Measuring and Predicting Enhancer-Promoter Communication

Measuring and predicting enhancer-promoter communication

A dissertation presented

by

Charles Perry Fulco

to

The Committee on Higher Degrees in Systems Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Systems Biology

Harvard University

Cambridge, Massachusetts

April 2019

Dissertation Advisor: Professor Eric S. Lander Charles P. Fulco

Measuring and predicting enhancer-promoter communication

Abstract

Mammalian genomes harbor millions of noncoding elements called enhancers that quantitatively regulate gene expression, but it remains unclear which enhancers regulate which genes. A key bottleneck in understanding the regulatory wiring that connects noncoding regulatory elements to specific target genes has been that we have lacked scalable experimental approaches for perturbing enhancers in the genome and determining their effects on gene expression. To address this challenge, we developed new experimental approaches to systematically quantify the effects of enhancers regulating a gene of interest in a given cell type. Applying these approaches to dozens of genes uncovered complex networks of regulatory connections that could not be predicted by any existing approach. Strikingly, a simple equation based on a mechanistic model for enhancer function performed remarkably well at predicting the complex patterns of regulatory connections we observed in our datasets. This Activity-by-Contact (ABC) model involves multiplying measures of enhancer activity and enhancer-promoter 3D contacts, and can predict enhancer-gene connections based on chromatin state maps. Together, these experimental and computational approaches provide a systematic framework to understand gene regulation by enhancers and will catalyze efforts to interpret human genetic variation and manipulate gene expression for therapeutic purposes.

Table of Contents

Acknowledgements

I am extremely grateful to all the people who shaped my experiences in graduate school:

- Eric Lander for his insistence to "go for the jugular" and "speak with precision" and for building a community that makes big discoveries routine.
- Jesse Engreitz for his generous lessons in how to plan, do, lead, interpret, and describe science.
- My advisory and exam committees, Phil Sharp, Martha Bulyk, John Rinn, and Karen Adelman, for their constant support, advice, and enthusiasm.
- Joe Nasser and Ray Jones for the expert and careful analysis, and occasionally heated discussions.
- Glen Munson, Rockwell Anyoha, and Drew Bergman for leading the change in the lab.
- Vidya Subramanian and Patrick McDonel for being the voices of reason.
- Brian Cleary, Shari Grossman, Mathias Munschauer, Tim Wang, Ben Doughty, Tung Nguyen, Philine Guckelberger, Liz Perez, Michael Kane, Tejal Patwardhan, Andrew Katznelson, Floriane Ngako Kameni, Kathryn Lawrence, Atray Dixit, Oren Parnas, Rebecca Herbst, Christoph Muus, Jacob Ulirsch, John Ray, Carl de Boer, Ryan Tewhey, Steve Reilly, Chris Vockley, Ludwig Leif, Klara Sirokman, Erez Lieberman Aiden, Suhas Rao, Aviv Regev, Russell Ryan, Travis Saari, Jenny Chen, John Doench, Rajiv Khajuria, Biyu Li, Jonathan Hsu, Luca Pinello, Shirlee Wohl, Anna Green, Omar Abudayyeh, and Jonathan Gootenberg for making it all fun.
- John Calarco, Dave Shechner, Tim Nilsen, Pat Maroney, Trinh Tat, Lin He, Erich Sabio, Tony Jack, Ruth Siegel, Lisa Leon, and Mark Smith for providing me the early opportunities to explore science.
- Nicole Brellenthin, Kate Mulherin, Sam Reed, Liz Pomerantz, Erin Rist, Kristen Zarrelli, Katie Liguori, Sira Dooley Fairchild, and Geoffrey Shamu for making it all run smoothly.
- My parents Chuck and Cory, brother Cam, and in-laws Bruce and Liz for their love and support.
- My wife Sarah, for always making me smile.

Chapter 1: Introduction

Overview

Each cell in a metazoan organism contains the same genes, but genes are expressed at different levels and in different combinations in different the cell types (*1*). Regulation can occur at any step in expression, from transcribing a gene into mRNA, to mRNA splicing, export, and stability, to the efficiency of translating the mRNA into protein and the protein's stability (*2-4*). In practice, the predominant point of regulation is the selection of which genes to transcribe and how highly to transcribe them (*1*).

Transcription is controlled by *cis*-regulatory sequences in the genome that recruit *trans* factors to promote the production of the mRNA corresponding to a given gene. Each cell type expresses a distinct set of *trans* factors to orchestrate the expression of the correct subset of genes. Promoters, the *cis*-regulatory sequences in the immediate vicinity of the start of each gene, recruit and position RNA polymerase, the enzyme that carries out transcription (*5*). In most eukaryotes, transcription also depends on enhancers, elements separated from promoters by potentially large genomic distances and that can further refine gene expression across cell types and through developmental time (*6-8*). It is thought that the physical separation between enhancers and promoters provides multicellular organisms the flexibility needed to differentially regulate the same complement of genes in each cell type (*9*).

The genomic separation between enhancers and their target genes raises a central question for the understanding of gene regulation: how do enhancers regulate specific target genes?

In this thesis, I present our work toward understanding how enhancers regulate specific target genes by developing perturbation-based tools to characterize the regulatory functions of noncoding elements in their native genomic contexts. To frame this work, I will review the discovery and characterization of enhancers and our current understanding for how enhancers regulate specific target genes.

Discovery and definition of enhancers

As early as the nineteen-sixties, the conceptual paradigm for understanding transcriptional regulation was already established (*10*). Transcription factors bind sequences in the to promote or inhibit binding of the necessary factors for transcription, including RNA polymerase, the enzyme responsible for transcribing DNA into RNA (*11*). In this way, the transcription of each gene results from the integration of positive and negative signals that reflect cell state.

These principles are largely the result of work in prokaryotes and bacteriophages. In these systems, the sequences that control expression are in the immediate vicinity of the transcriptional start site of a gene. It was widely assumed that transcriptional regulation in metazoans would mirror the situation in prokaryotes, albeit with more complicated regulation. However, it has become clear that in addition to promoters, transcription in metazoans is regulated by enhancers, non-coding regulatory that influence gene expression across large genomic distances (*12*).

The first enhancer was discovered when a group led by Walter Shaffner cloned the rabbit beta-globin gene and promoter into two different plasmid vectors to introduce into HeLa cells. One of the vectors expressed beta-globin at dramatically higher levels, despite the only difference being that it included a portion of the genome of the SV40 virus that was expected to be functionally inert. Deletion analysis localized the enhancer element to a small portion of the viral SV40 genome (*13*).

Subsequent experiments revealed the "enhancing effect" remained when inverting the orientation of the beta-globin gene in the plasmid, when increasing the distance between the enhancer and the gene, but not when the gene and enhancer were on separate plasmids in the same cells (*13-16*). These properties came to constitute the classical definition of an enhancer, and to distinguish enhancers from other non-coding regulatory elements such as promoters (*17*).

Upon this discovery, Walter Shaffner speculated presciently "that cellular 'enhancers' are activating the genes within each chromosomal domain, and that classes of different 'enhancers' are involved in the developmental, as well as tissue-specific, expression of genes" (*13*), and quickly applied the same plasmid-based assay approach to identify enhancers in a mammalian genome.

The first mammalian enhancer was discovered from the immunoglobin heavy chain (IgH) locus (*18, 19*). Unlike the SV40 enhancer, which stimulated expression in any cell type, the IgH enhancer appeared active only in B-cells, the cell type in which IgH is highly expressed. The discovery of the IgH enhancer revealed two additional properties of enhancers that have proved widespread. First, enhancers tend to be cell type specific, driving expression in some cell types but not others. Indeed, the IgH enhancer was the "first component identified in the then enigmatic phenomenon of cell type-specific expression" (*20*). Second, the IgH enhancer is several thousand base pairs from the IgH promoter, the nearest and presumed target promoter, which suggested that enhancers may activate transcription over long distances in their endogenous locations in the genome.

Mechanisms of enhancer function

Platforms for transcription factor binding

Enhancers are DNA sequences 50-1000 base pairs long functionally defined as elements able to activate transcription from a promoter over long distances and in an orientation-dependent manner. Biochemically, enhancers act through the binding of transcription factors that are recruited to enhancer DNA through interactions with short sequence motifs. These sequence-specific transcription factors in turn recruit transcriptional co-activators through protein-protein interactions. The transcription factors and co-activators assembled at enhancers can activate transcription at any of its rate-limiting steps (*21*), including recruitment of the per-initiation complex (PIC) (*22*), initiation (*23*), polymerase pause release (*24*), and elongation (*25*).

It was initially thought that enhancers activate transcription by facilitating recruitment of RNA polymerase or the pre-initiation complex (PIC) to the target promoter (*26*). This was by analogy to the situation in prokaryotes and yeast (which largely lack enhancers), in which recruitment of polymerase is the major rate-limiting and regulated step in transcription (*27*). Indeed, enhancers do interact with general transcription factors and Mediator, protein factors critical for the formation of the pre-initiation complex and recruitment of RNA polymerase (*26*). However, the situation in metazoans appears to be more complex. Recruitment of polymerase to promoters is not sufficient to activate transcription in mammalian cells (*28*), and polymerase binding does not appear to be a key control point in mammalian transcription (*29*).

The emerging view is that an important (even predominant) mode of enhancer function is that they regulate the release of polymerase from promoter-proximal pausing (*30-33*). After transcriptional initiation, the polymerase pauses for many minutes after transcribing just~40 nucleotides (*34*).

Release from this promoter-proximal pausing into productive elongation is a major rate-limiting step in transcription, and requires the positive elongation factor P-TEFb (*35-38*). It is thought that factors bound at enhancers play a key role in bringing P-TEFb to promoters, such as in one noteworthy example in which BRD4 is required to recruit P-TEFb to the *FOSL1* promoter and for polymerase paused there to progress into productive elongation (*39*).

Enhancer-associated histone modifications

Many of the coactivators recruited to enhancers covalently modify the histone proteins that bind the DNA flanking the enhancers. There are >100 described histone modifications, potentially forming a vast combinatorial "histone code" that conveys regulatory signals and retains regulatory information through DNA replication, when most transcription factors dissociate from their binding sites (*40, 41*). Indeed, in some cases recruitment of histone modifying complexes has been observed to affect gene expression (*42-44*), and mutations in histones at modifiable positions that mimic or prevent specific modifications are transforming events in cancer (*45*).

Histone modifications can affect enhancer function and gene expression in two ways. First, the modifications on histones alter the biophysical properties of the chromatin fiber to influence chromatin compaction. *In vitro* studies have shown that histones form a less compact confirmation when acetylated (*46*), and this is thought to facilitate the ability of transcription factors or polymerase to access binding sites on the DNA (*47, 48*). Second, these modifications also provide a binding platform for additional layers of transcriptional cofactors. For example, BRD4 is recruited to enhancers by acetylated histones such as H3K27ac and promotes pause release and transcriptional elongation by recruiting p-TEFb to target promoters (*49*).

However, it not known which marks are causally related to gene expression, and it is thought that many marks may be incidental consequences of cofactor recruitment or transcriptional activity (*50*). For example, H3K4me1 is a mark associated with distal enhancers, but cells engineered to express catalytically inactive versions of the writers of this mark (the proteins Mll3 and Mll4) have only modest changes in gene expression. Notably, cells lacking these proteins completely have dramatic changes in expression, suggesting the histone modifying activity of these factors is not central to their role as transcriptional coactivators (*51*).

While their mechanistic roles in enhancer function are not clear, these histone modifications distinguish putative enhancers from other non-coding elements, and the presence of certain histone modifications, especially H3K27ac, is predictive for gene expression and for the ability of an element to act as an enhancer in plasmid based reporter assays (*52, 53*).

Physical contact between enhancers and target genes

It was initially unclear how to reconcile two defining observations about enhancers: (i) enhancers can activate at long distances (many kilobases) from their target promoters; and (ii) enhancers do not activate promoters on separate plasmids when co-transfected into the same cell. That is, enhancers can be far from their target promoters, but must nevertheless be connected.

Experiments in which an enhancer and a promoter were placed on separate DNA fragments revealed that enhancers can communicate when joined non-covalently through a protein bridge (e.g. biotin/streptavidin) (*54*) or by interweaving the plasmids (*55*). This suggested that enhancers can function when in physical proximity to a target promoter but do not require a continuous tether of DNA, ruling out models in which factors recruited to enhancers reach distant prompters by sliding across the DNA or by initiating a chain of adjacent binding events (*56*). These observations supported what has become the dominant model: that enhancers function by physically contacting target promoters and looping out the intervening sequence.

Direct evidence for physical association between enhancers and promoters within the nucleus first came from experimental measurements that quantify the contact frequency between fragments of DNA, such as capturing chromosome confirmation (3C) (*57*). In one seminal study, 3C at the mouse beta-globin locus found that in fetal liver, which expresses beta-globin, the LCR was in more frequent contact with the beta-globin promoter than with intervening sequences (*58*). Notably, in brain tissue, which does not express beta-globin, the inactive LCR did not have higher contact with the beta-globin promoter, which suggested that the looping between enhancers might itself be regulated. Similarly, DNA FISH demonstrated that in some cases enhancers and promoters are localized particularly in cell types in which they are active (*59*).

Recently, single molecule imaging approaches that directly reveal the physical interactions between enhancers and promoters as well as transcriptional output have provided strong evidence that contact is required for enhancers to regulate activate transcription, and that transcription ceases immediately when the enhancer disengages from the promoter (*60*). At least in this experimental system, enhancer-promoter contact is necessary for transcriptional activation.

Together, these observations indicate that enhancers physically contact their target promoters by looping out potentially large amounts (up to 2Mb) of the intervening DNA. It is important to note that these looping interactions are not necessarily stable interactions. Rather, loops between enhancers and promoters can form and dissociate dynamically (*60*). This dynamic movement allows many loci in a region to interact (*58, 61*).

Genome-wide enhancer maps

In order to understand how the genome integrates developmental and environmental cues in the control of gene expression, it is necessary to recognize enhancer elements in the genome, to discern in which cell types enhancers are active and, eventually, to know which genes they control.

One hallmark of regulatory elements was accessibility to digestion by DNase (assayed genomewide by DNase-seq), as transcription factor binding displaces the histones proteins that would otherwise bind enhancer DNA (*62*). As described above, enhancers are also characterized by histone modifications such as H3K27ac and H3K4me1 (*63*) and the binding of transcriptional cofactors such as P300 (*64*). Elements bearing these features have been found to correspond, albeit imperfectly, with elements observed to have enhancer activity in plasmid based reporter assays (*65*). Efforts such as the ENCODE project have leveraged these insights to identify and catalogue functional genomic elements in multiple cell types and species (*66*). Currently the catalogue of human putative enhancer elements contains over 1 million distinct genomic regions (*67*).

These systematic maps of regulatory elements across cell types and species have provided several insights into the functions of non-coding regulatory elements including enhancers. First, elements with the biochemical signatures of enhancers (e.g. H3K27ac) have highly cell-type specific activity, apparently active in a minority of cell types and quiescent in others, consistent with a role for enhancers in creating cell type specific phenotypes (*50*). Second, genetic variants associated to common diseases by GWAS studies are highly enriched in putative enhancers, suggesting disordered gene regulation plays a central role in pathophysiology (*68, 69*). Finally, these maps highlight the central importance of non-coding regulatory elements in development, as highly regulated developmental genes tend to be located in gene deserts with few genes and many highly conserved, putative enhancers (*70*). For example the proto-oncogene and transcriptional regulator *MYC* is encoded within a 3Mb region containing no other protein coding genes, but dozens of putative enhancers that are thought to precisely control *MYC* expression (*71, 72*).

While genome-wide approaches have enumerated putative enhancers, the vast majority have not been functionally tested, and we do not know which genes in which cell types they regulate (if any). There are very few cases in which natural or experimental evidence has demonstrated that a given enhancer actually regulates a specific target gene in the genome. These few cases have suggested that the network connecting enhancers and targets genes is staggeringly complex; there are examples of enhancers regulating one or more target genes across large genomic distances and instances of enhancers appearing to "skip" over a proximal to gene to regulate a more distant one (*73, 74*). For example, mice with activating insertions into an enhancer (termed the ZRS) within the intron of the *Lmbr1* gene have limb malformations caused by increased expression of the *Shh* gene nearly 1 Mb away (*75*). Moreover, numerous studies (including the work presented in the thesis) have demonstrated that multiple enhancers can contribute the expression of a single gene in the same cell type (*76, 77*) or across cell types (*78*). In the face of this complexity, we currently lack a systematic, mechanistic understanding of how enhancers achieve specificity for their target genes.

Mechanisms of enhancer specificity

Enhancers are broadly compatible with many promoters

One potential mechanism ("biochemical compatibility") to explain the specificity of enhancers for certain promoters is that enhancers can only regulate promoters with complementary combinations of compatible transcription factors, that the specificity of an enhancer for its target promoter(s) is encoded in the sequences of the enhancer and promoter. This mechanism is tested in experiments that remove the enhancer and promoter from their genomic context to test the ability of enhancers to activate different promoters in plasmid-based assays.

