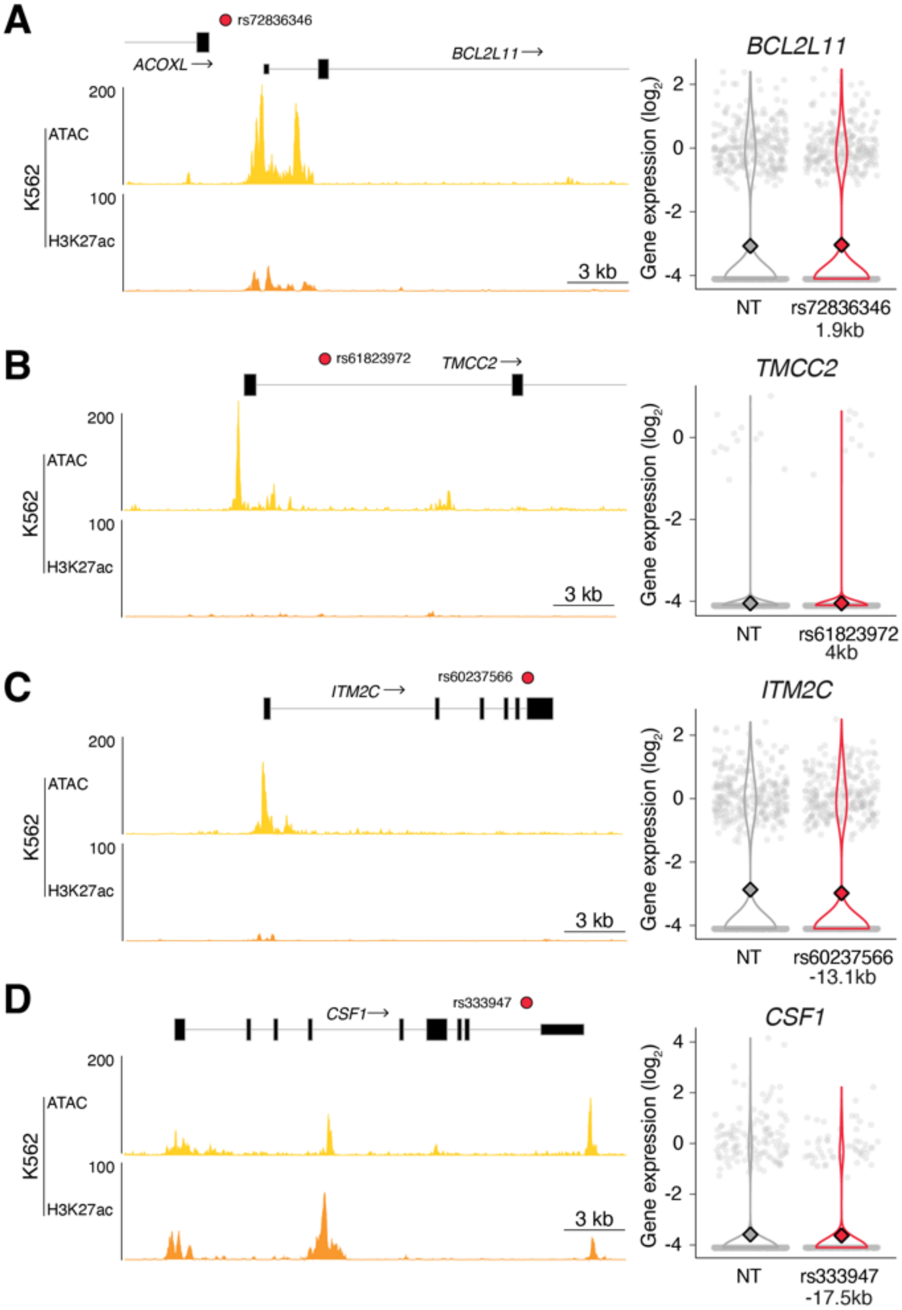


Fig. S8. Plausibly causal variants without biochemical hallmarks of enhancers.
We targeted four independent loci where the lead variant was the sole fine-mapped variant but did not map to called peaks for biochemical hallmarks of enhancers. We did not identify any *cis*-target genes, including the genes with the closest transcription start sites for: (**A**) rs72836346 and *BCL2L11*, (**B**) rs61823972 and *TMCC2*, (**C**) rs60237566 and *ITM2C*, and (**D**) rs333947 and *CSF1*. Each panel shows the locus with biochemical hallmarks of enhancers and closest gene (*left*) and the expression of the closest gene for cells receiving the GWAS-CRE perturbation or a non-targeting (NT) gRNA (*right*).

