nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-019-0538-0

In the format provided by the authors and unedited.

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*] and Jesse M. Engreitz[1,8,10*]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [3]Department of Biology, MIT, Cambridge, MA, USA. [4]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX, USA. [5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [6]Department of Computer Science and Department of Computational and Applied Mathematics, Rice University, Houston, TX, USA. [7]Center for Theoretical Biological Physics, Rice University, Houston, TX, USA. [8]Harvard Society of Fellows, Harvard University, Cambridge, MA, USA. [9]These authors contributed equally: Charles P. Fulco, Joseph Nasser. [10]These authors jointly supervised this work: Eric S. Lander, Jesse M. Engreitz. *e-mail: lander@broadinstitute.org; engreitz@broadinstitute.org

Supplementary Notes

Supplementary Note 1. CRISPRi-FlowFISH advantages and limitations

In this study, we set out to develop and apply an approach to comprehensively survey the regulatory elements for several genes.

We and others have recently developed high-throughput methods that use CRISPR to perturb noncoding elements in their native genomic locations and measure their effect on a target gene of interest[9,13,37-39,50]. However, these methods have had two major limitations: (i) they cannot be readily applied to any target gene (they require that a gene has a phenotype that is well suited for multiplex screening, such as affecting cell proliferation, or is engineered to facilitate such screening, for example by introduction of a reporter construct under the control of its promoter in the genome) and (ii) they do not directly read out RNA levels.

To overcome these limitations, we use CRISPRi in combination with RNA FISH and FACS to perturb hundreds of noncoding elements in parallel and quantify their effects on the expression of an RNA of interest (Fig. 1a; Extended Data Fig. 1). In this approach, we design a library of guide RNAs (gRNAs) targeting a large collection of candidate regulatory elements, transduce the library into a population of cells expressing KRAB-dCas9 (on average 1 gRNA per cell), and induce KRAB-dCas9 expression for 48 hours. To measure the effects of candidate elements on the expression of a gene of interest, we: (i) use fluorescence in situ hybridization (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 expression; (iii) use high-throughput sequencing to determine the abundance of each gRNA in each bin; and (iv) use this information to infer the effect of each gRNA on RNA expression. To assess quantitative effects and statistical significance, we average the effects of all gRNAs within each candidate element (Fig. 1c, Extended Data Fig. 2a,b) and compare to hundreds of negative control gRNAs in the same screen. We note that these experiments do not distinguish between *cis* and *trans* effects.

This CRISPRi-FlowFISH approach is well suited to comprehensively survey all the putative regulatory elements in the vicinity of a gene of interest (i.e. “gene-centric” studies). Alternative approaches, such as those based on single-cell[15,36] RNA-seq, are well suited to survey all the genes in the vicinity of elements of interest (i.e. “enhancer-centric” studies).

We review below some of the considerations and advantages of the CRISPRi approach.

Mechanism of KRAB-dCas9 inhibition. CRISPRi uses catalytically inactive Cas9 to recruit a KRAB domain to genomic sites of interest using a single gRNA. KRAB inhibits regulatory elements by recruiting epigenetic repressors to induce histone methylation, deacetylation, and heterochromatin formation[51,52]. CRISPRi has already been used to successfully perturb at least hundreds of promoters and enhancers[9,13,36]. Delivery of KRAB-dCas9 to gene promoters typically leads to 85-95% reduction in gene expression[8], and the quantitative effects on gene expression observed upon enhancer perturbation with CRISPRi agree with the effects observed upon genetic deletion of the same elements[7,9,36,53]. Despite this evidence, we note that it is currently unknown whether use of CRISPRi might fail to discover certain regulatory elements, for example due to differential sensitivity to KRAB-mediated inhibition.

Resolution of KRAB-dCas9 inhibition. We estimate that the resolution of our KRAB-dCas9 tiling screens is at most 200-500 bp — that is, when tiling gRNAs around an element, only gRNAs within ~200-500 bp of the element score in the screen (*e.g.*, see[9]). In some cases, two elements are located close to one another, but only one element scores in the screen. For example, we identify a strong enhancer for *GATA1* just 1.5 kb from another element (the promoter of *HDAC6*) that does not regulate *GATA1* (Fig. 1c, top panel). This is not attributable to differences in gRNA efficacy, as the gRNAs targeting the *HDAC6* promoter indeed have strong effects on *HDAC6* expression (Fig. 1c, lower panel). This resolution is consistent with a prior study that found by ChIP-seq that the H3K9me3 modification induced by CRISPRi is limited to within 500-1000 bp of the targeted site[7].

Supplementary Note 2. Regulatory effects of promoters on nearby genes

In addition to distal element-gene (DE-G) pairs, our CRISPR dataset in K562 cells included 1228 distal promoter-gene (DP-G) pairs (where the CRISPR-targeted element is located <500 bp from a TSS, and excluding elements which are the promoter of the assayed gene).

We explored whether, beyond DE-G pairs, the ABC model did a good job of predicting DP-G connections – that is, regulatory effects of one promoter on the promoter of another nearby gene. In fact, it did not. Our dataset in K562 cells included 61 significant DP-G pairs (out of 1228 total tested), and the ABC score was only moderately predictive of these effects (AUPRC = 0.15, Supplementary Fig. 10). Importantly, the DP-G pairs in our dataset behaved qualitatively differently from the DE-G pairs: promoters more frequently had repressive effects (33 of 61 DP-G pairs, 54%, versus 23% for DE-G pairs, Fisher's exact $p < 10^{-4}$).

Promoters are known have the ability to affect the expression of neighboring genes through several mechanisms, including: activation of nearby genes in *cis,* for example by acting as an enhancer[26,49]; second-order, downstream effects of the promoter's protein product; promoter-promoter competition, in which two promoters are proposed to compete for nearby regulatory elements[54]; and transcriptional interference, in which transcription of one gene physically blocks transcription of another[55]. We observe likely instances of each of these in our CRISPR dataset, detailed below.

Cis activation

We and others have shown that many gene promoters activate a neighboring gene in *cis* through DNA-mediated functions of their promoters[26,49,56]. In this dataset, promoters that activated a nearby gene indeed had higher 3D contact with their target genes compared to other nearby genes (rank-sum $p < 10^{-3}$).

Transcriptional interference

We identified 4 promoters where CRISPR perturbation caused an increase (10-21%) in the expression of a convergently transcribed neighboring gene. In each of these cases, precision run-on sequencing (PRO-seq) showed that the transcriptional units of these genes overlap (Supplementary Fig. 10c), suggesting that these promoters might repress the neighboring gene via transcriptional interference[55].

Second-order *trans* effects

Effects on nearby genes observed when inhibiting a promoter may be second-order effects mediated by functions of the RNA or protein product, rather than first-order, *cis* effects of the promoter itself. We examined the 4 promoters whose inhibition affected multiple and at least 25% of nearby genes that were not convergently transcribed (*GATA1*, *KLF1*, *LYL1*, and *PPP1R15A*). Of these, 3 encode transcription factors and 1 encodes a regulator of translation, consistent with these genes having widespread effects on gene expression. For 3 of these genes, we found additional evidence to support that these effects on nearby genes did not result from direct *cis* effects of the promoter: inhibiting distal elements that regulate these genes had directionally consistent effects on other genes. The promoters of these 4 genes also more often had repressive effects than other promoters we found to affect the expression of nearby genes (median 3 repressed genes vs 1, rank-sum test $p = 0.02$). Based on this evidence, we expect that the effects of these promoters on nearby genes are likely due to second-order, downstream effects of their protein products in *trans.*

For example, inhibiting the promoter of *GATA1* with CRISPRi led to increased expression of 3 nearby genes, and we confirmed through siRNA knockdown experiments that these effects are likely to result from *trans* functions of the GATA1 protein (Supplementary Fig. 8d).

Promoter competition

In addition to acting through a *trans* function of its product, promoters may inhibit nearby genes by competing for enhancers or other activating signals. Our dataset included 23 promoters that appeared to repress a nearby gene. Notably, these included 2 promoters near *HBE1* and 1 near *MYC* that have been previously shown to compete with *HBE1* or *MYC* for activating signals in the genome[57,58].

Supplementary Note 3. Additional mechanisms of distal regulatory elements

We considered two situations in which distal elements might have effects on gene expression through mechanisms distinct from or above that of enhancers: indirect effects and CTCF-bound elements. In addition to explaining some of the activating effects of distal elements (7 out of 109), these two situations also account for many of the DE-G pairs with repressive effects (16 out of 32). (“Activating effects” are those where perturbation of the element leads to a decrease in gene expression; “repressive effects” are those where perturbation leads to an increase.)

Indirect regulatory effects of distal elements

The first situation involves indirect regulatory effects. For example, an enhancer that activates gene A might appear to repress B in the event that activation of A represses B. The 32 significant DE-G pairs involving repressive effects included 28 unique DEs. 6 of these 28 have activating effects on at least one other nearby gene (Supplementary Fig. 8a). In one case, we verified that apparent repressive effects of an element on *PLP2* expression are due to that element activating *GATA1*, which in turn represses *PLP2* via a *trans*-acting function of the GATA1 protein product (Supplementary Fig. 8b-d).

CTCF Sites

The second situation involves elements bound by CTCF, a protein that affects gene regulation by shaping 3D genomic architecture[59] (29% of tested DEs bind CTCF in K562s by ChIP-seq). Notably, some CTCF sites appear to be coincident with enhancer elements (in that they are strongly marked by H3K27ac), while others appear to be separate. When we divided CTCF-bound distal DHS sites into H3K27ac^{high} and H3K27ac^{low} elements, we found clear differences between the two classes (Supplementary Fig. 7). H3K27ac^{high} CTCF elements were far more often activating rather than repressive (25 vs. 4, chi-squared test $p < 10^{-4}$), consistent with these elements primarily affecting gene expression as enhancers. The ABC model accurately predicts the effects of the perturbation of H3K27ac^{high} CTCF elements (AUPRC = 0.52, Supplementary Fig. 7b). In contrast, H3K27ac^{low} CTCF elements had balanced effects on gene expression (5 activating and 3 repressive, chi-squared test $p = 0.5$), and the ABC model performed less well (AUPRC = 0.15, Supplementary Fig. 7c).

Supplementary Note 4. Activity by Contact (ABC) model

We designed the Activity by Contact (ABC) score to represent a mechanistic model in which enhancers contact target promoters to activate gene expression. In a simple conception of such a model, the quantitative effect of an enhancer depends on the frequency at which it contacts a promoter ("Contact") multiplied by the strength of the enhancer ("Activity", *i.e.*, the ability of the enhancer to activate transcription upon contacting a promoter)[9]. Moreover, 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. We note that the precise biochemical basis of enhancer "Activity" is unknown, but we presume that it depends on the transcription factors and cofactors that can be recruited.

To extend this conceptual framework to enable computing the quantitative effects of enhancers on the expression of any gene, we formulated the ABC score:

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

Operationally, Activity (A) is defined 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, and elements are defined as ~500 bp regions centered on DHS peaks.

This model has the following characteristics or assumptions:

1. The effect of an element on gene expression is linearly proportional to contact frequency and enhancer Activity.
2. A given enhancer has equal Activity for all genes — that is, it does not model the potential for biochemical specificity that could allow certain enhancers to regulate only certain promoters.
3. Different enhancers contribute additively and independently to the expression of a gene.
4. The sum in the denominator includes the gene's own promoter, which is considered a potential enhancer with Activity calculated in the same manner as other enhancers.
5. The model computes the relative effect of an enhancer on gene expression, but does not estimate the absolute effect.
6. The model aims to predict the functions of enhancers, but not the functions of elements that act through other mechanisms.

Supplementary Note 5. Alternative methods to estimate Contact in the ABC score

We explored alternative methods to estimate Contact in the ABC score in order to understand which features of genome architecture — such as loops and domains — are important for good prediction.

Because >70% of the variance in Hi-C contact frequencies across a chromosome can be explained by modeling chromatin as a featureless, uniform polymer in the condensed (globular) state[60], we tested simply using the theoretical contacts expected from extrusion globule and fractal globule models ($\text{Contact}_{\textit{Globule}}$ is proportional to $\text{Distance}^{-\gamma}$, with $\gamma = 0.7$ and 1, respectively)[60]. Both scores performed nearly as well as the ABC score based on Hi-C data (AUPRC = 0.63 for $\text{Contact}_{\textit{Extrusion}}$, AUPRC = 0.64 for $\text{Contact}_{\textit{Fractal}}$ versus 0.65 for ABC, Supplementary Fig. 6a,c). In comparison, Activity x Loop, Activity x Domain, Activity x Genomic Distance, and Activity x $\text{Contact}_{\textit{Globule}}$ models with more extreme values of γ performed less well (Supplementary Fig. 6a-c). These results show that the ABC model can predict DE-gene regulation reasonably well even without using information about locus-specific or cell-type specific features of the 3D genome. This yields a useful rule of thumb: 10-fold greater genomic distance between an enhancer and promoter leads to approximately 10-fold lower contact frequency and 10-fold smaller predicted effects on gene expression.

