

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. Jones¹, Glen Munson¹, Drew T. Bergman¹, Vidya Subramanian¹, Sharon R. Grossman^{1,3}, Rockwell Anyoha¹, Benjamin R. Doughty¹, Tejal A. Patwardhan¹, Tung H. Nguyen¹, Michael Kane¹, Elizabeth M. Perez¹, Neva C. Durand^{1,4,5,6}, Caleb A. Lareau¹, Elena K. Stamenova¹, Erez Lieberman Aiden^{1,4,5,6,7}, Eric S. Lander^{1,2,3,10*} and Jesse M. Engreitz^{1,8,10*}

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 diseases¹⁻⁴. Yet, we still do not know how enhancers regulate specific genes, and we lack general rules to predict enhancer-gene connections across cell types^{5,6}. We developed an experimental approach, CRISPRi-FlowFISH, to perturb enhancers in the genome, and we applied it to test >3,500 potential enhancer-gene connections for 30 genes. We found that a simple activity-by-contact model substantially outperformed previous methods at predicting the complex connections in our CRISPR dataset. This activity-by-contact model allows us to construct genome-wide maps of enhancer-gene connections in a given cell type, on the basis of chromatin state measurements. Together, CRISPRi-FlowFISH and the activity-by-contact 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 by using a library of guide RNAs (~1 gRNA per cell). KRAB-dCas9 has previously been shown to repress many promoters and enhancers, and affects elements within ~200–500 base pairs (bp) of the gRNA⁷⁻⁹; see Supplementary Note 1. To measure the effects of candidate elements on the expression of a gene of interest, we: (1) use RNA FISH to quantitatively label single cells according to their expression of an RNA of interest; (2) sort labeled cells with fluorescence-activated cell sorting (FACS) into six bins based on RNA abundance; (3) use high-throughput sequencing to determine the abundance of each gRNA in each bin; and (4) 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 this 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 five 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 (108–277 elements per gene for a total of 884 unique elements). The 30 genes included some with erythroid lineage-specific expression (for example, *GATA1*) and some that are ubiquitously expressed (for example, *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 quantitative PCR with reverse transcription (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 (DE)-gene (DE-G) pairs (where the targeted element is located >500 bp from a transcriptional start site (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 five genes, individual

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

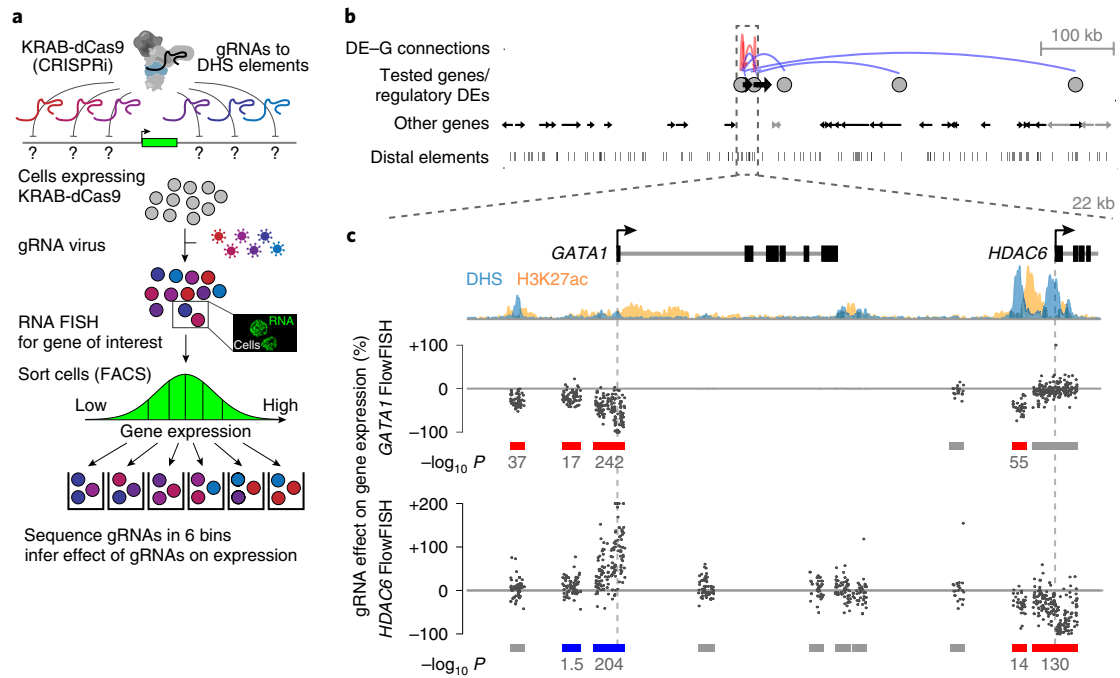


Fig. 1 | CRISPRi-FlowFISH identifies regulatory elements for *GATA1* and *HDAC6*. **a**, CRISPRi-FlowFISH method for identification of gene regulatory elements. Cells expressing KRAB-dCas9 are infected with a pool of gRNAs targeting DHS elements near a gene of interest, labeled using RNA FISH against 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 gray. Red/blue arcs: perturbation of a DE resulted in a significant decrease/increase in the expression of the tested gene. Gray 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 Fig. 2a for the full tested region spanning 4 Mb. **c**, Close-up of region containing *GATA1* and *HDAC6*. Points represent the effect on gene expression of a single gRNA. *HDAC6* vertical axis capped at 200%. Gray, red and blue bars: DHS elements in which CRISPRi leads to either no detectable change (gray) 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 transcription⁹. Such elements are included in analyses for other genes, as shown for the elements overlapping *GATA1*.

