Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2020 May 29.

Published in final edited form as:

Nat Genet. 2019 December ; 51(12): 1664–1669. doi:10.1038/s41588-019-0538-0.

Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations

Charles P. Fulco[1,2,9], **Joseph Nasser**[1,9], **Thouis R. Jones**[1], **Glen Munson**[1], **Drew T. Bergman**[1], **Vidya Subramanian**[1], **Sharon R. Grossman**[1,3], **Rockwell Anyoha**[1], **Benjamin R. Doughty**[1], **Tejal A. Patwardhan**[1], **Tung H. Nguyen**[1], **Michael Kane**[1], **Elizabeth M. Perez**[1], **Neva C. Durand**[1,4,5,6], **Caleb A. Lareau**[1], **Elena K. Stamenova**[1], **Erez Lieberman Aiden**[1,4,5,6,7], **Eric S. Lander**[1,2,3,10,*], **Jesse M. Engreitz**[1,8,10,*]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

[2]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA.

[3]Department of Biology, MIT, Cambridge, Massachusetts, USA.

[4]The Center for Genome Architecture, Baylor College of Medicine, Houston, Texas, USA.

[5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA.

[6]Department of Computer Science and Department of Computational and Applied Mathematics, Rice University, Houston, Texas, USA.

[7]Center for Theoretical Biological Physics, Rice University, Houston, Texas, USA.

[8]Harvard Society of Fellows, Harvard University, Cambridge, Massachusetts, USA.

[9]These authors contributed equally.

[10]These authors jointly supervised the work.

Abstract

Enhancer elements in the human genome control how genes are expressed in specific cell types and harbor thousands of genetic variants that influence risk for common diseases[1–4]. Yet, we still do not know how enhancers regulate specific genes, and we lack general rules to predict enhancer-

* engreitz@broadinstitute.org, lander@broadinstitute.org.

Author Contributions

C.P.F., E.S.L, and J.M.E. designed the study. C.P.F., V.S., G.M., and J.M.E. developed experimental methods. J.N., C.P.F., T.R.J., T.A.P., B.R.D., and J.M.E. developed computational methods. G.M., D.T.B., R.A., T.H.N., M.K., E.M.P., and E.K.S. performed experiments. C.P.F., J.N., T.R.J., S.R.G., C.A.L., N.C.D., E.L.A., E.S.L., and J.M.E. contributed to data analysis and interpretation. C.P.F., J.N., E.S.L., and J.M.E. wrote the manuscript with input from all authors. E.S.L. and J.M.E. supervised the work. E.S.L. obtained funding.

Competing Interests Statement

E.S.L. serves on the Board of Directors for Codiak BioSciences and Neon Therapeutics, and serves on the Scientific Advisory Board of F-Prime Capital Partners and Third Rock Ventures; he is also affiliated with several non-profit organizations including serving on the Board of Directors of the Innocence Project, Count Me In, and Biden Cancer Initiative, and the Board of Trustees for the Parker Institute for Cancer Immunotherapy. He has served and continues to serve on various federal advisory committees. C.P.F., E.S.L., and J.M.E. are inventors on a patent application (WO2018064208A1) filed by the Broad Institute related to this work.

gene connections across cell types[5,6]. We developed an experimental approach, CRISPRi-FlowFISH, to perturb enhancers in the genome and applied it to test >3,500 potential enhancer-gene connections for 30 genes. We found that a simple Activity-by-Contact (ABC) model substantially outperformed previous methods at predicting the complex connections in our CRISPR dataset. This ABC model allows us to construct genome-wide maps of enhancer-gene connections in a given cell type based on chromatin state measurements. Together, CRISPRi-FlowFISH and the ABC model provide a systematic approach to map and predict which enhancers regulate which genes, and will help to interpret the functions of the thousands of disease risk variants in the noncoding genome.

We developed an approach called CRISPRi-FlowFISH to perturb hundreds of noncoding elements in parallel and quantify their effects on the expression of an RNA of interest, combining CRISPR interference, RNA fluorescence *in situ* hybridization (FISH), and flow cytometry (Fig. 1a and Extended Data Fig. 1). In this approach, we deliver KRAB-dCas9 to many candidate regulatory elements in a population of cells using a library of guide RNAs (gRNAs; ~1 gRNA per cell). (KRAB-dCas9 has previously been shown to repress many promoters and enhancers, and affects elements within ~200–500 bp of the gRNA; Supplementary Note 1)[7–9]. To measure the effects of candidate elements on the expression of a gene of interest, we: (i) use RNA FISH to quantitatively label single cells according to their expression of an RNA of interest; (ii) sort labeled cells with fluorescence-activated cell sorting (FACS) into 6 bins based on RNA abundance; (iii) use high-throughput sequencing to determine the abundance of each gRNA in each bin; (iv) and use this information to infer the effect of each gRNA on gene expression. To assess quantitative effects and statistical significance, we calculate the average effect of all gRNAs within each candidate element (Fig. 1c) and compare to hundreds of negative control gRNAs in the same screen.

To generate a large enhancer perturbation dataset, we used CRISPRi-FlowFISH in K562 human erythroleukemia cells to test a total of 4,662 candidate regulatory element-gene pairs. We performed CRISPRi-FlowFISH screens for 30 genes in 5 genomic regions (spanning 1.1–4.0 Mb) and tested all DNase I hypersensitive (DHS) elements in K562 cells within 450 kb of any of the genes (108 to 277 elements per gene for a total of 884 unique elements). The 30 genes included some with erythroid lineage-specific expression (*e.g.*, *GATA1*) and some that are ubiquitously expressed (*e.g.*, *RAB7A*) and were selected to have FlowFISH probesets that met stringent criteria for both specificity and statistical power (Supplementary Fig. 1; see Methods). Replicate screens produced highly correlated estimates for the effect sizes of each element (Pearson $R = 0.94$, Extended Data Fig. 2f), and we confirmed that the effects on gene expression estimated from CRISPRi-FlowFISH agreed with RT-qPCR measurements (Pearson $R = 0.81$, Extended Data Fig. 2e). As expected, these screens identified the three previously identified elements for *GATA1* (Fig. 1b,c)[9].

We analyzed these CRISPRi-FlowFISH data together with data from an additional 429 candidate regulatory element-gene pairs from previous CRISPR-based experiments in K562 cells[7,9–17]. In total, our dataset included 3,863 candidate distal element-gene (DE-G) pairs (where the targeted element is located >500 bp from a TSS) and 1,228 distal promoter-gene

(DP-G) pairs (where the targeted element is located <500 bp from a TSS). Here we focused on DE-G pairs, and analyzed DP-G pairs separately (Supplementary Note 2).

These perturbation-based maps uncovered complex connections wherein individual enhancers regulated up to 5 genes, individual genes were regulated by up to 14 distal elements, and some enhancers appeared to "skip" over proximal genes to regulate more distant ones (Fig. 2 and Supplementary Figs. 2 and 3). Of the 3,863 DE-G pairs tested, 141 involved a significant effect on gene expression at a false discovery rate (FDR) < 0.05. DE perturbation led to a decrease in expression in 77% of cases and increase in 23% of cases (109 vs. 32), with absolute effect sizes ranging from 3%–93% (median: 22%).

Using these data, we sought to identify generalizable rules to explain which enhancers regulate which genes in the genome. To do so, we compared predictors to our experimental results by means of a precision-recall plot (Fig. 3a) — where true regulatory connections are the 109 DE-G pairs where perturbation of the element led to a significant decrease in gene expression (*i.e.*, the element activates gene expression in the genome), and the non-regulatory connections are the 3,754 pairs where no decrease was detected despite >80% power to detect 25% effects. (For analysis of repressive effects, see Supplementary Note 3).

We first examined existing methods that are commonly used to predict functional enhancer-gene connections and found these had only modest predictive value (Fig. 3a):

1. Predictions based solely on distance thresholds along the genome performed poorly. For example, while 84% of regulatory DEs were located within 100 kb of their target promoter, only 13% of DEs within 100 kb of an expressed gene promoter had a regulatory effect (precision = 13%, recall = 84%). Assigning each DE to the closest expressed gene yielded 47% precision and 37% recall.
2. Predictions based solely on features of the 3D genome also performed poorly. Assigning each DE to promoters based on the presence of Hi-C peaks ("loops"[18]) yielded 29% precision and 4% recall, and assigning each DE to each promoter in the same contact domain yielded 7% precision and 72% recall.
3. Predictions based on prior machine learning approaches, including correlating chromatin marks with gene expression across cell types, were similarly unsuccessful (see Supplementary Methods)[19,20].

Given the limitations of existing methods, we developed the Activity-by-Contact (ABC) model to predict enhancer-gene connections. This model is based on the simple biochemical notion that an element's quantitative effect on a gene should depend on its strength as an enhancer ("Activity") weighted by how often it comes into 3D contact with the promoter of the gene ("Contact"), and that the *relative* contribution of an element on a gene's expression (as assayed by the proportional decrease in expression upon CRISPR-inhibition) should depend on the element's effect divided by the total effect of all elements. Under this model (Fig. 3b), the fraction of regulatory input to gene G contributed by element E is given by:

$$\text{ABC Score}_{\text{E,G}} = \frac{\text{A}_E \times \text{C}_{E,G}}{\sum_{e \text{ within 5Mb of } G} \text{A}_e \times \text{C}_{e,G}}$$

Operationally, we estimated Activity (A) as the geometric mean of the read counts of DHS and H3K27ac ChIP-seq at an element E, and Contact (C) as the KR-normalized Hi-C contact frequency between E and the promoter of gene G at 5-kb resolution (see Supplementary Note 4 and Supplementary Figs. 4 and 5).