Notably, however, locus-specific Hi-C data did appear to yield better predictions for some DE-G pairs, including for long-range enhancer-gene connections in the *MYC* locus that coincide with the anchors of 3D loops (Supplementary Fig. 6g,h)[9]. These and other 3D loops are present across many cell types[9,18]. Accordingly, we tested estimating Contact for a given pair of loci using the average contact frequency for those loci in Hi-C data from 10 human cell types (see Supplementary Methods). We found that an Activity x Contact$_{Average}$ model did a better job at predicting connections in the *MYC* locus than the Activity x Contact$_{Globule}$ models, and had similar performance to using K562 specific Hi-C data in the full K562 CRISPR dataset (AUPRC = 0.66 versus 0.65; Supplementary Fig. 6a).

Together, these results indicate that cell-type specific features of the 3D genome are not required for good predictions, and that the relationship between genomic distance and quantitative contact frequency — more so than loops or domains — contains important information about regulatory enhancer-gene connections. These observations allow us to calculate ABC scores in a given cell type even without Hi-C data from that cell type.

(We note that the ABC model predictions are highly cell-type specific, even without incorporating cell-type specific features of the 3D genome. The cell-type specificity of the ABC scores is driven by the cell-type specificity of H3K27ac and DNase-accessibility signals.)

Supplementary Note 6. Comparison to capture Hi-C

Having found that assigning each DE to promoters based on the presence of Hi-C peaks (HICCUPS "loops" [18]) performs poorly at predicting functional DE-G connections (Fig. 3a), we considered whether this performance might be improved by examining capture Hi-C data.

Accordingly, we investigated two capture Hi-C (CHi-C) datasets in mouse embryonic stem cells (mESCs) from Atlasi *et al.*[61] and Sahlen *et al.*[62] We downloaded the focal interactions provided in these studies.

First, we compared our ABC predictions to these CHi-C datasets. Specifically, we asked whether statistically significant physical interactions between promoter and non-promoter elements ("loops") called in this CHi-C data could identify the ABC-predicted enhancer-gene connections in mESCs. Considering all expressed genes in mESCs, only 13% (Atlasi *et al.* 2019) or 9% (Sahlen *et al.* 2015) of ABC-predicted DE-G connections corresponded to a CHi-C loop. Conversely only 13% (Atlasi *et al.* 2019) or 4% (Sahlen *et al.* 2015) of CHi-C loops corresponded to ABC-predicted DE-G connections. As such, this CHi-C data does not match the ABC predictions in mESCs.

Next, we compared the CHi-C loops to 15 functional DE-G connections identified by CRISPR experiments in mESCs from our study or from previous publications (see Supplementary Methods, Supplementary Table 6). Of these, CHi-C identifies only 5 (Atlasi *et al.* 2019) or 3 (Sahlen *et al.* 2015). For comparison, at a threshold corresponding to 70% recall in our K562 dataset, the ABC model correctly predicts all of these connections.

Further work will be required to explore whether even higher resolution data on 3D chromatin contacts might be leveraged to better predict regulatory connections.

Supplementary Methods

Tissue Culture

We maintained K562 (ATCC) cells at a density between 100K and 1M per ml in RPMI-1640 (Thermo Fisher Scientific, Waltham, MA) with 10% heat-inactivated FBS (HIFBS, Thermo Fisher Scientific), 2mM L-glutamine, and 100 units/ml streptomycin and 100 mg/ml penicillin. We maintained HEK293Ts between 20 and 80% confluence in DMEM with 1 mM Sodium Pyruvate, 25mM Glucose (Thermo Fisher Scientific) and 10% HIFBS. CRISPRi-FlowFISH and qPCR experiments used K562 cells expressing KRAB-dCas9-IRES-BFP from a third generation tet-inducible promoter (Addgene # 85449).

Individual gRNA qPCR

We generated stable cell lines expressing single gRNAs (Supplementary Table 2) by lentiviral transduction in 8 μg/ml polybrene by centrifugation at 1200 x g for 45 minutes with 200K cells per well in 24-well plates. After 24 hours, we selected for transduction with 1 μg/ml puromycin (Gibco) for 72 hours then maintained cells in 0.3 μg/ml puromycin. For each gRNA, we generated 2 independent polyclonal cell populations through duplicate infections. We isolated RNA, made cDNA, and performed RT-qPCR as previously described[9] using primers listed in Supplementary Table 2.

Gene and TSS Annotation

We downloaded the UCSC RefSeq track (refGene, version 2017-03-08). This track contains multiple isoforms per gene symbol. We selected one TSS for each gene in the genome. To make this selection, we used the TSS used by the largest number of coding isoforms.

For the genes we studied experimentally, we manually confirmed the predominant TSS based on CAGE and PRO-seq data. In one case we determined that a different TSS was used in our cells of interest and adjusted our annotation for this gene (*PVT1*).

When making genome-wide predictions, we removed genes corresponding to small RNAs (gene symbol contains 'MIR' or 'RNU', or gene body length <300 bp), as well as very long RNAs (gene body >2 Mb), which appear to correspond to artifactual UCSC transcript alignments.

For dividing DE-G and DP-G connections, we defined a distal element as a promoter if it overlapped any TSS in the annotation, even if that TSS was not the selected one.

We provide the gene annotations we used in Supplementary Table 5.

Because KRAB-dCas9 can decrease gene expression when delivered anywhere within a gene body, we excluded from our analysis any DE-G pairs where the DE landed within the body of gene G. To further confirm that none of the remaining DE-G pairs involved upstream, unannotated TSSs, we performed one additional analysis. We inspected paired-end K562 RNA-seq data (ENCFF412EYU) to identify all cases in which an RNA-seq read-pair mapping to the assayed gene spanned the region containing an enhancer. For the 87 regulatory DE-G pairs identified through CRISPRi, we observed no such spanning reads for 66 DE-G pairs, and for the remaining 21 DE-G pairs such reads add up to at most 2% of the total expression of the gene. Thus, DE-G pairs involving upstream, unannotated TSSs cannot explain the effects we observed when inhibiting enhancers with CRISPRi (median 22% effects).

Enhancer perturbation data from other sources

To complement the data from our FlowFISH dataset, we curated results from previous experiments involving perturbations to accessible elements and precise measurements of the effects on gene expression. These included experiments involving a variety of perturbation methods (CRISPRi, 2-guide deletion, or other genome or epigenome editing) and methods of measuring the effect on gene expression (RNA-seq, allele-specific RNA-seq, CRISPR screens, or RT-qPCR), and included six cell lines (K562, GM12878, NCCIT, LNCaP, hepatocytes, and mES cells). In cases where the same element-gene pair had been characterized in the same cell type by more than one group or by more than one assay, we included it only once in assessing the performance of predictors such as the ABC model. We did not consider element-gene pairs where the element was that gene's own promoter. Sources and study-specific details are annotated in Supplementary Table 6. Additional details are included below, and in the following section (Power calculations).

Fulco 2016. We previously used CRISPRi (KRAB-dCas9) to tile gRNAs across a large region around *GATA1* and *MYC* in K562 cells and measured the effects using a proliferation assay[9]. We used RT-qPCR data from this study to represent the effect sizes for the 2 distal enhancers that significantly affected *GATA1* expression. For significant elements in the *MYC* locus, we quantified the effects on *MYC* expression using a small-scale FlowFISH screen (see below). For other elements, we estimated their effect sizes on gene expression based on the linear relationship between *MYC* expression and proliferation[9].

Klann 2017. Klann *et al.* used dCas9-KRAB to target gRNAs to DHS elements in a large region around *HBE1* in K562 cells and measured the effects by FACS sorting on an integrated HBE1-mCherry reporter[13]. We downloaded the raw count file from this study (GSE96875) and filtered for gRNAs with a minimum total 50 reads across the high and low mCherry bins. We calculated the mean log2 fold-change across all replicates, and estimated effect sizes according to the linear relationship between this value and qPCR experiments for individual enhancers (Supplementary Figure 3b in Klann *et al*. 2017).

Ulirsch 2016. Ulirsch *et al.* used Cas9 and one gRNA per enhancer to introduce small deletions at each of 3 enhancers in K562 cells[11]. We obtained the original qPCR data from the authors and assessed expression differences between homozygous knockout and wild-type clones using a two-sided *t*-test.

Wakabayashi 2016. Wakabayashi *et al*. used Cas9 and one gRNA per enhancer to introduce small deletions at each of 5 enhancers in K562 cells[12]. We obtained the original qPCR data from the authors and assessed expression differences between homozygous knockout and wild-type clones using a two-sided *t*-test.

Thakore 2015. Thakore *et al.* used dCas9-KRAB to inhibit an enhancer (HS2) in the globin locus in K562 and performed RNA-seq[7]. We downloaded RNA-seq count matrices from GEO (GSE71557) and used DESeq2 to compute differential expression between biological replicate experiments using CR4 (the most effective guide RNA used in this study) versus no-guide controls. Genes within 1 Mb of the enhancer with FDR < 0.05 were considered true positives for downstream analysis; only genes within this range and with sufficiently high expression (>1 sample with read count >= 5) were considered in the multiple hypothesis correction.

Liu 2017. Liu *et al.* used KRAB-dCas9 to inhibit the promoters of several lncRNAs in K562 cells and performed RNA-seq[14]. We downloaded the raw data from GSE85011 and quantified transcript abundance with kallisto (v. 0.43.0)[63]. A total of 19 RNA-seq experiments were performed; we removed one outlier (k562-LINC00910-1). We used DESeq2 to call differentially expressed genes for each of the 5 lncRNAs where two or more replicates were performed (EPB41L4A-AS1, LINC00263, LINC00909, MIR142, XLOC-042889). We compared the samples for a given promoter to all of the other samples (in which other lncRNA promoters were targeted) because there were no negative control samples. Genes within 1 Mb of the enhancer with FDR < 0.05 were considered positives for downstream analysis; only genes within this range and with sufficiently high expression (>1 sample with read count >= 5) were considered in the multiple hypothesis correction.

Engreitz 2016. We previously generated homozygous and heterozygous knockout clones of 12 lncRNA and 6 mRNA promoters in mESCs on a 129S1/*Castaneus* hybrid genetic background, and measured the effects on gene expression using allele-specific RNA-seq[26]. We calculated the average effects on the allelic expression of each gene within 1 Mb of the deleted promoter and included these in our perturbation database for this study. We assessed significance using DESeq2 to calculate the marginal effect of genotype (promoter knockout) after controlling for allele and sample (design formula = "~0 + Genotype + Allele + SampleName"). This effectively combines the allele-specific expression information across heterozygous and homozygous clones and leverages the statistical power of the empirical Bayes approach in DESeq2. We performed multiple hypothesis correction using the Benjamini-Hochberg method considering all genes within 1 Mb of the deleted promoter. This approach proved more powerful than the permutation-based method we previously used to analyze this data[26], and identified several additional nearby genes that showed significant allele-specific effects on expression. In Supplementary Table 3 for this analysis, "nCtrl" and "nKO" refer to the number of wild-type and knockout *chromosomes* for each locus.

mESC enhancer deletions (this study). We also included data from new experiments in which we deleted two putative enhancers in mESCs via transfection of multiple gRNAs and measured the effects on nearby genes using allele-specific RNA sequencing, as previously described[26] (see Supplementary Table 2 for gRNA and genotyping primer sequences). These two enhancers were selected on the basis of previous plasmid reporter assays showing enhancer activity for these elements[64] and are named “Chen2008-1” and “Chen2008-25” according to their number assignment from this previous study. We performed hybrid selection RNA-seq and produced allele-specific count tables as previously described[26]. We assessed statistical significance using DESeq2 as described above. See Supplementary Table 3.

Moorthy 2017. Moorthy *et al.* generated enhancer knockouts in mESCs on a 129S1/*Castaneus* hybrid genetic background, and measured the effects on gene expression using allele-specific RNA-seq as well as RT-qPCR[29]. For the RNA-seq data, we calculated the average effects on the allelic expression of each gene within 1 Mb of the deleted element and assessed significance using DESeq2, considering allele-specific read counts in both heterozygous and homozygous clones as described above[26]. This study generated a variety of heterozygous and homozygous deletions, including of multiple elements in different combinations in the same clones. We considered only the loci where at least one clone carried the deletion on the 129 allele and at least one clone carried the deletion on the *Castaneus* allele. For each deletion, we averaged the allele-specific effects across all clones. We looked for genes that showed >5% change in allele-specific expression with FDR < 0.25, but did not identify any significantly affected genes beyond those identified by the authors’ analysis.

Xie 2017. Xie *et al*. used KRAB-dCas9 and single-cell RNA-seq to identify 12 enhancers in K562 that significantly affect the expression of a neighboring gene[15]. We used the log2 fold-change reported in the paper for genes whose expression was significantly affected by enhancer perturbations according to the authors’ analysis.