The largest and most compelling studies of this type tested fragments from the entire *Drosophila* genome for enhancer activity when paired with one of several exemplar promoters (*79*), or as promoters paired with several exemplar enhancers (*80*). In both formats, there appear to be just two classes of enhancers and promoters: housekeeping genes and developmental genes. Housekeeping gene promoters tended to be strongly activated only by enhancer elements very near or overlapping housekeeping gene promoters, while developmental gene promoters were activated by intergenic elements distal to transcription start sites. Moreover, the enhancers within a class activated all promoters in the class with a consistent fold change. Other smaller studies suggest that this broad compatibility (*20*) and limited core-promoter-dictated specificity (*81*) may also extend to mammalian cells. Thus, enhancers appear to be broadly compatible with many, but not all promoters. At least in *Drosophila*, there appears to be a clear distinction between the enhancers that strongly activate housekeeping promoters versus developmental promoters.

The notion that enhancers are compatible with many target promoters is also supported by laboratory experiments or experiments of nature that juxtaposed enhancers with non-native targets in the genome. For example, Burkitt's lymphoma, a subtype of B-cell lymphoma, is characterized by chromosomal rearrangements that bring the highly active IgH enhancer on chromosome 14 into close proximity with the oncogene *c-MYC* on chromosome 8 (*82*). Similarly, the human beta-globin LCR is able to activate other genes when experimentally inserted into a distant location in the mouse genome (*83*).

The pattern of 3D contacts in the genome constrains enhancer-promoter communication.

An additional mechanism to provide specificity for enhancers is that the three dimensional structure of the genome constrains the set of promoters with which an enhancer has the opportunity to interact. Indeed, it has long been observed that some genetic elements are capable of insulating promoters from the effects of enhancers on the opposite side of the element (*7, 84, 85*). It has only recently been understood that many insulators act by shaping the three dimensional structure of the genome, which in turn constrains enhancer-promoter contacts (*86*).

Hi-C, an assay that characterizes the chromatin confirmation of the entire genome (*87*), has revealed that the genome is packaged into domains of approximately 100 kb to 1 Mb in length termed topologically associating domain (TADs). These domains are defined as regions in which loci within the domain exhibit increased contact relative to regions outside the domain at the same linear distance (*88*). Globally, TADs appear to be functional units of genome regulation. Both gene expression and chromatin state are more correlated across cell types for genes within the same TAD than genes in different TADs at the same linear distance (*89, 90*), suggesting genes within a TAD may be influenced by common regulatory signals.

The boundaries of many mammalian TADs are characterized by the binding of the insulator factor CTCF, and are thought to insulate the promoters in one domain from the effects of enhancers in another (*91*). Ablation of these TAD boundary insulator elements can allow enhancers in one TAD to interact with genes in another (*92, 93*). For example, one study perturbed a single CTCF element at the boundary between two adjacent TADs, and observed that a gene in one TAD, *PDGFRA*, increased contact with an apparently highly active putative enhancer in the other, resulting in activation of *PDGFRA* (*89*).

Based on the observation that enhancers physically contact target promoters, it has been hypothesized that enhancer specificity is encoded in the specific looping structures adopted by chromatin in a given cell type. Indeed, genome-wide characterizations of chromatin contacts by Hi-C have revealed frequent focal loops between promoters and elements with enhancer-associated histone modifications, and the genes associated with a loop are higher expressed in cell types where the loop is present than in cell types where the loop is not (*94*).

Consistent with an instructive role for specific chromatin loops in enhancer specificity, deletion studies in *Drosophila* have identified "tethering elements" near the promoters of some genes required to permit activation from distal enhancers (*95, 96*). These elements are hypothesized to function by stabilizing the looping structure between an enhancer and promoter. Moreover, recent experiments in which artificial transcription factors were used to create a chromatin loop between enhancers and promoters found that forced looping is sufficient for an enhancer to activate a promoter (*61, 97-99*).

Emerging evidence complicates the view that insulation by TAD boundaries or the existence of focal chromatin loops represent predominant means of dictating enhancer-gene interactions.

Degradation of CTCF or cohesin abolishes the demarcation of the genome into TADs as well as most focal loops, but has only modest effects on transcriptional regulation (*100, 101*). This indicates that even without TAD boundaries or loops, enhancers do not generally activate inappropriate targets and are often still able to regulate their native targets, suggesting these structures are not the primary determinants of enhancer-gene communication.

Contributions of this thesis

Despite the tremendous progress over the last four decades in understanding how enhancers tune gene expression, a critical question remains: how do enhancers target specific promoters? A key bottleneck in understanding enhancer-promoter communication has been that we have lacked the tools to characterize the functions of large numbers of enhancers in the genome.

In Chapter 2, I describe our work to develop a high-throughput approach based on CRISPR interference (CRISPRi) to characterize the functions of gene regulatory elements in their native genomic contexts (*102*). In this work, we leveraged pooled CRISPR screens (*103, 104*) in combination with CRISPR interference (CRISPRi)—which alters chromatin state at targeted loci through recruitment of a KRAB effector domain fused to catalytically dead Cas9 (*105-107*) — to characterize the functions of 1.29 Mb of genomic sequence around the essential transcription factors *GATA1* and *MYC* in K562 erythroleukemia cells. This method allowed us for the first time to systematically define the enhancers quantitatively tuning the expression of a gene in a given cell type.

The work described in Chapter 3 extended the CRISPRi-based functional mapping approach to develop a more comprehensive understanding of enhancer-promoter connectivity.

First, we expanded the CRISPR functional genomics toolkit by combining pooled CRISPR screens with RNA FISH and flow cytometry. This approach enables high-throughput, quantitative genetic screens for any arbitrary phenotype that can be linked to expression of one or several genes. Our FlowFISH approach precludes the need to knock in reporter genes or develop complicated cellular assays for pooled screens, and can be readily applied for other studies of non-coding regulatory elements or protein coding genes.

Next, we systematically mapped the enhancers controlling the expression of 28 genes in K562 cells. This perturbation dataset precisely quantifies the effect size of each enhancer and allowed us to evaluate, for the first time, existing models of enhancer-gene specificity in a general way across many genomic loci. No existing predictive tool explained the complex patterns of connections we observed.

In order to predict the observed functional connections, we developed the "Activity-by-Contact" (ABC) model 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"). The ABC model represents the first means to accurately predict enhancer-gene connections based on epigenetic data.

Moreover, the ABC model provides a new conceptual understanding for how enhancers regulate specific genes in the genome. Quantitative contact frequency of enhancers with target genes — rather than the presence of specific loops and domains — predicts enhancer-gene regulation. The success of the ABC model demonstrates that functional specificity can arise from the precise arrangement and activities of enhancers in the genome, even in the absence of biochemical specificity between enhancers and promoters.

Finally, in Chapter 3, I conclude with my perspective on the remaining questions in how enhancers and promoters collaborate to control transcription and the outlook for applying our growing understanding of transcriptional regulation to dissect the contributions of noncoding genetic variation to human disease and to manipulate gene expression for therapeutic purposes.

References

1. M. Levine, R. Tjian, Transcription regulation and animal diversity. *Nature* **424**, 147-151 (2003).

2. M. J. Moore, From birth to death: the complex lives of eukaryotic mRNAs. *Science* **309**, 1514-1518 (2005).

3. W. C. Merrick, Mechanism and regulation of eukaryotic protein synthesis. *Microbiol Rev* **56**, 291-315 (1992).

4. T. Platt, Transcription termination and the regulation of gene expression. *Annu Rev Biochem* **55**, 339-372 (1986).

5. R. G. Roeder, Transcriptional regulation and the role of diverse coactivators in animal cells. *FEBS Lett* **579**, 909-915 (2005).

6. M. Bulger, M. Groudine, Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev Biol* **339**, 250-257 (2010).

7. G. A. Maston, S. K. Evans, M. R. Green, Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**, 29-59 (2006).

8. M. Bulger, M. Groudine, Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339 (2011).

9. M. Levine, C. Cattoglio, R. Tjian, Looping back to leap forward: transcription enters a new era. *Cell* **157**, 13-25 (2014).

10. K. Struhl, Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98**, 1-4 (1999).

11. F. Jacob, J. Monod, Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-356 (1961).

12. W. Schaffner, Enhancers, enhancers - from their discovery to today's universe of transcription enhancers. *Biol Chem* **396**, 311-327 (2015).

13. J. Banerji, S. Rusconi, W. Schaffner, Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).

14. M. Fromm, P. Berg, Simian virus 40 early- and late-region promoter functions are enhanced by the 72-base-pair repeat inserted at distant locations and inverted orientations. *Mol Cell Biol* **3**, 991-999 (1983).

15. B. Wasylyk, C. Wasylyk, P. Augereau, P. Chambon, The SV40 72 bp repeat preferentially potentiates transcription starting from proximal natural or substitute promoter elements. *Cell* **32**, 503-514 (1983).

16. P. Moreau *et al.*, The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res* **9**, 6047-6068 (1981).

17. P. Gruss, Magic enhancers? *DNA* **3**, 1-5 (1984).

18. J. Banerji, L. Olson, W. Schaffner, A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729-740 (1983).

19. S. D. Gillies, S. L. Morrison, V. T. Oi, S. Tonegawa, A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33**, 717-728 (1983).

20. M. Kermekchiev, M. Pettersson, P. Matthias, W. Schaffner, Every enhancer works with every promoter for all the combinations tested: could new regulatory pathways evolve by enhancer shuffling? *Gene Expr* **1**, 71-81 (1991).

21. E. M. Blackwood, J. T. Kadonaga, Going the distance: a current view of enhancer action. *Science* **281**, 60-63 (1998).

22. R. G. Roeder, The complexities of eukaryotic transcription initiation: regulation of preinitiation complex assembly. *Trends Biochem Sci* **16**, 402-408 (1991).

23. J. W. Conaway, J. P. Hanley, K. P. Garrett, R. C. Conaway, Transcription initiated by RNA polymerase II and transcription factors from liver. Structure and action of transcription factors epsilon and tau. *J Biol Chem* **266**, 7804-7811 (1991).

24. E. B. Rasmussen, J. T. Lis, In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci U S A* **90**, 7923-7927 (1993).

25. R. J. Sims, 3rd, R. Belotserkovskaya, D. Reinberg, Elongation by RNA polymerase II: the short and long of it. *Genes & development* **18**, 2437-2468 (2004).

26. H. Szutorisz, N. Dillon, L. Tora, The role of enhancers as centres for general transcription factor recruitment. *Trends Biochem Sci* **30**, 593-599 (2005).

27. M. Ptashne, Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci* **30**, 275-279 (2005).

28. D. R. Dorris, K. Struhl, Artificial recruitment of TFIID, but not RNA polymerase II holoenzyme, activates transcription in mammalian cells. *Mol Cell Biol* **20**, 4350-4358 (2000).

29. C. R. Bartman *et al.*, Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. *Molecular cell* **73**, 519-532 e514 (2019).

30. L. J. Core, J. J. Waterfall, J. T. Lis, Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848 (2008).

31. J. Zeitlinger *et al.*, RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39**, 1512-1516 (2007).

32. G. W. Muse *et al.*, RNA polymerase is poised for activation across the genome. *Nat Genet* **39**, 1507-1511 (2007).

33. M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, R. A. Young, A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77-88 (2007).

34. B. Li, J. A. Weber, Y. Chen, A. L. Greenleaf, D. S. Gilmour, Analyses of promoter-proximal pausing by RNA polymerase II on the hsp70 heat shock gene promoter in a Drosophila nuclear extract. *Mol Cell Biol* **16**, 5433-5443 (1996).

35. J. T. Lis, P. Mason, J. Peng, D. H. Price, J. Werner, P-TEFb kinase recruitment and function at heat shock loci. *Genes & development* **14**, 792-803 (2000).

36. N. F. Marshall, D. H. Price, Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J Biol Chem* **270**, 12335-12338 (1995).

37. N. F. Marshall, D. H. Price, Control of formation of two distinct classes of RNA polymerase II elongation complexes. *Mol Cell Biol* **12**, 2078-2090 (1992).

38. T. Wada, T. Takagi, Y. Yamaguchi, D. Watanabe, H. Handa, Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. *EMBO J* **17**, 7395-7403 (1998).

39. A. Zippo *et al.*, Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation. *Cell* **138**, 1122-1136 (2009).

40. Y. Zhao, B. A. Garcia, Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb Perspect Biol* **7**, a025064 (2015).

41. T. Jenuwein, C. D. Allis, Translating the histone code. *Science* **293**, 1074-1080 (2001).

42. E. M. Mendenhall *et al.*, Locus-specific editing of histone modifications at endogenous enhancers. *Nat Biotechnol* **31**, 1133-1136 (2013).

43. J. Szulc, M. Wiznerowicz, M. O. Sauvain, D. Trono, P. Aebischer, A versatile tool for conditional gene expression and knockdown. *Nature methods* **3**, 109-116 (2006).

44. I. B. Hilton *et al.*, Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol* **33**, 510-517 (2015).

45. K. F. V. Tabar, Histone Mutations in Cancer. *Annual Review of Cancer Biology* **2**, (2018).

46. M. Garcia-Ramirez, C. Rocchini, J. Ausio, Modulation of chromatin folding by histone acetylation. *J Biol Chem* **270**, 17923-17928 (1995).

47. P. Tessarz, T. Kouzarides, Histone core modifications regulating nucleosome structure and dynamics. *Nat Rev Mol Cell Biol* **15**, 703-708 (2014).

48. J. L. Workman, R. E. Kingston, Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu Rev Biochem* **67**, 545-579 (1998).

49. G. E. Winter *et al.*, BET Bromodomain Proteins Function as Master Transcription Elongation Factors Independent of CDK9 Recruitment. *Molecular cell* **67**, 5-18 e19 (2017).

50. E. Calo, J. Wysocka, Modification of enhancer chromatin: what, how, and why? *Molecular cell* **49**, 825-837 (2013).

51. K. M. Dorighi *et al.*, Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Molecular cell* **66**, 568-576 e564 (2017).

52. R. Karlic, H. R. Chung, J. Lasserre, K. Vlahovicek, M. Vingron, Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**, 2926-2931 (2010).

53. E. M. Mendenhall, B. E. Bernstein, Chromatin state maps: new technologies, new insights. *Curr Opin Genet Dev* **18**, 109-115 (2008).

54. H. P. Mueller-Storm, J. M. Sogo, W. Schaffner, An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge. *Cell* **58**, 767-777 (1989).

55. M. Dunaway, P. Droge, Transactivation of the Xenopus rRNA gene promoter by its enhancer. *Nature* **341**, 657-659 (1989).

56. M. Ptashne, Gene regulation by proteins acting nearby and at a distance. *Nature* **322**, 697-701 (1986).

57. J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing chromosome conformation. *Science* **295**, 1306-1311 (2002).

58. B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, W. de Laat, Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell* **10**, 1453-1465 (2002).

59. T. Amano *et al.*, Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell* **16**, 47-57 (2009).

60. H. Chen *et al.*, Dynamic interplay between enhancer-promoter topology and gene activity. *Nat Genet*, (2018).

61. C. R. Bartman, S. C. Hsu, C. C. Hsiung, A. Raj, G. A. Blobel, Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Molecular cell* **62**, 237-247 (2016).

62. D. S. Gross, W. T. Garrard, Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**, 159-197 (1988).

63. A. Rada-Iglesias *et al.*, A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283 (2011).

64. Q. Wang, J. S. Carroll, M. Brown, Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Molecular cell* **19**, 631-642 (2005).

65. A. Visel *et al.*, ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858 (2009).

66. E. P. Consortium *et al.*, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

67. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

68. M. T. Maurano *et al.*, Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).

69. A. Visel, E. M. Rubin, L. A. Pennacchio, Genomic views of distant-acting enhancers. *Nature* **461**, 199-205 (2009).

70. K. Lindblad-Toh *et al.*, Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).

71. J. Sotelo *et al.*, Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci U S A* **107**, 3001-3005 (2010).

72. D. Levens, You Don't Muck with MYC. *Genes Cancer* **1**, 547-554 (2010).

73. D. Shlyueva, G. Stampfel, A. Stark, Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-286 (2014).

74. J. van Arensbergen, B. van Steensel, H. J. Bussemaker, In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol* **24**, 695-702 (2014).

75. L. A. Lettice, A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12**, 1725-1735 (2003).

76. M. Osterwalder *et al.*, Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239-243 (2018).

77. J. W. Hong, D. A. Hendrix, M. S. Levine, Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314 (2008).

78. S. A. Vokes, H. Ji, W. H. Wong, A. P. McMahon, A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes & development* **22**, 2651-2663 (2008).

79. M. A. Zabidi *et al.*, Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556-559 (2015).

80. C. D. Arnold *et al.*, Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077 (2013).

81. F. C. Wefald, B. H. Devlin, R. S. Williams, Functional heterogeneity of mammalian TATA-box sequences revealed by interaction with a cell-specific enhancer. *Nature* **344**, 260-262 (1990).

82. J. Erikson, A. ar-Rushdi, H. L. Drwinga, P. C. Nowell, C. M. Croce, Transcriptional activation of the translocated c-myc oncogene in burkitt lymphoma. *Proc Natl Acad Sci U S A* **80**, 820-824 (1983).

83. D. Noordermeer *et al.*, Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region. *PLoS Genet* **4**, e1000016 (2008).

84. P. K. Geyer, V. G. Corces, DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes & development* **6**, 1865-1873 (1992).