The ABC model performed remarkably well, and much better than alternatives, at predicting DE-G connections in our CRISPR dataset. The quantitative ABC score correlated with the experimentally measured relative effects of candidate elements on gene expression (Spearman ρ for regulatory DE-G pairs = –0.63; Fig. 3c). Binary classifiers based on thresholds on the ABC score substantially outperformed existing predictors of enhancer-gene regulation. For example, when we used an ABC threshold corresponding to 70% recall, the predictions had 59% precision. The area under the precision-recall curve (AUPRC) was 0.65, compared to 0.39 for predictions from genomic distance (Fig. 3a). The ABC score also outperformed using either Activity or Contact individually (AUPRC = 0.22 and 0.29, respectively; Extended Data Fig. 3a).

Given the ability of the ABC model to make predictions in K562 cells based on epigenomic data from that cell type, we explored whether the ABC model could generalize to predict enhancer-gene connections in other cell types.

To do so, we first identified alternative ways to estimate Contact in the ABC model; although maps of chromatin accessibility and histone modifications are available in many cell types, maps of 3D contacts are not. Because contact frequencies in Hi-C data correlate well across cell types and are largely determined by 1D genomic distance[21,22], we compared versions of the ABC model in which we estimated Contact for each DE-G pair using either K562 Hi-C data, the average of Hi-C data from 10 human cell types, or a function of distance (Contact ~ Distance^{-1}) (Supplementary Note 5). All three approaches performed similarly at predicting our CRISPR data in K562 cells (AUPRC = 0.65, 0.66, and 0.64 respectively; Supplementary Fig. 6a). Thus, the ABC model can make predictions in a given cell type without cell-type specific Hi-C data and minimally requires: (i) a measure of chromatin accessibility (DHS or ATAC-seq) and (ii) a measure of enhancer activity (ideally, H3K27ac ChIP-seq) (Extended Data Fig. 3).

Using this approach, we evaluated the ability of the ABC model to predict 997 measured DE-G pairs in 5 additional human and mouse cell types beyond our initial K562 dataset (see Supplementary Methods)[23–33]. We generated genome-wide predictions of functional enhancer-gene connections in each of these 5 cell types and compared them to the functional data in the corresponding cell type. The ABC scores correlated with the quantitative effects on gene expression (Spearman ρ for regulatory DE-G pairs = –0.30, Fig. 4a), and had 70% precision at an ABC threshold corresponding to 70% recall (AUPRC = 0.73, Fig. 4b). As expected, the predictions of the ABC model were highly cell-type specific; when we used ABC scores computed using epigenetic data in K562 cells to predict DE-G pairs measured in other human cell types, the AUPRC dropped from 0.73 to 0.11.

We next examined the 16 DE-G pairs in our dataset that involved enhancers that harbor noncoding genetic variants known to influence human traits and to regulate specific genes.

At a threshold corresponding to 70% recall in our K562 dataset, the ABC model correctly connected these DEs to their target gene(s) in 13 of 16 cases (81% recall, compared to 56% for assigning DEs to the closest expressed gene). For example, a variant associated with coronary artery disease and plasma low-density lipoprotein cholesterol (NC_000001.10:g.109817590G>T, rs12740374) has been shown to be an eQTL for *SORT1* in liver tissue, and CRISPR edits in the corresponding element affect *SORT1* expression in primary hepatocytes[33,34]. ABC maps in liver tissue correctly connected this enhancer to *SORT1* (Fig. 4c). Thus, the ABC model can predict enhancer-gene connections based on cell-type specific epigenomic data and may be widely useful for interpreting the functions of noncoding genetic variants associated with human diseases.

Finally, toward further improving predictions, we identified situations in which the ABC model failed to accurately predict DE-G connections.

We first compared predictions for tissue-specific versus ubiquitously expressed genes (see Supplementary Methods) and found that the ABC model performed dramatically better for tissue-specific than for ubiquitously expressed genes (AUPRC = 0.73 vs. 0.18; Extended Data Fig. 4). The ubiquitously expressed genes were affected by very few enhancers: for the 32 genes for which we had data for all nearby DEs, tissue-specific genes ($n = 24$) had an average of 2.5 distal enhancers per gene, while ubiquitously expressed genes had only 0.4 (3 enhancers across 8 ubiquitously expressed genes; rank-sum test $P = 0.007$). We conclude that the ABC model applies well to tissue-specific genes (97% of all genes) but not to ubiquitously expressed genes, which appear to be largely insensitive to the effects of distal enhancer perturbations for reasons that remain to be explored[35].

We next examined our CRISPR dataset for DE-G pairs that likely represent effects due to mechanisms other than the *cis*-acting functions of enhancers (Supplementary Note 3). We identified effects of distal CTCF sites, which may regulate gene expression by affecting 3D contacts (8 significant pairs, Supplementary Fig. 7), and likely indirect effects, such as an enhancer regulating one gene that in turn affects a second nearby gene in *trans* (15 pairs, Supplementary Fig. 8, see Supplementary Methods). Because these DE-G pairs do not represent direct effects of enhancers, we reasoned that removing them from the CRISPR dataset should provide a better estimate of the ability of the ABC model to predict enhancer-gene connections. The AUPRC rose from 0.64 to 0.67 for all genes and to 0.76 for tissue-specific genes (Supplementary Fig. 9). These results suggest a strategy to refine our predictions of DE-G connections by using CRISPRi tiling to identify exceptions to the ABC model, characterizing their molecular mechanisms, and developing new models to predict these effects.

In summary, our work reveals key properties of enhancer-gene connections and provides an important foundation for future studies of regulatory elements and noncoding genetic variants. Our perturbation data, consistent with the predictions of the ABC model, indicate that enhancers often regulate more than one gene (Fig. 2d), that most enhancers with detectable effects are located within 100 kb of their target promoters (Fig. 2e), and that enhancers can have a wide range of quantitative effects on gene expression — including many elements with small effects (Fig. 3c).

Our results raise the intriguing possibility that the ABC model reflects an underlying biochemical principle: that enhancer "specificity" for particular genes may often be controlled by quantitative factors including enhancer activity and enhancer-promoter contact frequency, rather than by qualitative logic involving particular combinations of transcription factors at the enhancers and promoters. The ABC model, CRISPRi-FlowFISH, and other approaches to map enhancer function[9,13,36–39] provide a means to test this principle and to further refine our understanding of noncoding regulatory elements by mapping and modeling promoter-promoter regulation, functions of CTCF sites, and combinatorial effects of multiple enhancers in a locus.

Beyond its conceptual implications concerning gene regulation, the ABC model has important practical applications. Because it can make genome-wide predictions in a given cell type based on easily obtained epigenomic datasets, the ABC model provides a framework for mapping enhancer-gene connections across many cell types — including cell types that are difficult to directly manipulate with CRISPR. This suggests a systematic approach to decode transcriptional regulatory networks and to interpret the functions of noncoding genetic variants that influence human traits.

Methods

CRISPRi-FlowFISH screens

gRNA selection for CRISPRi-FlowFISH screens—We designed gRNAs within K562 candidate elements, and evaluated the specificity of gRNAs by exhaustively evaluating all potential off-target sites in the human genome (up to 5 mismatches), and selected only gRNAs that exceeded a specificity score >50 as previously described[41] and lacked homopolymer stretches of more than 7 As, Gs, or Cs, or 4 Ts (Supplementary Table 1). We targeted each element with many independent gRNAs (median = 55, Extended Data Fig. 2a), and required significant connections to show a consistent and significant effect across many gRNAs (see Analysis of CRISPRi-FlowFISH screens below).

Gene selection for CRISPRi-FlowFISH screens—We used a series of filters for each probeset and screen to ensure robust, comprehensive, and quantitative discovery of regulatory elements for each gene (Supplementary Fig. 1). We initially tested PrimeFlow probesets for genes expressed at >20 RPKM in K562 cells (GSE87257) in five genomic loci (Supplementary Fig. 2). We first screened probesets by flow cytometry and selected those with >2-fold signal vs unstained cells. We next performed a tiling CRISPRi-FlowFISH screen (see below) and focused our analysis on the screens that showed the following characteristics: (i) maximum unscaled knockdown among 20-gRNAs windows within 500 bp of the TSS >50%; and (ii) >80% power to detect a 25% effect in at least 80% of elements (see below). Based on these filters, we performed and analyzed CRISPRi-FlowFISH screens for 30 genes.

As a practical note, the signal in stained versus unstained cells appears to be a good predictor of successful CRISPRi-FlowFISH screens (Supplementary Fig. 1d). For example, of the 16 probesets for which we attempted screens with signal between 2 and 3, 7 (44%)

did not yield successful screens due to lack of specificity or power. Of the 27 probesets with signal >3, only 6 (22%) failed.

CRISPRi-FlowFISH screens—We cloned gRNA libraries purchased from CustomArray (now GenScript) for each of 5 genomic loci (Supplementary Fig. 2). We transduced these libraries (consisting of a single genomic locus and non-targeting gRNAs in the same pool) at low multiplicity of infection (MOI ≈ 0.3) into K562 cells harboring KRAB-dCas9, and selected for transduced cells as previously described[9]. In order to limit indirect effects or other changes in expression due to the expression of KRAB-dCas9, we used a dox-inducible system, inducing KRAB-dCas9 expression with 1 μg/ml doxycycline for 48 hours. We used 30 million cells for each screen.

We used the PrimeFlow RNA Assay Kit (Thermo Fisher; Catalog number: 88–18005) according to the manufacturer's instructions with some modifications. Specifically, we used 10 million cells per reaction (three reactions per screen) and performed five total washes with 35 °C wash buffer following the staining protocol. We stained each sample for the gene of interest with an Alexa Fluor 647 (AF647, "Type 1") probeset and against a positive control housekeeping gene with Alexa Fluor 488 (AF488, "Type 4"). For most screens we used control gene *RPL13A*, but because *BAX*, *BCAT2, FTL*, *FUT1, NUCB1*, and *PPP1R15A* are <700 kb from *RPL13A*, we used *ACTB* for these. Probesets used are listed in Supplementary Table 2.