Blinka 2016, *Huang 2018, Li 2014, Mumbach 2017, Musunuru 2010, Qi 2018, Rajagopal 2016, Spisak 2015, Tewhey 2016, Wang 2018, Xu 2015, Zhou 2014*. For experiments from these studies, we estimated effect sizes and standard errors from figures in these studies, and assigned significance according to the authors’ analysis[10,16,17,23-25,27,28,30,32-34].

Fuentes 2018. Fuentes et al. used CARGO to deliver an array of 12 gRNAs with dCas9-KRAB to simultaneously perturb LTR5HS, LTR5A, and LTR5B repeat elements (of which there are 910 annotated in the genome) in the NCCIT cell line, and measured the resulting changes in gene expression using RNA-seq[31]. Because all elements were perturbed simultaneously (in each individual cell) in this study, the nature of the data is distinct from other data we analyzed, where only a single element was perturbed in any given experiment (or in any given cell in our CRISPRi screens). Accordingly, the data from Fuentes *et al.* required special analysis to identify DE-G pairs where effects on gene expression are likely to be due to the direct effects of an individual nearby DE/LTR.

We first identified the elements that were potentially targeted by Fuentes *et al.*: we considered 910 LTR5HS, LTR5A and LTR5B elements in the RepeatMasker (v4.0.5) database as well as 1194 dCas9 ChIP-seq peaks (see below for ChIP-seq analysis). We merged overlapping regions, resulting in 1427 candidate elements.

As different instances of the LTR5 repeats have high sequence similarity, we next determined how accurately we could measure the epigenetic profile (and thus the Activity component of the ABC score) of each LTR element. To determine the mappability of each element, we (i) simulated reads in each LTR region by tiling the region with 150 bp paired-end reads of insert sizes between 150 bp and 400 bp (in increments of 10 bp), (ii) mapped the simulated reads to the hg19 genome using BWA, and (iii) computed the fraction of reads from each LTR that map uniquely to that LTR (mapq >30). We considered the 1073 regions in which >95% of simulated reads mapped uniquely as sufficiently mappable for the purposes of the ABC score calculation.

In order to consider only the elements that were successfully perturbed in the CRISPRi condition, we further limited our analysis to the 1057 elements that displayed sufficient reduction in H3K27ac signal in the CRISPRi condition (>2-fold decrease in CRISPRi vs control condition, and less than 1 read per million in total H3K27ac ChIP-seq signal in the CRISPRi condition).

We next identified the set of genes that had exactly one nearby targeted LTR element (within 500 kb, not within the gene body). To assess changes in gene expression, we re-analyzed the RNA-seq data from Fuentes *et al.* (GSE111337). We quantified gene abundances using kallisto[63] and computed differential expression with DESeq2 as described in Fuentes *et al.*[31]. We considered a gene significantly differentially expressed if its Benjamini-Hochberg adjusted *p*-value was <0.05. We calculated the statistical power to detect effects as described in the following section.

In order to reduce the contribution of *trans* effects, we applied a filter similar to that described in Fuentes *et al.*[31]: we limited our analysis to genes that have concordant effects in the CRISPRi and CRISPRa conditions. Specifically, we only analyzed genes that were significantly down-regulated in the CRISPRi condition and up-regulated in the CRISPRa condition, or genes that were not significant in both conditions and that had sufficient power in both conditions.

To summarize, we applied the following to filters to the dataset generated by Fuentes *et al*:

We only considered LTR elements that:

- Had sufficient decrease in H3K27ac signal upon CRISPRi perturbation
- Had sufficiently high simulated mappability
- Were at least 500 kb from the closest other LTR element.
- Did not overlap a gene promoter

We only considered genes that:

- Did not have an LTR within the gene body
- Had concordant effects under perturbations by CRISPRi and CRISPRa
- Had exactly one LTR within 500 kb

This resulted in a set of 22 positive and 872 negative LTR-gene pairs at the lenient power threshold (see below), and 22 positive and 0 negative LTR-gene pairs at the stringent power threshold (Supplementary Table 6). We additionally considered 5 LTR-gene pairs where Fuentes *et al*. deleted the LTR and quantified the effect on the target gene by qPCR. The deletion of the

LTR proximal to *EPHA7* was not included as this LTR element did not have sufficiently high simulated mappability to calculate Activity.

Power calculations for differential expression.

Enhancers are known to have a wide range of effect sizes on gene expression (including examples as low as 10%)[9], and so we designed our experimental and computational analysis of enhancer-gene connections to precisely estimate effect sizes and carefully estimate the power to detect certain effect sizes. For all datasets (including in our FlowFISH data and from other sources), we assigned each tested element-gene pair into one of four categories: (i) statistically significant decrease on gene expression ("positive" for precision-recall analysis); (ii) statistically significant increase on gene expression ("negative" for precision-recall analysis); (iii) >80% power to detect a 25% effect on gene expression, but no significant effect detected ("negative" for precision-recall analysis); or (iv) <80% power to detect a >25% effect on gene expression (not considered in our analysis of element-gene connections due to lack of power). As this stringent power cutoff permitted only 21 negative DE-G pairs for analysis of the perturbation data in other cell types (Supplementary Fig. 11), we also tested using a lenient threshold of >80% power to detect a 50% effect on gene expression (Fig. 4), which increased the number of negative pairs in other cell types to 947.

Power calculations for FlowFISH experiments. For each candidate element, we used a 2-sided *t*-test (equal variances) to compare the MLE effects of the gRNAs in that element to the MLE effects of 828-3858 negative controls (non-targeting gRNAs), and applied the Benjamini-Hochberg correction across the set of tests in each screen. We used summary statistics from these experiments (standard error of the mean and *n* for cases and controls) to analytically solve for the power to detect >25% changes in gene expression. We removed screens without 80% power to detect a 25% effect in at least 80% of elements, and additionally any tested E-G connections with insufficient power.

Power calculations for qPCR datasets. We used a 2-sided *t*-test (equal variances) to evaluate differences in gene expression for RT-qPCR datasets. We used summary statistics from these experiments (standard error of the mean and *n* for cases and controls) to analytically solve for the power to detect >25% or >50% changes in gene expression. *P*-value cutoffs for power calculations were determined using the multiple hypothesis correction methods used in the original studies.

Power calculations for RNA-seq datasets. We used DESeq2 to calculate differences in gene expression between cases (enhancer perturbation) and controls[65]. DESeq2 uses a series of empirical Bayes steps to estimate the mean, variance, and log-fold-change for each gene. We cannot compute the power for this test analytically and instead used a simulation-based procedure to estimate the power to detect changes in the expression of each gene in each enhancer perturbation:

1. We considered the real RNA-seq data for each test, for example consisting of several replicates of case and control conditions.

2. We removed genes where fewer than two samples had five or more reads.
3. We estimated the mean and dispersion parameters for each gene using the DESeq2 empirical Bayes procedure.
4. Based on these parameters, we simulated 100 random datasets across all genes with the same total read counts as the original experiments. For each gene within 1 Mb of the perturbation, we reduced the mean parameter by 25% or 50% for these simulations.
5. We used the DESeq2 pipeline on each simulated dataset to compute the *p*-value for every gene in the genome. For each gene within 1 Mb of the perturbation, we computed the FDR by performing multiple hypothesis correction with the Benjamini-Hochberg method using the *p*-value of each gene in the simulated dataset together with the *p*-values of other genes within 1 Mb derived from the real data.
6. We computed power based on the fraction of the 100 simulations in which FDR < 0.05.

We used an identical procedure for power calculations for allele-specific RNA-seq, with the only difference being the inclusion of additional variables (representing allele and sample) in the DESeq2 design matrix (as described above).

Computing the effects of large deletions: In some cases, certain genomic perturbations (*e.g.*, from Moorthy *et al.* 2017) involved large genomic deletions that spanned multiple ABC model elements. In these cases, we predicted the effect of the deletion as the sum of the ABC score of all overlapping elements, and assigned it to the "promoter" category if it overlapped a promoter element.

Stringent and lenient power filters for data in other cell types

We analyzed the enhancer perturbation data collated in other cell types at two different power thresholds, the "stringent" threshold we used for analysis of the K562 data (80% power to detect 25% effects on gene expression), and a "lenient" threshold of 80% power to detect 50% effects on gene expression because the experiments in other cell types were not as well powered as our CRISPRi-FlowFISH experiments, and thus assigned fewer non-regulatory DE-G pairs.

In the stringently-filtered dataset, applying the threshold on the ABC score corresponding to 70% recall and 59% precision in our initial K562 dataset could identify DE-G connections in other cell types with 86% recall and 80% precision (Supplementary Fig. 11).

When we relaxed the power requirements for data in other cell types to include more non-regulatory DE-G pairs (from 80% power for detecting 25% effects to detecting 50% effects), we found that the ABC model performed similarly in the K562 and cross-cell-type datasets (AUPRC = 0.65 vs 0.75, respectively; Fig. 4).

Epigenomic datasets, processing, and analysis.

DNase I hypersensitivity sequencing (DHS), ChIP-seq, and Expression datasets

We downloaded bam files for DNase I hypersensitivity sequencing (DHS), ChIP-seq for several chromatin marks including H3K27ac, and several transcription factors from a variety of sources including ENCODE (see Supplementary Table 4)[31,66-69]. We generated our own H3K27ac ChIP-seq data in F1 129/Castaneus hybrid mESCs grown in 2i media as previously described[26], and our own ATAC data in NCCIT cells as described below (available from GSE118912).

Hi-C

We analyzed K562 and GM12878 *in situ* Hi-C maps described previously (GSE63525)[18]. We also generated new *in situ* Hi-C maps of male mouse V6.5 embryonic stem cells grown in 2i conditions as previously described[18], and sequenced 4 technical replicates to a combined depth of 1.17 billion reads (available from GSE118912). Hi-C loop and contact domain annotations were computed using the Juicer suite of tools[70].

NCCIT ATAC

We performed ATAC-seq on 10K NCCIT cells in duplicate according to the protocol described by Buenrostro *et al.*[71] with some modifications. Specifically, we used Sigma Nuclei EZ lysis buffer for lysis for 10 minutes while centrifuging 500 x g at 4°C, resuspended with the lysis buffer, and spun again for 3 minutes. We then resuspended the nuclei pellet with a tagmentation buffer containing 12.5 μL of TD buffer, 1.25 μL of Tn5 transposase, 7.5 μL of PBS and 2.75 μL of water. After 15 cycles of PCR we cleaned the products with Agencourt XP (SPRI) beads and sequenced to a depth of at least 30M reads per sample with 100 and 200 bp paired-end reads on a HiSeq 2500.

NCCIT ChIP-seq processing

For analysis of CRISPRi and CRISPRa data from Fuentes *et al.*, we downloaded dCas9-GFP ChIP-seq data from GSE111337 and obtained H3K27ac ChIP-seq data directly from the authors[31]. We aligned reads using BWA (v0.7.17)[72], removed PCR duplicates using the MarkDuplicates function from Picard (v1.731), and removed reads with mapq < 30. We used MACS2 (v2.1.1)[73] to call peaks on Cas9 ChIP-seq using the non-targeting conditions as controls as described previously[31].

K562 H3K27ac HiChIP

We downloaded triplicate K562 H3K27ac HiChIP fastq files from GSE101498 and used the workflow outlined in hichipper[74]. Here, valid paired-end tags were identified from each individual replicate and subsequently combined to generate the H3K27ac HiChIP contact matrix using uniform 5 kb bins genome-wide. We downloaded HiCCUPS “loop” calls from Supplemental Table 2 of Mumbach *et al.*[30].

Alternative methods to estimate Activity and Contact in the ABC model

Estimating Activity using alternative epigenomic datasets.

Various epigenomic features (including histone modifications and enhancer RNA transcription) have been proposed to correlate with enhancer activity. Accordingly, we tried replacing H3K27ac ChIP-seq in the ABC score with read counts from various other datasets: ChIP-seq for P300, H3K9ac, H3K4me1, H3K4me3, and H3K27me3; and CAGE and PRO-seq from K562 cells (Extended Data Fig. 3c).

Approximating Hi-C contact frequency with the average Hi-C data

To evaluate the performance of the ABC model using a non-cell-type-specific Hi-C dataset, we generated locus specific Hi-C profiles from an average of 10 human Hi-C datasets (Supplementary Table 4). These averaged profiles were created as follows:

1. For each gene in the genome, we extracted the row corresponding to the TSS of the gene from each cell type's Hi-C matrix (KR normalized, at 5 kb resolution).
2. Each of these profiles was then scaled using the cell-type specific power law parameters relative to the K562 power law parameters (see below).
3. Finally, the total Hi-C signal in each cell-type specific profile was normalized to sum to one and then averaged across cell types to create the average profile anchored at a given TSS.

Normalizing Hi-C Profiles Using the Power-Law Fit

We found that different Hi-C datasets have slightly different power-law parameters. To weight all cell types equally in generating an average Hi-C profile, we scale the Hi-C profile in a given cell type by the cell-type specific gamma parameter from the power law relationship in that cell type (see below). The scaling factor at distance *d* is given by d ^ $(gamma_{\mathrm{ref}} - gamma_{celltype})$, where $gamma_{\mathrm{ref}}$ is the reference gamma parameter. For this study we used K562 as a reference (gamma = 1.024).