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% (109 versus 32), with absolute effect sizes of 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 (that is, the element activates gene expression in the genome), and the nonregulatory 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 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 three-dimensional (3D) genome also performed poorly. Assigning each DE to promoters according to the presence of all-versus-all (Hi-C) peaks (loops¹⁸) 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 previous machine learning approaches, including correlation of 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 following CRISPR-inhibition) should depend on that 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:

$$\text{ABC score}_{E,G} = \frac{A_E \times C_{E,G}}{\sum_{\text{all elements } e \text{ within 5 Mb of } G} A_e \times C_{e,G}}$$

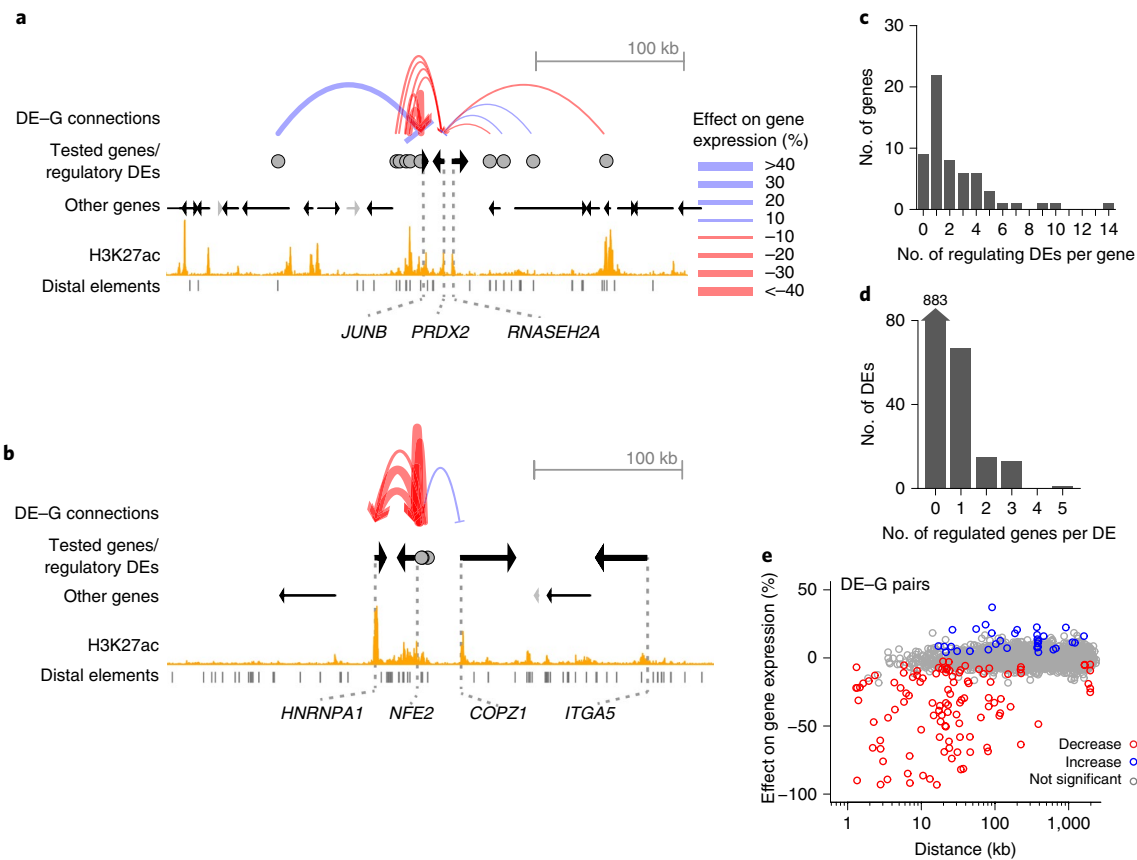


Fig. 2 | CRISPRi-FlowFISH produces regulatory maps of DE-G connections in multiple loci. a, Example of CRISPRi-FlowFISH screen data. DE-G connections are elements affecting the expression of *JUNB*, *PRDX2* and *RNASEH2A* in CRISPRi-FlowFISH screens in K562 cells. Red/blue arcs: perturbation of a DE resulted in a significant decrease/increase in the expression of the tested gene; arc width corresponds to effect size. Distal elements are DHS peaks. Tested genes refer to genes for which we performed CRISPRi-FlowFISH experiments. See Supplementary Fig. 2b for the full tested region spanning 1.4 Mb. **b**, Same as in **a**, for the genes *HNRNPA1*, *NFE2*, *COPZ1* and *ITGA5*. See Supplementary Fig. 2c for the full tested region spanning 1.2 Mb. **c**, Histogram of the number of distal elements affecting each gene in our dataset. **a-e** include both FlowFISH data from this study and tested pairs from other studies. See Supplementary Fig. 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 to observed changes in gene expression following 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. Gray dots: no significant effect.

Operationally, we estimated Activity (A) as the geometric mean of the read counts of DHS and H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) at 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 the use of 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 according to 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 for 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 one-dimensional 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 ten human cell types or a function of distance (Contact \approx 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 for a given cell type without cell-type-specific Hi-C data and minimally requires: (1) a measure of chromatin accessibility (DHS or assay for transposase-accessible chromatin using sequencing (ATAC-seq)) and (2) 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 five 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 five cell types and compared them to the functional data in the corresponding cell type.