Cell sorting—We diluted the stained cells in PBS with 0.5% BSA to a concentration of 2×10^7 cells/ml and filtered using a 30-μm filter (CellTrics, Catalog number 04–004-2326). We sorted 30 million cells for each screen into six bins based on the fluorescence intensity of target genes using the Astrios EQ Sorter (Beckman Coulter B25982). To control for differences in staining efficiency for each cell, we normalized the fluorescence associated with the gene of interest to that of the control gene (Extended Data Fig. 1c,d). Specifically, we used the color compensation tool in the Astrios control software (Summit v6.3.1) to subtract a portion of each cell's AF647 signal based on the intensity of its AF488 signal. This portion was selected such that the mean AF488 signal in the top and bottom 25% of cells based on AF647 was within 10%. If necessary, we then reduced the level of compensation until the fraction of cells with AF647 signal equal to 0 was no more than 5%. We set the gates for each bin on the compensated signal to capture 10% of the cells according to the percentiles (i) 0–10%, (ii) 10–20%, (iii) 35–45%, (iv) 55–65%, (v) 80–90%, and (vi) 90–100% (Extended Data Fig. 1e).

Genomic DNA extraction and gRNA sequencing—We collected the sorted cells by centrifugation at 800*g* for 5 minutes, resuspended cells in 100 μL of lysis buffer (50 mM Tris-HCl, pH 8.1, 10 mM EDTA, 1% SDS), and incubated at 65 °C for 10 minutes for reverse crosslinking. Once the samples cooled to 37 °C, we added 2 μl of RNase Cocktail (Invitrogen, catalog number AM2286), mixed well, and incubated the mixture at 37 °C for 30 minutes. Finally, we added 10 μl Proteinase K (NEB, catalog number P8107S), mixed well, and incubated the mixture at 37 °C for 2 hours followed by incubation at 95 °C for 20 min. We extracted genomic DNA using Agencourt XP (SPRI) beads (Beckman Coulter). We sequenced gRNA integrations as previously described[9].

Analysis of CRISPRi-FlowFISH screens—To determine the effects of each gRNA on fluorescence, we used a maximum likelihood estimation (MLE) method. First, we normalized gRNA frequencies in each bin by dividing each gRNA count by the total read count for all gRNAs in that bin and summed normalized counts across PCR replicates. Next, we used the limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm MLE method in the R stats4 package to fit the read counts in each fluorescence bin to the log-normal distribution that would have most likely produced the observed counts in the bins. The effect size is from the mean of the log-normal fit for a given gRNA divided by the mean of the log-normal fits across all negative control gRNAs. We assumed the gRNAs targeting the TSS of the assayed gene have a "true" effect size of 85% (based on previous observations that show CRISPRi effects of 80–90% across a panel of genes[8]), but that some portion of the FlowFISH signal is due to non-specific binding of the probe. Accordingly, we scaled the effect size of each gRNA within each screen linearly so that the strongest 20-gRNA window within 500 bp of the target gene's TSS has effect size 85%. We then averaged the effect sizes of individual gRNAs across replicates.

To identify elements affecting the expression of the assayed gene, we used a two-sided *t*-test to determine whether the mean effect size of the gRNAs in each candidate element deviated significantly from the mean of scrambled-sequence, control gRNAs contained in the same population of cells. We computed the FDR for elements using the Benjamani-Hochberg method applied per gene, and used an FDR threshold of 0.05 to call significant E-G connections. We report the result for each E-G connection in Supplementary Table 3.

We excluded certain E-G pairs measured with CRISPRi-FlowFISH from further analysis. E-G pairs were excluded if the pair met any of the below criteria:

- **i.** There was less than 80% power to detect a 25% effect for this E-G pair.
- **ii.** The element overlapped the gene's promoter.
- **iii.** The element was within the gene body or was within 2 kb of the 3' end of the gene.

Activity by Contact (ABC) model

Defining candidate elements—We defined candidate regulatory elements in 5 human cell types (K562, GM12878, NCCIT, LNCaP, liver tissue), and 1 mouse cell type (embryonic stem cells; mESCs).

For K562, we concatenated all peaks called by ENCODE in both replicate DNase-seq experiments (Supplementary Table 4). Given that the ENCODE peaks were initially 150 bp in length, we extended each of these peaks 175 bp to arrive at candidate elements that were 500 bp in length. We then removed any peaks overlapping regions of the genome that have been observed to accumulate anomalous number of reads in epigenetic sequencing experiments ('blacklisted regions'[42,43] downloaded from https://sites.google.com/site/anshulkundaje/projects/blacklists). To this peak list, we added 500 bp regions centered on the transcription start site of all genes. Any overlapping regions resulting from these

additions or extensions were merged. In total, this procedure resulted in 162,181 candidate regions in K562 whose average length is 576 bp (Extended Data Fig. 2b).

For GM12878, NCCIT, LNCaP, liver tissue, and mESCs, we called peaks using MACS2 on the first replicate of either DNase-seq or ATAC-seq as a measure of chromatin accessibility (Supplementary Table 4). We initially considered all peaks with $P < 0.1$ and removed peaks overlapping blacklisted regions. In order to approximately match the number of candidate elements considered in K562, we then counted DNase-seq (or ATAC-seq) reads overlapping these peaks and kept the 150,000 peaks with the highest number of read counts. We then resized these peaks to be 500 bp in length centered on the peak summit. To this peak list, we added 500-bp regions centered on the transcription start site of all genes. Any overlapping regions resulting from these additions or extensions were merged.

We define these extended and merged peaks as *candidate elements.* We classified each candidate element as a promoter, genic, or intergenic element. Promoter elements are those that are within 500 bp of any annotated TSS (see Supplementary Methods). Genic elements are those contained within any annotated gene body. Intergenic elements are all other candidate elements. We denote any genic or intergenic element as a 'distal' element ("DE"). For the elements we or others studied experimentally, we manually confirmed the classification by inspecting CAGE and PROseq data, and in 11 cases we adjusted the annotation based on transcriptomic data and to match the previously reported annotations (Supplementary Table 5).

Calculating enhancer activity from DHS and H3K27ac ChIP-seq signals—We estimated enhancer activity of candidate elements using a combination of quantitative DNase-seq and H3K27ac ChIP-seq signals. DNase accessibility and acetylation of H3K27 are commonly used to identify enhancer elements[44,45], and are predictive of the expression of nearby genes and enhancer activity in plasmid based reporter assays[46–48]. Quantile normalization of epigenetic signals is used to facilitate comparison of ABC Scores across cell types (see Supplementary Methods).

DNase peaks were extended 175 bp because H3K27ac ChIP-seq signals are strongest on the nucleosomes flanking the nucleosome-free DHS peak. We computed the geometric mean of DNase-seq and H3K27ac ChIP-seq signals because we expect that strong enhancers should have strong signals for both, and that elements that have only one or the other likely represent other types of elements. (Elements with strong DNase-seq signal but no H3K27ac ChIP-seq signal might be CTCF-bound topological elements. Elements with strong H3K27ac signal but no DNase-seq signal might be sequences that are close to strong enhancers but do not themselves have enhancer activity, due to the spreading H3K27ac signal over hundreds to thousands of bp.) We report sources of epigenomic data in Supplementary Table 4. Where replicate experiments are listed, we averaged the signal in each element across the replicates unless otherwise stated.

We note that this calculation of enhancer activity is the same for a given element across all genes. This means that the model assumes that an enhancer has the same "Activity" for every promoter (*i.e.*, no differences due to biochemical specificity).

Calculating contact frequency from cell-type specific Hi-C data—In our initial analysis in K562 cells, we obtained the Contact component of the ABC score for E-G pairs from Hi-C data in K562 cells, using the quantitative signal observed in the 5-kb x 5-kb bin containing the center of E and TSS of G.

Specifically, we used KR-normalized Hi-C contact maps at 5 kb resolution, and processed these maps in two steps:

1. For rows and columns corresponding to KR normalization factors less than 0.1 we did not use KR normalization (these typically correspond to 5 kb bins with very few reads). We instead linearly interpolated the Hi-C signal in these bins from the neighboring bins (with KR normalization factors > 0.1)
2. Each diagonal entry of the Hi-C matrix was replaced by the maximum of its four neighboring entries. The diagonal of the Hi-C contact map corresponds to the measured contact frequency between a 5-kb region of the genome and itself. The signal in bins on the diagonal can include restriction fragments that self-ligate to form a circle, or adjacent fragments that re-ligate, which are not representative of contact frequency. Empirically, we observed that the Hi-C signal in the diagonal bin was not well correlated with either of its neighboring bins and was influenced by the number of restriction sites contained in the bin.

We then computed Contact for an E-G pair by rescaling the data as follows:

1. We extracted the row of the processed Hi-C matrix that contains the TSS of G. For convenience, the row is rescaled so that the maximum value is 100.
2. We set the Contact of the E-G pair to the Hi-C signal at the bin of this row corresponding to the midpoint of E.
3. We added a small adjustment ("pseudocount") to ensure that the contact frequency for each E-G pair is non-zero. For E-G pairs within 1 Mb, the adjustment is equal to the expected contact frequency at 1Mb (as predicted by the power-law relationship between contact frequency and genomic distance; Supplementary Methods), and for E-G pairs at distance d (d > 1 Mb), the adjustment is equal to the expected contact at distance d. In each case the adjustment was scaled to be in the same units as described in (i). Adding the adjustment sometimes results in a quantitative Contact greater than 100; in such cases, the Contact is reduced to 100.