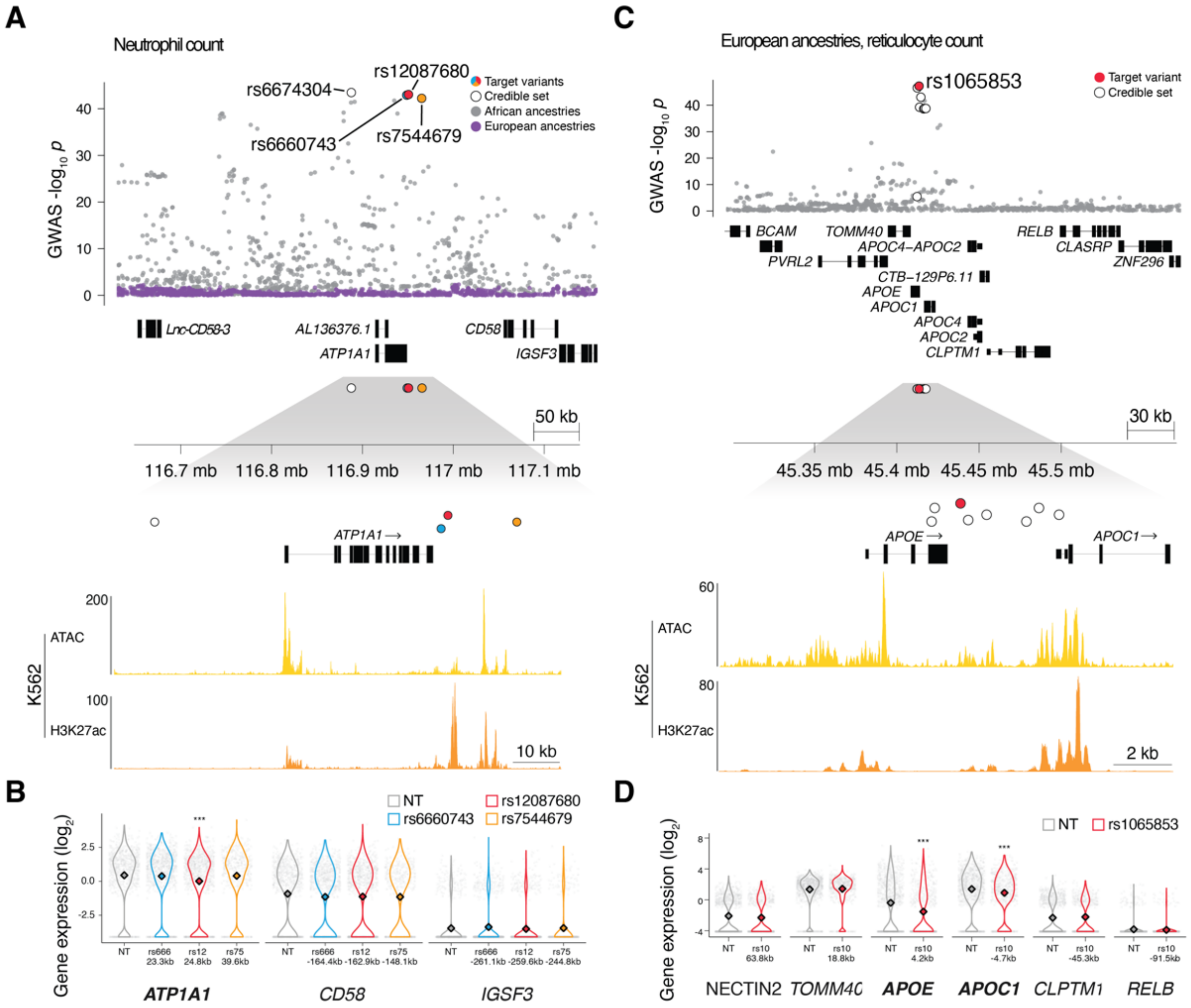


Fig. S9. Additional GWAS loci where STING-seq identifies target genes.
(**A**) An ancestry-specific neutrophil count locus, only detected in individuals with African ancestries (gray) and not in individuals with European ancestries (purple). The lead variant, rs6674304, did not map to hallmarks of enhancers, therefore we targeted the remaining three variants in the credible set that did: rs6660743 (blue), rs12087680 (red), and rs7544679 (orange). (**B**) Single-cell gene expression for cells bearing NT and targeting gRNAs. Only rs12087680 had a *cis*-target gene, *ATP1A1*. (**C**) A European ancestries reticulocyte count locus. One fine-mapped variant was targeted, the lead variant, rs1065853 (red). (**D**) Single-cell gene expression for cells bearing NT and targeting gRNAs. *APOE* and *APOC1* were both identified as *cis*-target genes. Asterisks denote q-values, Benjamini-Hochberg adjusted SCEPTRE p-values (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$).