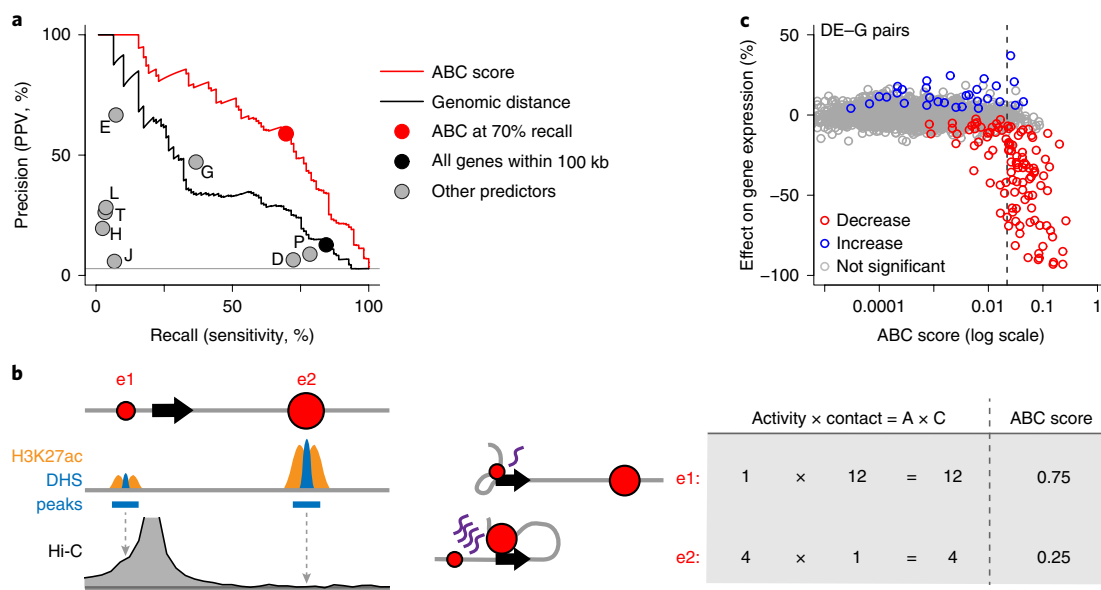


Fig. 3 | The ABC model predicts the target genes of enhancers. a, Precision-recall plot for classifiers of DE–G pairs. Positive DE–G pairs are those for which perturbation of the distal element significantly decreases expression of the gene. Curves represent the performance for predicting significant decreases in expression for DE–G pairs according to thresholds on the ABC score (red) and genomic distance between the DE and TSS of the gene (black). Circles represent the performance of various predictors to which DEs are assigned: 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). PPV, positive predictive value. **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. e1 and e2 (red circles) represent two arbitrary enhancers for the gene (black arrow). **c**, Comparison of ABC scores (predicted effect) to observed changes in gene expression following perturbations. Each dot represents one tested DE–G pair. Red/blue dots: connections for which perturbation resulted in a significant decrease/increase in the expression of the tested gene. Gray dots: no significant effect. Dotted black line denotes 70% recall, corresponding to the red dot in **a**.

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 expression quantitative trait locus (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 on the basis of 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 versus 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 (three enhancers across eight 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 explored³⁵.

We next examined our CRISPR dataset for DE–G pairs that are likely to represent effects due to mechanisms other than the *cis*-acting functions of enhancers (Supplementary Note 3). We identified effects of distal CCCTC-binding factor (CTCF) sites, which may regulate gene expression by affecting 3D contacts (eight significant pairs; Supplementary Fig. 7), and are likely to be indirect effects, such as an enhancer regulating one gene that in turn affects a second nearby gene in *trans* (15 pairs; Supplementary Fig. 8, and 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