Calculating the contribution of one candidate element relative to others in the region—To calculate the relative effect of each element to the expression of a gene, we normalized the Activity by Contact of one element for a given gene to the sum of the Activity by Contact of other nearby elements. We included all elements within 5 Mb of the gene's promoter in this calculation, and found that the performance of the model was not sensitive to this parameter (Supplementary Methods and Supplementary Fig. 5). We also included each gene's own promoter as an element in the denominator of the ABC score. This is because the promoters of genes are known to have the potential to act as enhancers for other genes and are frequently bound by activating TFs[26,49]. Thus, the ABC score

considers that the element near the TSS can have enhancer activity that contributes to the total regulatory signals relevant for that gene. We note that this normalization encodes the simplifying assumption that each element contributes independently and additively to gene expression. Based on the performance of the model in distinguishing significant DE-G pairs, this assumption appears sufficient for practical performance of the model. This first-order ABC model provides a foundation for incorporating higher-order effects such as the potential for nonlinear effects of multiple enhancers in a locus.

ABC scores are provided for all tested connections in Supplementary Table 6.

Genome build

All coordinates in the human genome are reported using build hg19, and all coordinates in the mouse genome are reported using build mm9.

Code Availability Statement

Code to calculate the ABC model is available at https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction.

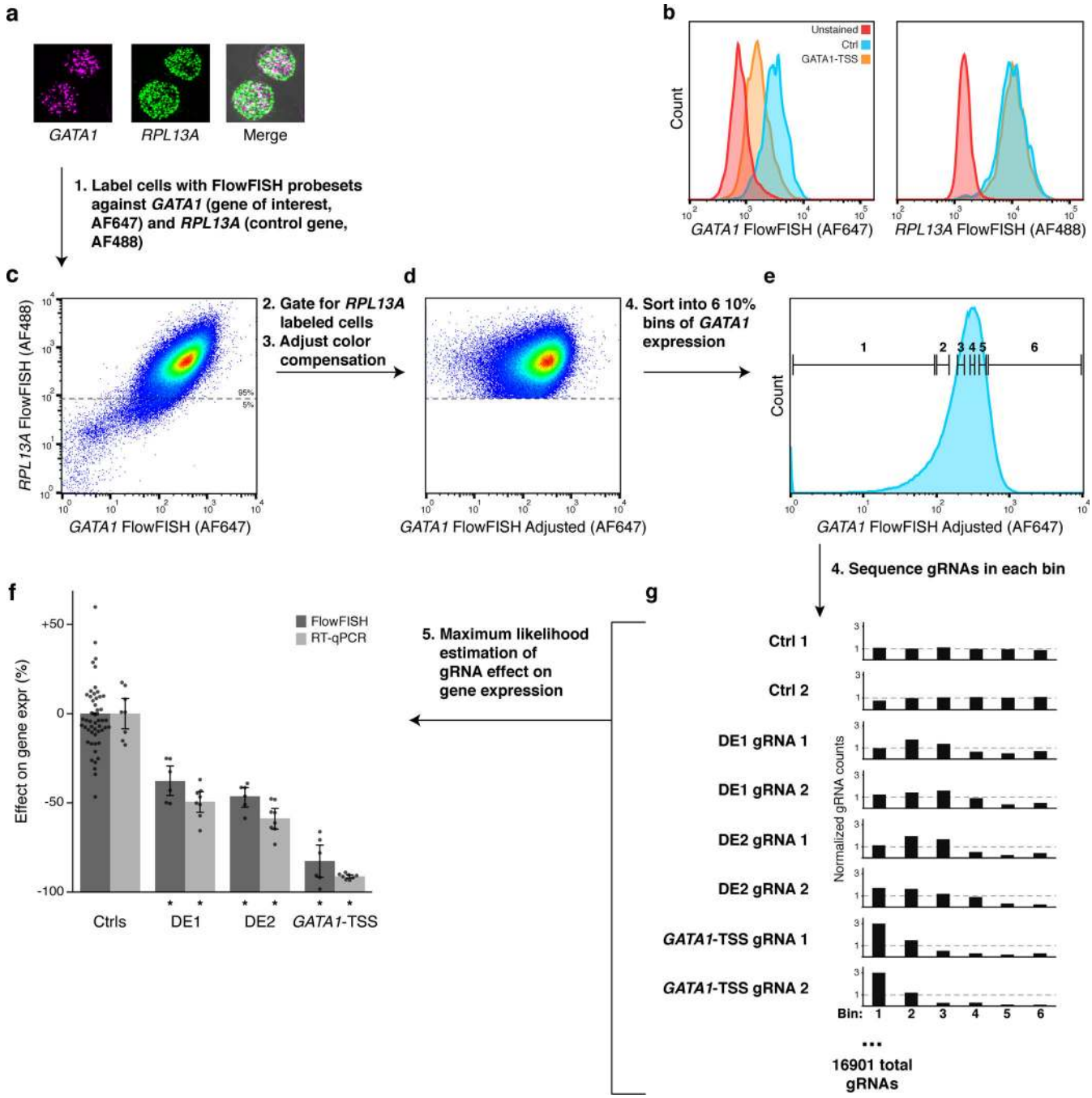
Reporting Summary

Further information on research design is available in the Reporting Summary linked to this study.

Data Availability Statement

Genome-wide ABC predictions for the six cell types considered in this study (K562, mESC, GM12878, NCCIT, LNCAP, liver tissue) and raw counts from CRISPRi-FlowFISH are available on the Open Science Framework https://osf.io/uhnb4/. ChIP-seq, ATAC-seq, Hi-C, and RNA-seq data from this study are available at GSE118912.

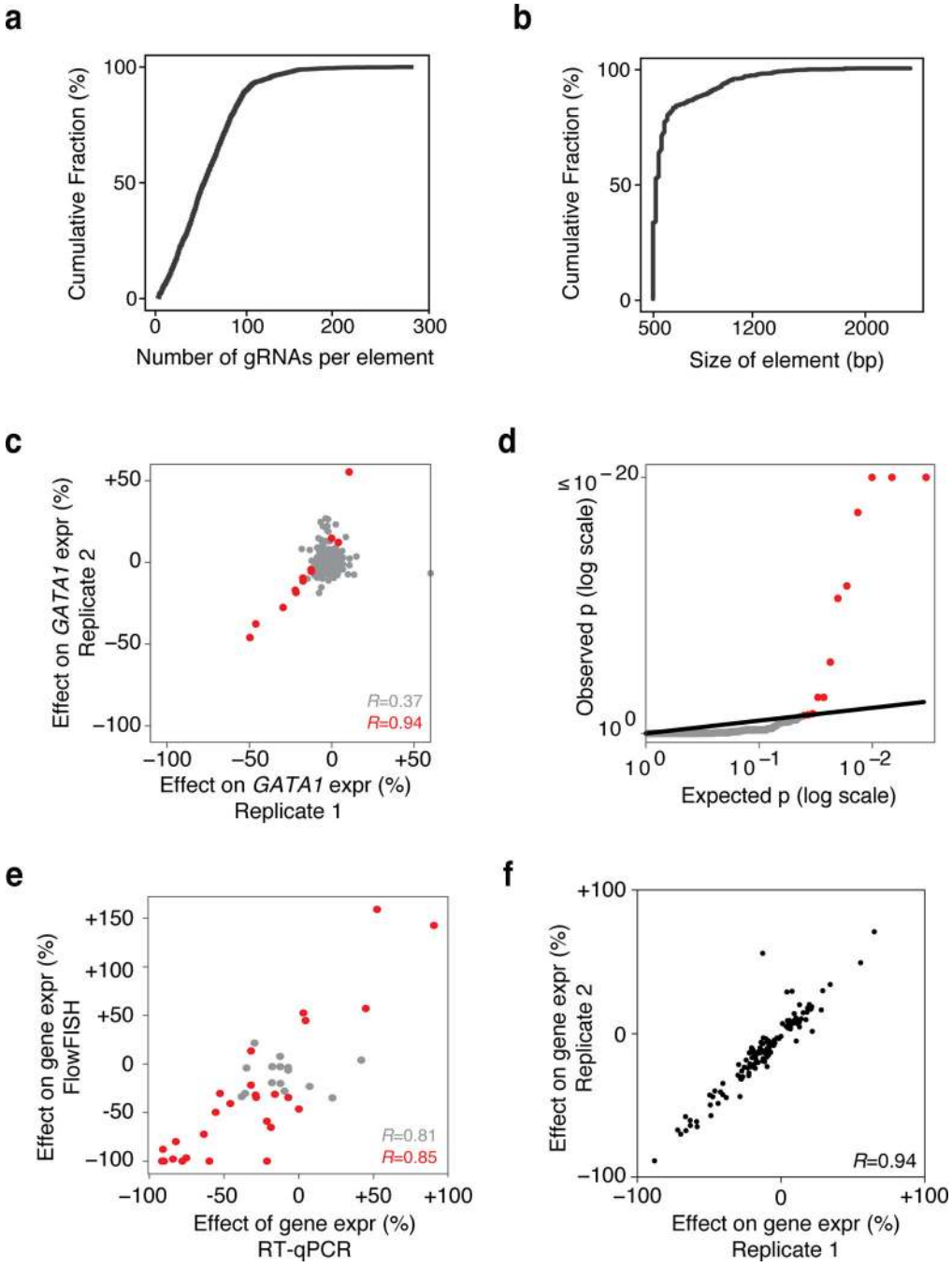
Extended Data



Extended Data Fig. 1. Sorting and sequencing strategy for CRISPRi-FlowFISH Screens

a, K562 cells labeled with FlowFISH probesets against *RPL13A* (control gene) and *GATA1* (gene of interest) imaged by fluorescence microscopy. **b,** Histograms of FlowFISH signal (arbitrary units of fluorescence) for *GATA1* (left) and *RPL13A* (right) in unlabeled K562s (red), K562s stained for *GATA1* expressing a gRNA against the *GATA1*-TSS (orange), or a non-targeting Ctrl gRNA (blue). Results typical of cells across 2 independent samples (**a,b**). **c,** Scatterplot of FlowFISH fluorescent signal for *RPL13A* versus *GATA1*. **d,** Cells in **c** with cells unstained for *RPL13A* (below dotted line in **c**) removed and using the color