Fitting a power-law relationship to Hi-C data

We fit a power-law relationship to the Hi-C data in a given cell type as follows:

1. We aggregated all entries of the Hi-C matrix located less than 1 Mb from all gene promoters (KR normalized at 5 kb resolution)
2. We then performed a linear regression of the Hi-C signal in these bins on genomic distance in log-log space. The slope of this line is the *gamma* parameter.

Approximating Hi-C contact frequency with polymer globule models

To compute the variance in Hi-C contact frequencies (KR-normalized contacts) explained by a polymer globule model (and relevant to enhancer-gene regulation), we examined all gene TSSs and their contacts with loci at distances between 10 kb and 5 Mb in K562. The fractal globule model explained 69% of the variance in Hi-C contact frequency and the extrusion globule model explained 71% of the variance.

Jointly estimating Activity and Contact with HiChIP

The HiChIP assay combines chromatin immunoprecipitation with DNA proximity ligation to identify 3D chromatin contacts between genomic sites associated with a factor of interest[75]. H3K27ac HiChIP has recently been used to identify enhancer-promoter physical interactions[30]. From the perspective of the ABC Model, H3K27ac HiChIP may be considered a method to jointly measure the Activity of an element and its Contact to a gene promoter in one experiment.

Accordingly, we evaluated a version of the ABC model using H3K27ac HiChIP to jointly estimate Activity × Contact (Extended Data Fig. 3d).

In order to compute an ABC score using H3K27ac HiChIP, we first defined the quantitative H3K27ac HiChIP signal for a DE-G pair as follows:

1. We extracted the portion of the HiChIP counts matrix with row corresponding to the TSS of G and columns containing DHS peaks within 5 Mb of the TSS of G.
2. We then normalized this vector to sum to one.
3. We then divided all entries in this vector by the maximum value of the vector.
4. We then extracted the entry of this vector corresponding to the element E.

We then tested 2 versions of ABC scores based on this quantitative signal. We considered all distal candidate elements, defined based on DNase peaks as described above. For a given DE-G pair, we calculated:

- $ABC_{H3K27ac\ HiChIP}$. ABC score computed using H3K27ac HiChIP as a combined measure of Activity and Contact. $\mathrm{ABC}_{\mathrm{H3K27ac\ HiChIP}}$(E,G) = Quantitative H3K27ac HiChIP (E,G) / sum(Quantitative H3K27ac HiChIP (E,G)) where the sum is over all candidate elements E within 5 Mb of the TSS of G.
- $ABC_{DHS\ x\ H3K27ac\ HiChIP}$. ABC score computed using H3K27ac HiChIP and quantitative DHS signal. $\mathrm{ABC}_{\mathrm{DHS\ x\ H3K27ac\ HiChIP}}$(E,G) = DHS(E) x Quantitative H3K27ac HiChIP (E,G) / sum(DHS(E) x Quantitative H3K27ac HiChIP (E,G)) where the sum is over all candidate elements E within 5 Mb of the TSS of G.

Comparison of ABC predictions across cell types

Quantile normalization of epigenomic data

In order to facilitate a comparison of epigenomic datasets across cell types (and across assays, *e.g.*, DNase-seq vs ATAC-seq), we quantile normalized the read counts in candidate elements from other cell types to the read counts in the corresponding assays in K562. Specifically, for each data type (H3K27ac ChIP-seq and DNase-seq or ATAC-seq) and for each class of element (promoter-proximal and distal), we quantile normalized the signal (in RPM) from this data-type and element-class to the signal in K562 (ATAC-seq signal in other cell types was quantile normalized to DNase-seq signal in K562). We then computed genome-wide ABC scores using these normalized epigenomic profiles as described above. If Hi-C data was not available in the cell type, we used the average Hi-C profile described above. We used VC normalization to make ABC predictions for chromosome 9 in K562 since KR normalization is not available on this chromosome.

Identifying expressed genes for ABC predictions

When using the ABC model to predict functional enhancer-gene connections genome-wide (available at https://osf.io/uhnb4/), we made predictions only for genes that are “expressed”. For cell types where RNA-seq data was available, we defined expressed genes as those with RNA-seq RPKM (for K562; GSE87257) or transcripts per million (TPM; see Supplementary Table 4) >1.

For cell types where RNA-seq data was not available (LNCaP, primary liver) we defined expressed genes as those whose promoters had chromatin states consistent with active transcription. Specifically, we calculated a promoter score as the product of DHS (or ATAC-seq) reads and H3K27ac ChIP-seq reads on a 1 kb region centered at the gene's transcription start site, and then defined expressed genes and those with the top 60% of promoter scores.

Sensitivity of the ABC score to chosen parameters.

An attractive feature of the ABC model is its simplicity: at its core, the formula involves counting reads in DHS, H3K27ac, and Hi-C experiments, and performing a few addition and multiplication operations. We designed this ABC model based on the conceptual model of enhancer function described. Notably, there are no free parameters that need to be fit. While the model contains no free parameters, there are certain choices that need to be made in data processing. We made these choices based on known properties of epigenomic datasets. Specifically:

- We set the size extension of DHS peaks to 175 bp to include the nucleosome signal neighboring the DHS peak, and, together with the 150 bp DHS peaks in ENCODE data, to yield extended elements with a convenient size (500 bp).
- We chose a genomic distance cutoff of 5 Mb based on this including all confirmed cases of *cis* regulation by enhancers — the longest of which is ~2 Mb.
- We regularized the Hi-C data by adding an adjustment factor ("pseudocount"), equal to the average contact at d = 1 Mb (as described above).
- We included the promoter of each gene as a regulatory element and assigned its "Contact" (with itself) according to the diagonal Hi-C signal as described above.

To determine if the performance of the ABC score was sensitive to these choices, we varied the size of extension of DHS peaks (range: 0 to 1000 bp; our choice was 175 bp), the genomic distance over which elements were included in the model (range: 500 kb to 10 Mb; our choice was 5 Mb), the Hi-C adjustment factor (range: average signal at 100 kb to 10 Mb; our choice was 1 Mb), and the signal at the diagonal bin of the Hi-C matrix relative to its neighboring bins (range: 0 to 500%; our choice was 100%). A broad range of parameter choices gave nearly identical performance (Supplementary Fig. 5). The parameter that appeared most important was the size extension of DHS peaks, where either much lower or much higher extensions led to somewhat lower accuracy. This appears to be because at lower extension values, the H3K27ac signal is not properly captured, while at higher values the merging of nearby elements results in poor ability to distinguish between the functions of adjacent DHS peaks. These observations suggest that the ABC score is robust to our initial choices in data processing.

Comparison to other published enhancer-gene prediction methods

We evaluated the performance of the following published enhancer-gene prediction methods in predicting functional DE-G connections in our dataset:

JEME enhancer-gene predictions from from Cao *et al.* 2017. The Joint Effects of Multiple Enhancers (JEME) method first computes correlations between gene expression and various enhancer features (*e.g.*, DNase1, H3K4me1) across multiple cell types to identify a set of putative enhancers. Then a sample-specific model is used to predict the enhancer gene connections in a given cell type[20]. We downloaded the lasso-based JEME predictions in K562 (ID 121) from http://yiplab.cse.cuhk.edu.hk/jeme/. For each E-G pair in our dataset, we searched to see if the element and gene TSS overlapped two interacting regions listed in this file. If so, the pair received a score of 1, otherwise it received 0.

K562 ChIA-PET loops from Li *et al.* 2012. We downloaded the K562 saturated PET clusters from supplementary table 2 of https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3339270[40]. For each E-G pair in our dataset, we searched to see if the element and gene TSS overlapped two interacting regions listed in this file. If so, the pair received a score of 1, otherwise it received 0.

TargetFinder enhancer-promoter predictions from Whalen *et al.* 2016. We downloaded the TargetFinder K562 predictions from https://github.com/shwhalen/targetfinder[19]. We used the GBM classifier including Enhancer and Promoter windows (EPW). For each DE-G pair in our dataset, we searched to see if the element and gene TSS overlapped an enhancer and promoter loop listed in this file. If so, we assigned the pair a score corresponding to the 'prediction' column from this file, otherwise it received 0.

HiChIP loops from Mumbach *et al.* 2017. We downloaded the HiCCUPS high-confidence loop calls from K562 cells from supplementary table 2 of https://www.nature.com/articles/ng.3963#supplementary-information[30]. For each DE-G pair in our dataset, we searched to see if the element and gene TSS overlapped a loop listed in these files. If so, we assigned the pair a score of 1, otherwise it received 0.

FlowFISH to study enhancers and promoters in the *MYC* locus.

In our previous study, we identified 7 *MYC* enhancers that quantitatively tuned *MYC* expression (by 9-60%)[9]. We studied the effects of these 7 enhancers on two other genes in the locus (*PVT1* and *CCDC26*, both noncoding RNAs) to examine the potential for these enhancers to specifically regulate certain genes. We designed a pool containing 2 gRNAs per gene and 13 negative control gRNAs. We used CRISPRi-FlowFISH for *MYC*, *PVT1*, and *CCDC26* to measure the effects of these 7 enhancers on the expression of each of these genes (Supplementary Fig. 6h).

siRNA-mediated knockdown of GATA1

We transfected 200K K562 CRISPRi cells (from the same population of cells that was used in the CRISPRi-FlowFISH screens) with siRNAs (from Ambion, Thermo Fisher Scientific, Supplementary Table 2) using the Amaxa Nucleofector 96-well Shuttle (Lonza, program: 96-FF-120) following the manufacturer's protocol. We transfected each siRNA in quadruplicate. We harvested cells in buffer RLT (Qiagen, Germantown, MD) 48 hours after transfection and estimated target gene expression relative to cells transfected with non-targeting siRNAs by RNA sequencing.

For RNA-seq, we followed version 2 of a 3' cDNA-enriched bulk RNA barcoding and sequencing (BRB-seq) protocol[76] with minor modifications. Specifically, we isolated RNA from 100K cells in RLT with 2.2X volume Agencourt RNAClean XP SPRI beads (Beckman Coulter, Danvers, MA). We used 125 ng RNA input per sample (as measured by the RNA Qubit High

Sensitivity Kit, Thermo Fisher Scientific) during first strand synthesis with a barcoded RT primer. We then pooled 7-12 barcoded first-strand cDNA samples together. After an overnight second-strand synthesis, we split each pool (containing multiple samples indexed during first strand synthesis) into 4-8 tagmentation replicates. We tagmented 5 ng of cDNA using 1 µL Nextera Tagment DNA Tn5 transposase (Illumina, San Diego, CA, 15027916) in a 10 µL tagmentation mix for 10 minutes at 55 °C.

Using the custom P5 primer and a standard Nextera i7 indexing primer, we used qPCR to optimize the number of PCR amplification cycles by choosing the cycle number that produced half the maximal fluorescent signal. We cleaned up the reaction twice using 0.8X volume Agencourt Ampure XP SPRI solution (Beckman Coulter, Danvers, MA). We sequenced the resulting libraries on a HiSeq 2500 (Illumina) with 35 bp reads.

We trimmed reads using BRB-seqTools v1.3, aligned reads to hg19 using STAR (v2.5.2b), and used BRB-seqTools v1.3 to count UMIs in RefSeq gene exons. We used DESeq2 to compute differential expression of siRNAs against *GATA1* versus non-targeting controls with the design formula "~Perturbation + Dose" (to control for the doses of siRNAs). Genes within 1 Mb of *GATA1* with Benjamini-Hochberg-corrected *p*-value < 0.05 were considered differentially regulated; only genes within this range and with sufficiently high expression (>1 sample with read count >= 5) were considered in the multiple hypothesis correction.

Analysis of ubiquitously expressed genes

To define the set of ubiquitously expressed genes for human, we intersected 4 published lists of ubiquitously expressed genes from studies enumerating genes with detectable[77] or uniform expression across many tissues[78,79] for 847 total ubiquitously expressed genes (Supplementary Table 5). For mouse, we used the list of 4781 uniformly expressed genes provided in Li *et al.*[80]. We refer to all other genes as "tissue-specific".

To compute the number of enhancers per tissue-specific or ubiquitously expressed gene, we focused on the subset of our data where we had comprehensive CRISPRi tiling data testing all elements near a gene, including 30 genes from this study and 2 genes (*MYC* and *HBE1*) from previous studies[9,13]. In this subset of the data, we found 60 regulatory DE-G pairs for the 24 tissue-specific genes and 3 regulatory DE-G pair for the 8 ubiquitously expressed genes (Fisher's exact $p < 10^{-3}$), as reported in the main text. We note that the same trend holds in the full CRISPR dataset across all cell types (including DE-G pairs where we do not necessarily have comprehensive mapping of all DEs for that gene): we find more significant regulatory DE-G pairs for tissue-specific genes (151 significant pairs out of 3068 tested) than for ubiquitously expressed genes (8 significant pairs out of 873 tested, Fisher's exact $p < 10^{-8}$).