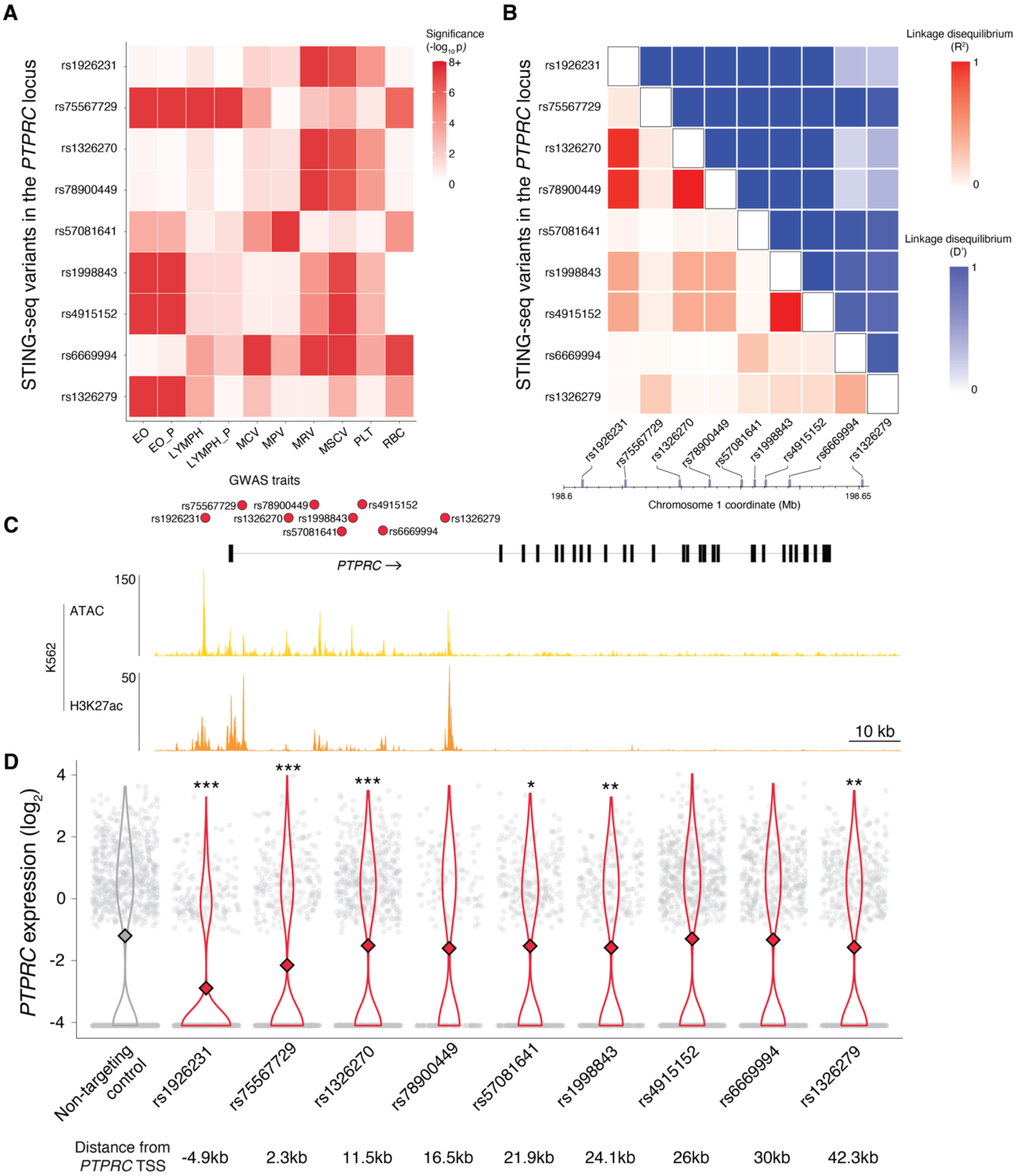


Fig. S10. STING-seq of nine GWAS variants at the *PTPRC* locus.
(**A**) Heatmap of *p*-values from 10 blood trait GWASs for nine variants mapping to cCREs proximal to *PTPRC*. The maximal color value indicates genome-wide significance (6.6x10^{-9}). (**B**) Pairwise linkage disequilibrium matrix (R^2 and D') for the nine targeted variants using the 1000 Genomes CEU and GBR populations. (**C**) Targeted GWAS-cCRE locations located at least 1 kb distal to the *PTPRC* TSS. (**D**) Normalized single cell *PTPRC* expression for the top targeting gRNAs. *PTPRC* was differentially expressed upon perturbation of six out of nine variants, identifying six

significant CREs (5% FDR). Two variants (rs1926231 and rs75567729) were located closest to the TSS of the *PTPRC* gene and had the strongest impact on gene expression. However, they were not in LD and had different GWAS significance patterns. Two other variants (rs78900449 and rs4915152) were not significant but were in strong LD ($R^2 \geq 0.95$) with variant-identified CREs for *PTPRC* (rs1326270 and rs1998843, respectively), suggesting they may be non-functional LD proxy variants. Asterisks denote *q*-values, Benjamini-Hochberg adjusted SCEPTRE *p*-values (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$).