85. R. Kellum, P. Schedl, A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Mol Cell Biol* **12**, 2424-2431 (1992).

86. T. Sexton, G. Cavalli, The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049-1059 (2015).

87. E. Lieberman-Aiden *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

88. J. R. Dixon *et al.*, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).

89. W. A. Flavahan *et al.*, Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110-114 (2016).

90. E. P. Nora *et al.*, Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385 (2012).

91. M. W. Vermunt, D. Zhang, G. A. Blobel, The interdependence of gene-regulatory elements and the 3D genome. *J Cell Biol* **218**, 12-26 (2019).

92. M. Franke *et al.*, Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265-269 (2016).

93. D. Hnisz, D. S. Day, R. A. Young, Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167**, 1188-1200 (2016).

94. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

95. V. C. Calhoun, A. Stathopoulos, M. Levine, Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proc Natl Acad Sci U S A* **99**, 9243-9247 (2002).

96. V. C. Calhoun, M. Levine, Long-range enhancer-promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex. *Proc Natl Acad Sci U S A* **100**, 9878-9883 (2003).

97. W. Deng *et al.*, Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244 (2012).

98. W. Deng *et al.*, Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849-860 (2014).

99. S. L. Morgan *et al.*, Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nat Commun* **8**, 15993 (2017).

100. S. S. P. Rao *et al.*, Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324 (2017).

101. E. P. Nora *et al.*, Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922 (2017).

102. C. P. Fulco *et al.*, Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773 (2016).

103. O. Shalem *et al.*, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).

104. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).

105. L. A. Gilbert *et al.*, CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451 (2013).

106. P. I. Thakore *et al.*, Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature methods* **12**, 1143-1149 (2015).

107. L. A. Gilbert *et al.*, Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647-661 (2014).

Chapter 2: Systematic mapping of functional enhancer-promoter connections with CRISPR interference

Most of this chapter was originally published as: Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*. 2016 Nov 11.

Preface

In this chapter, I describe a collaborative project within Eric Lander's lab to develop a high-throughput approach based on CRISPR interference (CRISPRi) to characterize the functions of gene regulatory elements in their native genomic contexts (*1*). This method allowed us for the first time to systematically define the enhancers quantitatively tuning the expression of a gene in a given cell type.

This effort grew out of a shared vision in Eric's lab that perturbing noncoding elements such as lncRNAs and enhancers in their native genomic locations would be critical for understanding genome function. Jesse Engreitz, Mathias Munschauer, and I set out to develop high throughput genetic screening tools to perturb the noncoding genome, including enhancers, promoters, and lncRNAs. We worked together to generate and validate the K562 CRISPRi cell line. We initially designed tiling CRISPR gRNA libraries to cover the region containing the essential transcription factor *MYC* and the lncRNA *PVT1* (which we mused might play a role in *MYC* regulation). Jesse worked with Shari Grossman with advice from Russell Ryan to expand the library to cover additional sites of interest throughout the region.

I worked side-by-side with Rockwell Anyoha to clone the CRISPRi library, perform the pooled essentiality screen in K562 cells, and validate the identified enhancers using single sgRNA qPCR. Jesse and I wrote the code to analyze the screens. Michael Kane cloned the enhancers into a plasmid and quantified the enhancers' activities in luciferase assays. Liz Perez performed ChIP-qPCR to confirm that even at a site we found not to affect *MYC* expression or proliferation, CRISPRi reduced H3K27ac occupancy.

At the time there were no examples of an enhancer discovered using CRISPRi, so it was not known how many such cases would be validated by more conventional, orthologous approaches. That is, the false positive rate of the approach was completely unknown. To validate these enhancers and the CRISPRi screening approach, Jesse and I devised a strategy to test these enhancers by genetic deletion and read out the direct, *cis*-effects on *MYC* expression using ddPCR. Glen Munson engineered the cell lines and Liz performed the ddPCR with some help from Rockwell.

Based on conversations with Shari, Brian Cleary, Eric, and me, Jesse developed a heuristic model to predict *MYC*-regulating enhancers across cell types and applied it analyze GWAS variants near *MYC*. This model evolved into the ABC model (see Chapter 3).

Abstract

Gene expression in mammals is regulated by noncoding elements that can impact physiology and disease, yet the functions and target genes of most noncoding elements remain unknown. We present a high-throughput approach that uses CRISPR interference (CRISPRi) to discover regulatory elements and identify their target genes. We assess >1 megabase (Mb) of sequence in the vicinity of 2 essential transcription factors, *MYC* and *GATA1*, and identify 9 distal enhancers that control gene expression and cellular proliferation. Quantitative features of chromatin state and chromosome conformation distinguish the 7 enhancers that regulate *MYC* from other elements that do not, suggesting a strategy for predicting enhancer-promoter connectivity. This CRISPRi-based approach can be applied to dissect transcriptional networks and interpret the contributions of noncoding genetic variation to human disease.

Introduction

A fundamental goal in modern biology is to identify and characterize the noncoding regulatory elements that control gene expression in development and disease, yet we have lacked systematic approaches to do so. Studies of individual regulatory elements have revealed principles of their function, such as the ability of enhancers to recruit activating transcription factors, modify chromatin state, and physically interact with target genes (*2, 3*). From these insights, systematic mapping of chromatin state and chromosome conformation across cell types has been used to identify putative regulatory elements (*4-7*). However, these measurements do not determine which (if any) genes are regulated or assess the quantitative effects on gene expression. Indeed, the rules that connect regulatory elements with their target genes in the genome appear to be complex.

Regulatory elements do not necessarily affect the closest gene, but instead may act across long distances (*8, 9*). It remains unclear how many regulatory elements control any given gene, or how many genes are regulated by any given element (*3, 4, 9*).

We developed a high-throughput approach that utilizes the programmable properties of CRISPR/Cas9 to characterize the regulatory functions of noncoding elements in their native contexts. We use pooled CRISPR screens in combination with CRISPR interference (CRISPRi) — which alters chromatin state at targeted loci through recruitment of a KRAB effector domain fused to catalytically dead Cas9 (dCas9) (*10-13*) — to simultaneously characterize the regulatory effects of up to 1 Mb of sequence on a gene of interest (Figure 2-1A) (See Appendix A).

Results

We studied two gene loci, *GATA1* and *MYC*, that affect proliferation of K562 erythroleukemia cells in a dose-dependent manner (Figure A-1). This allowed us to search for regulatory elements that quantitatively tune *GATA1* or *MYC* expression using a proliferation-based pooled assay (Figure 2-1A). Importantly, *GATA1* and *MYC* are not located near other strongly essential genes (Figure A-1); thus, proliferation defects caused by sgRNAs targeted to sequences near these genes can be attributed to elements regulating *GATA1* or *MYC*. We designed a library containing 98,000 sgRNAs tiling across a total of 1.29 Mb of genomic sequence around *GATA1* and *MYC* as well as 85 kb of control noncoding regions (See Appendix A). We infected K562 cells expressing KRAB-dCas9 under a doxycycline-inducible promoter with a lentiviral sgRNA library and sequenced the representation of sgRNAs before and after growing cells in doxycycline for 14 population doublings (Figure 2-1A). As expected, internal control sgRNAs targeting the promoters of known essential genes (*11*) were depleted (Figure A-2A) and correlated across biological replicates (*R* = 0.91, Figure A-2B).

We examined the quantitative depletion of sgRNAs in a 74 kb region surrounding *GATA1*, which encodes a key erythroid transcription factor (Figure 2-1B). Because the efficiency of different sgRNAs for CRISPRi can vary dramatically (*11*), we used a sliding window approach, averaging the scores of 20 consecutive sgRNAs and assessing the false discovery rate (FDR) of this metric through comparison to negative control, non-essential regions (See Appendix A, Figure A-3). Because the average spacing between consecutive sgRNAs was 16 bp, the regions targeted by 20 consecutive sgRNA spanned an average of 314 bp (Figure A-3C,D). With this approach, the window with the highest score (strongest depletion) overlapped the *GATA1* TSS itself (Figure 2-1B, Figure A-3F). In addition, we identified 3 distal elements that significantly affected cellular proliferation (FDR < 0.05, Figure 2-1B, See Appendix A). One such element (e-GATA1) is located ~3.6 kb upstream of *GATA1* and corresponds to a DNase I hypersensitive site (DHS) marked by H3K27ac (Figure 2-1C); notably, this element shows high sequence conservation among vertebrates, and the syntenic sequence in mouse is required for proper *Gata1* expression in murine erythroid progenitor cells (*14*). The second distal element (e-HDAC6) corresponds to a conserved DHS located ~1.5 kb upstream of *HDAC6* (Figure 2-1C). A third significant element is located at a DHS near the promoter of *GLOD5*, which itself is not essential and only weakly expressed in K562 cells. The first two elements overlap GATA1 ChIP-Seq peaks and sequence motifs (Figure 2-1C), consistent with known auto-regulatory loops in which GATA1 activates its own expression (*15*). All three elements reside in close linear and spatial proximity to GATA1 (Figure A-4A). Finally, multiple regions in the gene body of *GATA1* scored as significantly depleted in the screen (Figure 2-1B), but, because recruitment of KRAB-dCas9 to these sites may directly interfere with transcription (*10*), we focused on distal regulatory elements in subsequent analysis.

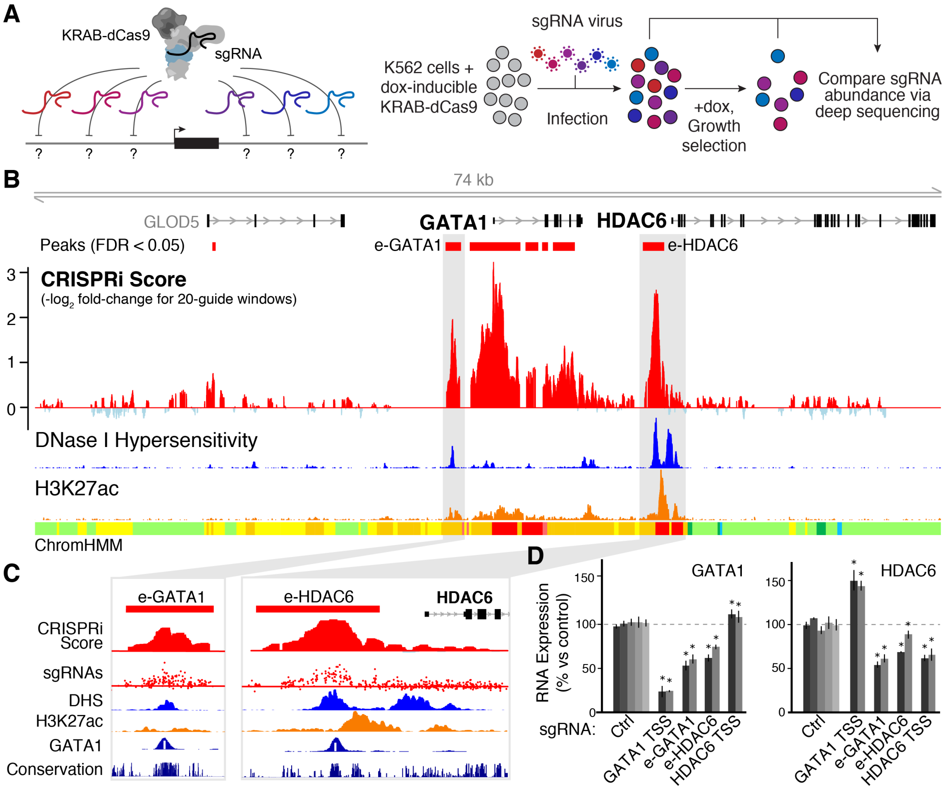


Figure 2-1. Systematic mapping of noncoding elements that regulate *GATA1*. (A) CRISPRi method for identifying gene regulatory elements. Cells expressing KRAB-dCas9 from a dox-inducible promoter are infected with a pool of single guide RNAs (sgRNAs) targeting every possible site across a region of interest. In a proliferation-based screen, cells expressing sgRNAs that target essential regulatory elements will be depleted in the final population. **(B)** CRISPRi screen results in the *GATA1* locus. A high CRISPRi score indicates strong depletion over the course of the screen. Red boxes: Windows showing significant depletion compared to negative control sgRNAs (See Appendix A). DNase I hypersensitivity, H3K27ac ChIP-Seq, and histone modification annotations (ChromHMM) in K562 cells are from ENCODE (*4*). **(C)** Close-up of e-GATA1 and e-HDAC6. sgRNA track shows CRISPRi scores for each individual sgRNA in the region. White bar in GATA1 ChIP-seq track represents the GATA1 motif. **(D)** qPCR for GATA1 and HDAC6 mRNA in cells expressing individual sgRNAs. KRAB-dCas9 expression was activated for 24 hours before measurement. Gray bars: different sgRNAs for each target. Ctrl: negative control sgRNAs without a genomic target. Error bars: 95% confidence intervals (CI) for the mean of 3 biological replicates (See Appendix A). *: p < 0.05 in T-test versus Ctrl.

To characterize these elements, we measured *GATA1* expression using quantitative PCR in cell lines stably expressing individual sgRNAs (See Appendix A). As expected, targeting KRAB-dCas9 to the *GATA1* TSS reduced *GATA1* expression (76% reduction, Figure 2-1D). sgRNAs targeting e-GATA1 or e-HDAC6 reduced *GATA1* expression by 44% and 33%, respectively (Figure 2-1D), and affected the expression of genes known to be regulated by the GATA1 transcription factor (Figure A-4B), confirming that these enhancers regulate *GATA1*. In contrast, sgRNAs targeting the *HDAC6* TSS did not reduce *GATA1* expression despite reducing *HDAC6* expression (Figure 2-1D), indicating that (i) the pooled screen accurately predicted that this region does not reduce *GATA1* expression and (ii) the effects seen for the e-GATA1 and e-HDAC6 sgRNAs are not due to general effects of targeting KRAB-dCas9 to the gene neighborhood. Additionally, both e-GATA1 and e-HDAC6 can activate the expression of a plasmid-based reporter gene (Figure A-4C). Together, these results support the specificity of this CRISPRi-based approach and demonstrate that e-GATA1 and e-HDAC6 quantitatively control *GATA1* expression in K562 cells.

Considering the close proximity of *GATA1* to *HDAC6* (Figure 2-1B, S4A), we tested whether this pair of enhancers also regulates *HDAC6*. sgRNAs targeting e-GATA1 and e-HDAC6 reduced *HDAC6* expression by 42% and 22%, respectively, comparable to their effects on *GATA1* (Figure 2-1D). Intriguingly, inhibition of the *GATA1* promoter led to an increase in *HDAC6* expression (+47%, Figure 2-1D), and inhibition of the *HDAC6* promoter modestly activated *GATA1* (+9%, Figure 2-1D); this suggests that *GATA1* and *HDAC6* may compete for these shared enhancers, similar to observations for other pairs of neighboring genes (*16, 17*). Interestingly, histone deacetylases are required for erythropoiesis (*18*) and HDAC6 has been implicated in cellular proliferation in multiple cancers (*19*). Thus, although HDAC6 does not score as essential in proliferation assays in K562 cells, it is possible that proliferative defects observed upon inhibition of e-GATA1 or e-HDAC6 result from the combined effects on both *GATA1* and *HDAC6* expression, and the genomic proximity of these genes may be important for coordinating their expression *in vivo*. These observations indicate a complex connectivity between enhancers and promoters in their native genomic contexts (Figure A-4D).

We next investigated the *cis* regulatory architecture of *MYC*, a critical transcription factor encoded within a 3-Mb topological domain that contains hundreds of putative enhancers. Several enhancers in this domain regulate *MYC* in other cell types (See Appendix A), but chromatin state varies dramatically across cell types and it is unclear which of these elements regulate *MYC* in a given cell type. Notably, the domain contains over 60 genetic haplotypes associated (through genome-wide association studies) with human phenotypes, including cancer susceptibility (*20*).

To identify elements that regulate *MYC* in K562 cells, we tiled sgRNAs across ~1.2 Mb of sequence in this topological domain (Figure 2-2A). A sliding window analysis identified several regions whose inhibition reproducibly reduced cellular proliferation, including a known promoter-proximal element located 2 kb upstream of the *MYC* TSS (Figure A-5A)(*21*), the transcribed region of the *MYC* gene body (Figure A-5A), and seven distal regions (labeled e1 through e7) located between 0.16 and 1.9 Mb downstream of *MYC* (Figure 2-2A, A-5B,C). We also identified two regions that significantly *increased* cell proliferation (r1 and r2), and thus may repress *MYC* expression (Figure 2-2A, Figure A-5D,E)(See Appendix A).