Analysis of CTCF sites

We considered that CRISPRi perturbation of CTCF-bound elements may affect gene expression through effects on 3D genome contacts rather than that through disruption of enhancer elements[81]. We downloaded CTCF ChIP-seq peak calls generated by ENCODE (Supplementary Table 4) and labeled a distal element as a CTCF-bound if the element overlapped a CTCF ChIP-seq peak. We further classified each CTCF site as H3K27ac^{High} or H3K27ac^{Low}, corresponding to elements with H3K27ac signal above or below the median H3K27ac signal for all tested distal elements in K562s.

Estimating the performance of the ABC score at predicting enhancer-gene connections

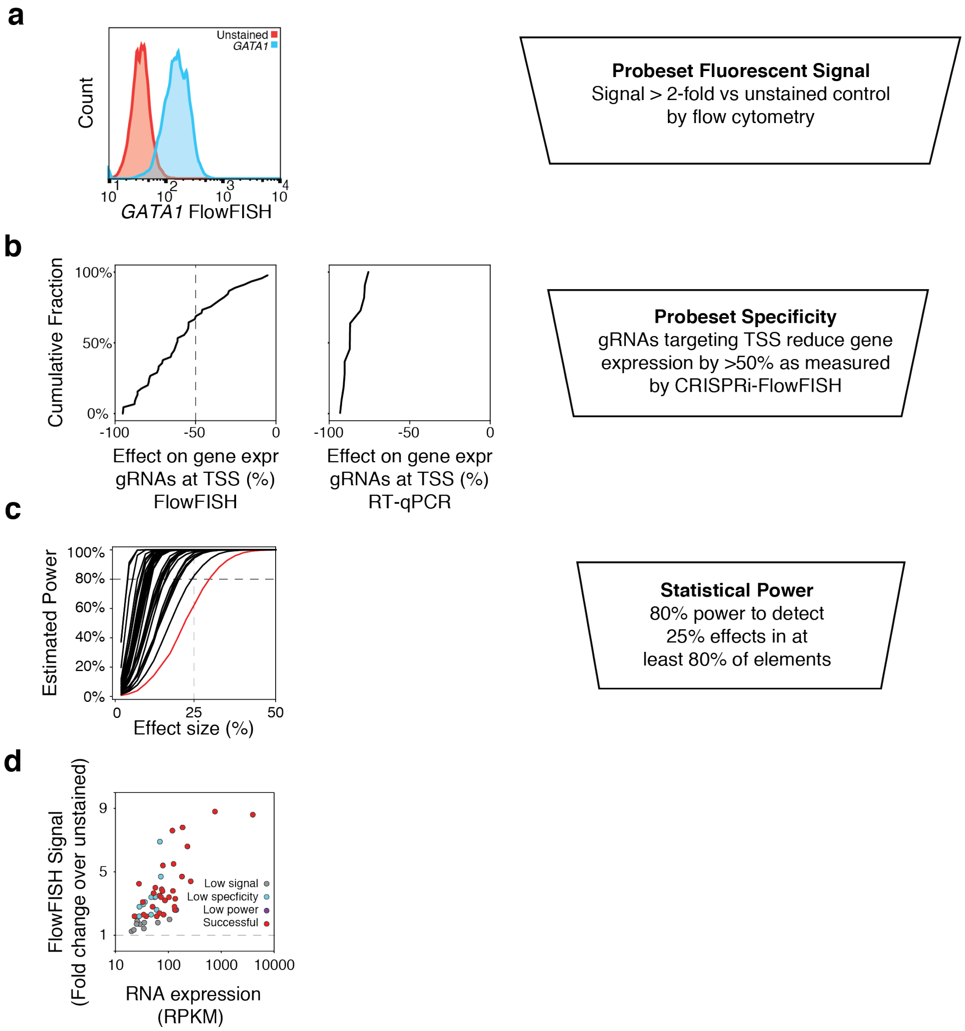
To estimate the performance of the ABC score on a dataset measuring only the direct *cis*-effects of enhancers, we removed 827 total DE-G pairs that involved (i) CTCF-bound elements unlikely to function as enhancers (H3K27ac^{Low}, 812 DE-G pairs), or (ii) DE-G pairs likely to result from indirect effects (15 DE-G pairs).

The latter category was defined as follows: We first identified genes (A) where the effects of promoter inhibition on nearby genes (G) are likely to be explained by second-order, indirect effects of the protein product (as described above). Enhancers that regulate gene A may also have indirect effects on gene G. Accordingly, we removed the 15 DE-G pairs where the element activates gene A and also affects gene G in a direction consistent with effect of promoter A on gene G.

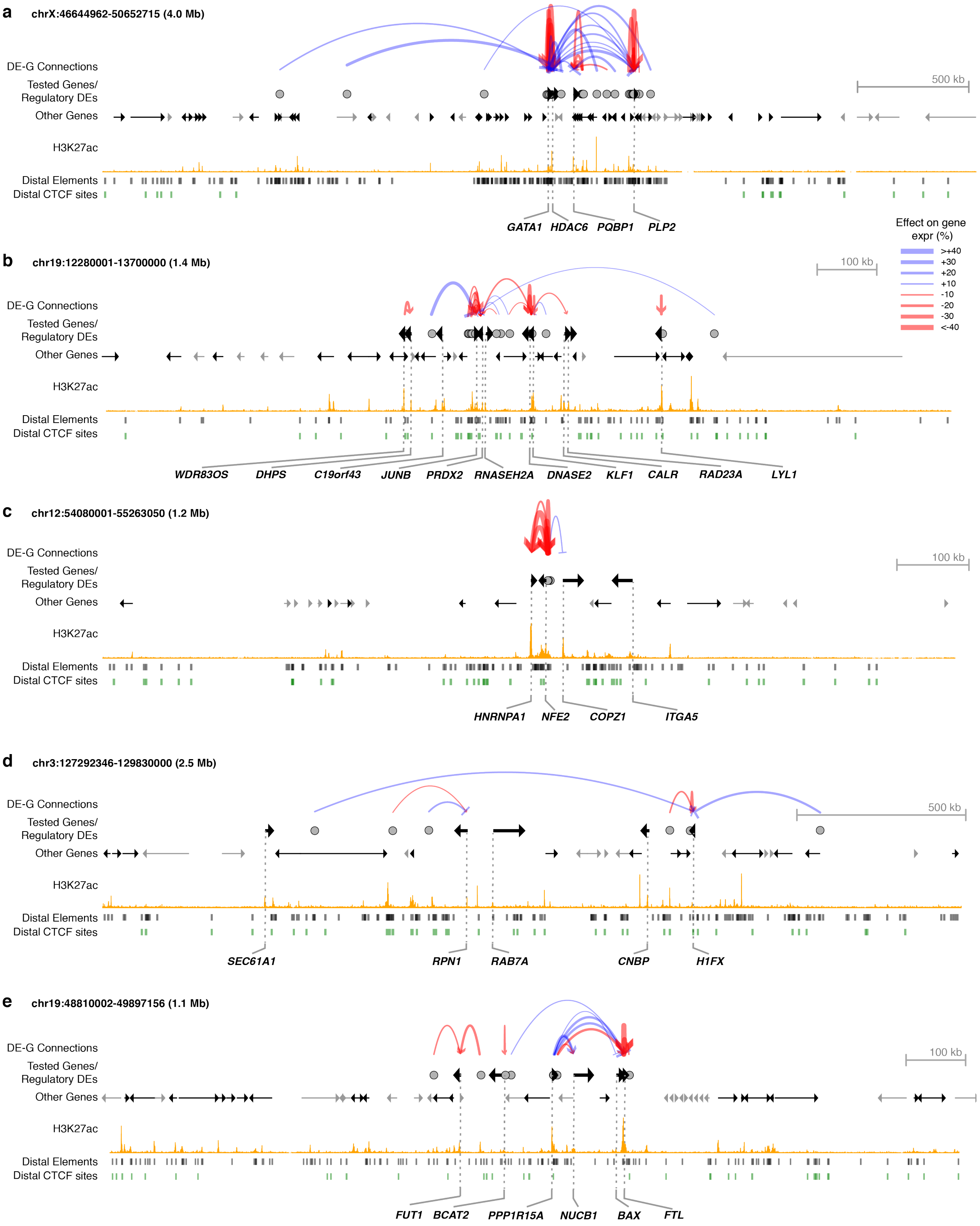
On this filtered dataset, the AUPRC of the ABC score rose from 0.73 to 0.76 for tissue-specific genes and 0.65 to 0.67 for all genes (Supplementary Fig. 9b). We note that all analyses presented in the paper use the full, unfiltered dataset in K562 cells unless otherwise specified.

Software for data analysis and graphical plots

We used the following software for data analysis and graphical plots: R (3.1.1) with Bioconductor (3.0)[82], Python (3.4.2), matplotlib (1.5.3), numpy (1.15.2), Pandas (0.23.4), Pybedtools (0.7.8), pyBigWig (0.3.2), pysam (0.13), scikit-learn (0.18.2), scipy (0.18.1), seaborn (0.7.1).

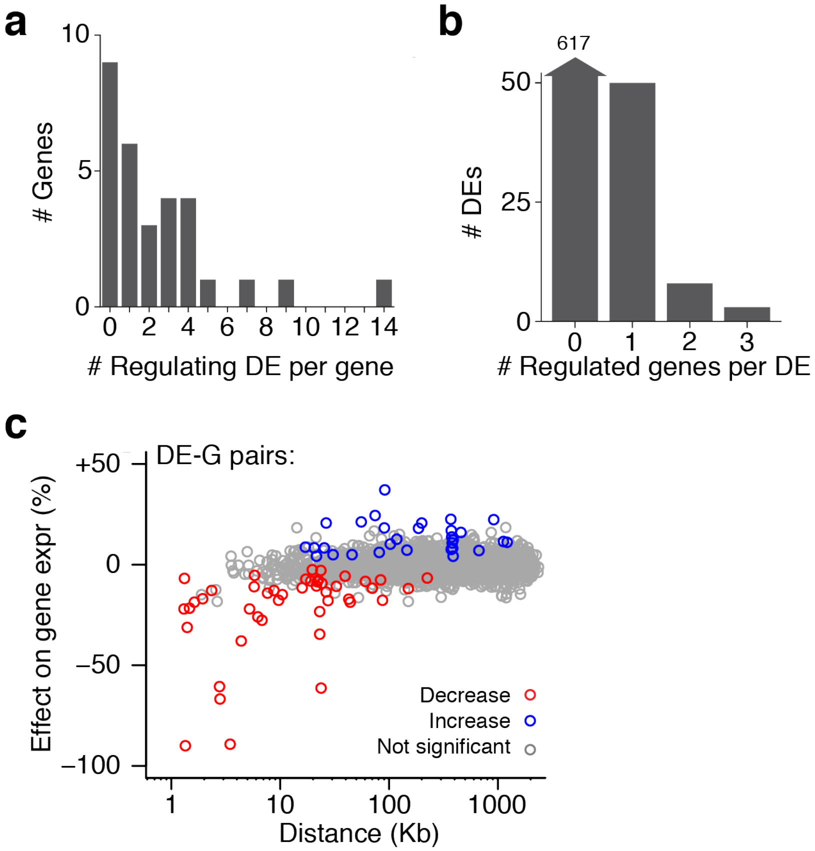


Supplementary Fig. 1. Quality filters for CRISPRi-FlowFISH probesets and screens. (a) Histogram of FlowFISH signal, as measured by flow cytometry, for K562 cells stained with *GATA1* probes compared to unstained, negative-control cells. Representative result from one of three *GATA1* CRISPRi-FlowFISH screens. We required probesets to have >2-fold mean fluorescent signal in stained versus unstained control. **(b)** Percent expression remaining in gRNAs targeting the TSS estimated from CRISPRi-FlowFISH screening. In all cases where we assessed CRISPRi knockdown by gRNAs at a TSS by qPCR, we observed >75% knockdown (right). However, some FlowFISH probesets reported <50% knockdown for gRNAs at their TSSs (left); we expect that some of the signal detected by these probesets results from off-target binding. Accordingly, we excluded these probesets from further analysis. **(c)** Power to detect a given effect size in 80% of E-G pairs for each gene. We analyzed screens with at least 80% power to detect a 25% effect for at least 80% of tested elements. Red line represents the screen that did not meet this power threshold. **(d)** Relationship between RNA expression, FlowFISH fold change over unstained control, and screen success. Red points denote 30 probesets yielding successful CRISPRi-FlowFISH screens, grey points denote 9 probesets with insufficient fluorescent signal, blue points denote 11 probesets with insufficient specificity, purple denotes 1 screen with insufficient statistical power. Pearson *R*=0.74 between RNA expression and FlowFISH signal.