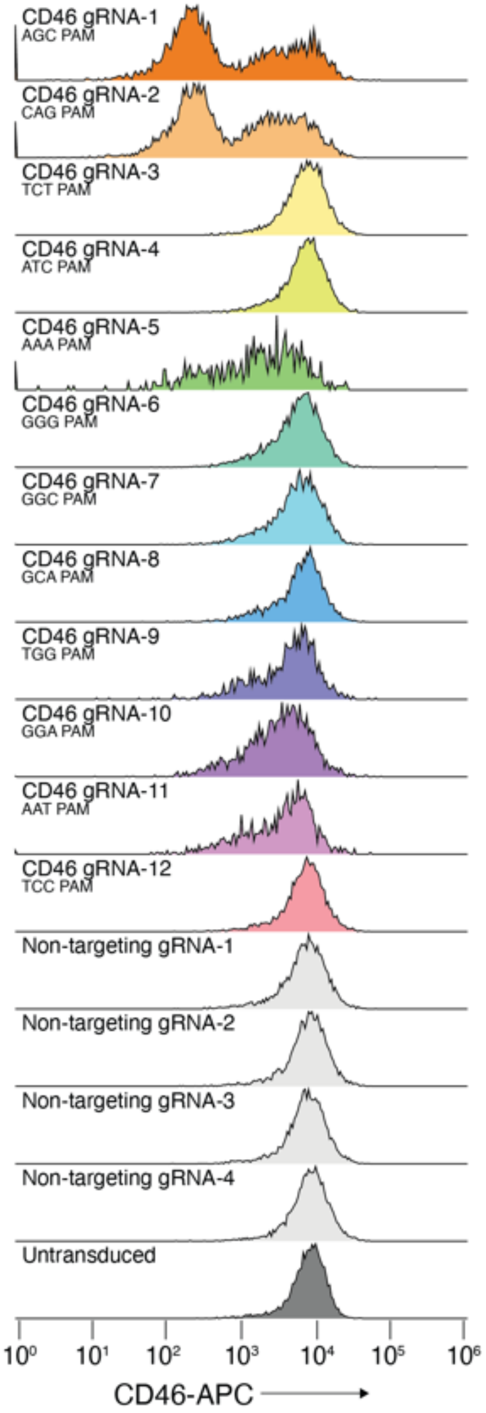


Fig. S11. Cytosine base editing of *CD46* splice donor sites.
We designed 12 *CD46* targeting gRNAs to engineer C>T mutations predicted to disrupt splice donor sites and used flow cytometry to measure CD46 protein depletion, compared to non-targeting gRNAs and untransduced cells (negative controls). gRNAs were designed for assorted non-canonical PAMs.

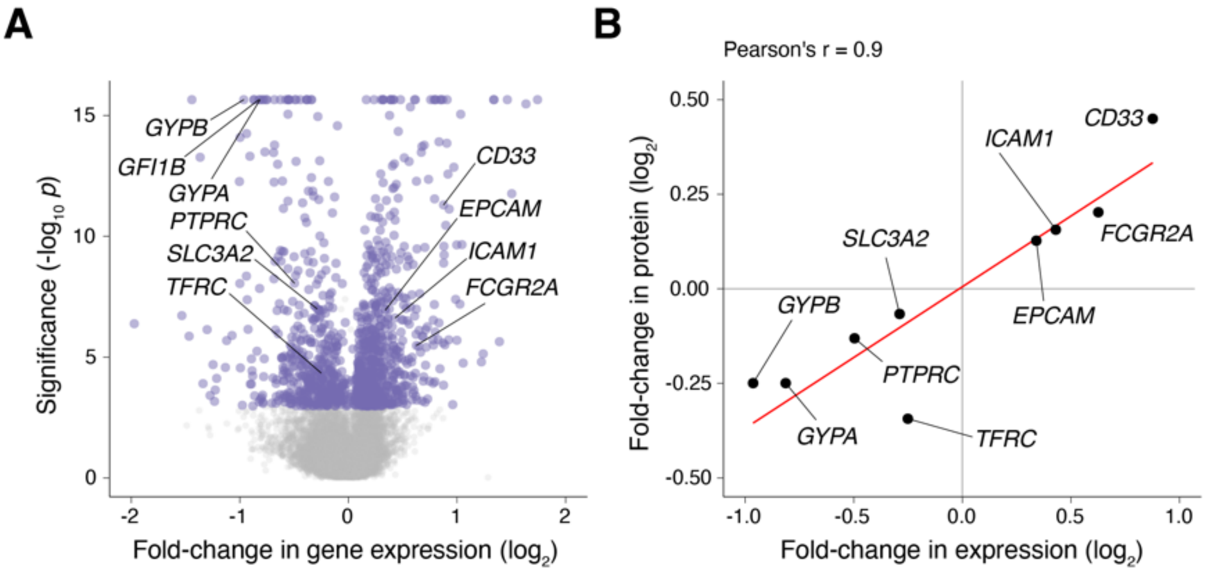


Fig. S12. *Trans*-effects of the rs524137-CRE for GFI1B on transcript and protein expression. (**A**) Volcano plot of the transcriptome-wide effects on differential expression upon inhibiting the rs524137-CRE for *GFI1B*. We labeled nine additional genes that had significant changes in gene expression (1% FDR) and found that their changes in expression were highly correlated with changes in protein measured with oligo-tagged antibodies (**B**).