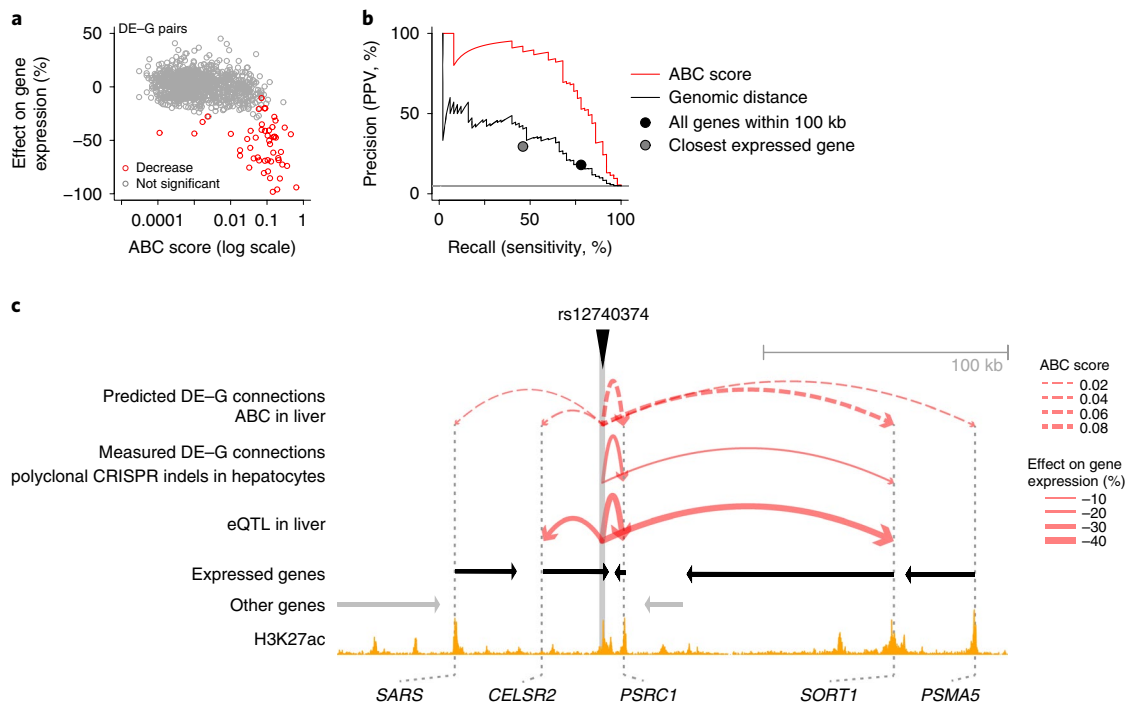


Fig. 4 | The ABC model generalizes across cell types. **a**, Comparison of ABC scores (predicted effect) and observed changes in gene expression following perturbations in GM12878 cells, LNCaP cells, NCCIT cells, primary human hepatocytes and mouse ES cells. Each dot represents one tested DE-G pair. Red dots: connections where perturbation resulted in a significant decrease in the expression of the tested gene. Gray dots: no significant effect. **b**, Precision-recall plot for classifiers of DE-G pairs shown in **a**. Positive DE-G pairs are those for which perturbation of the distal element significantly decreases expression of the gene. Curves represent the performance for prediction of significant decreases in expression for DE-G pairs according to 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 according to various criteria: pair lies within 100 kb (black), and DEs are assigned to regulate the nearest expressed gene (gray). PPV, positive predictive value. **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 derived from previous experiments in which CRISPR was used to introduce indels near rs12740374 in primary hepatocytes³³, and an eQTL study in human liver³⁴.

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 mapping of 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 of 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 that are based on readily obtained epigenomic datasets, the ABC model provides a framework for mapping enhancer–gene connections across many cell types—including those in which direct manipulation with CRISPR is problematic. This suggests a systematic approach to decoding of transcriptional regulatory networks and to interpret the functions of noncoding genetic variants that influence human traits.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests, and statements of code and data availability are available at <https://doi.org/10.1038/s41588-019-0538-0>.

Received: 23 January 2019; Accepted: 25 October 2019;

Published online: 29 November 2019

References

- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
- Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
- Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
- van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer–promoter interaction specificity. *Trends Cell Biol.* **24**, 695–702 (2014).
- Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327–339 (2011).
- Thakore, P. I. et al. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).
- Gilbert, L. A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).

9. Fulco, C. P. et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
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).
11. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
12. Wakabayashi, A. et al. Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *Proc. Natl Acad. Sci. USA* **113**, 4434–4439 (2016).
13. Klann, T. S. et al. CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).
14. Liu, S. J. et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, aah7111 (2017).
15. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol. Cell* **66**, 285–299.e5 (2017).
16. Huang, J. et al. Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat. Commun.* **9**, 943 (2018).
17. Qi, Z. et al. Tissue-specific gene expression prediction associates vitiligo with SUOX through an active enhancer. Preprint at *bioRxiv* <https://doi.org/10.1101/337196> (2018).
18. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
19. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
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).
21. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).
22. Yardımcı, G. et al. Measuring the reproducibility and quality of Hi-C data. *Genome Biol.* **20**, 57 (2019).
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).
24. Zhou, H. Y. et al. A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* **28**, 2699–2711 (2014).
25. Blinka, S., Reimer, M. H. Jr., Pulakanti, K. & Rao, S. Super-enhancers at the Nanog locus differentially regulate neighboring pluripotency-associated genes. *Cell Rep.* **17**, 19–28 (2016).
26. Engreitz, J. M. et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
27. Rajagopal, N. et al. High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).
28. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
29. Moorthy, S. D. 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).
30. Mumbach, M. R. et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49**, 1602–1612 (2017).
31. Fuentes, D. R., Swigut, T. & Wysocka, J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *eLife* **7**, e35989 (2018).
32. Spisak, S. et al. CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. *Nat. Med.* **21**, 1357–1363 (2015).
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).
34. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
35. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell. Biol.* **19**, 621–637 (2018).
36. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390.e19 (2019).
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).
38. Sanjana, N. E. et al. High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545–1549 (2016).
39. Canver, M. C. et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).
40. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

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 five mismatches), and selected only gRNAs that exceeded a specificity score >50 as previously described⁴¹ and lacked homopolymer stretches of more than seven As, Gs or Cs or four 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 probe set and screen to ensure robust, comprehensive and quantitative discovery of regulatory elements for each gene (Supplementary Fig. 1). We initially tested PrimeFlow probe sets for genes expressed at >20 reads per kilobase per million mapped in K562 cells (GSE87257) in five genomic loci (Supplementary Fig. 2). We first screened probe sets by flow cytometry and selected those with greater than twofold signal versus unstained cells. We next performed a tiling CRISPRi-FlowFISH screen (see below) and focused our analysis on the screens that showed the following characteristics: (1) maximum unscaled knockdown among 20-gRNA windows within 500 bp of the TSS > 50%; and (2) >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 probe sets 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 probe sets with signal >3, only 6 (22%) failed.

CRISPRi-FlowFISH screens. We cloned gRNA libraries purchased from CustomArray (now GenScript) for each of five genomic loci (Supplementary Fig. 2). We transduced these libraries (consisting of a single genomic locus and nontargeting gRNAs in the same pool) at a low multiplicity of infection (~0.3) into K562 cells harboring KRAB-dCas9, and selected for transduced cells as previously described⁴. To limit indirect effects or other changes in expression due to the expression of KRAB-dCas9, we used a dox-inducible system by inducing KRAB-dCas9 expression with 1 µg ml⁻¹ doxycycline for 48 h. We used 30 million cells for each screen.

We used the PrimeFlow RNA Assay Kit (Thermo Fisher, no. 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') probe set 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. Probe sets 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 ml⁻¹ and filtered using a 30-µm filter (CellTrics, no. 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, no. 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, v.6.3.1) 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 (1) 0–10%, (2) 10–20%, (3) 35–45%, (4) 55–65%, (5) 80–90% and (6) 90–100% (Extended Data Fig. 1e).