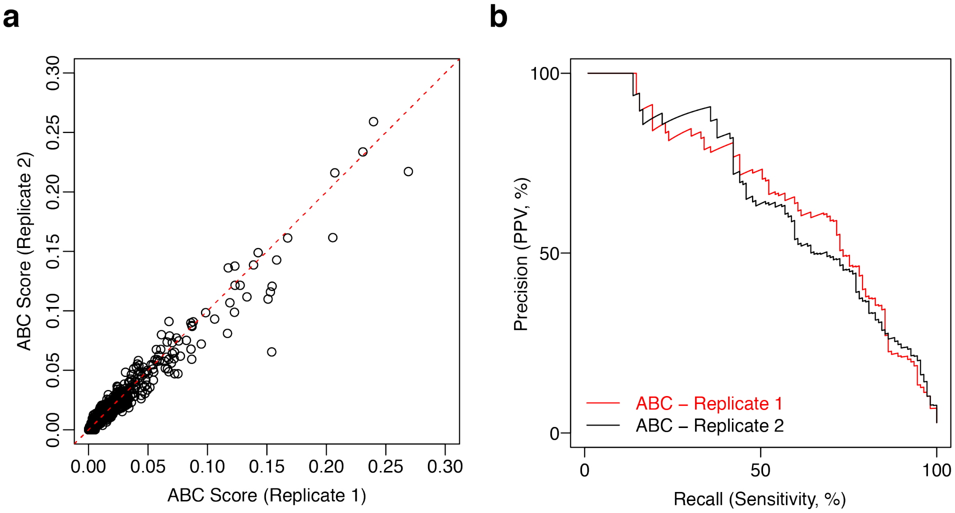


Supplementary Fig. 2 (Legend on next page)

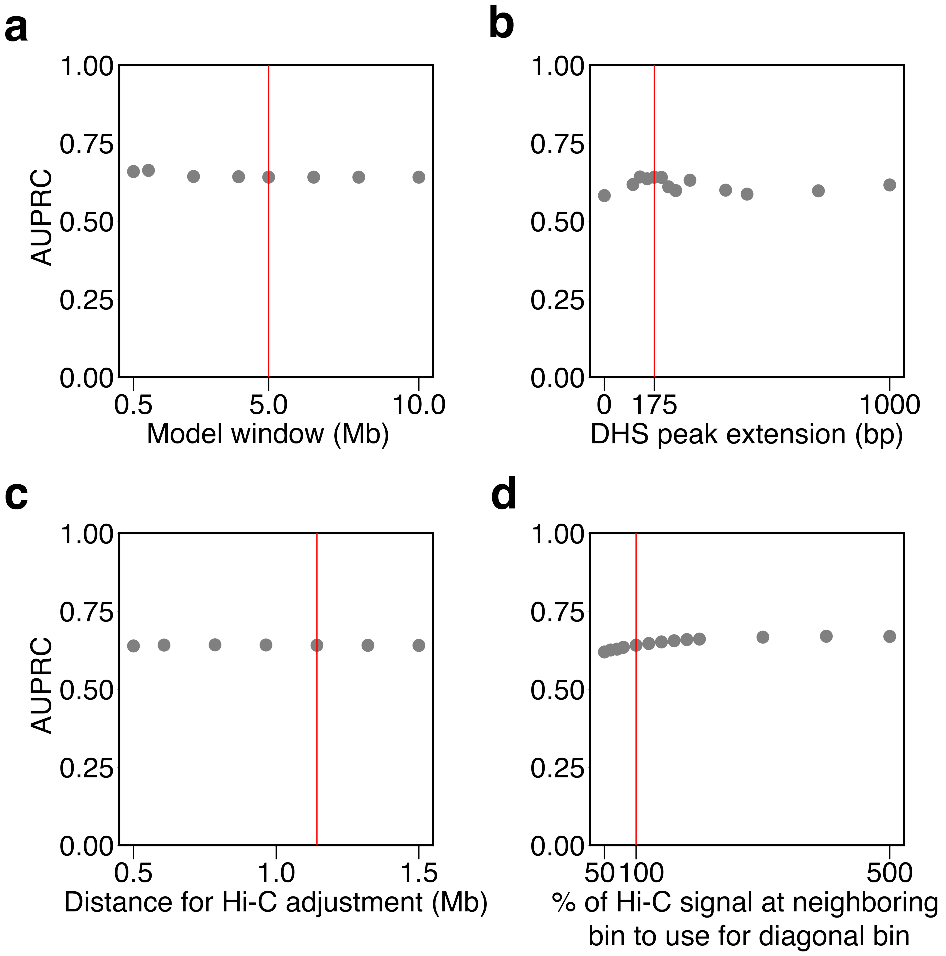
Supplementary Fig. 2. CRISPRi-FlowFISH enhancer perturbation dataset. DE-G connections are elements affecting the expression of the indicated gene in CRISPRi-FlowFISH screens in K562 cells. Red arcs denote activation, blue arcs denote repression. The width of the arc corresponds to the effect size. Distal elements (black) are tested DHS peaks. Distal CTCF elements (green) are CTCF ChIP-seq peaks within distal elements. Tested genes refer to genes for which we performed CRISPRi-FlowFISH experiments. Grey circles are DEs where perturbation with CRISPRi affects the expression of at least one tested gene as measured by CRISPRi-FlowFISH.



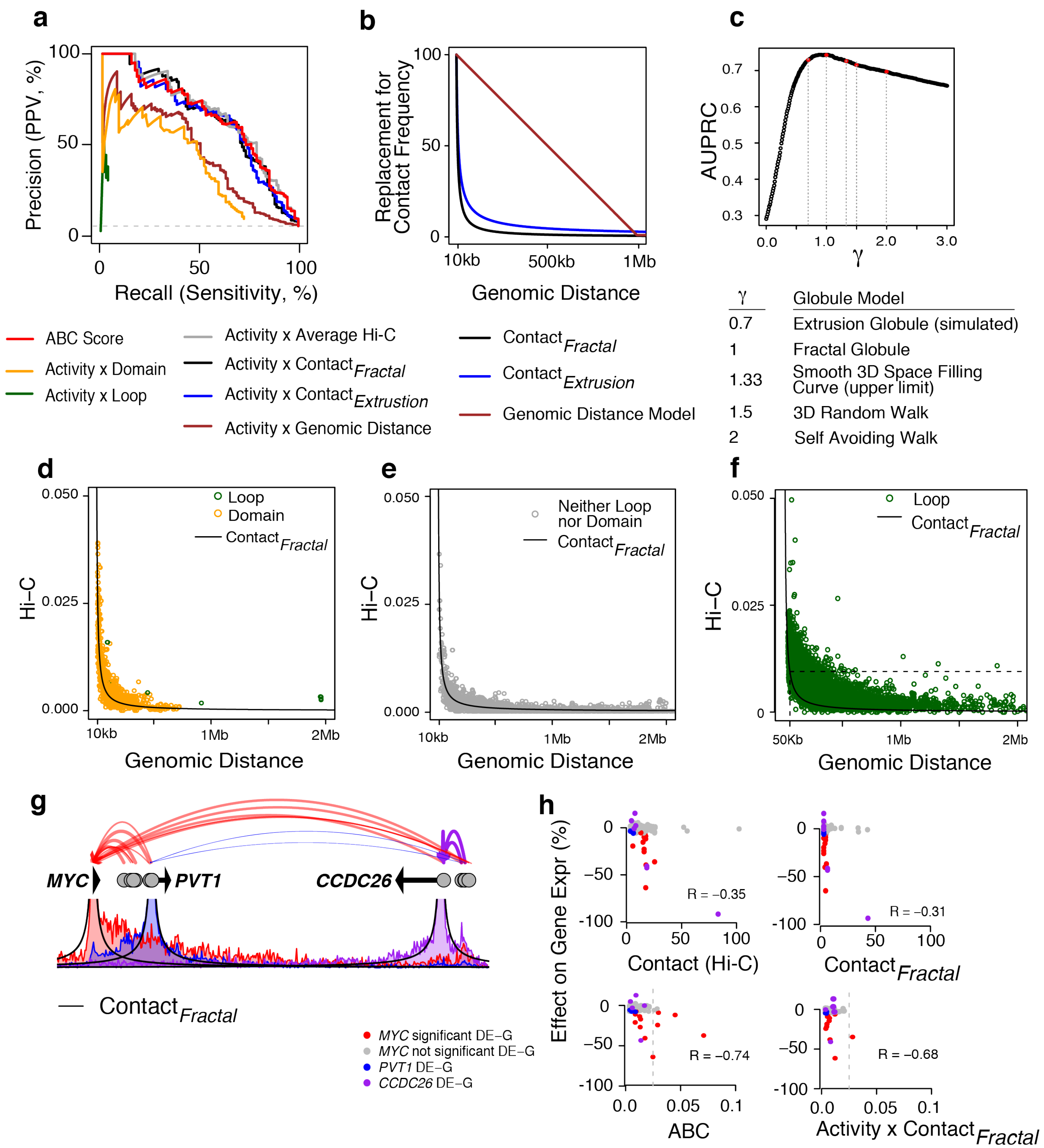
Supplementary Fig. 3. Properties of the CRISPRi-FlowFISH dataset. (a) Histogram of the number of distal elements affecting each gene in CRISPRi-FlowFISH experiments. **(b)** Histogram of the number of genes affected by each distal element tested in CRISPRi-FlowFISH experiments. **(c)** Comparison of genomic distance with observed changes in gene expression upon CRISPRi perturbation. 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.



Supplementary Fig. 4. The ABC Score is reproducible between replicates of the epigenomic datasets. (a) Scatter plot of ABC Score computed from two independent samples for DHS and H3K27ac (Pearson *R* = 0.98). N=3863 DEs. **(b)** Precision-recall curves for classifying regulatory DE-G pairs (Positive DE-G pairs are those where perturbation of element DE significantly reduces the expression of gene G) for the ABC Score using replicates 1 and 2 of DHS and H3K27ac.



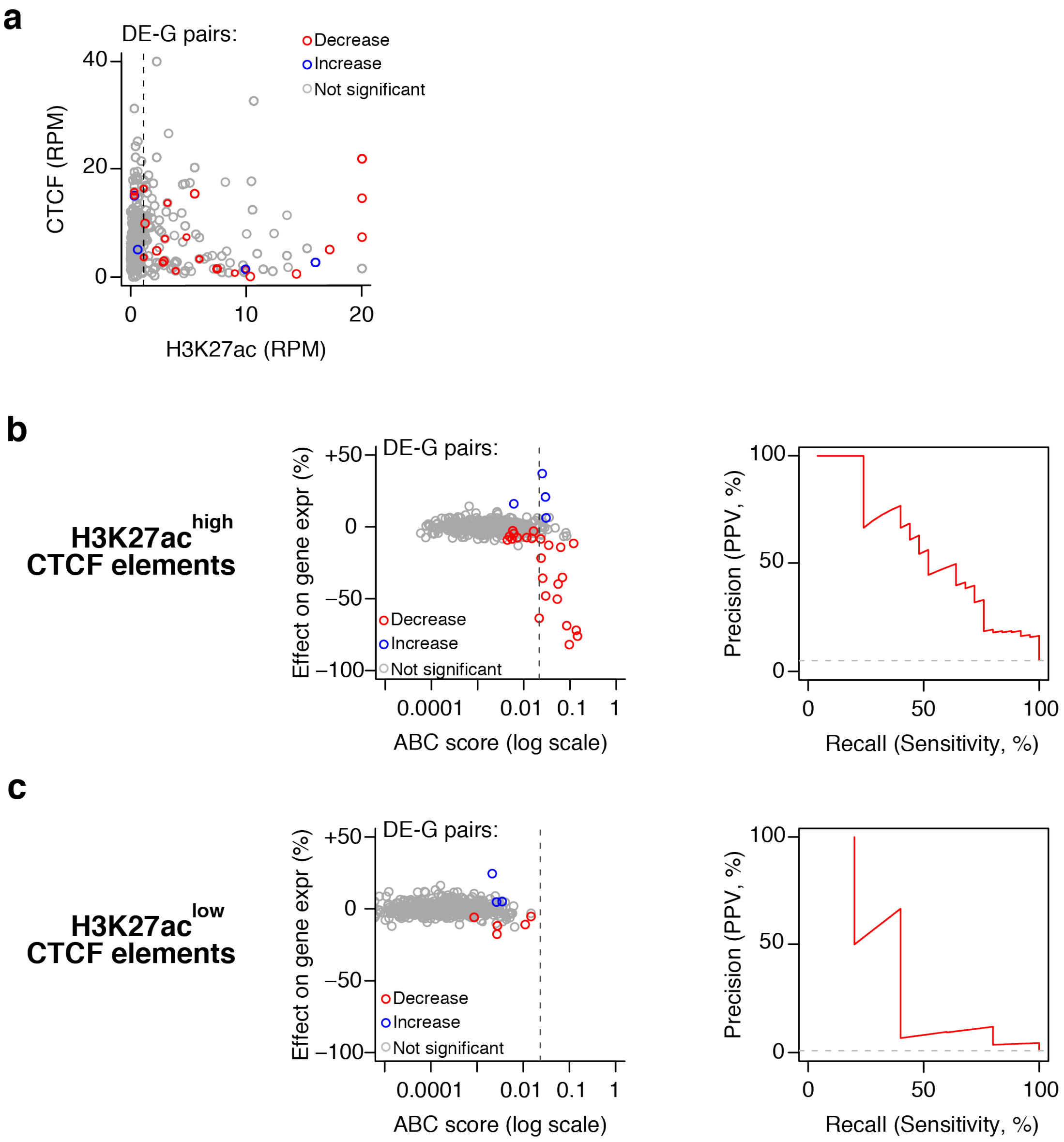
Supplementary Fig. 5. Sensitivity of ABC score performance to chosen parameters. Changing the parameters of the ABC score does not dramatically affect performance near the default values. Each panel presents the area under the precision recall curve (AUPRC) for the ABC score when changing the specified parameter. Red lines indicate the values used throughout this paper. **(a)** Genomic distance within which elements are included in the model. **(b)** Number of bases DHS peaks were extended on either side before merging to create candidate elements. **(c)** Genomic distance used to compute the pseudocount added to the Contact component (see Methods). **(d)** In processing Hi-C data, each diagonal entry of the Hi-C matrix is replaced by some percentage of the maximum of its four neighboring entries (This only affects DE-G pairs whose distance is <5 kb; see Methods).