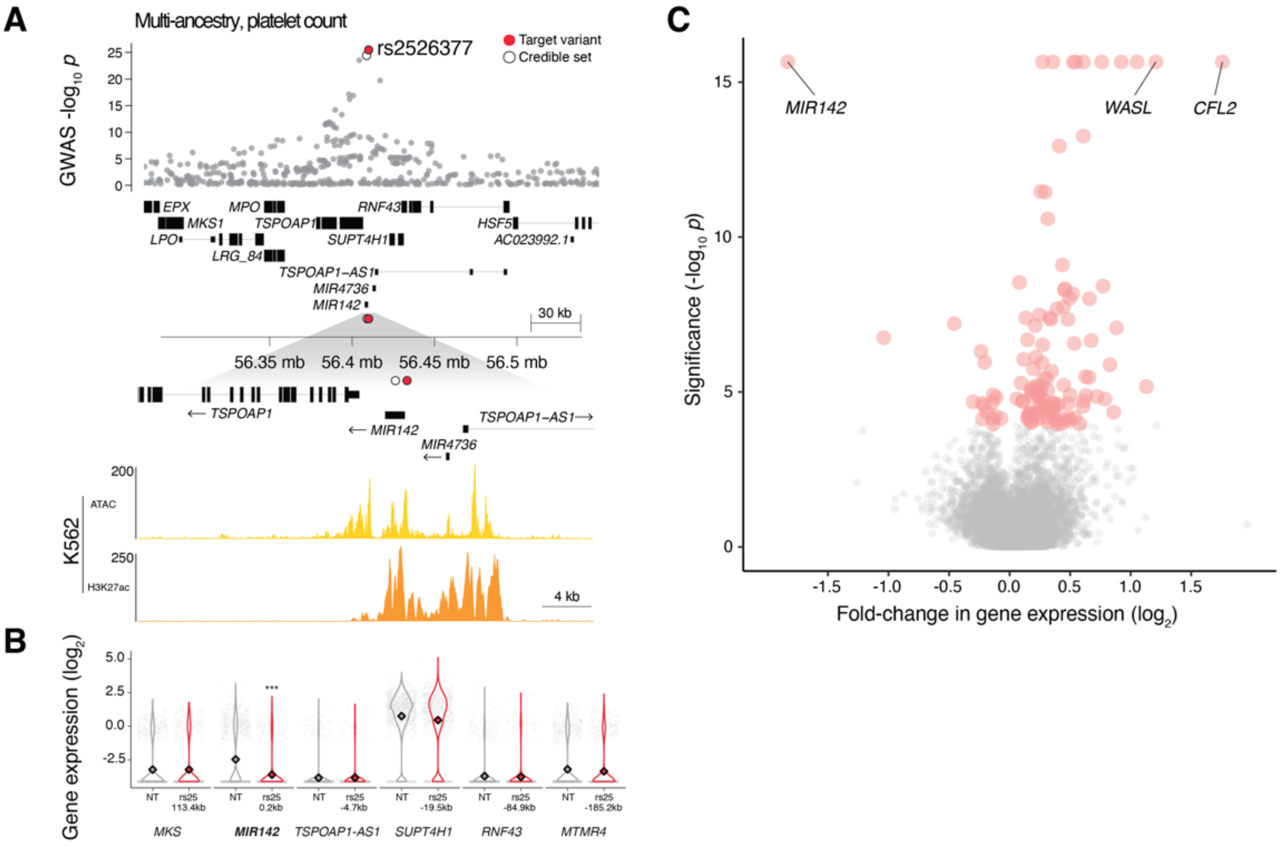


Fig. S13. A *trans*-regulatory network uncovered by perturbing a short RNA gene.
(**A**) A multi-ancestry platelet count locus, where we targeted a fine-mapped variant that was also the lead variant, rs2526377 (red). Rs2526377 does not map to any protein coding regions, however it does map to a promoter for a short, noncoding microRNA host gene for miR-142. (**B**) Single-cell gene expression for cells bearing NT and targeting gRNAs. The miR-142 host gene, a noncoding RNA gene, was identified as a *cis*-target gene. (**C**) Volcano plot of the transcriptome-wide effects of perturbing the miR-142 host gene, where the top two up-regulated genes, *WASL* and *CFL2*, were known targets of miR-142. Asterisks denote *q*-values, Benjamini-Hochberg adjusted SCEPTRE *p*-values (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$).