Genomic DNA extraction and gRNA sequencing. We collected the sorted cells by centrifugation at 800g for 5 min, resuspended them in 100 µl of lysis buffer (50 mM Tris-HCl, pH 8.1, 10 mM EDTA, 1% SDS) and incubated them at 65 °C for 10 min for reverse cross-linking. Once the samples had cooled to 37 °C, we added 2 µl of RNase Cocktail (Invitrogen, no. AM2286), mixed well and incubated the mixture at 37 °C for 30 min. Finally, we added 10 µl Proteinase K (NEB, no. P8107S), mixed well and incubated the mixture at 37 °C for 2 h 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 described⁴.

Analysis of CRISPRi-FlowFISH screens. To determine the effects of each gRNA on fluorescence, we used a maximum-likelihood estimation 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 algorithm maximum-likelihood method in the R stats4 package to fit the

read counts in each fluorescence bin to the log-normal distribution that would most probably have 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 that 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 genes⁴), but that some portion of the FlowFISH signal is due to nonspecific 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 Benjamini–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. Such pairs were excluded if they met any of the following criteria:

1. There was <80% power to detect a 25% effect for this *E–G* pair.
2. The element overlapped the gene's promoter.
3. The element was within the gene body or was within 2 kb of the 3' end of the gene.

ABC model. Defining candidate elements. We defined candidate regulatory elements in five human cell types (K562, GM12878, NCCIT, LNCaP, liver tissue) and one 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 by 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 was 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 those overlapping blacklisted regions. 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 with the highest number of read counts. We then resized these peaks to 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 that 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).

Calculation of 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 H3K27ac 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 by 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 expected that strong enhancers would have strong signals for both, and that elements that have only one or the other probably represent other types of element (elements with strong DNase-seq signal but no H3K27ac ChIP-seq signal might be CTCF-bound topological elements, and elements with strong H3K27ac signal but no DNase-seq signal might be sequences that are close to strong enhancers, but do not per se have enhancer activity due to the spreading H3K27ac signal over hundreds to thousands of base pairs). 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 an enhancer has the same Activity for every promoter (that is, 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 5×5 -kb² 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 Knight–Ruiz matrix balancing (KR) normalization factors < 0.1 , we did not use KR normalization (these typically correspond to 5-kb bins with very few reads). Instead, we 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 nonzero. For *E–G* pairs within 1 Mb, the adjustment is equal to the expected contact frequency at 1 Mb (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 (1). Adding the adjustment sometimes results in a quantitative Contact > 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.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Genome-wide ABC predictions for the six cell types considered in this study (K562, mESC, GM12878, NCCIT, LNCAP, hepatocytes) and raw counts from

CRISPRi-FlowFISH are available on the Open Science Framework at <https://osf.io/uhn4/>. CHIP-seq, ATAC-seq, Hi-C and RNA-seq data from this study are available at GSE118912.

Code availability

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

References

1. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
3. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
4. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
5. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
6. Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
7. Karlic, R., Chung, H. R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA* **107**, 2926–2931 (2010).
8. Mendenhall, E. M. & Bernstein, B. E. Chromatin state maps: new technologies, new insights. *Curr. Opin. Genet. Dev.* **18**, 109–115 (2008).
9. Dao, L. T. M. et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* **49**, 1073–1081 (2017).

Acknowledgements

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 (to E.S.L.) and by a 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 no. T32GM007753). E.L.A. was supported by an NSF Physics Frontiers Center Award (no. PHY1427654), the Welch Foundation (grant no. Q-1866), a USDA Agriculture and Food Research Initiative Grant (no. 2017-05741), a NIH 4D Nucleome Grant (no. U01HL130010) and a NIH Encyclopedia of DNA Elements Mapping Center Award (no. UM1HG009375).

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

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 nonprofit 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 (no. WO2018064208A1) filed by the Broad Institute related to this work.

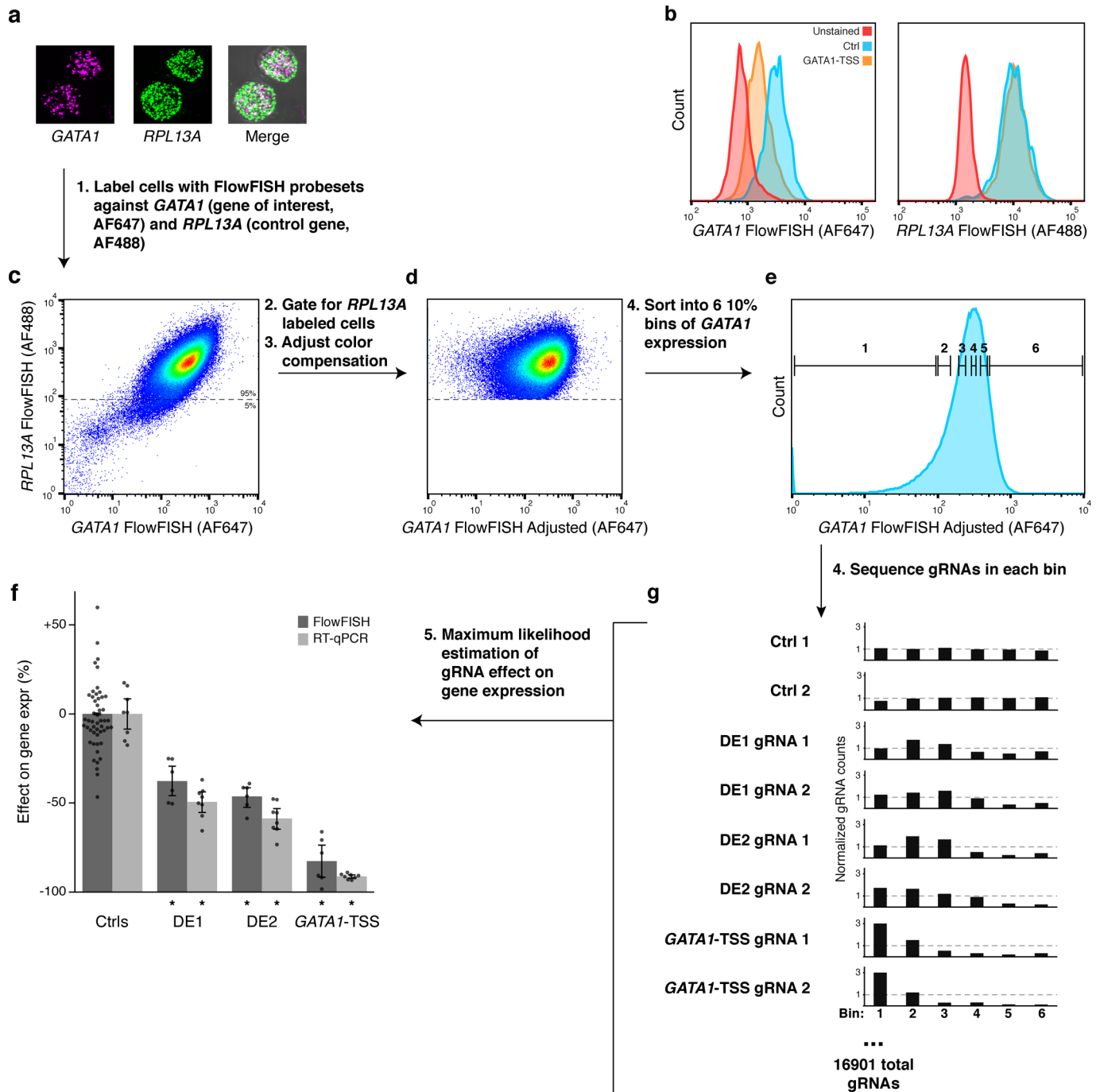
Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41588-019-0538-0>.