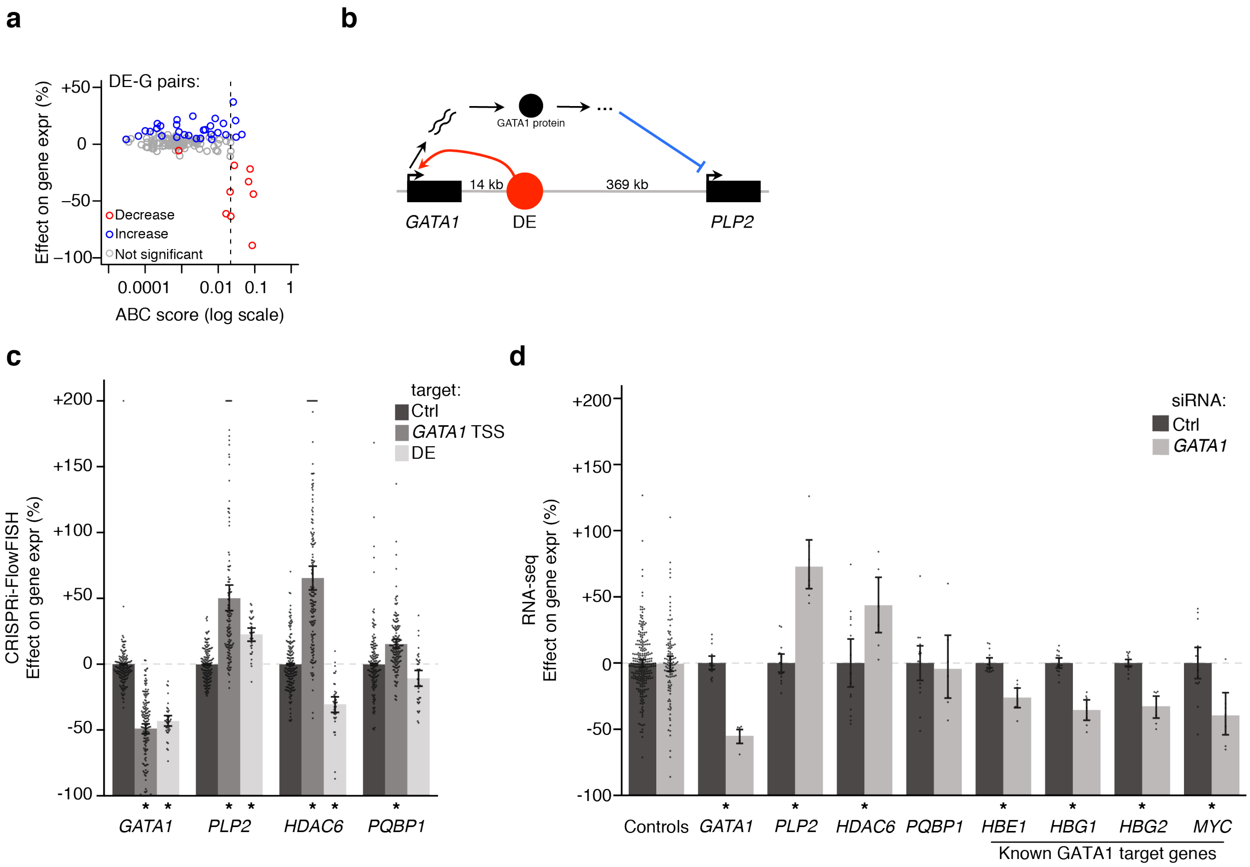
γ	Globule Model
0.7	Extrusion Globule (simulated)
1	Fractal Globule
1.33	Smooth 3D Space Filling Curve (upper limit)
1.5	3D Random Walk
2	Self Avoiding Walk

Supplementary Fig. 6. (Legend on next page)

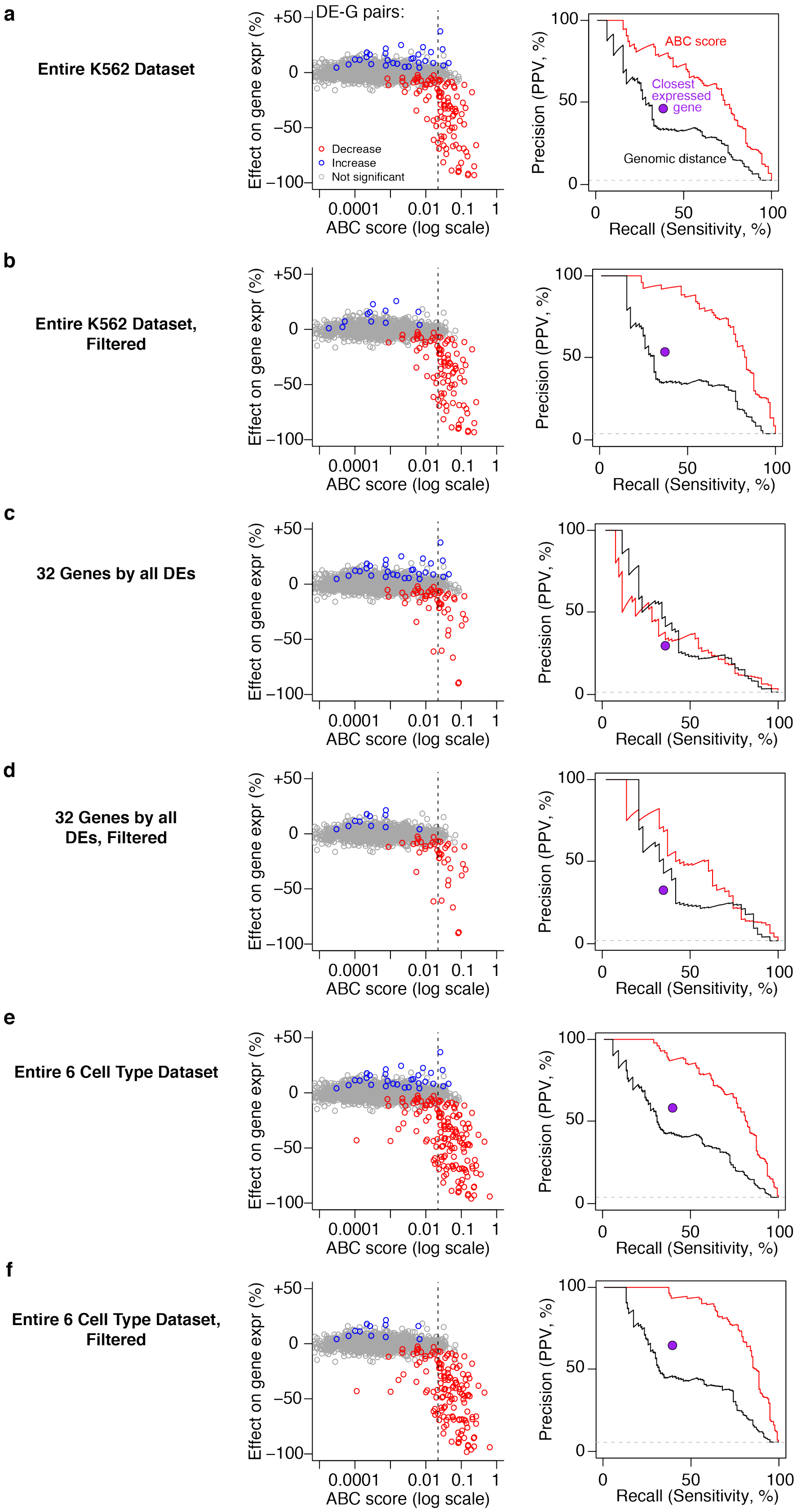
Supplementary Fig. 6. Testing other methods to estimate contact frequency for the ABC score. (a) Precision-recall curves comparing the ABC score to other models where the Contact component is replaced with binary Hi-C features (loops or domains) or decreasing functions of genomic distance (as visualized in panel (b)). *Activity x Genomic Distance:* Contact component is proportional to max(0.01, (1e6 - *Distance*)/1e6). *Activity x* $Contact_{Fractal}$, *Activity x* $Contact_{Extrusion}$*:* Contact component is proportional to $Distance^{-\gamma}$. $Contact_{Fractal}$ uses $\gamma = 1$, $Contact_{Extrusion}$ uses $\gamma = 0.7$. *Activity x Loop* and *Activity x Domain:* Contact component replaced by 1 if the element and gene TSS are located at the anchors of the same loop or within the same contact domain, respectively, or 0 otherwise. **(b)** Visualization of the quantitative functions used in (a) to replace contact frequency. Y-axis is in arbitrary units. In models of chromosome dynamics that assume chromatin is a featureless, uniform polymer in the globular state, Contact is inversely proportional to genomic distance raised to a fixed power (γ). Extrusion globule and fractal globule models ($\gamma = 0.7$ and 1) well represent the empirically observed Hi-C contacts at various distances[21]. **(c)** AUPRC for ABC models where the Contact component is replaced with $Distance^{-\gamma}$, with γ in the range [0, 3]. Values of γ corresponding to various polymer models are highlighted in red. The optimal values of γ as estimated from our CRISPRi data correspond to the values of γ that best predict Hi-C data (in the range of 0.7-1)[21]. **(d, e)** Scatterplot of genomic distance vs contact frequency (Hi-C) for K562 tested DE-G pairs whose distance is greater than 10 kb. Colors represent membership in the same contact domain (orange), Hi-C loop (green) or neither annotation (gray). These relationships explain why the ABC score performs similarly to the Activity x $Contact_{Fractal}$ model: the power law relationship explains 69% of the variance of Hi-C contact frequency. In contrast, the ABC score performs very differently from the Activity x Loop and Activity x Domain models because loops and domains are not predictive of contact frequency. Y-axis is KR-normalized Hi-C signal (and, for convenience, is not scaled on a per-gene basis as is used in ABC model, see Methods). **(f)** Scatterplot of genomic distance vs quantitative contact frequency (Hi-C) for all loops in K562[18]. Although Hi-C contact frequency at loops is higher than expected under the Fractal Globule model, the absolute increase in contacts is modest. For example, the loops with highest contact frequency at 500 kb have the expected contact frequency of non-loop loci at 50 kb (dotted line). **(g, h)** Comparison of DE-G predictions in the *MYC* locus using Hi-C vs the $Contact_{Fractal}$ model. (g) Visualization of Hi-C tracks anchored at the *MYC*, *PVT1* and *CCDC26* promoters (colored lines), compared to the $Contact_{Fractal}$ model (black lines). Arcs denote experimentally measured E-G connections (see Supplementary Methods). (h) Computation of the ABC score for 189 DE-G pairs in the *MYC* locus using Hi-C vs the $Contact_{Fractal}$ model. Using Hi-C data better predicts the quantitative effects of enhancers in this locus (Plots show Pearson R).



Supplementary Fig. 7. Analysis of CTCF-bound elements. (a) Scatterplot of CTCF signal (reads per million) vs. H3K27ac signal (reads per million) for all DE-G pairs where the DE is bound by CTCF (see Supplementary Methods). Dotted black line corresponds to the median H3K27ac signal for all distal elements in the dataset. We denote elements whose H3K27ac signal is greater than the median "H3K27ac^{High} CTCF elements" and those with H3K27ac signal less than the median "H3K27ac^{Low} CTCF elements". **(b)** 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 the DE is a H3K27ac^{High} CTCF element. Right: precision-recall curve for the ABC score in classifying regulatory DE-G pairs where each DE is a H3K27ac^{High} CTCF element. **(c)**: Same as (b) for H3K27ac^{Low} CTCF elements.