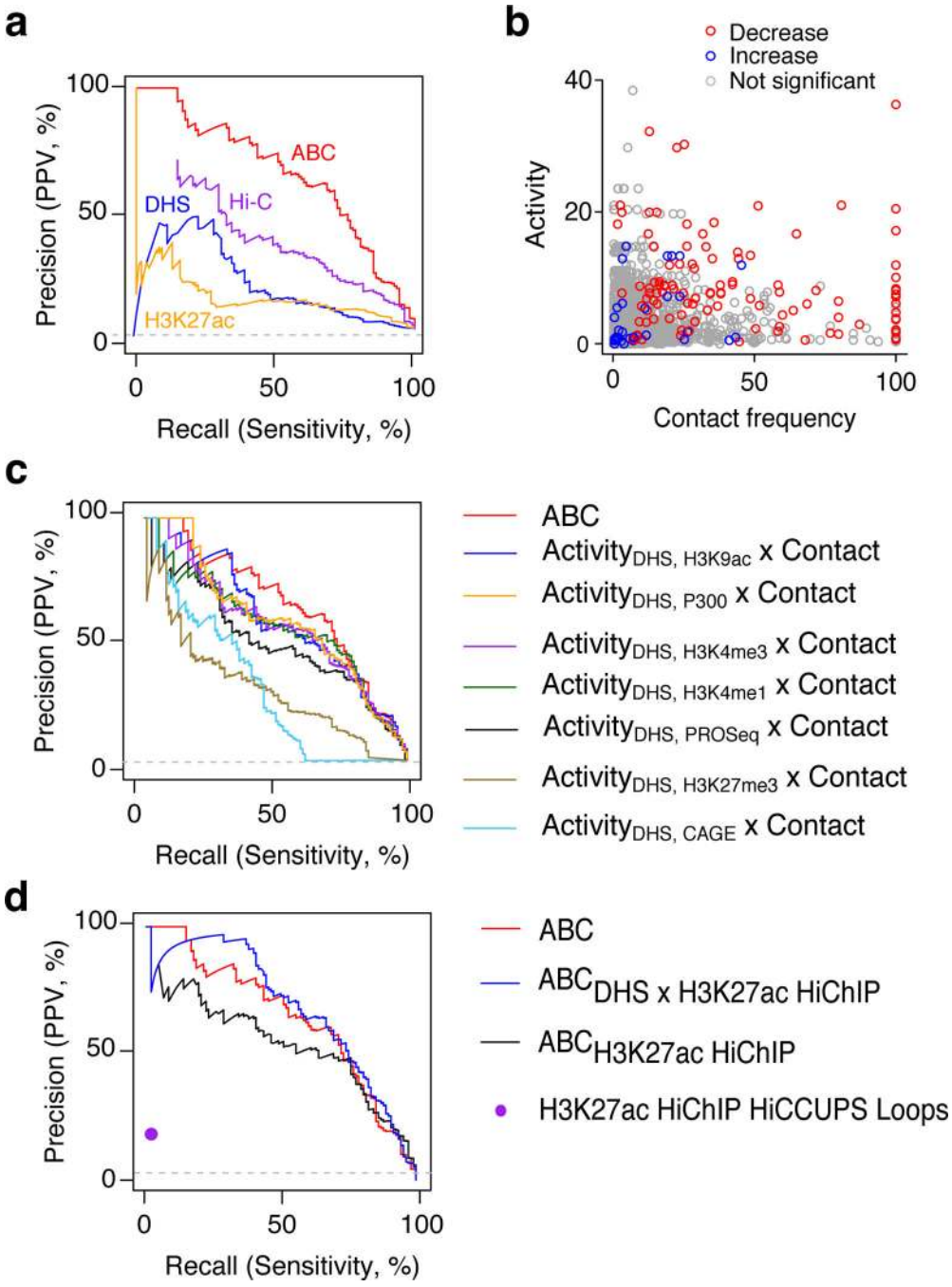
compensation tool to reduce the correlation between the control gene and gene of interest (see Methods). **e,** Binning strategy for sorting FlowFISH-labeled cells into 6 bins each containing 10% of the cells. Typical results from 3 independent *GATA1* CRISPRi-FlowFISH screens (**c-e**). **f,** Effect on gene expression as measured by CRISPRi-FlowFISH (dark grey) and RT-qPCR (light grey). Error bars: 95% confidence intervals for the mean of 2 gRNAs per target, 3505 Ctrl gRNAs for FlowFISH (a random 50 shown), and 6 Ctrl gRNAs for RT-qPCR. $n = 3$ independent experiments per gRNA for CRISPRi-FlowFISH screens. $n = 4$ independent samples per gRNA for RT-qPCR. $*P < 0.05$ in 2-sided *t*-test versus Ctrl. *P*-values, test statistics, confidence intervals, effect sizes, and degrees of freedom are available in Supplementary Table 3. **g,** Counts in each of the 6 bins for single gRNAs targeting the *GATA1* TSS, two *GATA1* enhancers (DE1 and DE2) identified in Fulco *et al.*, and representative negative controls (Ctrl).



Extended Data Fig. 2. CRISPRi-FlowFISH reproducibly quantifies effects of regulatory elements

a, Cumulative distribution plot of the number of gRNAs in each tested candidate element. **b,** Cumulative distribution plot of the width of each tested candidate element. **c,** Correlation between independent CRISPRi-FlowFISH screens for *GATA1*. Red points denote elements significantly affecting expression. Pearson $R = 0.94$ for significant elements, 0.37 for all elements. **d,** Quantile-quantile plot for *GATA1* CRISPRi-FlowFISH screen. Red points denote elements significantly affecting expression. Vertical axis capped at 10^{-20}. **e,** Pearson correlation between effect on gene expression as measured by CRISPRi-FlowFISH

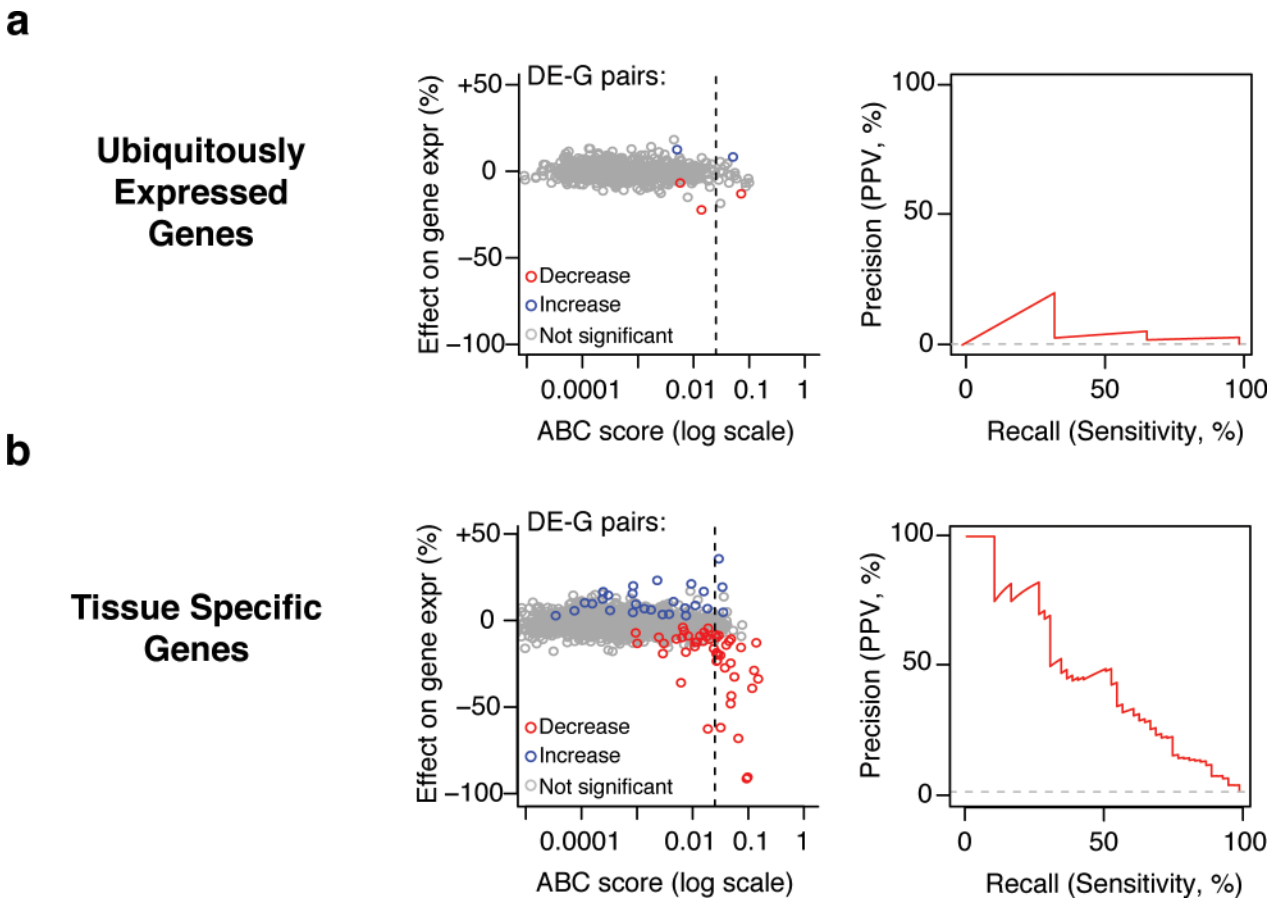
screening and RT-qPCR for 42 E-G pairs tested by both methods. Value is the mean effect of the two gRNAs for each element. **f,** Pearson correlation between effects on gene expression for all significant E-G pairs measured in biologically independent CRISPRi-FlowFISH screens. *P*-values, test statistics, confidence intervals, effect sizes, and degrees of freedom for all panels are available in Supplementary Table 3.