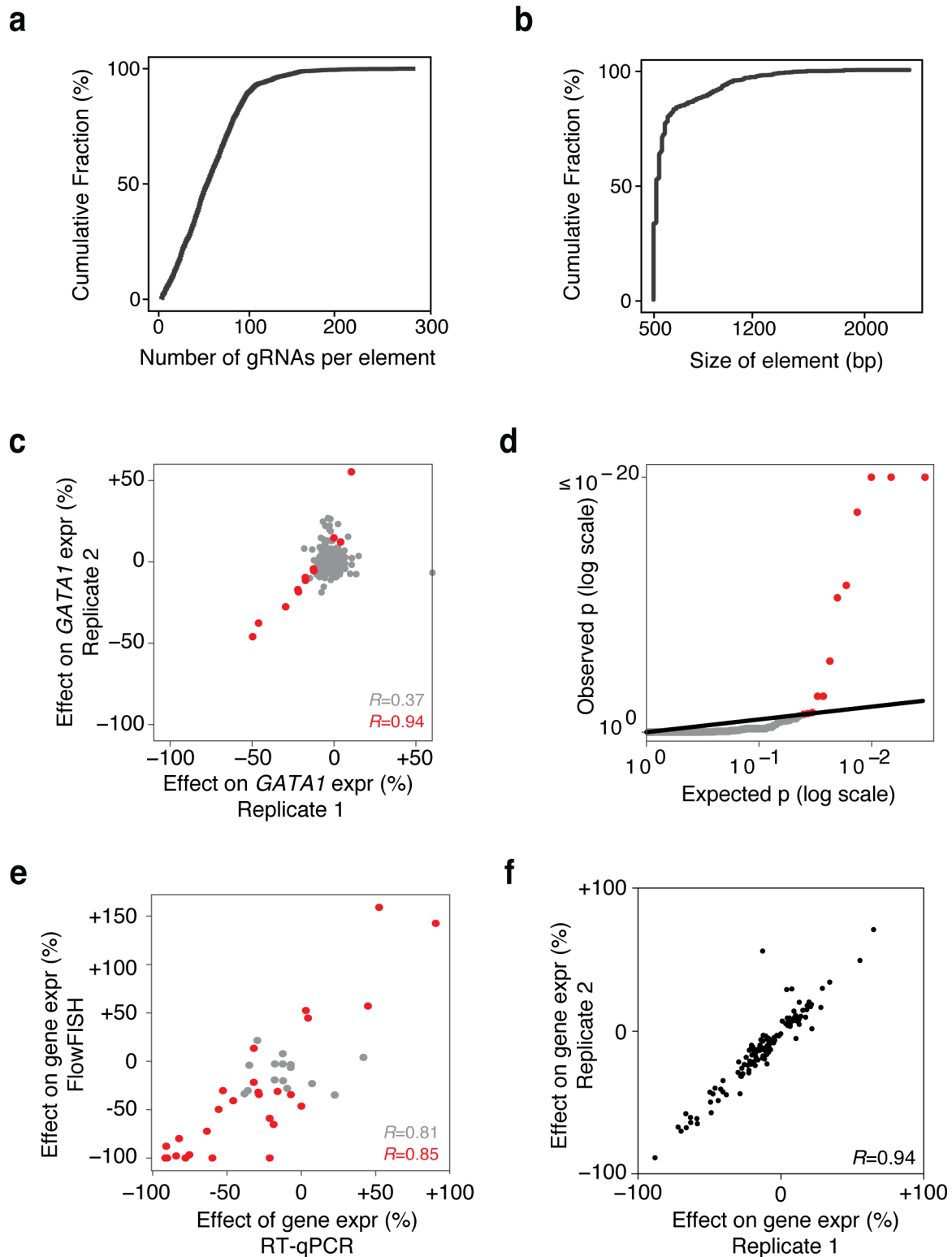
Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-019-0538-0>.

Correspondence and requests for materials should be addressed to E.S.L. or J.M.E.