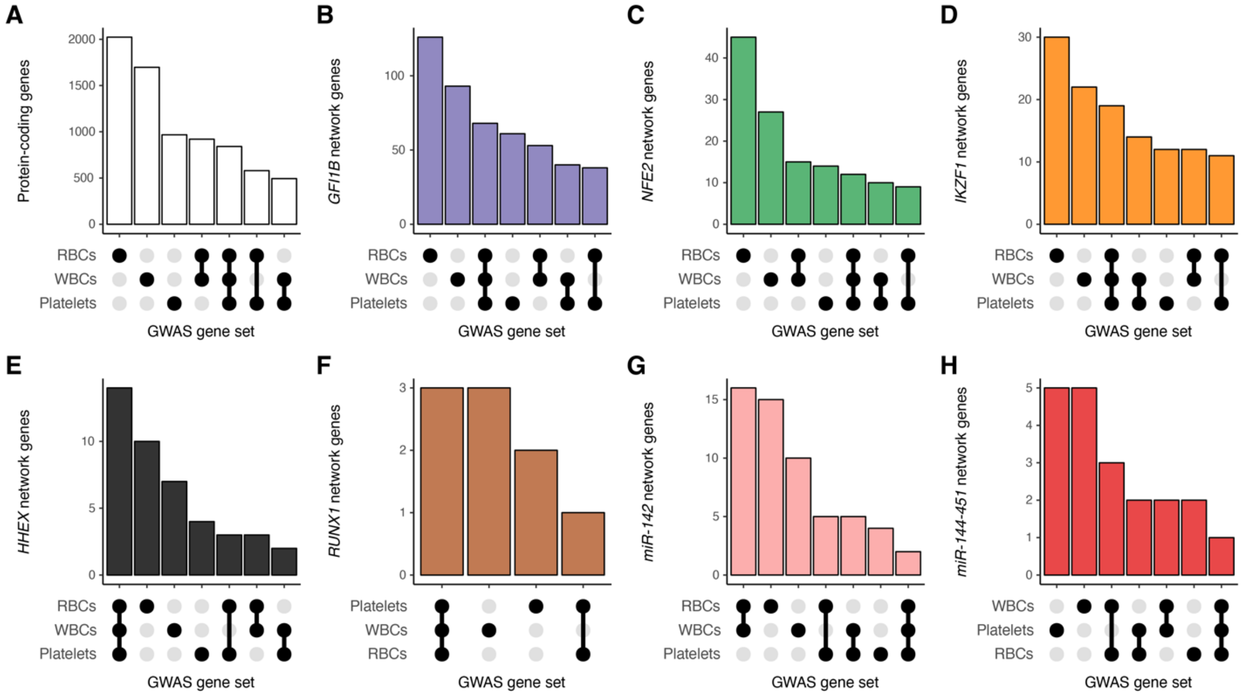


Fig. S14. Counts for the number of *trans*-regulatory network genes identified within each GWAS gene set for red blood cells (RBCs), platelets, and white blood cells (WBCs).
(**A**) All protein-coding genes and whether they were unique or shared across RBC, platelet, and WBC GWAS gene sets. We then inspected the *GFI1B* (**B**), *NFE2* (**C**), *IKZF1* (**D**), *HHEX* (**E**), *RUNX1* (**F**), *miR-142* (**G**), and *miR-144-451* (**H**) networks and whether they were found in our GWAS gene sets.