Extended Data Fig. 3. Investigating components of the ABC score

a, Precision-recall curves for classifying regulatory DE-G pairs, comparing each of the components of the ABC score. **b,** Scatterplot of Activity and Contact frequency for each tested DE-G pair. KR-normalized Hi-C contact frequencies are scaled for each gene so that the maximum score of an off-diagonal bin is 100 (see Methods). **c,** Precision-recall curves comparing different measures of Activity. $Activity_{Feature1,Feature2}$ = sqrt(Feature1 RPM x Feature2 RPM). (ABC score corresponds to $Activity_{DHS,H3K27ac}$ x Contact). **d,** Precision-recall curves for the ABC model using H3K27ac HiChIP. $ABC_{DHS\ x\ H3K27ac\ Hi\text{-}ChIP}$

corresponds to a predictive model whose score is proportional to the DHS signal at the candidate element multiplied by the H3K27ac Hi-ChIP signal between the element and gene promoter (see Supplementary Methods). $ABC_{H3K27ac\ Hi\text{-}ChIP}$ is the same as above but only uses the existence of the DHS peak as opposed to the quantitative signal in the DHS peak. H3K27ac HiChIP HiCCUPS Loops is the HiCCUPS loop calls derived from the H3K27ac HiChIP experiment (see Supplementary Methods). ABC corresponds to $ABC_{sqrt(DHS\ x\ H3K27ac)\ x\ Hi\text{-}C}$. These results suggest that the ABC score computed using H3K27ac HiChIP data is an effective predictor of regulatory enhancer-gene connections.



Extended Data Fig. 4. Tissue-specific genes have more distal enhancers than ubiquitously expressed genes

a, Left: Comparison of ABC scores (predicted effect) with observed changes in gene expression upon CRISPR perturbations. Each dot represents one tested DE-G pair where G is a ubiquitously expressed gene. Right: precision-recall curve for ABC score in classifying regulatory DE-G pairs where each G is a ubiquitously expressed gene. **b,** Same as **a** for tissue-specific genes. All panels include only the subset of our dataset for which we have CRISPRi tiling data to comprehensively identify all enhancers that regulate each gene (30 genes from this study, 2 from previous studies; see Supplementary Methods).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J. Chen, A. Chow, B. Cleary, C. de Boer, A. Dixit, M. Guttman, R. Herbst, K. Mualim, S. Rao, J. Ray, S. Reilly, R. Tewhey, J. Ulirsch, and C. Vockley for discussions and J. Marshall for assistance with fluorescence microscopy. FACS sorting was performed at the Broad Institute FACS Core by P. Rogers, S. Saldi, and C. Otis. This work was supported by funds from the Broad Institute (E.S.L.) and by NIH NHGRI Grant No. 1K99HG009917-01 to J.M.E. J.M.E. is supported by the Harvard Society of Fellows. S.R.G. is supported by National Institute of General Medical Sciences grant T32GM007753. E.L.A. was supported by an NSF Physics Frontiers Center Award (PHY1427654), the Welch Foundation (Q-1866), a USDA Agriculture and Food Research Initiative Grant

(2017-05741), an NIH 4D Nucleome Grant (U01HL130010), and an NIH Encyclopedia of DNA Elements Mapping Center Award (UM1HG009375).

References

1. Maurano MT et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–5 (2012). [PubMed: 22955828]
2. Visel A, Rubin EM & Pennacchio LA Genomic views of distant-acting enhancers. Nature 461, 199–205 (2009). [PubMed: 19741700]
3. Spitz F & Furlong EE Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13, 613–26 (2012). [PubMed: 22868264]
4. Shlyueva D, Stampfel G & Stark A Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15, 272–86 (2014). [PubMed: 24614317]
5. van Arensbergen J, van Steensel B & Bussemaker HJ In search of the determinants of enhancer-promoter interaction specificity. Trends Cell Biol 24, 695–702 (2014). [PubMed: 25160912]
6. Bulger M & Groudine M Functional and mechanistic diversity of distal transcription enhancers. Cell 144, 327–39 (2011). [PubMed: 21295696]
7. Thakore PI et al. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat Methods 12, 1143–9 (2015). [PubMed: 26501517]
8. Gilbert LA et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154, 442–51 (2013). [PubMed: 23849981]
9. Fulco CP et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science 354, 769–773 (2016). [PubMed: 27708057]
10. Xu J et al. Developmental control of polycomb subunit composition by GATA factors mediates a switch to non-canonical functions. Mol Cell 57, 304–316 (2015). [PubMed: 25578878]
11. Ulirsch JC et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell 165, 1530–1545 (2016). [PubMed: 27259154]
12. Wakabayashi A et al. Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. Proc Natl Acad Sci U S A 113, 4434–9 (2016). [PubMed: 27044088]
13. Klann TS et al. CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nat Biotechnol 35, 561–568 (2017). [PubMed: 28369033]
14. Liu SJ et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science 355(2017).
15. Xie S, Duan J, Li B, Zhou P & Hon GC Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. Mol Cell 66, 285–299.e5 (2017). [PubMed: 28416141]
16. Huang J et al. Dissecting super-enhancer hierarchy based on chromatin interactions. Nat Commun 9, 943 (2018). [PubMed: 29507293]
17. Qi Z et al. Tissue-specific gene expression prediction associates vitiligo with SUOX through an active enhancer. bioRxiv 337196 (2018).
18. Rao SS et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–80 (2014). [PubMed: 25497547]
19. Whalen S, Truty RM & Pollard KS Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet 48, 488–96 (2016). [PubMed: 27064255]
20. Cao Q et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat Genet 49, 1428–1436 (2017). [PubMed: 28869592]
21. Sanborn AL et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A 112, E6456–65 (2015). [PubMed: 26499245]
22. Yardimci G et al. Measuring the reproducibility and quality of Hi-C data. Genome Biol. 20, 57 (2019). [PubMed: 30890172]

23. Li Y et al. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. PLoS One 9, e114485 (2014). [PubMed: 25486255]
24. Zhou HY et al. A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. Genes Dev 28, 2699–711 (2014). [PubMed: 25512558]
25. Blinka S, Reimer MH Jr., Pulakanti K & Rao S Super-enhancers at the Nanog locus differentially regulate neighboring pluripotency-associated genes. Cell Rep 17, 19–28 (2016). [PubMed: 27681417]
26. Engreitz JM et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature 539, 452–455 (2016). [PubMed: 27783602]
27. Rajagopal N et al. High-throughput mapping of regulatory DNA. Nat Biotechnol 34, 167–74 (2016). [PubMed: 26807528]
28. Tewhey R et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell 165, 1519–1529 (2016). [PubMed: 27259153]
29. Moorthy SD et al. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. Genome Res 27, 246–258 (2017). [PubMed: 27895109]
30. Mumbach MR et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet 49, 1602–1612 (2017). [PubMed: 28945252]
31. Fuentes DR, Swigut T & Wysocka J Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. Elife 7, e35989 (2018). [PubMed: 30070637]
32. Spisak S et al. CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. Nat Med 21, 1357–63 (2015). [PubMed: 26398868]
33. Wang X et al. Interrogation of the atherosclerosis-associated SORT1 (Sortilin 1) locus with primary human hepatocytes, induced pluripotent stem cell-hepatocytes, and locus-humanized mice. Arterioscler Thromb Vasc Biol 38, 76–82 (2018). [PubMed: 29097363]
34. Musunuru K et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466, 714–9 (2010). [PubMed: 20686566]
35. Haberle V & Stark A Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol 19, 621–637 (2018). [PubMed: 29946135]
36. Gasperini M et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell 176, 377–390.e19 (2019). [PubMed: 30612741]
37. Gasperini M et al. CRISPR/Cas9-mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. Am J Hum Genet 101, 192–205 (2017). [PubMed: 28712454]
38. Sanjana NE et al. High-resolution interrogation of functional elements in the noncoding genome. Science 353, 1545–1549 (2016). [PubMed: 27708104]
39. Canver MC et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527, 192–7 (2015). [PubMed: 26375006]
40. Li G et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84–98 (2012). [PubMed: 22265404]

Methods-only references

41. Hsu PD et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol 31, 827–32 (2013). [PubMed: 23873081]
42. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]
43. Amemiya HM, Kundaje A & Boyle AP The ENCODE Blacklist: identification of problematic regions of the genome. Sci Rep 9, 9354 (2019). [PubMed: 31249361]
44. Gross DS & Garrard WT Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 57, 159–97 (1988). [PubMed: 3052270]
45. Rada-Iglesias A et al. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279–83 (2011). [PubMed: 21160473]

46. Visel A et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457, 854–8 (2009). [PubMed: 19212405]

47. Karlic R, Chung HR, Lasserre J, Vlahovicek K & Vingron M Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A 107, 2926–31 (2010). [PubMed: 20133639]

48. Mendenhall EM & Bernstein BE Chromatin state maps: new technologies, new insights. Curr Opin Genet Dev 18, 109–15 (2008). [PubMed: 18339538]