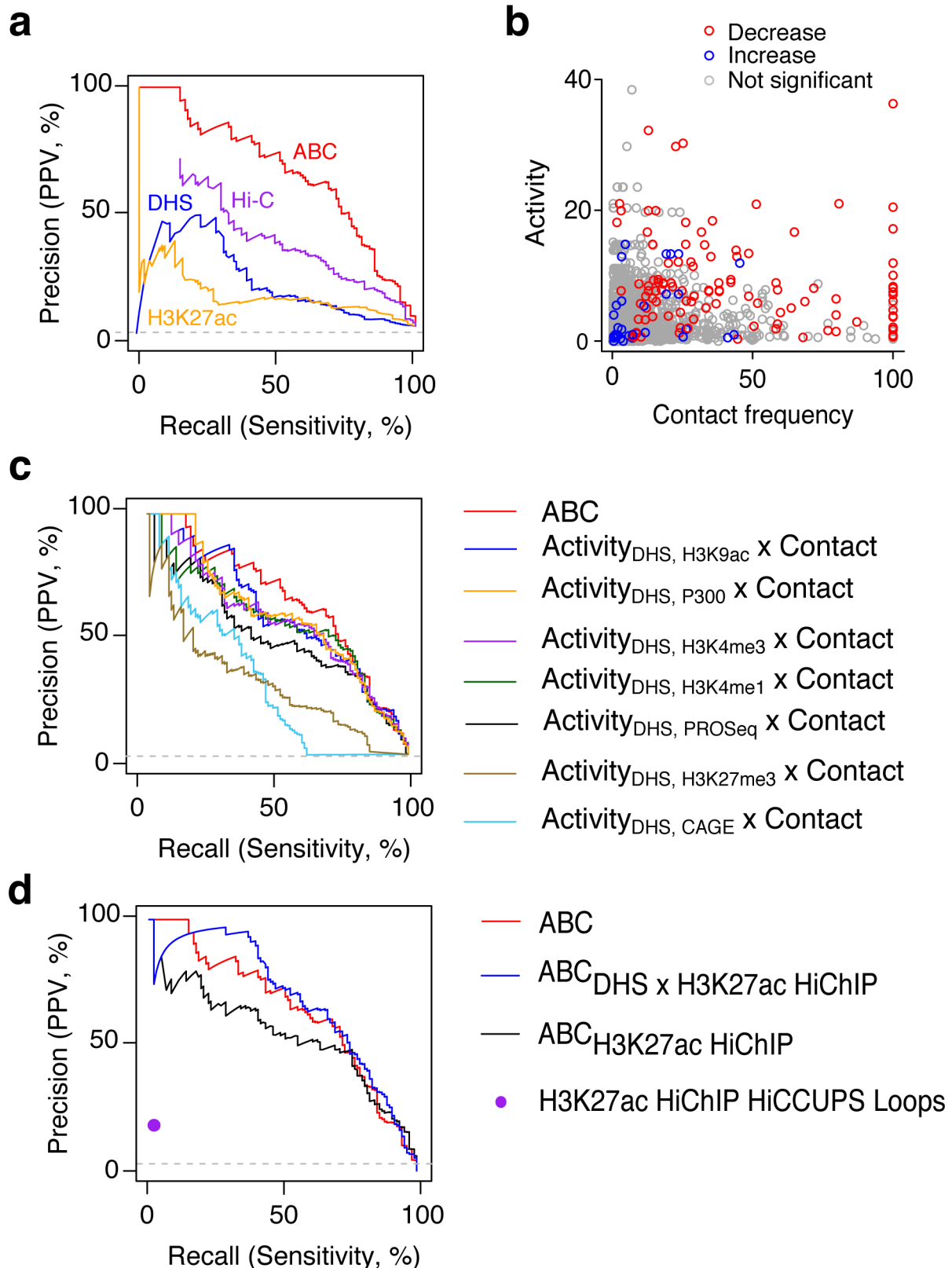
Reprints and permissions information is available at www.nature.com/reprints.



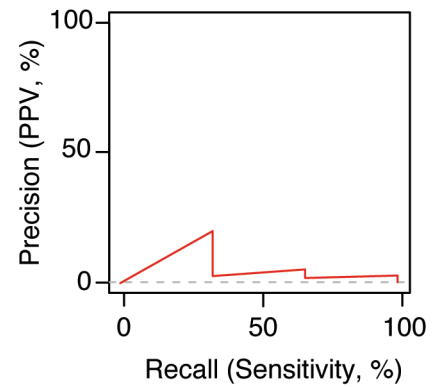
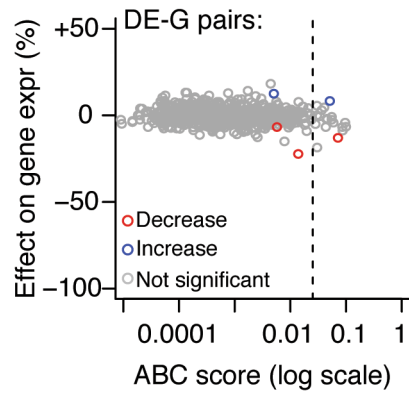
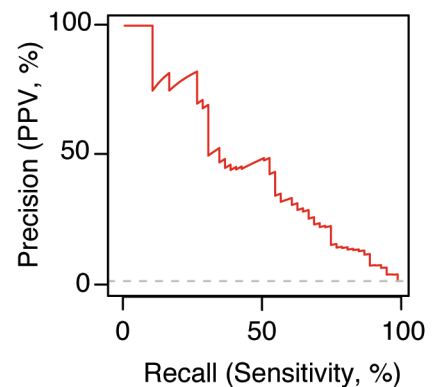
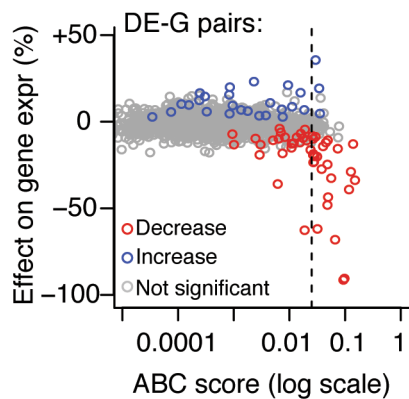
Extended Data Fig. 1 | Sorting and sequencing strategy for CRISPRi-FlowFISH Screens. **a**, K562 cells labeled with FlowFISH probesets against *RPL13A* (control gene) and *GATA1* (gene of interest) imaged by fluorescence microscopy. **b**, Histograms of FlowFISH signal (arbitrary units of fluorescence) 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. $\text{Activity}_{\text{Feature1, Feature2}} = \sqrt{\text{Feature1 RPM} \times \text{Feature2 RPM}}$. (ABC score corresponds to $\text{Activity}_{\text{DHS, H3K27ac}} \times \text{Contact}$). **d**, Precision-recall curves for the ABC model using H3K27ac HiChIP. $\text{ABC}_{\text{DHS} \times \text{H3K27ac HiChIP}}$ 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). $\text{ABC}_{\text{H3K27ac HiChIP}}$ 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 $\text{ABC}_{\text{sqr}(\text{DHS} \times \text{H3K27ac}) \times \text{Hi-C}}$. These results suggest that the ABC score computed using H3K27ac HiChIP data is an effective predictor of regulatory enhancer-gene connections.