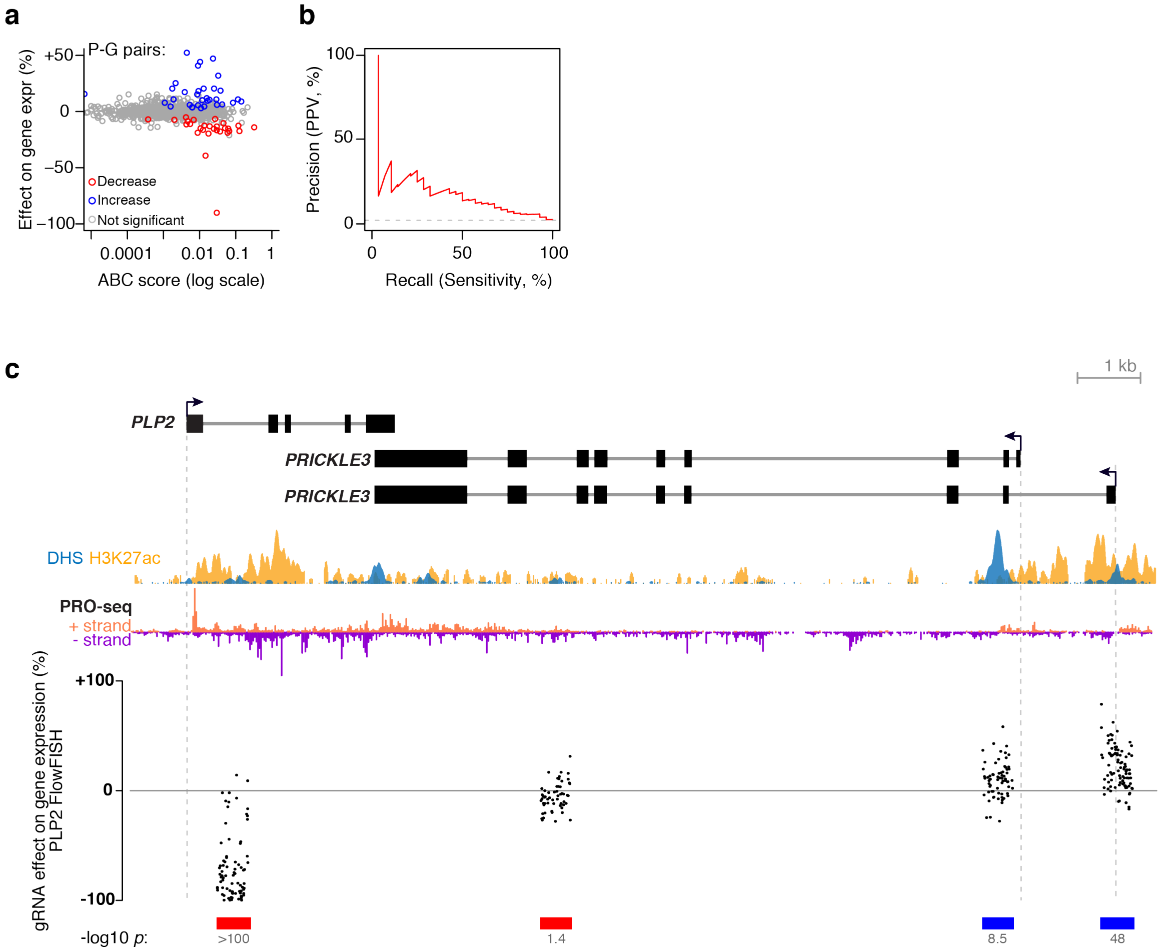


Supplementary Fig. 8. Elements that repress a distal gene are likely explained by indirect regulatory effects. (a) Comparison of ABC scores (predicted effect) with observed changes in gene expression upon CRISPR perturbations. Each dot represents one tested DE-G pair where the element represses at least one gene. **(b)** Summary of the effect of a *GATA1*-regulating DE on *PLP2*. The observed repressive effect of this DE on *PLP2* is consistent with this DE activating *GATA1* (red arc), which in turn represses *PLP2* via a trans-acting function of the GATA1 protein product (blue arc). **(c)** Effects of inhibiting the *GATA1* TSS or a *GATA1* enhancer (DE) with CRISPRi. mRNA expression measured by CRISPRi-FlowFISH. Error bars: 95% confidence intervals for the mean of all gRNAs within the target element (Supplementary Table 3). *: $p <$ 0.05 in BH-adjusted 2-sided t-test versus negative controls (see Supplementary Methods, Supplementary Table 3. A random 150 Ctrl points shown for each gene). **(d)** Effects of inhibiting *GATA1* with siRNAs on gene expression of *GATA1, PLP2, HDAC6, PQBP1*, and known GATA1 transcription factor targets[83-85] as measured by RNA sequencing of cells transfected with *GATA1* siRNA compared to non-targeting siRNAs (Ctrl). Control genes are the average of commonly used housekeeping genes (See Supplementary Table 3). Error bars: 95% confidence interval for the mean. *: p < 0.05 in BH-adjusted, 2-sided test from DESeq2 for *GATA1* siRNA (n=7 independent samples spanning 2 siRNAs) versus Ctrl (n=15 independent samples spanning 2 siRNAs) (see Supplementary Methods). P-values, test statistics, confidence intervals, effect sizes, and degrees of freedom are available in Supplementary Table 3.

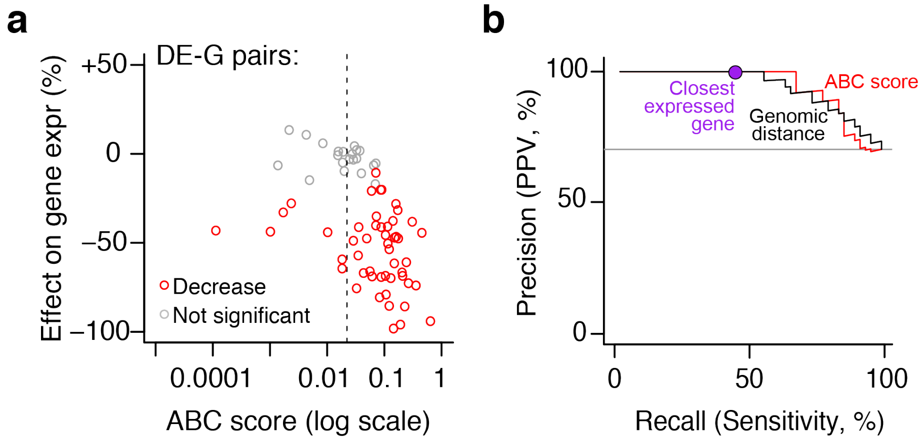


Supplementary Fig. 9 (Legend on next page)

Supplementary Fig. 9. Performance of the ABC score after filtering ubiquitously expressed genes, H3K27ac^{low} **CTCF elements, and indirect effects.** Performance of the ABC score on subsets of the CRISPR dataset. **(a)** Entire initial dataset in K562 cells (same as Fig. 3). **(b)** K562 dataset with H3K27ac^{low} CTCF elements, DE-G pairs likely to result from indirect effects, and ubiquitously expressed genes removed (see Supplementary Methods). **(c)** DE-G pairs in CRISPRi tiling experiments that, for a given gene, perturb and test the effects of all nearby DEs. **(d)** Subset described in (c) with H3K27ac^{low} CTCF elements, DE-G pairs likely to result from indirect effects, and ubiquitously expressed genes removed. **(e)** Entire dataset across 6 cell types. Includes cell types without Hi-C data, so the performance of Hi-C loops and domains cannot be evaluated. **(f)** Subset described in (e) with H3K27ac^{low} CTCF elements, DE-G pairs likely to result from indirect effects, and ubiquitously expressed genes removed. In each panel: Left plot is a comparison of ABC scores (predicted effect) with observed changes in gene expression upon CRISPR perturbations. Each dot represents one tested DE-G pair. Right plot is a set of precision-recall curves for classifying regulatory DE-G pairs (Positive DE-G pairs are those where perturbation of element DE significantly reduces the expression of gene G).



Supplementary Fig. 10. Effects of promoters on nearby genes. (a) Comparison of ABC scores (predicted effect) with observed changes in gene expression upon CRISPR perturbations in K562 cells. Each dot represents one tested DP-G pair (where the element itself is a promoter). **(b)** Precision-recall curve for classifying regulatory DP-G pairs (Positive DP-G pairs are those where perturbation of promoter P significantly reduces the expression of distal gene G). **(c)** Some promoters appear to affect expression of neighboring genes by transcriptional interference. One example is the effect of *PRICKLE3* on *PLP2*. Points represent the effect of gRNAs on *PLP2* expression, as measured by CRISPRi-FlowFISH. Red and blue bars: DHS elements in which CRISPRi leads to a significant decrease (red) or increase (blue) in *PLP2* expression. Transcription of *PRICKLE3* as measured by PRO-seq (negative strand, purple) extends into the gene body of *PLP2* (positive strand, salmon). Therefore, transcriptional interference may explain why CRISPRi inhibition of the *PRICKLE3* promoter leads to an increase in *PLP2* expression.



Supplementary Fig. 11. The ABC model generalizes across cell types.

Similar to Fig. 4, with a more stringent requirement for statistical power for measuring DE-G connections. **(a)** Comparison of ABC scores (predicted effect) with observed changes in gene expression upon perturbations in GM12878, LNCaP cells, NCCIT cells, primary human hepatocytes, and mouse ES cells. Each dot represents one tested DE-G pair. **(b)** Precision-recall plot for classifiers of DE-G pairs shown in (a). Positive DE-G pairs are those where 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). The purple circle represents the performance of assigning each DE to the closest expressed gene. DE-G pairs that were not significant are filtered for those that pass the same stringent power filter applied to the K562 dataset, requiring 80% power to detect a 25% effect on gene expression. (See Supplementary Methods. See Fig. 4 for data in these cell types using a lenient power filter of 80% power to detect a 50% effect on gene expression).

Supplementary References

50. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* **14**, 629-635 (2017).
51. Sripathy, S.P., Stevens, J. & Schultz, D.C. The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. *Mol Cell Biol* **26**, 8623-38 (2006).
52. Groner, A.C. *et al.* KRAB-zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading. *PLoS Genet* **6**, e1000869 (2010).
53. Carleton, J.B., Berrett, K.C. & Gertz, J. Multiplex Enhancer Interference Reveals Collaborative Control of Gene Regulation by Estrogen Receptor alpha-Bound Enhancers. *Cell Syst* **5**, 333-344 e5 (2017).
54. Martin, D.I., Fiering, S. & Groudine, M. Regulation of beta-globin gene expression: straightening out the locus. *Curr Opin Genet Dev* **6**, 488-95 (1996).
55. Shearwin, K.E., Callen, B.P. & Egan, J.B. Transcriptional interference--a crash course. *Trends Genet* **21**, 339-45 (2005).
56. Paralkar, V.R. *et al.* Unlinking an lncRNA from Its Associated cis Element. *Mol Cell* **62**, 104-10 (2016).
57. Cho, S.W. *et al.* Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell* **173**, 1398-1412.e22 (2018).
58. Townes, T.M. & Behringer, R.R. Human globin locus activation region (LAR): role in temporal control. *Trends Genet* **6**, 219-23 (1990).
59. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. *Cell* **137**, 1194-211 (2009).
60. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
61. Atlasi, Y. *et al.* Epigenetic modulation of a hardwired 3D chromatin landscape in two naive states of pluripotency. *Nat Cell Biol* **21**, 568-578 (2019).
62. Sahlen, P. *et al.* Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol* **16**, 156 (2015).
63. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-7 (2016).
64. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-17 (2008).
65. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
66. Core, L.J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-20 (2014).
67. Hazelett, D.J. *et al.* Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet* **10**, e1004102 (2014).
68. Corces, M.R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193-203 (2016).
69. Canver, M.C. *et al.* Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nat Genet* **49**, 625-634 (2017).
70. Durand, N.C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-8 (2016).

71. Buenrostro, J.D., Wu, B., Chang, H.Y. & Greenleaf, W.J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, 21 29 1-9 (2015).
72. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013).
73. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728-40 (2012).
74. Lareau, C.A. & Aryee, M.J. hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat Methods* **15**, 155-156 (2018).
75. Mumbach, M.R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919-922 (2016).
76. Alpern, D. *et al.* BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol* **20**, 71 (2019).
77. Zhu, J., He, F., Song, S., Wang, J. & Yu, J. How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* **9**, 172 (2008).
78. Forrest, A.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-70 (2014).
79. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet* **29**, 569-74 (2013).
80. Li, B. *et al.* A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. *Sci Rep* **7**, 4200 (2017).
81. Tarjan, D.R., Flavahan, W.A. & Bernstein, B.E. Epigenome editing strategies for the functional annotation of CTCF insulators. *Nat Commun* **10**, 4258 (2019).
82. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
83. Gong, Q. & Dean, A. Enhancer-dependent transcription of the epsilon-globin promoter requires promoter-bound GATA-1 and enhancer-bound AP-1/NF-E2. *Mol Cell Biol* **13**, 911-7 (1993).
84. Rylski, M. *et al.* GATA-1-mediated proliferation arrest during erythroid maturation. *Mol Cell Biol* **23**, 5031-42 (2003).
85. Woon Kim, Y., Kim, S., Geun Kim, C. & Kim, A. The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal gamma-globin genes. *Nucleic Acids Res* **39**, 6944-55 (2011).