49. Dao LTM et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. Nat Genet 49, 1073–1081 (2017). [PubMed: 28581502]

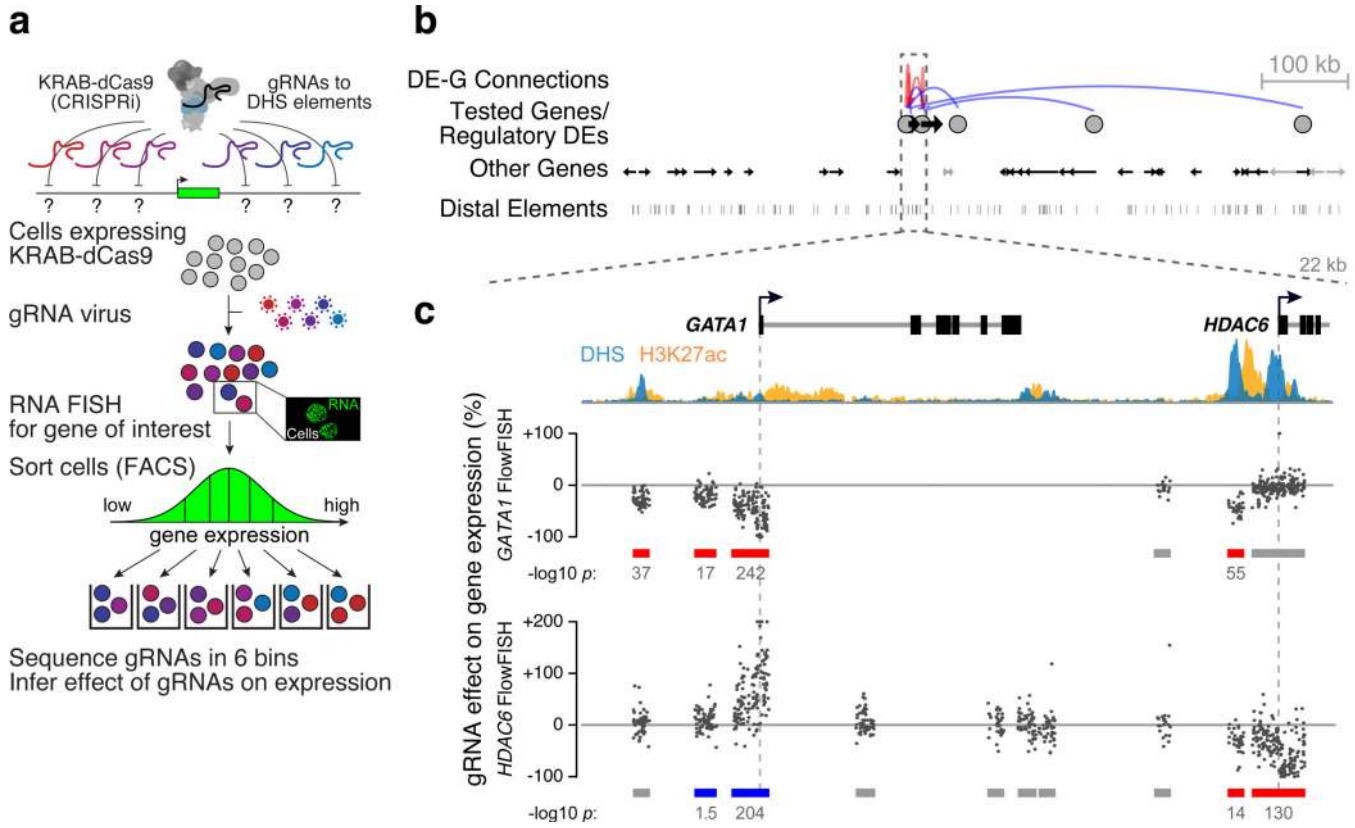


Fig. 1 |. CRISPRi-FlowFISH identifies regulatory elements for *GATA1* and *HDAC6*.

a, CRISPRi-FlowFISH method for identifying gene regulatory elements. Cells expressing KRAB-dCas9 are infected with a pool of gRNAs targeting DHS elements near a gene of interest, labeled using RNA FISH against that gene, and sorted into bins of fluorescence signal by FACS. The quantitative effect of each gRNA on the expression of the gene is determined by sequencing the gRNAs within each bin. Inset: example of K562 cells labeled for *RPL13A*. **b,** Distal elements affecting *GATA1* and *HDAC6* expression in K562 cells. Genes expressed in K562 cells are shown in black; those not expressed are shown in grey. Red/blue arcs: perturbation of a DE resulted in a significant decrease/increase in the expression of the tested gene. Grey circles are DEs where perturbation with CRISPRi affects the expression of at least one tested gene as measured by CRISPRi-FlowFISH. Distal Elements are DHS peaks. See Supplementary Figure 2a for the full tested region spanning 4 Mb. **c,** Close-up on region containing *GATA1* and *HDAC6*. Points represent the effect on gene expression of a single gRNA. *HDAC6* vertical axis capped at 200%. Grey, red, and blue bars: DHS elements in which CRISPRi leads to no detectable change (grey), or a significant decrease (red) or increase (blue) in expression. Elements overlapping the assayed gene are excluded from analyses because recruitment of KRAB-dCas9 in a gene body directly interferes with transcription[9]. Such elements are included in analyses for other genes, as shown for the elements overlapping *GATA1*.

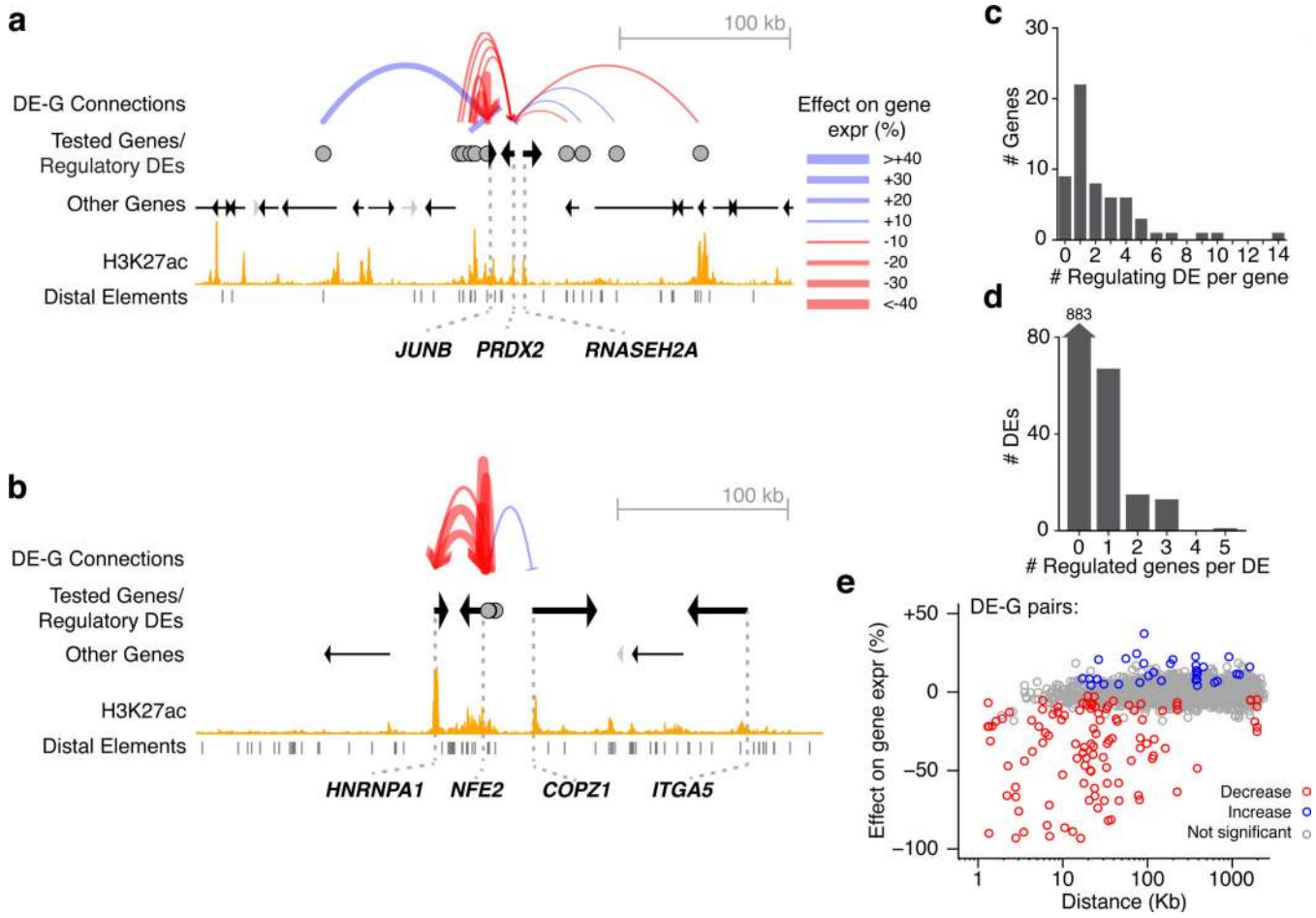


Fig. 2 |. CRISPRi-FlowFISH produces regulatory maps of DE-G connections in multiple loci.

a, Example of CRISPRi-FlowFISH screen data. DE-G connections are elements affecting the expression of *JUNB*, *PRDX2*, and *RNASEH2A* in CRISPRi-FlowFISH screens in K562 cells. Red/blue arcs: perturbation of a DE resulted in a significant decrease/increase in the expression of the tested gene. Width of arc corresponds to effect size. Distal elements are DHS peaks. Tested genes refer to genes for which we performed CRISPRi-FlowFISH experiments. See Supplementary Figure 2b for the full tested region spanning 1.4 Mb. **b,** Same as **a** for the genes *HNRNPA1, NFE2*, *COPZ1*, and *ITGA5.* See Supplementary Figure 2c for the full tested region spanning 1.2 Mb. **c,** Histogram of the number of distal elements affecting each gene in our dataset. Panels **a-e** include both FlowFISH data from this study and tested pairs from other studies. See Supplementary Figure 3 for plots including FlowFISH data only. **d,** Histogram of the number of genes affected by each distal element tested in our dataset. **e,** Comparison of genomic distance with observed changes in gene expression upon CRISPR perturbations. Each dot represents one tested DE-G. Red/blue dots: connections where perturbation resulted in a significant decrease/increase in the expression of the tested gene. Grey dots: no significant effect.