a**Ubiquitously
Expressed
Genes****b****Tissue Specific
Genes**

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).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Flow Cytometry data were collected using Summit (v6.3.1) and CytExpert (v2.2.0.97). Sequencing data were collected using HiSeq Control Software (v2.2.68).

Data analysis

Flow Cytometry data were visualized using FlowJo (10.4.2). We used the following software for data analysis and graphical plots in the context of custom scripts: R(v3.1.1) with Bioconductor(3.0), Python(v3.4.2), matplotlib(v1.5.3), numpy(v1.15.2), Pandas(v0.23.4), Pybedtools (v0.7.8), pyBigWig(v0.3.2), pysam(v0.13), scikit-learn(v0.18.2), scipy(v0.18.1), seaborn(v0.7.1); and the following software for sequencing read mapping and analysis: BRB-seqTools (v1.3), BWA(v0.7.17), Picard(v1.731), Kallisto(v0.43.0), Bowtie(v0.12.7), Bowtie2(v2.1.0), STAR(v2.5.2b), Bedtools(v2.26.0); as described in the supplementary methods. Code to run the ABC model is available at <https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated in this study are available from GEO (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118912>), from the Open Science Framework (<https://osf.io/u4nb4/>), and in the supplementary materials.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Statistical methods were not used to predetermine sample sizes. Sample sizes used were based on standard CRISPR screening practices, and were determined to be suitable based on statistical power calculated following the experiments.
Data exclusions	Data were not excluded. Genes were selected for inclusion based on criteria outlined in the supplemental methods.
Replication	Results were highly reproducible between independent biological replicates (e.g. Extended Data Fig. 2).
Randomization	CRISPRi-FlowFISH experiments are performed as pooled screens, with gRNAs randomly infected into cells in a large pool in which experimenters are blind to which cells harbor which perturbations.
Blinding	CRISPRi-FlowFISH experiments are performed as pooled screens, with gRNAs randomly infected into cells in a large pool in which experimenters are blind to which cells harbor which perturbations.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	K562 and HEK293T cells were a gift from D. Sabatini, F2-1 mESCs were a gift from K. Plath, V6.5 mESCs were a gift from A. Meissner, NCCIT cells were purchased from ATCC.
Authentication	Cells lines were not authenticated after receipt.
Mycoplasma contamination	Cultured cell lines were tested and confirmed negative for mycoplasma contamination monthly
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

ChIP-seq

Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118912> .

Files in database submission	GSM3523691 H3K27ac-ControlClone1 GSM3523692 H3K27ac-ControlClone2
Genome browser session (e.g. UCSC)	Not applicable

Methodology

Replicates	For H3K27ac ChIP-seq data in F2-1 mESCs, 2 Biological Replicates from different mESC clones were used.
Sequencing depth	14 M total reads, 11 M unique reads, 97% of reads alignment, 30x30 paired end reads
Antibodies	H3K27ac antibody: Monoclonal Active Motif #39685, Lot 35813005
Peak calling parameters	Mapping: bowtie2-2.1.0 with default parameters to mm9 reference genome Peaks were not called on this data.
Data quality	We confirmed that ChIP-seq reads were enriched within 1 kb of the TSS, as expected for H3K27ac profiles.
Software	We used this data to count reads in particular regions of the genome using BEDTools.

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells were prepared using the PrimeFlow RNA Assay Kit (ThermoFisher 88-18005) according to the manufacturer protocol with minor modifications described in the methods.
Instrument	FACS sorting was performed on a MoFlo Astrios EQ. Flow Cytometry was performed on a Cytoflex S.
Software	Flow Cytometry data were collected using Summit (v6.3.1) and CytExpert (v2.2.0.97), and visualized using FlowJo (v10.4.2).
Cell population abundance	We confirmed via next generation sequencing that CRISPRi gRNAs targeting the TSS of the assayed gene were strongly enriched in the bin corresponding to the lowest staining intensity, as expected (see supplemental figures).
Gating strategy	Intact, single cells were selected based on FSC-height/SSC-height and SSC-width/SSC-height. Successfully stained cells were selected based on the AF488 positive control gene. Gates for sorting were drawn as 6 10% bins of increasing AF647 gene-of-interest intensity (see Extended Data Fig. 1).

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.