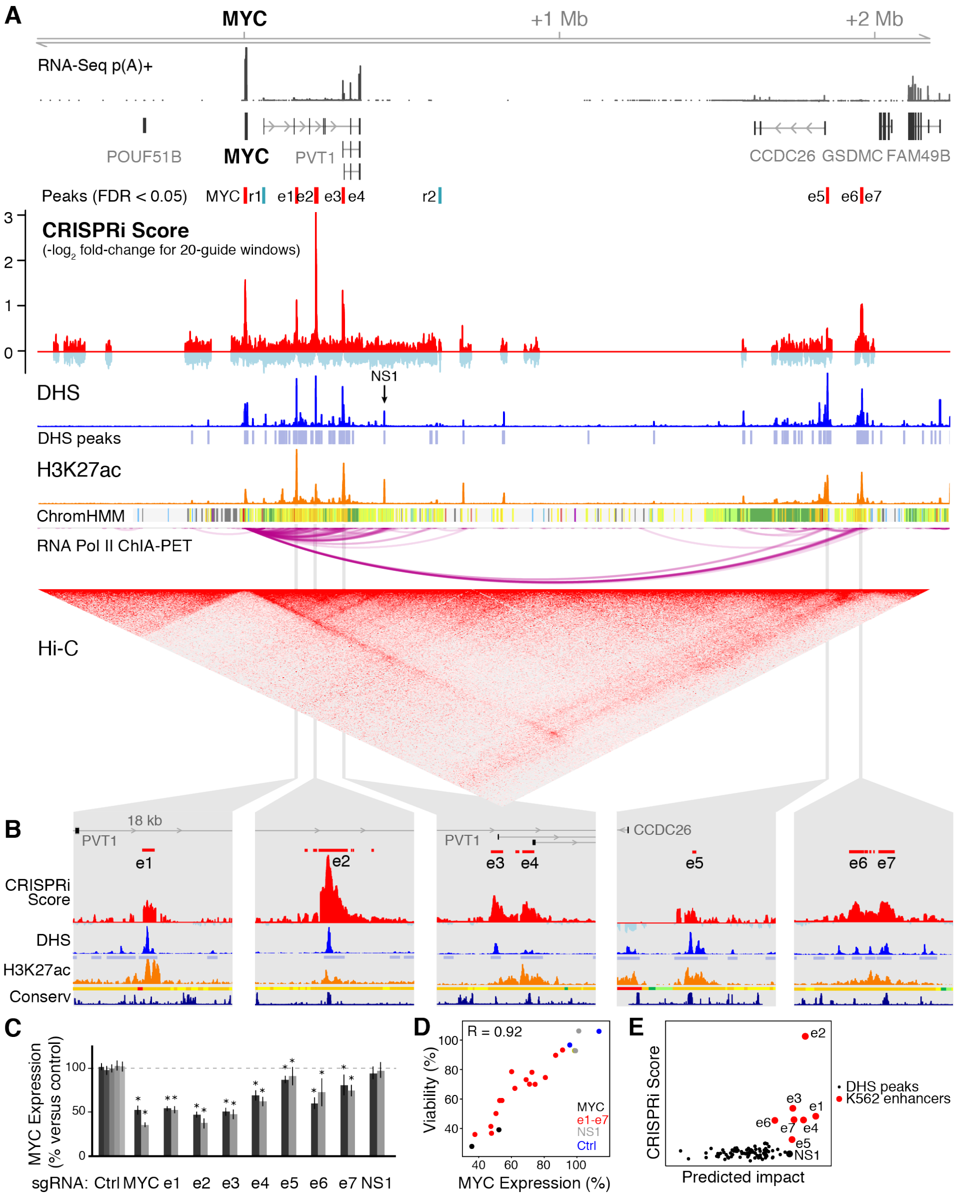


Figure 2-2. Identification and prediction of elements that regulate *MYC*. (A) CRISPRi screening identifies 7 distal enhancers (e1-e7) that activate *MYC* and two repressive elements (r1, r2) that may act to repress *MYC*. NS1: an element that does not score in the screen. **(B)** 18-kb windows around each of the 7 distal enhancers. *Y*-axis scales are equivalent between panels. **(C)** qPCR for MYC mRNA in cells expressing individual sgRNAs 24 hours after KRAB-dCas9 activation. Gray bars: 2 different sgRNAs per target, or 5 for non-targeting controls (Ctrl). Error bars: 95% CI for the mean of 12 biological replicates (See Appendix A). *: $p < 0.05$ in T-test versus negative controls. **(D)** Correlation between MYC expression and relative cell viability for e1-e7, MYC TSS, NS1, and Ctrl sgRNAs. Pearson's $R = 0.92$ includes e1-e7 sgRNAs only; with the others, $R = 0.95$. **(E)** Predicted impact of DHS elements on MYC expression (a function of quantitative DHS, H3K27ac, and Hi-C signal) versus their experimentally derived CRISPRi scores.

Each of the seven putative activating elements is marked by high levels of DNase I hypersensitivity (Figure 2-2A); is bound by multiple transcription factors (Figure A-6A); and shows patches of sequence conservation across mammals (Figure 2-2B). Each enhancer frequently contacts the *MYC* promoter in three dimensions as assayed by Hi-C and ChIA-PET in K562 cells (Figure 2-2A) (*4, 7*); elements e5 and e6/7 form very long-range (>1.8 Mb) loops to the *MYC* promoter and are located within 10 kb of CTCF ChIP-Seq peaks with motifs oriented toward *MYC* (Figure A-5B,C), consistent with the convergent rule for CTCF-mediated chromatin loops (*7*). Two elements (e3 and e4) correspond to alternative TSSs for the long noncoding RNA PVT1 (Figure 2-2A); knockdown experiments indicate that the mature PVT1 RNA transcript itself is likely not essential in K562 cells (Figure A-1) and so e3 and e4 likely affect cellular proliferation through direct regulation of *MYC*.

We experimentally characterized these seven activating elements to test whether they regulate *MYC*. CRISPRi inhibition of each of these elements with individual sgRNAs led to proliferation defects in a competitive growth assay (Figure A-6B) and led to a 9-62% reduction in *MYC* expression (Figure 2-2C). The magnitude of the change in gene expression correlated with the proliferation defect, consistent with a quantitative relationship between cell growth and precise *MYC* expression levels (Pearson R = 0.92, Figure 2-2D). In a plasmid-based reporter assay, each putative regulatory element led to >5-fold up-regulation of a reporter gene relative to a control sequence (Figure A-6C) (See Appendix A). For a subset of the elements (e2, e3, and e4), we generated clonal cell lines containing genetic deletions on one or two of the three chromosome 8 alleles (K562 cells are triploid) and measured the expression of *MYC* from each allele (See Appendix A). For each element, we found that genetic deletions reduced *MYC* expression from the corresponding allele(s), confirming our CRISPRi results (Figure A-7). Together, these data support the hypothesis that these seven elements, spanning 1.6 Mb of noncoding sequence, act as enhancers to control *MYC* expression and cellular proliferation.

In addition to e1-e7, we characterized one noncoding element (NS1) that did *not* score in the screen (Figure 2-2A). In K562 cells, NS1 displays strong DHS and H3K27ac occupancy, binds to multiple transcription factors (Figure A-6A), and participates in a long-range chromatin loop to the *MYC* promoter (Figure 2-2A). In a lung adenocarcinoma cell line, NS1 regulates *MYC* as assayed by CRISPRi inhibition with individual sgRNAs (*22*). Accordingly, we wondered whether NS1 regulates *MYC* in K562 cells despite not being detected as such in our CRISPRi screen. To explore this possibility, we targeted KRAB-dCas9 to NS1 with individual sgRNAs in K562 cells and found that CRISPRi successfully reduced H3K27ac occupancy to an extent similar to that observed when targeting other *MYC* enhancers (Figure A-6D). Despite affecting chromatin state at NS1 in K562 cells, these sgRNAs did not substantially impact cellular proliferation or *MYC* expression (Figure 2-2C,D), consistent with the results from the pooled screen. These observations support the ability of the CRISPRi screening approach to distinguish elements that do and do not regulate a given gene. However, we note that some regulatory elements, such as those that act redundantly with others in the locus, may not be discoverable by this method (See Note A1).

The ability to systematically test gene regulatory elements will help to train predictive models of functional enhancer-promoter connectivity. Notably, existing annotations and catalogs of enhancer-promoter predictions performed poorly at distinguishing e1-e7 from enhancers that do not impact *MYC* expression (See Appendix A). For example, ENCODE annotates 185 Kb of sequence in this domain as putative "strong enhancer" in K562 cells (Figure 2-2A), but only 8% of this sequence, corresponding to e1-e7, appears to regulate *MYC*. We sought to improve the ability to predict enhancers and connect them with genes that they regulate. When we examined chromatin state maps (including DHS, H3K27ac and Hi-C), we found that quantitative DHS or H3K27ac signal could distinguish most of the seven *MYC* enhancers but ranked them in the wrong order (Figure A-8A): for example, e5 shows the strongest DHS signal yet has the weakest effect on MYC expression (Figure 2-2). Accordingly, we considered a framework (Figure A-8B) wherein the impact of an enhancer on gene expression is determined both by its intrinsic activity level (for which we use quantitative DHS and H3K27ac levels as a proxy) and the frequency at which the enhancer contacts its target promoter (for which we use Hi-C data as a proxy) (See Appendix A). This metric correctly ranked 6 of the 7 distal enhancers as the most important of 93 DHS elements in K562 cells (Figure 2-2E) and provided a reasonable ordering of their relative effects (Spearman correlation = 0.79). We note that this approach did not perfectly distinguish between enhancers that do and do not regulate MYC: NS1 was ranked 7 and e6 was ranked 11. Nonetheless, quantitative measures of chromatin state and chromosome conformation are strongly predictive of enhancers that regulate *MYC* in K562 cells.

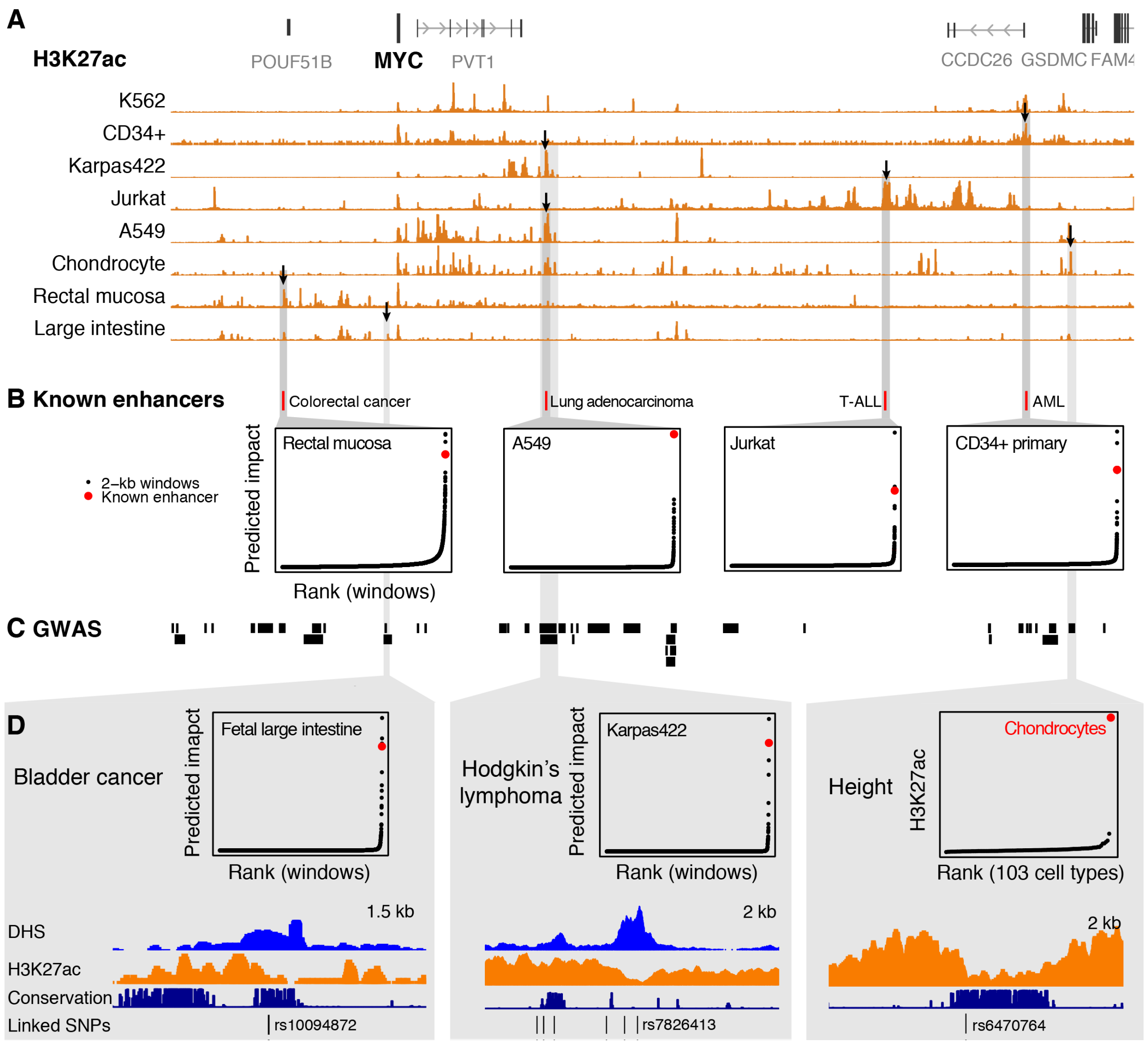


Figure 2-3. A heuristic model predicts disease-associated *MYC* enhancers across cell types. **(A)** H3K27ac occupancy around *MYC* varies among 8 cell types and primary tissues. Black arrows: -elements highlighted in panels below. **(B)** Locations of 4 enhancers previously shown to regulate MYC expression in other cell types and their predicted impact in a corresponding cell type. Points show predicted impact of 2-kb windows tiled in 100-bp increments across the *MYC* locus. T-ALL: T-cell acute lymphoblastic leukemia. AML: Acute myeloid leukemia. For each cell type, predicted impact is calculated based on available data. **(C)** Haplotype blocks of SNPs linked to human diseases and phenotypes ($R^2 > 0.8$ with index SNP in genome-wide association study). **(D)** SNPs associated with bladder cancer and Hodgkin's lymphoma overlap regulatory elements predicted by our metric to regulate *MYC* in a corresponding cell type or tissue. A SNP associated with height overlaps a conserved element that is active only in chondrocytes. Karpas422: diffuse large B cell lymphoma cell line.

To determine whether this approach might be applicable in other cellular contexts, we examined 4 *MYC* enhancers identified in other cell types (Figure 2-3A,B)(See Appendix A). In each case our metric ranked these known elements among the 3 most important in the corresponding cell type (Figure 2-3B). We also identified multiple instances where elements predicted to regulate *MYC* in one or more cell types harbor single nucleotide polymorphisms (SNPs) associated with human traits including cancer susceptibility and height (Figure 2-3C,D). Additional CRISPRi-based functional mapping in other cell types and gene loci might allow the derivation of general models to predict functional enhancer-promoter connections and help to understand noncoding genetic variation.

Discussion

In summary, CRISPRi screens can accurately identify and characterize the regulatory functions and connectivity of noncoding elements. In the *MYC* and *GATA1* loci, CRISPRi reveals complex and non-obvious dependencies between multiple genes and enhancers, including relationships that suggest regulation of multiple genes by the same enhancer, coordinated activity of multiple enhancers to control a single gene, and competition between neighboring promoters. Thus, learning the principles and connectivity of transcriptional networks requires dissecting putative regulatory elements in their native genomic contexts.

While we used cellular proliferation as a readout to investigate 2 essential genes, this CRISPRi approach can be applied to identify regulatory elements that control an arbitrary gene or phenotype of interest through alternative assays, for example by tagging an endogenous gene locus with green fluorescent protein (GFP) and sorting cells by GFP expression (*23*).

Together with complementary methods using catalytically active Cas9 (*23-25*), CRISPRi-based functional mapping provides a broadly applicable approach (See Appendix A) to dissect transcriptional networks and interpret the contributions of noncoding genetic variation in gene regulatory elements to human disease.

References

1. C. P. Fulco *et al.*, Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773 (2016).

2. M. Bulger, M. Groudine, Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339 (2011).

3. F. Spitz, E. E. Furlong, Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613-626 (2012).

4. G. Li *et al.*, Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98 (2012).

5. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

6. C. Roadmap Epigenomics *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).

7. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

8. D. Shlyueva, G. Stampfel, A. Stark, Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-286 (2014).

9. J. van Arensbergen, B. van Steensel, H. J. Bussemaker, In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol* **24**, 695-702 (2014).

10. L. A. Gilbert *et al.*, CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451 (2013).

11. L. A. Gilbert *et al.*, Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647-661 (2014).

12. N. A. Kearns *et al.*, Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. *Development* **141**, 219-223 (2014).

13. P. I. Thakore *et al.*, Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature methods* **12**, 1143-1149 (2015).

14. M. Suzuki, T. Moriguchi, K. Ohneda, M. Yamamoto, Differential contribution of the Gata1 gene hematopoietic enhancer to erythroid differentiation. *Mol Cell Biol* **29**, 1163-1175 (2009).

15. S. Nishimura *et al.*, A GATA box in the GATA-1 gene hematopoietic enhancer is a critical element in the network of GATA factors and sites that regulate this gene. *Mol Cell Biol* **20**, 713-723 (2000).

16. O. R. Choi, J. D. Engel, Developmental regulation of beta-globin gene switching. *Cell* **55**, 17-26 (1988).

17. S. Ohtsuki, M. Levine, H. N. Cai, Different core promoters possess distinct regulatory activities in the Drosophila embryo. *Genes & development* **12**, 547-556 (1998).

18. A. Fujieda *et al.*, A putative role for histone deacetylase in the differentiation of human erythroid cells. *Int J Oncol* **27**, 743-748 (2005).