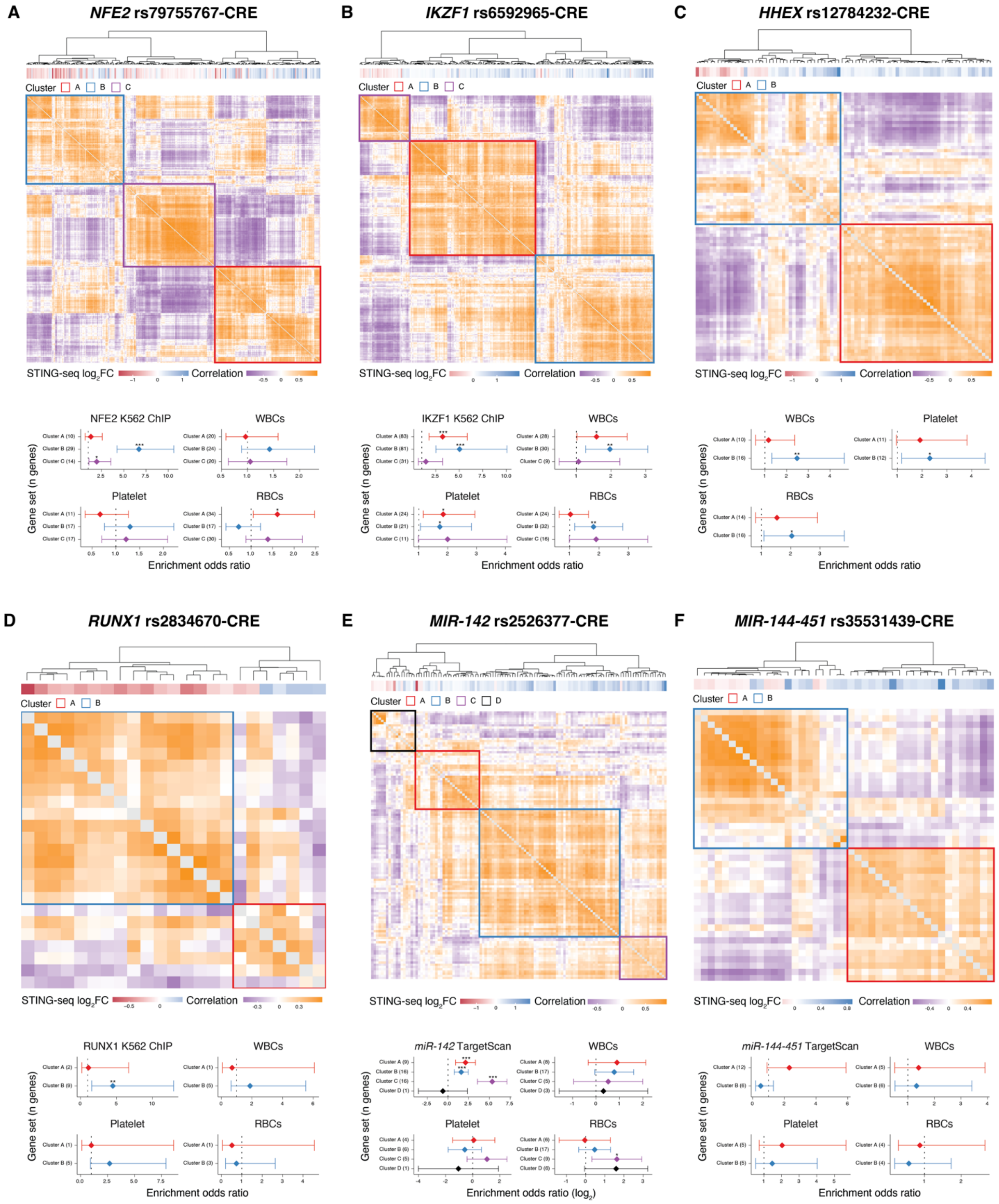


Fig. S15. Subnetworks have distinct gene set enrichment profiles.

Co-expression matrices of network genes with hierarchical clustering in K562 and their subnetwork cluster enrichments as odds ratios (*diamonds*) and 95% confidence intervals (*lines*) for direct targets and GWAS gene sets for the: (**A**) *NFE2* rs79755767-CRE, (**B**) *IKZF1* rs6592965-CRE, (**C**) *HHEX* rs12784232-CRE, (**D**) *RUNX1* rs2834670-CRE, (**E**) miR-142 rs2526377-CRE, and (**F**) miR-144-451 rs35531439-CRE. There were no *HHEX* ChIP-seq data in K562 for testing predicted direct target enrichments. Asterisks denote logistic regression *p*-values (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