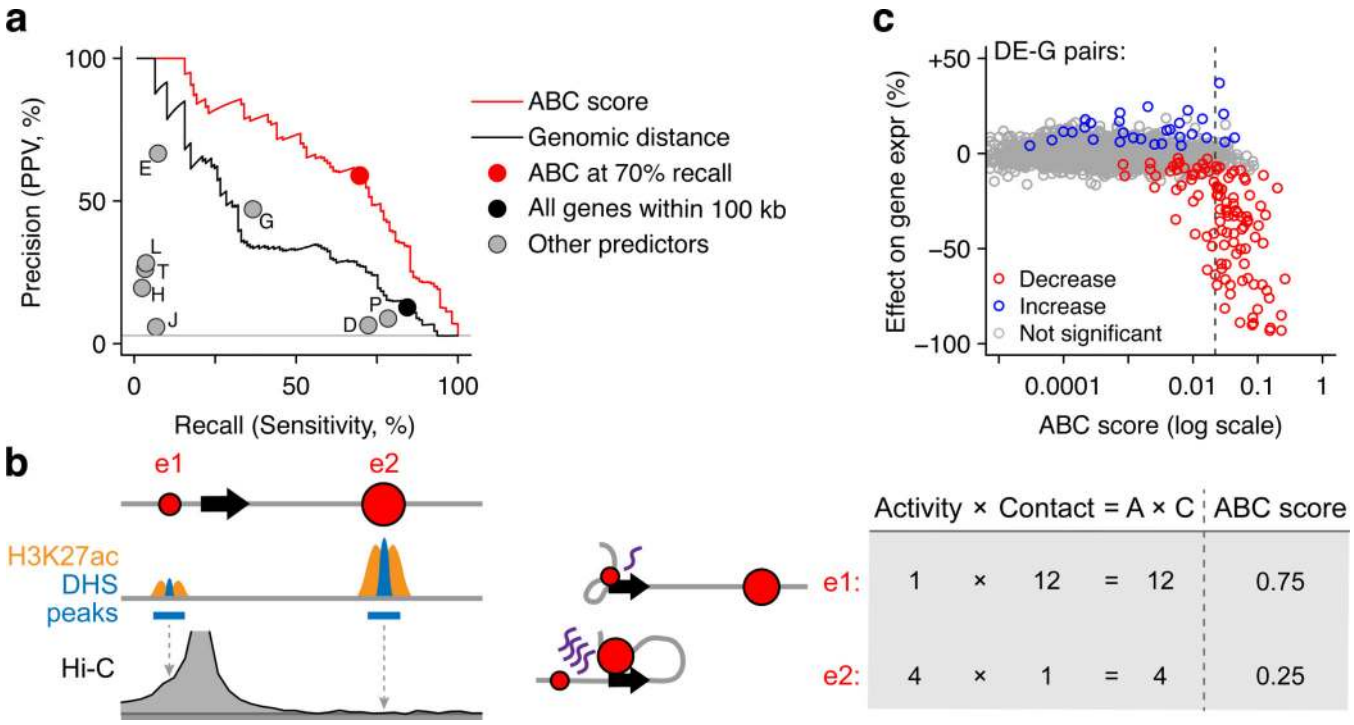


Fig. 3 |. The ABC model predicts the target genes of enhancers.

a, Precision-recall plot for classifiers of DE-G pairs. Positive DE-G pairs are those where perturbation of the distal element significantly decreases expression of the gene. Curves represent the performance for predicting significant decreases in expression for DE-G pairs based on thresholds on the ABC score (red) and genomic distance between the DE and the TSS of the gene (black). Circles represent the performance of various predictors in which DEs are assigned to: the TSS of the closest expressed gene ("G"); all promoters within 100 kb (black), genes predicted by the algorithms TargetFinder ("T")[19] or JEME ("J")[20]; promoters in same Hi-C contact domain ("D"); and promoters at the opposite anchors of Hi-C "loops" ("L"), RNA Polymerase II ChIA-PET loops ("P")[40], or H3K27ac HiChIP "loops" ("H")[30]; or assigning each expressed gene to the closest DE ("E"). **b,** Calculation of the ABC score (see Methods). Values for DHS, H3K27ac, and Hi-C are presented in arbitrary units and are not to scale. **c,** Comparison of ABC scores (predicted effect) with observed changes in gene expression upon perturbations. Each dot represents one tested DE-G pair. Red/blue dots: connections where perturbation resulted in a significant decrease/increase in the expression of the tested gene. Grey dots: no significant effect. Dotted black line marks 70% recall, corresponding to the red dot in **a**.

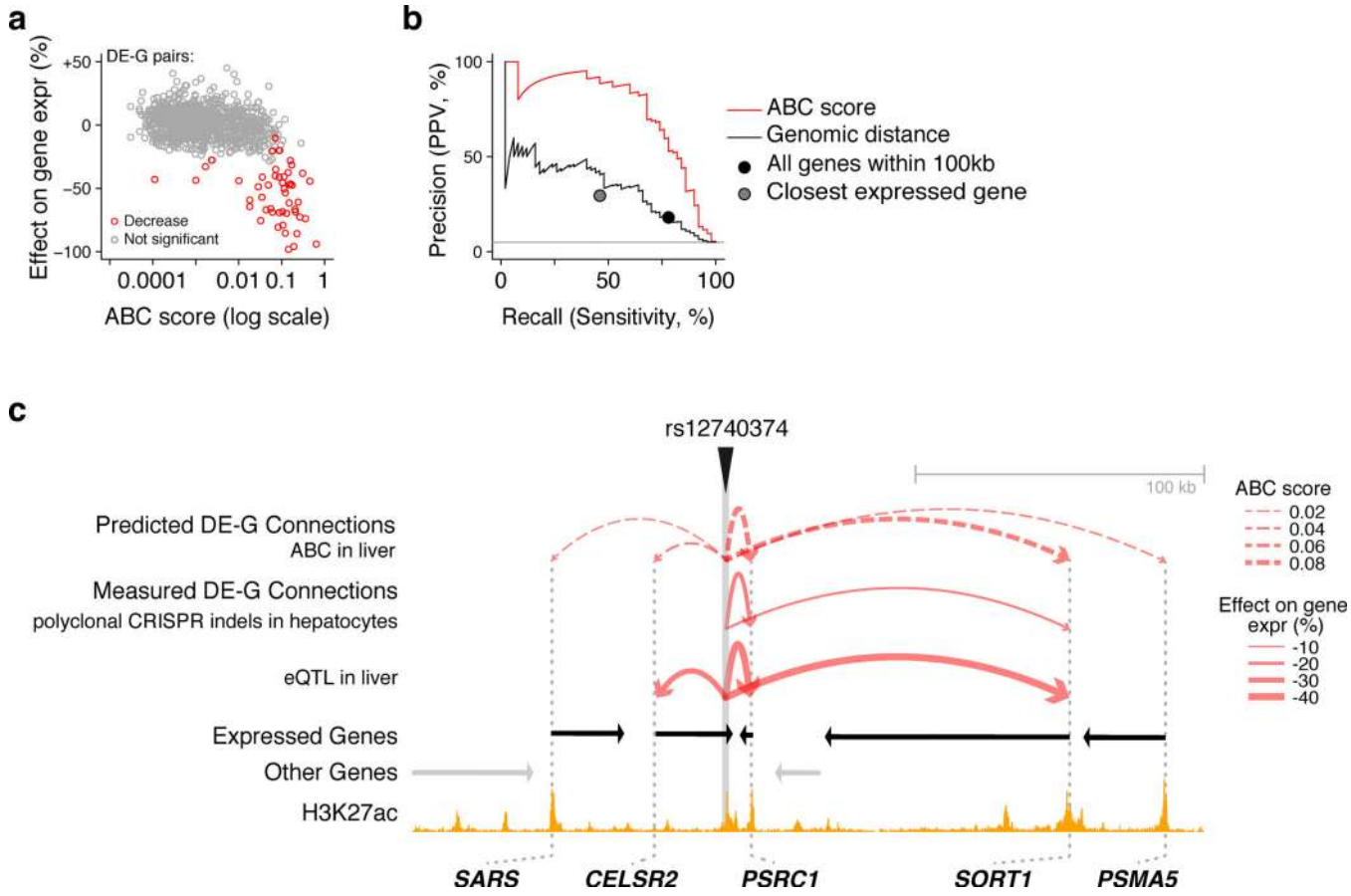


Fig. 4 |. The ABC model generalizes across cell types.

a, Comparison of ABC scores (predicted effect) with observed changes in gene expression upon perturbations in GM12878 cells, LNCaP cells, NCCIT cells, primary human hepatocytes, and mouse ES cells. Each dot represents one tested DE-G pair. Red dots: connections where perturbation resulted in a significant decrease in the expression of the tested gene. Grey dots: no significant effect. **b,** Precision-recall plot for classifiers of DE-G pairs shown in **a**. Positive DE-G pairs are those where perturbation of the distal element significantly decreases expression of the gene. Curves represent the performance for predicting significant decreases in expression for DE-G pairs based on thresholds on the ABC score (red) and genomic distance between the DE and the TSS of the gene (black). Circles represent the performance of models that predict significant regulation for DE-G pairs based on various criteria: pair lies within 100 kb (black), and DEs are assigned to regulate the nearest expressed gene (grey). **c,** Comparison of observed and predicted DE-G connections in the *SORT1* locus (chr1:109714926–109989926). Predicted DE-G connections (dotted red arcs) are based on ABC maps in primary human liver tissue. Observed DE-G connections (solid red arcs) are from previous experiments in which CRISPR was used to introduce indels near rs12740374 in primary hepatocytes[33] and an eQTL study in human liver[34].