19. K. J. Falkenberg, R. W. Johnstone, Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. *Nat Rev Drug Discov* **13**, 673-691 (2014).

20. T. Burdett, The NHGRI-EBI Catalog of published genome-wide association studies. (available at http://www.ebi.ac.uk/gwas).

21. W. M. Gombert, A. Krumm, Targeted deletion of multiple CTCF-binding elements in the human C-MYC gene reveals a requirement for CTCF in C-MYC expression. *PLoS One* **4**, e6109 (2009).

22. X. Zhang *et al.*, Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet* **48**, 176-182 (2016).

23. N. Rajagopal *et al.*, High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**, 167-174 (2016).

24. M. C. Canver *et al.*, BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192-197 (2015).

25. G. Korkmaz *et al.*, Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192-198 (2016).

Acknowledgements: We thank T. Wang and R. Issner for technical advice and reagents; and R. Ryan, B. Bernstein, N. Sanjana, J. Wright, and F. Zhang for discussions. This work was supported by funds from the Broad Institute (E.S.L.). C.P.F. is supported by the National Defense Science and Engineering Graduate Fellowship. J.M.E. is supported by the Fannie and John Hertz Foundation.

M.M. is supported by a DFG Research Fellowship. S.R.G. is supported by NIGMS T32GM007753. The Broad Institute, which E.S.L. directs, holds patents and has filed patent applications on technologies related to other aspects of CRISPR. Data presented in this chapter can be found in Appendix A and in GSE87257.

Chapter 3: Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations

Most of this chapter was originally published as: Fulco CP, Nasser J, Jones TR, Munson G, Bergman D, Subramanian V, Grossman SR, Anyoha R, Patwardhan TA, Nguyen TH, Kane M, Doughty B, Perez E, Durand NC, Stamenova EK, Lieberman Aiden E, Lander ES, Engreitz JM. Activity-by-Contact model for enhancer specificity from thousands of CRISPR perturbations. bioRxiv. 2019 Jan 26.

Preface

In this chapter, I describe our work to (i) expand the CRISPR functional genomics toolkit by combining pooled CRISPR screens with RNA FISH and flow cytometry, (ii) apply this new tool to systematically map the enhancers controlling the expression of 28 genes in K562 cells, and (iii) develop the ABC model as the first means to accurately predict enhancer-gene connections based on epigenetic data.

I developed the FlowFISH screening protocol initially with Vidya Subramanian. Vidya optimized the FISH protocol, we worked together to determine the flow cytometry sorting protocol, and I optimized the gRNA library preparation protocol. Vidya and I together performed our initial large-scale screen for *GATA1* enhancers using a pool cloned by Rockwell Anyoha. For subsequent screens, Glen Munson carried out the wet lab CRISPR pool cloning, cell culture, FlowFISH, and sequencing library preparation, and I sequenced the libraries with occasional assistance from Tung Nguyen.

Drew Bergman and I carried out validation experiments such as comparing the quantitative effect of single guide RNAs between qPCR and FlowFISH, and, with Tung Nguyen, using siRNAs and RNA-seq to evaulate potential *trans* effects. Tung Nguyen and Liz Perez also contributed to other validation experiments and data analysis.

I wrote the initial FlowFISH screen analysis pipeline, and provided minor assistance to Ben Doughty, Tejal Patwardhan, and Ray Jones in adapting it to be more scalable and accurate. Jesse curated additional enhancer-gene functional connection data from the literature with some help from me.

Joe Nasser, Ray Jones, and Jesse Engreitz wrote the majority of the code to operationalize the ABC model and to compare our compendium of functionally tested enhancer-gene connections to predictive models such as the ABC model, with some input from me. Shari Grossman was a source of critical discussions in the development, design, and evaluation of the ABC model.

Elena Stamenova, Neva Durand, and Erez Lieberman Aiden contributed Hi-C maps of chromatin contacts in mouse embryonic stem cells. Erez also helped Joe, Jesse, and me explore how the pattern of chromatin contacts predicted from simple globule models could explain our functional data.

Abstract

Mammalian genomes harbor millions of noncoding elements called enhancers that quantitatively regulate gene expression, but it remains unclear which enhancers regulate which genes. Here we describe an experimental approach, based on CRISPR interference, RNA FISH, and flow cytometry (CRISPRi-FlowFISH), to perturb enhancers in the genome, and apply it to test >3,000 potential regulatory enhancer-gene connections across multiple genomic loci. A simple equation based on a mechanistic model for enhancer function performed remarkably well at predicting the complex patterns of regulatory connections we observe in our CRISPR dataset. This Activity-by-Contact (ABC) model involves multiplying measures of enhancer activity and enhancer-promoter 3D contacts, and can predict enhancer-gene connections in a given cell type based on chromatin state maps. Together, CRISPRi-FlowFISH and the ABC model provide a systematic approach to map and predict which enhancers regulate which genes, and may help to interpret the functions of the thousands of disease risk variants in the noncoding genome.

Introduction

DNA elements in the human genome called enhancers control how different combinations of genes are expressed in different cell types and states, and harbor thousands of genetic variants that influence risk for common diseases. A major challenge in interpreting the functions of these variants is to map enhancer-gene connections: Which enhancers regulate which genes in which cell types, and with what quantitative effects?

Studies of individual enhancers and genes have shown that these connections can be complex: multiple enhancers can regulate a single gene, a single enhancer can regulate multiple genes across long genomic distances, and the network of enhancer-gene connections appears to rewire across cell types (*1, 2*). The mechanisms that give rise to this complexity remain poorly understood. One possibility ("biochemical specificity") is that a given enhancer can regulate only the promoters that have complementary combinations of compatible transcription factors (TFs) (*3-6*). In a few cases it has been shown that specific TF-TF interactions are required for an enhancer to regulate a promoter (*7, 8*). Another possibility is that enhancer-gene interactions depend primarily on the 3D architecture of the genome, such as topological domains (*9, 10*) or chromatin loops (*1, 11, 12*). In a few cases it has been shown that manipulating enhancer-promoter contacts can affect gene expression (*13-15*). Various studies have integrated aspects of transcription factor binding and 3D architecture to attempt to predict enhancer-gene regulation (*16-19*). Yet, it has been difficult to evaluate these models or discover new ones because we have lacked efficient ways to study the regulatory effects of large numbers of enhancers in the genome.

We set out to map the effects of many putative enhancers on gene expression and thereby learn general rules to predict enhancer-gene connections across many cell types. We and others have recently developed high-throughput methods that use CRISPR to perturb noncoding elements in their native genomic locations to measure their effect on a target gene (*17, 20-24*). 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.

Results

To overcome these limitations, 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 (Figure 3-1A; Figure B-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; (iv) and use this information to infer the effect of each gRNA on RNA expression. To assess quantitative effects and statistical significance, we calculate average the effects of all gRNAs within each candidate element (Figure B-2A,B) and compare to hundreds of negative control gRNAs in the same screen.

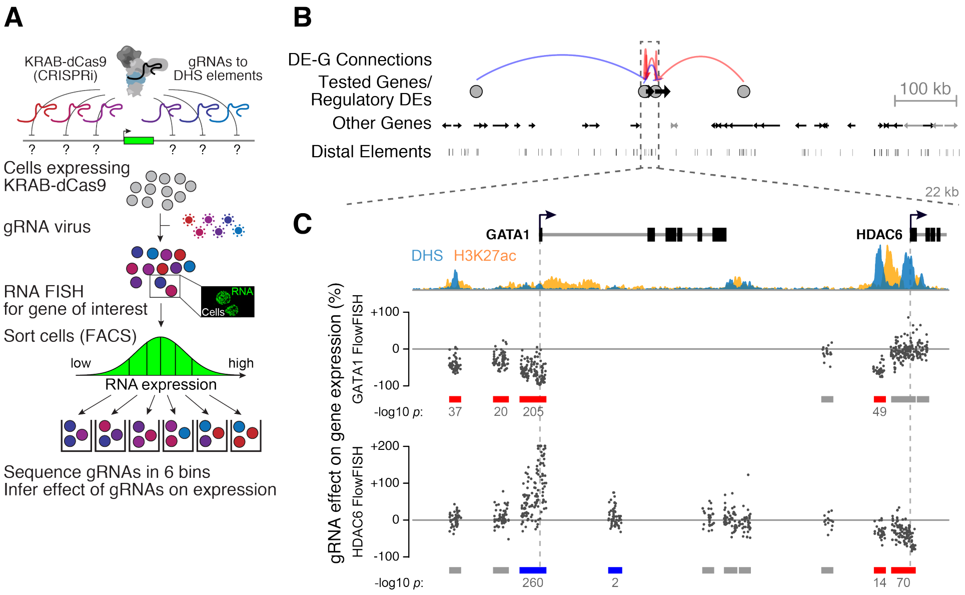


Figure 3-1. CRISPRi-FlowFISH identifies regulatory elements for *GATA1* and *HDAC6*
(A) CRISPRi-FlowFISH method for identifying gene regulatory elements. Cells expressing KRAB-dCas9 are infected with a pool of gRNAs targeting DHS elements near a gene of interest, labeled using RNA FISH against that gene, and sorted into bins of fluorescence signal by FACS. The quantitative effect of each gRNA on the expression of the gene is determined by sequencing the gRNAs within each bin. Inset: example of K562 cells labeled for *RPL13A*. **(B)** Distal elements affecting *GATA1* and *HDAC6* expression in K562 cells. Genes expressed in K562 cells are shown in black; those not expressed are shown in grey. Red arcs denote activation, blue arcs denote repression. Grey circles are DEs where perturbation with CRISPRi affects the expression of at least one tested gene as measured by CRISPRi-FlowFISH. See Figure B-3A for the full tested region spanning 4 Mb. **(C)** Close-up on region containing *GATA1* and *HDAC6*. Points represent the effect on gene expression of a single gRNA. *HDAC6* vertical axis capped at 200%. Grey, red, and blue bars: DHS elements in which CRISPRi leads to no detectable change (grey), a significant decrease (red) or increase (blue) in expression as measured by CRISPRi-FlowFISH. DHS elements in the gene body of the assayed gene are excluded from analyses because recruitment of KRAB-dCas9 to these sites may directly interfere with transcription.

To validate the approach, we first used CRISPRi-FlowFISH to identify elements that regulate the expression of *GATA1* in K562 human erythroleukemia cells. We performed replicate CRISPRi-FlowFISH screens using a probeset against *GATA1* and tested the functions of 127 candidate elements spanning 4 Mb (Figure 3-1B,C; Figure B-3A). Replicate screens produced highly correlated estimates for the effect sizes of each element on *GATA1* expression (Pearson *R* = 0.95 for significant elements, Figure B-2C). As expected, these screens identified the three elements that we previously found to regulate *GATA1* (*17*), and we confirmed for individual gRNAs that the effects on gene expression estimated from CRISPRi-FlowFISH agreed with RT-qPCR measurements (Pearson *R* = 0.93, Figure B-1F). We note that these experiments do not distinguish between *cis* and *trans* effects (see Appendix B).

To generate a large enhancer perturbation dataset, we used CRISPRi-FlowFISH in K562 cells to test a total of 3744 candidate regulatory element-gene pairs. Specifically, we designed FlowFISH assays for 28 genes in 5 genomic regions (spanning 1.1-4.0 Mb) and CRISPRi gRNAs against all DNase hypersensitive (DHS) elements in K562s within 450 kb of any of the genes (108 to 202 elements per gene for a total of 742 unique elements). The 28 genes included some with erythroid lineage-specific expression (*e.g.*, *GATA1*) and some that are ubiquitously expressed (*e.g.*, *RAB7A*), and were selected (after testing FlowFISH probesets for 51 genes) as those genes with probesets that met stringent criteria for both specificity and statistical power (Figure B-4, see Appendix B). We had >80% power to detect an effect on gene expression of 25% for all 28 genes and as low as 10% effects for 3 genes (Figure B-4C, see Appendix B). We analyzed these CRISPRi-FlowFISH data together with data from an additional 380 candidate regulatory element-gene pairs from previous CRISPR-based experiments in K562 cells, including our previous CRISPRi tiling proliferation screen in the *MYC* locus (*17, 23, 25-31*).

In total, our dataset included 3010 candidate distal element-gene (DE-G) pairs (where the targeted element is located >500 bp from a TSS) and 1114 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 because we and others have found promoters can affect the expression of nearby genes through a variety of mechanisms beyond that of *cis*-acting enhancers (see Appendix B).

These perturbation-based maps uncovered complex connections wherein individual enhancers regulated up to 5 genes, individual genes were regulated by up to 12 distal elements, and in some cases enhancers appeared to "skip" over proximal genes to regulate more distant ones (Figure 3-2; Figure B-3; Figure B-5). Of the 3010 DE-G pairs tested, 122 involved a significant effect on gene expression at a false discovery rate (FDR) < 0.05. The effect was activating in 80% of cases and repressing in 20% of cases (98 vs. 24), with absolute effect sizes ranging from 5%-93% (median: 24%). Of 818 distinct DEs studied, 79 (10%) detectably regulated at least one gene in our dataset.

Using this data, we sought to identify generalizable rules to explain which enhancers regulate which genes in the genome. To do so, we compared various predictors to our experimental results by means of a precision-recall plot (Figure 3-3A). (Precision refers to the proportion of positive predictions that are 'true positives' — where true regulatory connections are the 98 significant DE-G pairs where perturbation of the element led to a decrease in gene expression, and the 2912 non-regulatory connections are those where no decrease was detected despite >80% power to detect 25% effects. Recall refers to the proportion of true connections included in the predictions. For analysis of repressive effects, see Appendix B).

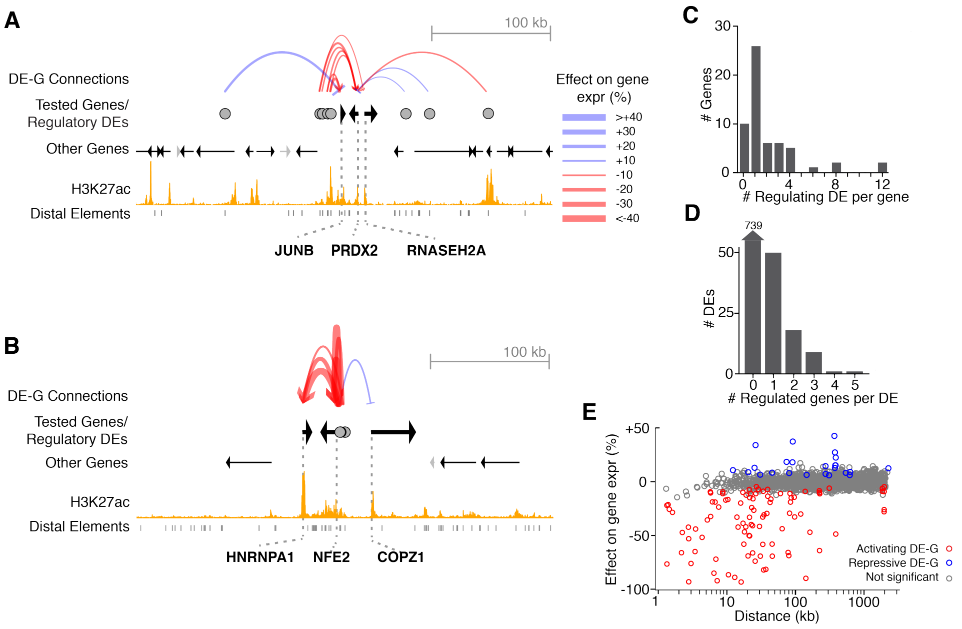


Figure 3-2. CRISPRi-FlowFISH produces regulatory maps of DE-G connections in multiple loci. (A) Example of CRISPRi-FlowFISH screen data. DE-G connections are elements affecting the expression of *JUNB*, *PRDX2*, and *RNASEH2A* in CRISPRi-FlowFISH screens in K562 cells. Red arcs denote activation, blue arcs denote repression. The width of the arc corresponds to the effect size. Distal elements are DHS peaks. Tested genes refer to genes for which we performed CRISPRi-FlowFISH experiments. See Figure B-3B for the full tested region spanning 1.4 Mb. **(B)** Same as (A) for the genes *HNRNPA1, NFE2*, and *COPZ1*. See Figure B-3C for the full tested region spanning 1.2 Mb. **(C)** Histogram of the number of distal elements affecting each gene in our dataset. **(D)** Histogram of the number of genes affected by each distal element tested in our dataset. **(E)** Comparison of genomic distance with observed changes in gene expression upon CRISPR perturbations. Each dot represents one tested DE-G. Red dots: connections where perturbation resulted in a decrease in the expression of the tested gene. Blue dots: perturbation resulted in an increase. Grey dots: had no significant effect. Panels (C-E) include both FlowFISH data from this study and tested pairs from other studies. See Figure B-5 for plots including FlowFISH data only.

We first examined three categories of methods that are commonly used to predict enhancer-gene connections, and found these had only modest predictive value (Figure 3-3A):

(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 14% of DEs within 100 kb of an expressed promoter had a regulatory effect (precision = 14%, recall = 84%). Assigning each DE to the closest expressed gene yielded 45% precision and 33% recall.

(2) Predictions based solely on features of the 3D genome also performed poorly. Assigning each DE to promoters based on the presence of Hi-C loops yielded 25% precision and 4% recall, and assigning each DE to each other promoter in the same Hi-C contact domain yielded 14% precision and 77% recall.