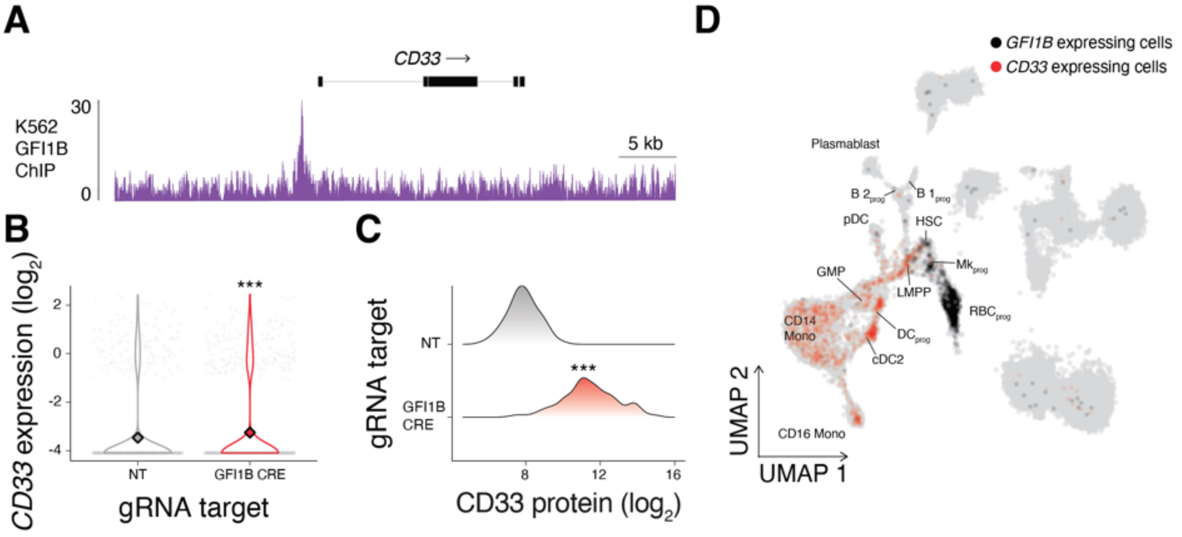
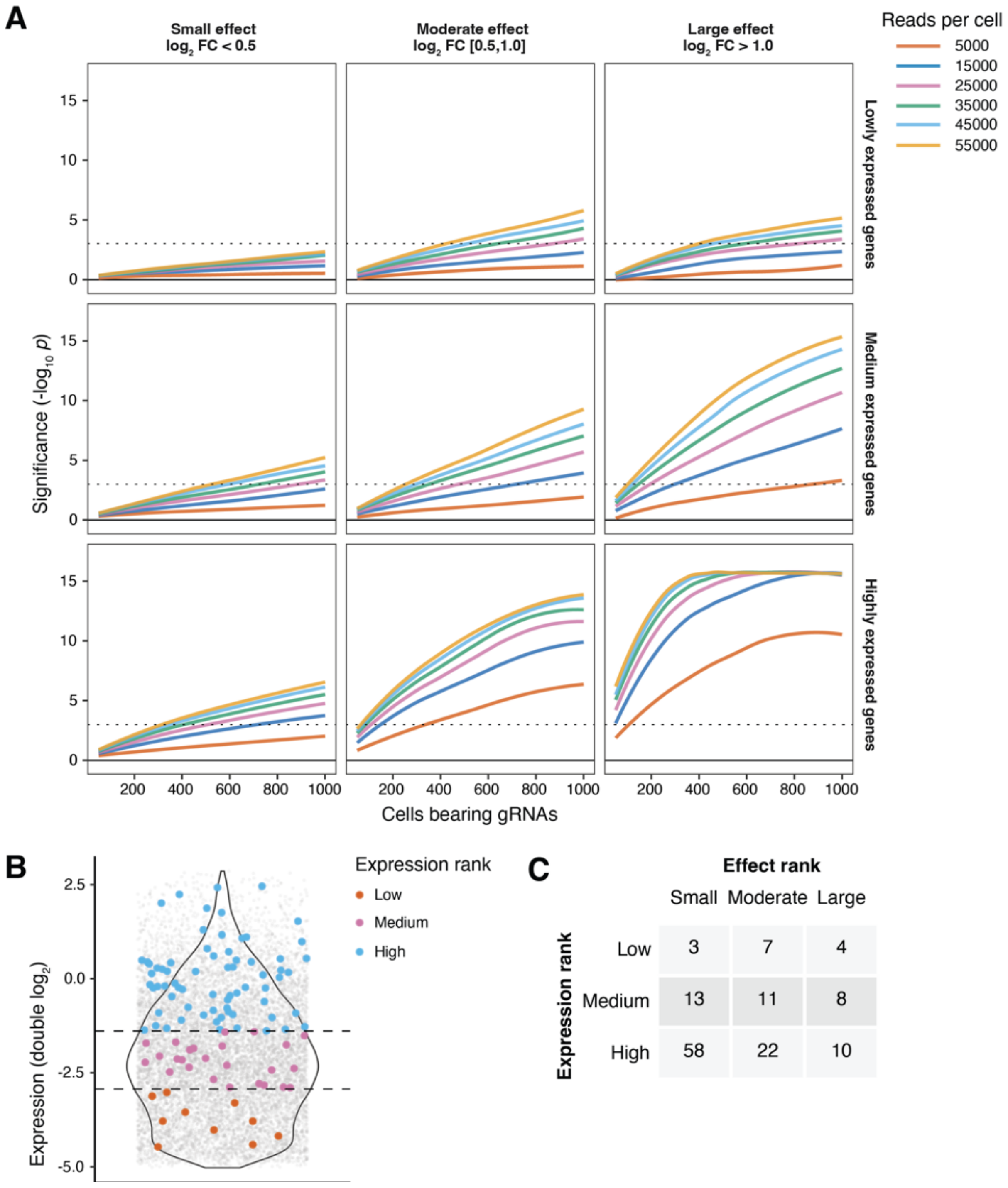


Fig. S16. A GFI1B target gene, *CD33*, is up-regulated upon *GFI1B* CRE inhibition.
(**A**) A GFI1B ChIP-seq peak was found directly upstream of *CD33* in K562. *GFI1B* is known to act as a transcription repressor, and we observed upon inhibiting the *GFI1B* CRE that *CD33* had increased expression (**B**) and protein (**C**). (**D**) These findings were consistent with *GFI1B* and *CD33* expression patterns in human bone marrow cells, where *GFI1B* was expressed in hematopoietic stem cells, RBC progenitors, and megakaryocyte progenitors, but not in WBC progenitors and differentiated myeloid cells. *CD33* is a marker of myeloid cells and was not expressed in cells where *GFI1B* was active. Asterisks denote q-values, Benjamini-Hochberg adjusted SCEPTRE p-values (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$).



Expression rank	Effect rank Small	Moderate	Large
Low	3	7	4
Medium	13	11	8
High	58	22	10

Fig. S17. Detection limits of *cis*-regulatory effects in STING-seq.
(**A**) CREs for target genes grouped by effect size and gene expression, and down-sampled from 1,000 cells bearing perturbations to 50 and from 55,000 to 5,000 sequencing reads per cell. (**B**) Distribution of genes with detected *cis*-regulatory effects and their expression rank relative to the full K562 transcriptome. (**C**) Number of genes for each corresponding down-sampling group.