(3) Predictions based on prior machine learning approaches were similarly unsuccessful, including supervised methods to predict enhancer-promoter interactions from epigenomic data and unsupervised methods based on correlations between chromatin marks and gene expression across cell types (Fig 3A, see Appendix B) (*18, 19*).

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

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

We defined Activity (A) as the geometric mean of the read counts of DHS and H3K27ac ChIP-Seq at an element E, and Contact (C) as the normalized Hi-C contact frequency between E and the promoter of gene G (see Appendix B). (The ABC score performed similarly across a range of data preprocessing parameters, and when defining Activity using other combinations of measurements of chromatin accessibility, histone modifications, and nascent transcription, see Appendix B, Figure B-6,B-7,B-8).

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 = –0.68 Figure 3-3C). 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 63% precision, and the area under precision-recall curve (AUPRC) was 0.66, compared to 0.36 for predictions based solely on genomic distance (Figure 3-3A).

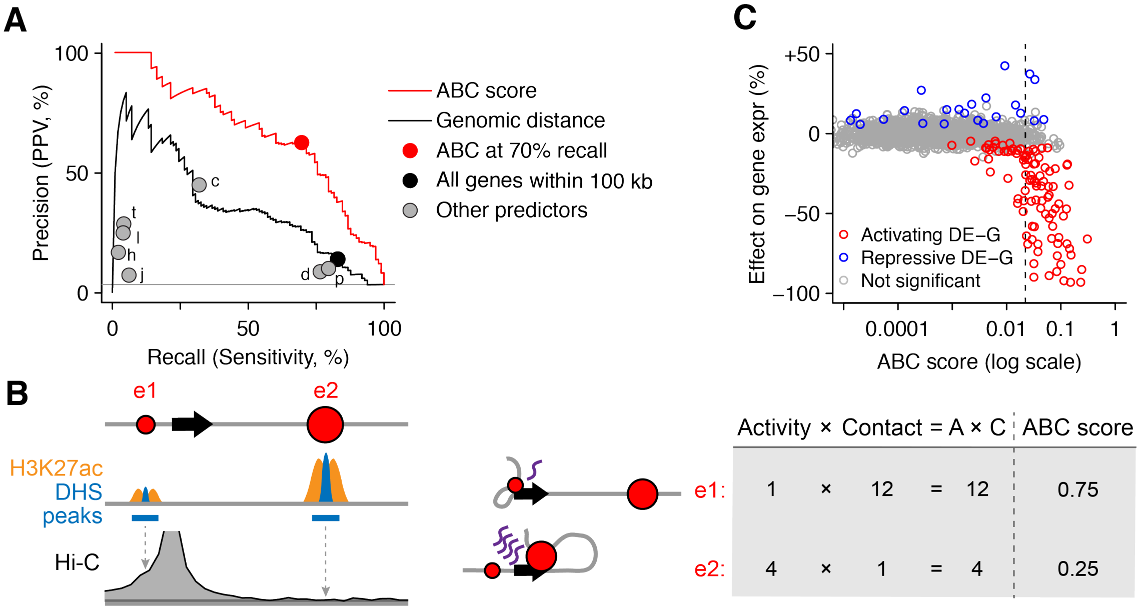


Figure 3-3. The ABC model predicts the target genes of enhancers. (A) Precision-recall plot for classifiers of DE-G pairs. Positive DE-G pairs are those where the distal element significantly decreases expression of the gene. Curves represent the performance for predicting significant decreases in expression for DE-G pairs based on thresholds on the ABC score (red) and genomic distance between the DE and the TSS of the gene (black). Circles represent the performance of various predictors in which DEs are assigned to: the closest expressed gene ("c"); all promoters within 100 kb (black), genes predicted by the algorithms TargetFinder ("t") (*18*) or JEME ("j") (*19*); 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") (*32*), or H3K27ac HiChIP loops ("h") (*33*). **(B)** Calculation of the ABC score (see Appendix B). Values for DHS, H3K27ac, and Hi-C are presented in arbitrary units. **(C)** Comparison of ABC scores (predicted effect) with observed changes in gene expression upon perturbations. Each dot represents one tested DE-G pair. Dotted black line marks 70% recall, corresponding to the red dot in panel A.

The ABC score far outperformed models based on either Activity (quantitative DHS or H3K27ac signal) or Contact (Hi-C contact frequency) alone (AUPRC = 0.25, 0.17 and 0.31, respectively; Figure 3-3A; Figure B-6A). This is because experimentally observed regulatory DE-G pairs varied substantially — with some having higher Activity and lower Contact, some showing higher Contact and lower Activity, and some having a balance of the two factors (Figure B-6B).

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

To do so, we first identified alternative ways to estimate Contact in the ABC model; although maps of chromatin accessibility and histone modifications are available in many cell types, maps of 3D contacts are not. Because contact frequencies in Hi-C data correlate well across cell types (see Appendix B) (*34, 35*), we compared versions of the ABC model in which we estimated Contact for each DE-G pair using either K562 Hi-C data or the average Hi-C contact frequency from 8 other human cell types. Both approaches performed similarly at predicting our CRISPR data in K562 cells (AUPRC = 0.66 and 0.68 respectively; Figure B-9A). Thus, the ABC model can make predictions in a given cell type without cell-type specific Hi-C data, and minimally requires: (i) a measure of chromatin accessibility (DHS or ATAC-seq) and (ii) a measure of enhancer activity (ideally, H3K27ac ChIP-seq).

Using this approach, we evaluated the ability of the ABC model to predict 968 measured DE-G pairs in 5 additional human and mouse cell types beyond our initial K562 dataset. These pairs included 940 from previous studies that inhibited DEs with epigenetic or genetic perturbations and measured the effects with RNA-seq or qPCR (*33, 36-45*), and 28 from new experiments in which we deleted enhancers in mouse embryonic stem cells and measured the effects using allele-specific RNA-seq (see Appendix B). We used epigenomic datasets to generate genome-wide predictions of enhancer-gene connections in each of these 5 cell types, and compared them to the CRISPR data in the corresponding cell type. The ABC scores correlated with the quantitative effects on gene expression (Spearman ρ for regulatory DE-G pairs = -0.38, Fig 4A), and at an ABC threshold corresponding to 70% recall, the predictions had 74% precision (AUPRC = 0.75, Fig 4B, see Appendix B). As expected, the predictions of the ABC model were highly cell-type specific: when we used ABC scores from K562 cells to predict DE-G pairs measured in other cell types, the AUPRC dropped from 0.75 to 0.12.

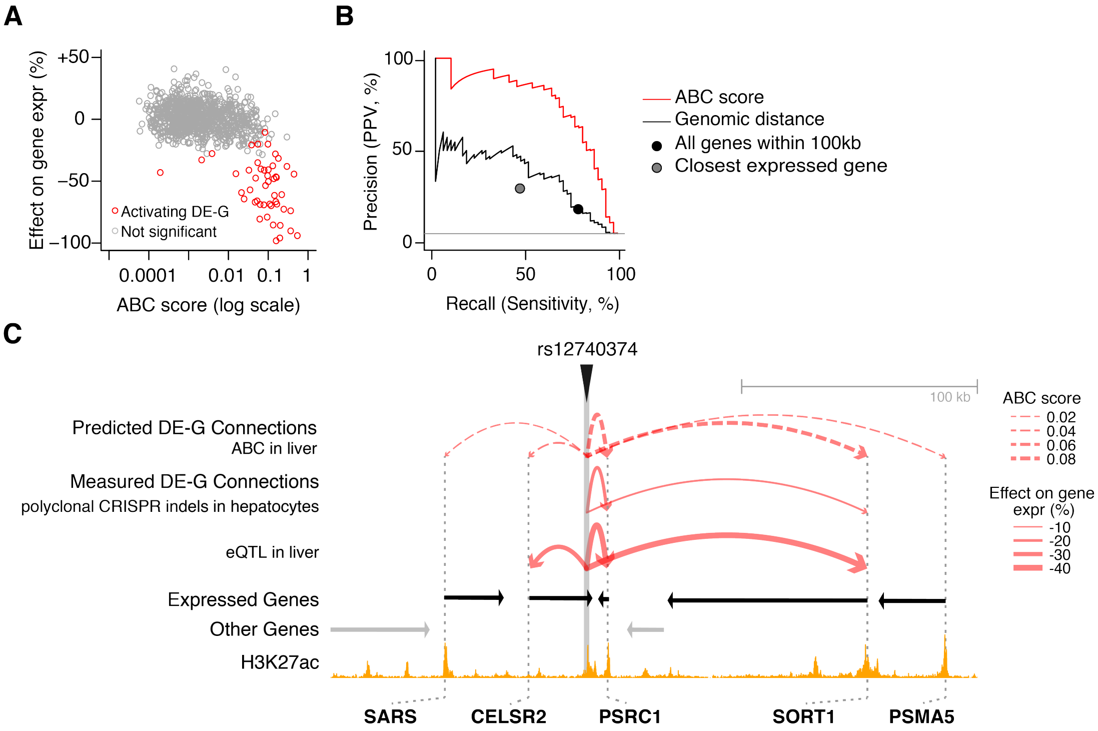


Figure 3-4. The ABC model generalizes across cell types. (A) Comparison of ABC scores (predicted effect) with observed changes in gene expression upon perturbations in GM12878 cells, LNCaP cells, NCCIT cells, primary human hepatocytes, and mouse ES cells. Each dot represents one tested DE-G pair. **(B)** 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). Circles represent the performance of models that predict significant regulation for DE-G pairs based on various criteria: pair lies within 100 kb (black), and DEs are assigned to regulate the nearest expressed gene (grey). **(C)** Comparison of observed and predicted DE-G connections in the *SORT1* locus (chr1:109714926-109989926). Predicted DE-G connections (dotted red arcs) are based on ABC maps in primary human liver tissue. Observed DE-G connections (solid red arcs) are from previous experiments in which CRISPR was used to introduce indels near rs12740374 in primary hepatocytes (*45*) and an eQTL study in human liver (*46*).

We next examined the 16 DE-G pairs in our dataset that involved enhancers that harbor noncoding genetic variants known to influence risk for human diseases or traits. 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). For example, a previous study identified 3 enhancers that contain noncoding variants associated with rare erythroid disorders, and found that introducing indels into these enhancers in K562 using CRISPR affected the expression of nearby genes involved in erythropoiesis (*28*). The ABC model correctly identified each of these 3 regulatory DE-G pairs in K562 cells, and, notably, also identified the same connections in primary human erythroid progenitor cells. As another example, a variant associated with coronary artery disease and plasma low-density lipoprotein cholesterol (rs12740374) has been shown to be an eQTL for *SORT1* in primary human liver tissue, and CRISPR edits in the corresponding element affect *SORT1* expression in primary hepatocytes (*45, 46*). ABC maps in primary human liver tissue correctly connected this enhancer to *SORT1* (Figure 3-4C). Thus, the ABC model can predict enhancer-gene connections based on cell-type specific epigenomic data, and may be widely useful for interpreting the functions of noncoding genetic variants associated with human diseases.

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

We first compared predictions for tissue-specific versus ubiquitously expressed genes (sometimes referred to as "housekeeping" genes, see Appendix B), and found that the ABC model performed dramatically better for tissue-specific than for ubiquitously expressed genes (AUPRC = 0.77 vs 0.12). This was because ubiquitously expressed genes had fewer enhancers: for the 30 genes for which we had data for all nearby DEs, tissue-specific genes (n=22) had an average of 2.6 distal enhancers per gene, while ubiquitously expressed genes had only 0.1 (only a single enhancer across 8 ubiquitously expressed genes; rank-sum test $p = 0.002$, Figure B-10). Interestingly, these observations are consistent with findings in *Drosophila*, where plasmid-based reporter assays have shown that ubiquitously expressed gene promoters are less sensitive to distal enhancers than are tissue-specific gene promoters (*6*). We conclude that the ABC model applies well to tissue-specific genes (97% of all genes, see Appendix B) but not to ubiquitously expressed genes, which appear to be largely insensitive to the effects of distal enhancer perturbations for reasons that remain to be explored.

We next examined our CRISPR dataset for DE-G pairs that likely represent regulatory effects due to mechanisms other than the *cis*-acting functions of enhancers (see Appendix B). We identified effects of distal CTCF sites, which may regulate gene expression by affecting 3D contacts (13 regulatory pairs, Figure B-11) and indirect effects, such as an enhancer regulating one gene that in turn affects a second nearby gene in *trans* (18 pairs, Figure B-12). 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. Indeed, the AUPRC rose from 0.66 to 0.72 for all genes and to 0.82 for tissue-specific genes (Figure B-13). These results suggest a strategy to iteratively 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.

Discussion

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

Our results raise the intriguing possibility that the ABC model reflects an underlying biochemical mechanism: that enhancer specificity may often be controlled by quantitative factors including enhancer activity and enhancer-promoter contact frequency, rather than by qualitative logic involving the particular combinations of transcription factors at the enhancers and promoters. CRISPRi-FlowFISH and the ABC model provide a means to test these mechanisms, and to further refine our understanding of noncoding regulatory elements by mapping and modeling promoter-promoter regulation, functions of CTCF sites, and combinatorial effects of multiple enhancers in a locus.

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

References

1. F. Spitz, E. E. Furlong, Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613-626 (2012).

2. D. Shlyueva, G. Stampfel, A. Stark, Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-286 (2014).

3. X. Li, M. Noll, Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo. *EMBO J* **13**, 400-406 (1994).

4. C. Merli, D. E. Bergstrom, J. A. Cygan, R. K. Blackman, Promoter specificity mediates the independent regulation of neighboring genes. *Genes & development* **10**, 1260-1270 (1996).

5. J. E. Butler, J. T. Kadonaga, Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes & development* **15**, 2515-2519 (2001).

6. M. A. Zabidi *et al.*, Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556-559 (2015).

7. W. Su, S. Jackson, R. Tjian, H. Echols, DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. *Genes & development* **5**, 820-826 (1991).

8. Q. Gong, A. Dean, Enhancer-dependent transcription of the epsilon-globin promoter requires promoter-bound GATA-1 and enhancer-bound AP-1/NF-E2. *Mol Cell Biol* **13**, 911-917 (1993).

9. R. Bernards, R. A. Flavell, Physical mapping of the globin gene deletion in hereditary persistence of foetal haemoglobin (HPFH). *Nucleic Acids Res* **8**, 1521-1534 (1980).

10. T. Sexton, G. Cavalli, The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049-1059 (2015).

11. M. Bulger, M. Groudine, Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339 (2011).

12. I. Krivega, A. Dean, Enhancer and promoter interactions-long distance calls. *Curr Opin Genet Dev* **22**, 79-85 (2012).

13. W. Deng *et al.*, Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244 (2012).

14. C. R. Bartman, S. C. Hsu, C. C. Hsiung, A. Raj, G. A. Blobel, Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Molecular cell* **62**, 237-247 (2016).

15. S. L. Morgan *et al.*, Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nat Commun* **8**, 15993 (2017).

16. B. He, C. Chen, L. Teng, K. Tan, Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci U S A* **111**, E2191-2199 (2014).

17. C. P. Fulco *et al.*, Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773 (2016).

18. S. Whalen, R. M. Truty, K. S. Pollard, Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48**, 488-496 (2016).

19. Q. Cao *et al.*, Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat Genet* **49**, 1428-1436 (2017).

20. M. C. Canver *et al.*, BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192-197 (2015).

21. N. E. Sanjana *et al.*, High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545-1549 (2016).

22. Y. Diao *et al.*, A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature methods* **14**, 629-635 (2017).

23. T. S. Klann *et al.*, CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat Biotechnol* **35**, 561-568 (2017).

24. M. Gasperini *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).

25. P. I. Thakore *et al.*, Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature methods* **12**, 1143-1149 (2015).

26. J. Xu *et al.*, Developmental control of polycomb subunit composition by GATA factors mediates a switch to non-canonical functions. *Molecular cell* **57**, 304-316 (2015).

27. J. C. Ulirsch *et al.*, Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530-1545 (2016).

28. A. Wakabayashi *et al.*, Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *Proc Natl Acad Sci U S A* **113**, 4434-4439 (2016).

29. S. J. Liu *et al.*, CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, (2017).

30. S. Xie, J. Duan, B. Li, P. Zhou, G. C. Hon, Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular cell* **66**, 285-299.e285 (2017).

31. J. Huang *et al.*, Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat Commun* **9**, 943 (2018).

32. G. Li *et al.*, Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98 (2012).

33. M. R. Mumbach *et al.*, Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* **49**, 1602-1612 (2017).

34. A. L. Sanborn *et al.*, Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-6465 (2015).

35. G. Yardımcı *et al.*, Measuring the reproducibility and quality of Hi-C data. *bioRxiv*, (2018).

36. Y. Li *et al.*, CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**, e114485 (2014).

37. H. Y. Zhou *et al.*, A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes & development* **28**, 2699-2711 (2014).

38. S. Blinka, M. H. Reimer, Jr., K. Pulakanti, S. Rao, Super-Enhancers at the Nanog Locus Differentially Regulate Neighboring Pluripotency-Associated Genes. *Cell reports* **17**, 19-28 (2016).

39. J. M. Engreitz *et al.*, Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452-455 (2016).

40. N. Rajagopal *et al.*, High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**, 167-174 (2016).

41. R. Tewhey *et al.*, Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529 (2016).

42. S. D. Moorthy *et al.*, Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome research* **27**, 246-258 (2017).

43. D. R. Fuentes, T. Swigut, J. Wysocka, Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *Elife* **7**, (2018).

44. S. Spisak *et al.*, CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. *Nat Med* **21**, 1357-1363 (2015).

45. X. Wang *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).

46. K. Musunuru *et al.*, From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-719 (2010).

Acknowledgements: We thank J. Chen, A. Chow, B. Cleary, C. de Boer, A. Dixit, M. Guttman, R. Herbst, C. Lareau, S. Rao, J. Ray, S. Reilly, R. Tewhey, J. Ulirsch, and C. Vockley for discussions and J. Marshall for assistance with fluorescent microscopy. FACS sorting was performed at the Broad Institute FACS Core by P. Rogers, S. Saldi, and C. Otis. This work was supported by funds from the Broad Institute (E.S.L.) and by NIH NHGRI Grant No. 1K99HG009917-01 to J.M.E. J.M.E. is supported by the Harvard Society of Fellows. S.R.G. is supported by National Institute of General Medical Sciences grant T32GM007753. E.S.L. serves on the Board of Directors for Codiak BioSciences and Neon Therapeutics, and serves on the Scientific Advisory Board of F-Prime Capital Partners and Third Rock Ventures; he is also affiliated with several non-profit organizations including serving on the Board of Directors of the Innocence Project, Count Me In, and Biden Cancer Initiative, and the Board of Trustees for the Parker Institute for Cancer Immunotherapy. He has served and continues to serve on various federal advisory committees. J.M.E., C.P.F., and E.S.L. are inventors on a patent application filed by the Broad Institute related to this work. Data presented in this chapter can be found in GEO accession GSE118912.

Chapter 4: Conclusion

Overview

In this thesis, I described our work toward understanding how enhancers regulate specific target genes. At the outset of this work, there were few examples where the target genes of an enhancer were known. These examples suggested a vast, complex network in which individual genes can be regulated by multiple enhancers and individual enhancers can regulate multiple genes across large distances in the genome (*1-3*). In the face of this complexity, we have lacked a systematic understanding of enhancer-promoter communication.

To address this challenge, we developed scalable, general, and quantitative perturbation-based tools to characterize the regulatory functions of noncoding elements in their native genomic contexts, and applied them to systemically map the enhancers regulating dozens of genes. This compendium allowed us to evaluate, in an unbiased manor across many loci, predictive models of enhancer-gene connections. No existing model explained the patterns of connections we observed.

In order to predict the observed functional connections, we developed the "Activity-by-Contact" (ABC) model 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"). The ABC model enabled us for the first time to accurately predict enhancer-gene connections based on epigenetic data.

Moreover, our comprehensive mapping of the enhancers regulating dozens of genes provides insight into the mechanisms of enhancer function. The appreciation that quantitative 3D contacts is a hallmark of functional regulatory connections unifies seemingly disparate observations, such as that in some cases enhancers-gene connections indeed correspond to chromatin loops (*4*), that disruption of TAD boundaries can alter enhancer regulation (*5-7*), and that many enhancers regulate the expression of genes in close linear proximity in the genome (*1*). Further, the success of the ABC model demonstrates that functional specificity can arise from the precise arrangement and activities of enhancers in the genome, even in the context of broad biochemical compatibility.

Opportunities for understanding enhancer function

Despite the success of the ABC model in predicting enhancer-gene connections, the regulatory maps we observed highlight two remaining questions.

First, how do multiple enhancers combine to regulate a single gene? Numerous studies have found that a single gene can be regulated by several enhancers even in the same cell type. While it is generally unclear if these enhancers combine additively, sub-additively, or synergistically, in several cases we find evidence that enhancers act synergistically. For example, the sum of the effect sizes observed upon individually perturbing the 7 enhancers for *MYC* in K562 cells is more than 100% (Chapter 2), and the ABC score shows a log-linear relationship with effect sizes across our full dataset (Chapter 3). Further work using combinatorial perturbations to enhancers will be required to address how multiple enhancers combine to regulate gene expression.

Second, how do enhancers activate their target genes biochemically? Enhancers are defined functionally as elements that can activate transcription, and we lack a unifying framework to describe how enhancers achieve this activation. For example, the ABC model relies on an estimate of activity based in part on H3K27ac ChIP-seq, but we lack a mechanistic understanding of why H3K27ac marks active enhancers. As described in Chapter 1, pioneering experiments based on the simultaneous single molecule imaging of promoters, enhancers, and RNA are enabling us to watch transcriptional activation as it unfolds (*8-10*). I anticipate that these approaches, in combination with targeted perturbations of enhancers, promoters, and/or transcriptional co-factors, will unlock the molecular basis of enhancer function. The observation that ubiquitously expressed genes are insensitive to enhancers may also provide a foothold into this problem; understanding why ubiquitously expressed gene promoters encode expression without distal enhancers may elucidate why tissue-specific gene promoters rely on enhancers. Notably, ubiquitously expressed genes have been observed to have highly efficient pause release (*11*), supporting an emerging hypothesis that many enhancers function through the release of promoter proximal pausing (*12, 13*).

An emerging view of transcriptional control is that the molecular factors required to regulate and carry out transcription may form phase-separated condensates, membraneless organelles created by networks of weak, transient, largely non-structured interactions between many multivalent factors (*14-17*). This view is prompting the field to rethink previous results and assumptions to consider how (and if) they fit into this new paradigm. The ABC model, though based on a molecular intuition separate from the phase-separated condensate model, is in fact highly compatible with it, albeit with a change in vocabulary. For example, it may be that "Contact" in the ABC model represents membership in the same phase-separated droplet rather than direct, structured protein-protein interactions between factors bound at the enhancer and promoter. The formation of phase-separated condensates depends synergistically on the concentrations of transcriptional co-factors recruited by enhancers (*14*), which may form part of the mechanism for the apparent synergy we observe in the combinatorial action of multiple enhancers.

Outlook

The ability of the ABC model to predict the effects of perturbing an enhancer suggests an immediate application: to predict the target genes of enhancers disrupted by genetic variants associated to human disease through GWAS. GWAS studies are unbiased surveys of the genome to identify common genetic variants correlated with a phenotype (*18*). The variants identified through GWAS are often located outside annotated genes and enriched in enhancer elements (*19-21*). Because we have lacked a predictive understanding of how variants in enhancers affect the expression of target genes, translating these observations into biological or therapeutic insight has proved challenging (*22*). Despite the thousands of disease-associated variants identified through GWAS (*23*), in only a few cases has the likely causal gene and cell type been identified experimentally (examples listed in (*22*)). The cell-type specific networks of enhancer-gene connections predicted by the ABC model present a path forward for narrowing the set of likely causal genes and cell types for each association.

More broadly, our understanding of the non-coding genome is approaching an inflection point. As our mechanistic and predictive models of regulatory genomics improve, our focus is shifting from describing to manipulating gene regulation. Already, therapies based on combining insights from human genetics and regulatory genomics are emerging, including approaches that target enhancer-associated factors (*24*) or even directly target enhancers (*25*). It is no longer hard to imagine a time when we understand how the genome encodes biology, and how to leverage this understanding therapeutically.

References

1. J. van Arensbergen, B. van Steensel, H. J. Bussemaker, In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol* **24**, 695-702 (2014).

2. F. Spitz, E. E. Furlong, Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613-626 (2012).

3. M. Bulger, M. Groudine, Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339 (2011).

4. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

5. W. A. Flavahan *et al.*, Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110-114 (2016).

6. M. Franke *et al.*, Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265-269 (2016).

7. D. Hnisz, D. S. Day, R. A. Young, Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167**, 1188-1200 (2016).

8. H. Chen *et al.*, Dynamic interplay between enhancer-promoter topology and gene activity. *Nat Genet*, (2018).

9. C. R. Bartman *et al.*, Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. *Molecular cell* **73**, 519-532 e514 (2019).

10. J. Rodriguez *et al.*, Intrinsic Dynamics of a Human Gene Reveal the Basis of Expression Heterogeneity. *Cell* **176**, 213-226 e218 (2019).

11. D. A. Gilchrist *et al.*, Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143**, 540-551 (2010).

12. A. Zippo *et al.*, Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation. *Cell* **138**, 1122-1136 (2009).

13. F. X. Chen *et al.*, PAF1 regulation of promoter-proximal pause release via enhancer activation. *Science* **357**, 1294-1298 (2017).

14. D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, P. A. Sharp, A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13-23 (2017).

15. A. Boija *et al.*, Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855 e1816 (2018).

16. B. R. Sabari *et al.*, Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, (2018).

17. W. K. Cho *et al.*, Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412-415 (2018).

18. J. N. Hirschhorn, M. J. Daly, Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108 (2005).

19. M. T. Maurano *et al.*, Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).

20. A. Visel, E. M. Rubin, L. A. Pennacchio, Genomic views of distant-acting enhancers. *Nature* **461**, 199-205 (2009).

21. M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, M. Snyder, Linking disease associations with regulatory information in the human genome. *Genome research* **22**, 1748-1759 (2012).

22. M. D. Gallagher, A. S. Chen-Plotkin, The Post-GWAS Era: From Association to Function. *Am J Hum Genet* **102**, 717-730 (2018).

23. A. Buniello *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).

24. A. Stathis, F. Bertoni, BET Proteins as Targets for Anticancer Treatment. *Cancer Discov* **8**, 24-36 (2018).

25. Y. Wu *et al.*, Highly efficient therapeutic gene editing of human hematopoietic stem cells. *Nat Med*, (2019).

Appendix A. Supplemental Material for Chapter 1

Supplemental Notes

Note A1. A generalizable method to discover and characterize gene regulatory elements

We set out to develop an approach to identify noncoding elements that regulate a given gene in its endogenous genomic context. A method to accomplish this would need to be able to (i) survey the regulatory function of many thousands of kilobases of genomic sequence, including regions not predicted to have regulatory function; (ii) sensitively identify and robustly quantify the effects of noncoding elements, and (iii) be generally applicable to study any gene of interest.

We designed our CRISPRi-based screening approach to address these goals. Our results in the *GATA1* and *MYC* loci demonstrate that this approach is scalable, sensitive, and specific. In the following sections we describe the conceptual and technical features that enable these characteristics and compare this method to similar approaches that use catalytically active Cas9 (*1-3*).

CRISPRi enables scalable functional characterization of gene regulatory elements.

Because noncoding regulatory elements can be located far from their target genes and a gene might be controlled by multiple elements (*4-6*), a method to dissect the regulatory architecture of a given gene must be able to interrogate, through loss-of-function experiments, large regions of genomic sequence. To develop a scalable method, we exploited the programmable CRISPR system in the setting of a pooled screen to simultaneously interrogate the functions of many noncoding regions. In this method, we synthesize a library of sgRNAs targeting noncoding regions of interest; generate a lentiviral library containing each of these sgRNAs; and establish a population of cells in which each cell expresses doxycycline-inducible KRAB-dCas9 and a single sgRNA. The effects of each sgRNA can be identified by using high-throughput sequencing to characterize the representation of sgRNAs in the cell population before and after a phenotypic selection (*7, 8*). This approach enables high-throughput interrogation of noncoding elements: in this study, we assay 1.29 Mb of sequence around *GATA1* and *MYC* in a single pooled experiment.

CRISPRi robustly identifies gene regulatory elements.

A method for characterizing the regulatory network for a given gene needs to be able to robustly identify regulatory elements, even when their effects on gene expression are relatively small in magnitude. Several features of our approach help to provide high sensitivity and specificity for regulatory elements.

First, the pooled screening format provides numerous advantages that help to identify small effects. Specifically, pooled screens include contributions of many individual cells for each sgRNA; assess the functions of different sgRNAs in the same experimental context (in the same plate); and measure changes in sgRNA representation using count-based statistics.

Second, the use of the KRAB-dCas9 system enables independent assessments of the function of the same regulatory element with multiple adjacent sgRNAs. This property stems from the fact that KRAB-dCas9 appears to disrupt the functions of regulatory elements across distances on the order of hundreds of base-pairs (*9*), such that in the *MYC* and *GATA1* loci we observe regions where dozens of sgRNAs are consistently depleted (Figure 2-1B, 2A). This is advantageous for quantifying the impact of an element because the efficacy of individual sgRNAs varies for reasons inherent to the CRISPR system, such as the effect of the targeting sequence on sgRNA transcription or stability (*7*). Thus, the degree to which an individual sgRNA affects gene expression reflects not only the importance of the disrupted element but also the potency of the sgRNA itself. To address this issue, we average the scores across multiple consecutive sgRNAs, providing a more robust estimate of the effect of an individual element. We note that this property appears to differ qualitatively from previous approaches using catalytically active Cas9 to perform mutagenesis of noncoding regions (*1-3*). Cas9-mediated mutagenesis relies on non-homologous end-joining to disrupt critical sequence motifs, and so – because the resulting indels are on the order of tens of bases or smaller – only the few sgRNAs very close to critical sequence motifs appear to disrupt the function of any given regulatory element (*1-3*). These properties may be important in determining the power of screens using each approach and may have different trade-offs for positive versus negative selection screens.

Supporting the specificity and sensitivity of this approach, we find that each of the elements identified by our CRISPRi screens (e-GATA1, e-HDAC6, and e1-e7), do in fact affect the expression of the intended gene, including effects on gene expression as small as 10%. We note that the sensitivity of this approach for even smaller effects might be accomplished by assaying more cells per sgRNA.

CRISPRi-based screening is general and can be applied to study other genes or phenotypes.

A general method for identifying gene regulatory elements should be applicable to any gene of interest. While we looked for effects on survival and proliferation in K562 cells in order to characterize multiple gene loci in a single screen, we note that this CRISPRi-based approach could be applied to study an arbitrary gene of interest through fluorescence-based readouts of cells with a gene tagged in its endogenous locus with GFP (*1*). This strategy for mapping regulatory elements can also be applied in the context of other functional readouts, including other FACS-based assays (*2, 10*) or drug or toxin resistance phenotypes (*8, 11*).

Together, these properties provide a scalable, sensitive, and general method for mapping the functions of gene regulatory elements. This CRISPRi-based approach appears to have complementary properties to Cas9-mediated mutagenesis approaches (*1-3*): CRISPRi can robustly identify gene regulatory elements and provides non-mutagenic inhibition that is consistent across individual alleles and cells, while mutagenesis-based approaches appear to provide high resolution for identifying specific motifs. Further work will be required to determine how to best leverage these complementary features to dissect the networks of noncoding elements controlling gene expression. Finally, we note that in theory neither approach will be able to identify elements that act redundantly with other elements in a given locus, or elements that reside in repetitive genomic regions that cannot be uniquely targeted with CRISPR. Although we found several instances in which promoters repress neighboring genes, perhaps by a competition mechanism, it remains unclear whether CRISPRi can identify other types of repressive elements that are not promoters. Similarly, its utility in identifying intronic enhancers within the body of the assayed gene is unclear, as recruitment of KRAB-dCas9 to these sites may directly interfere with transcription. Further technical advances will be required to characterize and explore the functions of these elements.

Note A2. Essentiality of noncoding RNAs in the MYC locus.

Previous CRISPR screens have established that the protein coding genes expressed in the vicinity of *MYC* are not essential in K562 cells (Figure A-1). We further considered whether noncoding RNA genes in this region — including *PVT1*, *CCDC26*, and 5 microRNAs — are also essential and thus might explain the effects on cell proliferation conferred by the enhancers we discover in the *MYC* locus. In each case, we found that these noncoding RNAs either do not affect cell proliferation in K562 cells (*PVT1* and *CCDC26*) or are not detectably expressed (microRNAs) and thus that e1-e7 likely control cell proliferation through regulation of *MYC*.

Two of the *MYC* enhancers we identified (e3 and e4) correspond to promoters that produce short alternative isoforms of the long noncoding RNA (lncRNA) PVT1 (Figure 2-2A). Because PVT1 has previously been reported to affect cellular proliferation in *trans* based on siRNA-mediated knockdown experiments in mammary and ovarian cell lines (*12, 13*), we investigated whether a *trans* function of the PVT1 transcript could be responsible for its promoters affecting cellular proliferation in K562 cells. We performed competition assays between K562 cells transfected with control siRNAs and cells transfected with siRNAs against PVT1 or, as positive controls, MYC or GATA1 (see Methods). Knockdown of MYC or GATA1 (27% or 52% reduction, respectively) led to a reduction in cellular proliferation relative to cells transfected with control siRNAs, as expected (Figure A-1C,D). In contrast, knockdown of PVT1 (66% reduction for the best siRNA) did not lead to detectable changes in proliferation (Figure A-1C,D). This indicates that reduction of the mature PVT1 lncRNA does not affect the proliferation of K562 cells.

In contrast, we found that CRISPRi targeting e3 (corresponding to a TSS of PVT1), which led to a ~77% reduction in PVT1 RNA levels (Figure A-1E), *did* affect cellular proliferation in competition assays (Figure 2-2C). Thus, the proliferative defect observed upon inhibition of these elements in K562 cells appears to reflect their functions in the *cis* regulation of *MYC* rather than previously reported *trans* functions of the PVT1 RNA transcript itself. This is consistent with previous findings that gene promoters (including promoters of lncRNAs) can act as enhancers for neighboring genes (*14, 15*). Indeed, we show that both e3 and e4 activate expression of a plasmid-based reporter gene (Figure A-5B, see Methods), indicating that these elements can act as enhancers. Further work will be required to investigate the possibility that other mechanisms associated with PVT1 transcription might also quantitatively contribute to controlling *MYC* expression in *cis*.

In addition to *PVT1*, the *MYC* region also contains the lncRNA *CCDC26* (a pseudogene), which is expressed from a TSS 7.2 Kb distal to e5. Although e5 scored in our screen and affected *MYC* expression, we did not observe depletion of sgRNAs targeting the *CCDC26* TSS or promoter despite an abundance of sgRNAs in these regions (Figure A-5B). Thus, e5 and other enhancers likely affect cell proliferation through regulation of *MYC* rather than through regulation of *CCDC26*. We note that it is technically possible that depletion of *CCDC26* or *PVT1* contributes to affecting cell proliferation *in the context of MYC suppression*, but our data are inconsistent with them having strong effects on cell proliferation independent of changes in *MYC*.

The genetic region around also *MYC* harbors five putative miRNA genes previously described in several cancer cell lines (miR1204-1208). To determine if these miRNAs are expressed in K562s, we inspected ENCODE short RNA sequencing data (wgEncodeCshlShortRnaSeqK562CellShortAln.bam) and found that 0 reads (out of >29 million reads) overlap the RefSeq-annotated putative miRNAs in the region. Because regulation by miRNAs is thought to be highly dependent on miRNA abundance (*16*), miR1204-1208 do not likely have important functions in K562 cells.

Note A3. Repressive elements in the MYC locus.

We identified 2 elements in the *MYC* locus (r1 and r2, Figure 2-2A, A-5) whose inhibition by CRISPRi led to *increased* proliferation of K562 cells in our screen, suggesting that these elements may act to repress *MYC* expression. Both of these elements have smaller absolute effect sizes in the screen data than the weakest detected enhancer (e5, 10% reduction in *MYC* expression), suggesting that these repressive elements may have even smaller quantitative effects on *MYC* expression. Interestingly, one of these elements corresponds to the promoter of a minor PVT1 isoform (Figure 2-2A), consistent with a model wherein this promoter of *PVT1* competes with the *MYC* promoter for regulatory signals.

Note A4: Conceptual framework for predicting enhancer function.

Our heuristic approach for comparing the relative activity of enhancers is based on a classic model in which an enhancer affects gene expression by recruiting transcription factors and activating gene expression upon physical contact ("looping") between the enhancer and a target promoter (*17, 18*). In this model, the quantitative impact of an enhancer might depend on (i) its intrinsic activity (*i.e.*, the complement of transcription factors recruited to the element and their effects on a target promoter) and (ii) the frequency at which the enhancer physically contacts its target promoter in the nucleus. We note that this model does not represent all of the possible mechanisms by which regulatory elements might regulate their target genes (*17*), but does provide a simple framework with which to combine these two aspects of enhancer function.

To represent the intrinsic activity of an enhancer, we used quantitative measures of DHS and H3K27ac occupancy (see Methods) based on previous evidence that they correlate with various measures of activity. For example, DHS signal at regulatory elements in the genome correlates with transcription factor occupancy (*19, 20*) and with the activity of those elements in plasmid-based reporter assays (*21*). H3K27ac occupancy correlates with expression of neighboring genes across cellular contexts (*22, 23*) as well as with on-plasmid enhancer activity (*21*).

To represent the contact frequency between an enhancer and promoter, we used genome-wide measurements based on Hi-C (*24*) (see Methods), a method that requires physical contact and crosslinking in order to produce a signal linking two regions of genomic DNA. Physical contacts between enhancers and promoters correlate with gene activation (*6, 17, 18, 25*), and in a few cases increasing the frequency of enhancer-promoter contact has been shown to activate gene expression (*26, 27*).

These observations provide a conceptual foundation for this heuristic approach to comparing the relative impact of enhancers on gene expression. Further work will be necessary to determine whether this approach in fact reflects the mechanisms by which these enhancers regulate *MYC*. Regardless of the underlying mechanisms, this simple heuristic can distinguish elements that regulate *MYC* in K562 cells from those that do not and may be more broadly useful for connecting regulatory elements with their target genes.

Note A5: Guidelines for design of additional CRISPRi screening libraries.

We sought to determine how to best design CRISPRi screening libraries using fewer sgRNAs per gene and thus enabling the interrogation of more genes. We analyzed our data by down-sampling the number of sgRNAs to every 2nd, 4th, 5th, or 10th sgRNA within each 20-sgRNA window. We found that, as expected, this reduces the reproducibility of estimates of the quantitative effects of elements and thus reduces power to detect elements with small effects (Figure A-9A).

An alternative strategy for designing smaller libraries is to focus on the subset of regions that are likely to score. All of the elements detected in our screen are centered on DHS sites (Figure A-9B) and every significantly depleted or enriched 20-sgRNA window is located within 1 kb of a DHS peak (the union of wgEncodeUwDnaseK562PkRep1.narrowPeak and wgEncodeUwDnaseK562PkRep2.narrowPeak). Designing a screen against only DHS sites could reduce the size of the library by approximately a factor of 5. However, it remains unclear whether there are regulatory elements in other loci that are not DHS sites.

Methods

Selection of targets for sgRNA library

To develop this CRISPRi screening approach (see Note A1), we focused on two genes — *MYC* and *GATA1* — that play critical roles in human development and disease and that are known to affect cellular proliferation in K562 cells (*28*). We determined by consulting a genome-wide catalog of gene essentiality in K562 cells (*28*) as well as Hi-C data in K562 cells (*25*) that *MYC* and *GATA1* are not located in close linear (500 Kb) or spatial proximity (within the same topological domain) to other genes expressed in K562 cells that strongly affect cell proliferation (Figure A-1). We also examined the potential effects of several noncoding RNAs in the *MYC* locus on cell proliferation, but determined that none are likely to contribute (see Supplemental text).

We designed an sgRNA library containing guides targeting several loci as well as internal controls, for a total of 98,599 sgRNAs. We dedicated most of the sgRNAs in the library to studying the *MYC* locus, due to the apparent complexity of its regulatory architecture (*e.g.*, see Figure 2-3A) (*29*) and its importance in many human cancers. To identify the elements that regulate *MYC*, we examined the 3-Mb topological domain and selected a ~666 Kb region that contained *MYC* itself, many elements with strong DHS and H3K27ac signal in K562 cells, and all intervening regions. We selected additional regions throughout the domain to cover other strong H3K27ac peaks downstream of *MYC* (including the regions surrounding e5-e7 that from Hi-C can be observed to form long-range loops to the *MYC* promoter), as well as additional regions upstream of *MYC* that are marked by active chromatin in other cell types but not in K562s (*e.g.*, see Figure 2-3A). In each case, we included at least 5 kb of sequence surrounding the ENCODE "broadPeak" annotations. We note that performing similar experiments with larger libraries — for example including all possible sgRNAs in the 3-Mb topological domain containing *MYC* — is possible and would require increasing the scale of the experiment (number of cells and reads) accordingly.

For *GATA1*, we tiled a 74 kb region containing the *GATA1* gene body as well as several putative enhancer elements nearby, including 17 kb annotated as "weak enhancer" and 19.4 kb annotated as "strong enhancer" by ENCODE ChromHMM (Figure 2-1B). We note that we do not rule out the possibility that additional regulatory elements beyond this span may regulate *GATA1*.

We included several additional sets of sgRNAs as internal positive and negative controls for the screen. As negative controls, we included 4,082 scrambled-sequence sgRNAs, selected to include all 20- or 21-nucleotide sgRNAs from the previous genome-wide CRISPRi screening library designed by the Weissman lab (*11*), subject to the filters described below. We also included sgRNAs targeting the promoters of 600 protein-coding genes – including 535 that are expressed in K562 cells (fragments per kilobase per million >1) and 65 that are not expressed – as internal standards in the screen to compare to previous genome-wide screens assessing gene essentiality (*11, 28*). We selected these genes to span the range of potential effects on cellular proliferation, including the 52 most essential genes reported previously (*28*).

Finally, because sgRNAs tiling across a noncoding region might be subject to different biases than scrambled-sequence sgRNAs (*e.g.*, due to specific sequence motifs, repetitive regions, or general toxic effects of targeting KRAB-dCas9 to chromatin), we selected additional negative control regions that are not close to genes known to be strongly essential but nonetheless do have putative regulatory elements marked by DHS and H3K27ac. We used these negative control regions (85 kb total) to estimate an empirical false discovery rate for elements in the *GATA1* and *MYC* loci (see below).

sgRNA design for tiling noncoding sequences

To design sgRNAs for tiling across noncoding sequences, we generated a list of all possible targeting sites with an NGG PAM. We calculated a specificity score based on potential off-target sites using a previously described algorithm (http://crispr.mit.edu, (*30*)), and removed guides with specificity scores <20. We note that this means that certain noncoding regions, including regions containing repetitive elements, are not tested by this screen. For cloning sgRNAs into sgOpti, we added a "G" base to the beginning of the 20-nucleotide sequence if the first base was not already a "G". We note that we applied additional filters to the sgRNAs considered during analysis of the screen (see below).

sgRNA design for targeting promoters

Because CRISPRi has a ~200-bp window of efficacy surrounding the TSS (Note A1) (*31*), we used capped analysis of gene expression (CAGE) data from K562 cells (*32*) to precisely define TSS locations (10-bp resolution) and designed sgRNAs targeting the regions immediately proximal to this site. In cases where genes showed multiple TSSs (as judged by the second-strongest TSS having >20% of the CAGE signal of the stronger TSS), we designed sgRNAs against both of these TSSs. To design sgRNAs targeting these sites, we used an algorithm based on a previous approach (*11*). We first generated all possible guides of length 18-24 where the first position in the genome corresponds to a "G", filtering out those with potential for off-target effects based on their specificity score. We defined prioritized windows around the TSS corresponding to (-30 to +45 bp), (-30 to +95 bp), and (-200 to +200 bp). We selected sgRNAs from these regions in order until we obtained 20 sgRNAs per promoter. For each window, we chose as many sgRNAs as possible that were spaced at least 5 bp apart, and then moved to the next priority window.

Tissue Culture

We maintained K562 (ATCC) cells 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 unless otherwise noted.

Constructs for CRISPRi

We expressed sgRNAs from sgOpti, a modification of pLenti-sgRNA (Addgene #71409) with the sgRNA scaffold replaced with the sgRNA-(F+E)-combined optimized scaffold previously described (*33*). We generated constructs expressing inducible KRAB-dCas9 by replacing the SFFV promoter with a TRE3G promoter and the P2A-mCherry cassette with an IRES-GFP or IRES-BFP cassette in pHR-SFFV-KRAB-dCas9-P2A-mCherry (Addgene #60954) (*11*).

CRISPRi line generation

We generated the inducible CRISPRi cell lines by (i) transducing K562 cells with a construct expressing rtTA linked by IRES to a neomycin resistance cassette expressed from an EF1α promoter (ClonTech, Mountain View, CA) and selecting with 200 μg/mL G418 (Thermo Fisher), then (ii) transducing these rtTA-expressing K562 cells with one of the KRAB-dCas9 constructs described in the section above. We selected for cells expressing GFP or BFP by fluorescence activated cell sorting (FACS).

sgRNA library cloning

We synthesized an oligo pool corresponding to the sgRNA library with PCR tags (purchased from CustomArray, Bothell, WA). We amplified the pool by PCR with primers sgRNA Library Fwd/Rev to add homology arms for Gibson assembly, and purified the product with an equal volume (1×) AMPure XP SPRI beads (Beckman Coulter, Danvers, MA). We prepared the vector backbone by digesting sgOpti with BsmBI (New England Biolabs (NEB), Ipswich, MA) followed by purification with 0.75× AMPure XP SPRI. We assembled 70 ng amplified library into 500 ng digested vector in a 50 μL Gibson reaction (NEB), cleaned these by 0.75× AMPure XP SPRI, eluted in 15 μL H_2O and electroporated the entire volume into Endura competent cells (Lucigen, Middleton, WI). We expanded the cells in liquid culture for 18 hours at 30 °C and purified the pooled library plasmid with the Endotoxin-Free Plasmid Maxiprep Kit (Qiagen, Hilden, Germany).

Lentivirus production

We plated 700,000 HEK293T cells on 6-well plates (Corning, Corning, NY) and 24 hours later transfected with 1 μg dVPR, 300 ng VSVG, and 1.2 μg transfer plasmid using XtremeGene9 (Roche Diagnostics, Indianapolis, IN). For pools, the cell number and plasmid mass were scaled proportionally to 14 million cells on a 15 cm plate (Corning). 16 hours post-transfection we changed media to DMEM with 20% HIFBS. At 48 hours post-transfection, we harvested viral supernatants and filtered them through a 0.45 μM syringe filter before use.

Pooled CRISPRi screens for essentiality

We transduced K562 harboring a doxycycline-inducible KRAB-dCas9 at an multiplicity of infection (MOI) of 0.3 at a coverage of 1,000 transduced cells per sgRNA as previously described (*28*). Starting 36 hours after transduction, we selected for successfully transduced cells with 1 μg/mL puromycin for 72 hours and collected 150 million cells as a reference sample. After maintaining cells at 1,000× coverage in 0.2 μg/mL puromycin and 0.5 μg/mL doxycycline for 14 population doublings, we collected 150 million cells of the final cell population. We extracted genomic DNA from both the reference and final cell populations using the QIAamp DNA Blood Maxi kit (Qiagen) according to the manufacturer's instructions. We amplified sgRNAs integrations from 900 μg genomic DNA by PCR with indexed sgRNA sequencing library primers containing Illumina adaptors and sequenced them on a HiSeq 2500 using custom Illumina sequencing and index primers to an average depth of >350 reads per sgRNA. We used Bowtie (*34*) to align the resulting sequences to the sgRNA library allowing perfect matches only.

Analysis of sgRNA depletion in proliferation-based screen

To evaluate the potential of off-target sgRNA-mediated toxicity to affect cellular proliferation, we inspected the depletion of the set of sgRNAs in the tiled negative control regions (where we expect no on-target sgRNA depletion) and noted that the frequency of sgRNAs more than 2-fold depleted across the screen is higher (2-proportion Z-test $p<0.0001$) in sgRNAs with specificity scores below 50 (9%) than those with a score of 50 or above (5%). We considered only the sgRNAs with specificity scores >50 in the subsequent analysis. We also ignored sgRNAs with more than 10 "G" bases in the targeting sequence, which also lead to an increased frequency of off-target toxicity based on analysis of the negative control sgRNAs. These filters retain >90% of sgRNAs. To ensure robust calculation of sgRNA scores, we examined only sgRNAs with at least 50 raw reads in the initial timepoints for both replicates (retains 98% of sgRNAs). We assessed the depletion of the remaining sgRNAs as described below.

CRISPRi score

The "CRISPRi score" represents the $-\log_2$ depletion between the beginning and end of the proliferation screen (14 population doublings). We calculated the CRISPRi score for each of two replicates and report the mean of these scores as the CRISPRi score for each sgRNA. To identify significant regions by integrating information from multiple sgRNAs, we used a sliding window approach, averaging the mean CRISPRi score across N consecutive guides. To choose N, we compared the correlation of the window CRISPRi scores between the two replicates as a function of N (Figure A-2A). We found that using $N = 20$ yielded a Pearson's correlation of 0.80 between the two replicates (Figure A-2B). As the sgRNAs were spaced on average every ~16 bp (Figure A-2C), windows of 20 consecutive sgRNAs spanned on average 314 bp (median = 237 bp, Figure A-2D). We note that this resolution is on the same order as the size of scoring regions in our CRISPRi screen (hundreds of bp), indicating that choosing a smaller window size would not necessarily increase the resolution of the approach. Because some regions are covered sparsely due to repetitive sequence, we considered windows only if they contained 20 guides within a span of 1000 bp (Figure A-2D). We note that the enhancers we identify (e-GATA1, e-HDAC6, e1-e7) are robust to the precise choice of window size.

To identify significant windows, we required first that the CRISPRi score for the window had an irreproducible discovery rate < 0.05 when comparing the two replicate screens (*35*). Second, we tested whether the mean of the sgRNAs in each window deviated significantly from the mean of the negative controls, using sgRNA CRISPRi scores averaged across duplicate screens. Specifically, we calculated a T-test statistic by comparing the CRISPRi scores of the 20 sgRNAs with those of the scrambled-sequence, negative control sgRNAs. We assessed the empirical false discovery rate (FDR) of windows in the GATA1 and MYC loci by comparing these T statistics to those generated from sliding windows across three negative control regions that are located far from known essential genes expressed in K562 (see Selection of targets for sgRNA library), and selected a threshold based on a FDR of 0.05. This threshold corresponded to a Benjamini-Hochberg-corrected T-test *p*-value of 0.032. We considered significant elements with an absolute effect size of >25%.

The final reported CRISPRi scores for 20-sgRNA windows in figures represent the average of the two replicate screens normalized to the average of the scrambled-sequence negative-control sgRNAs.

Sources for epigenomics data

We downloaded data generated by the ENCODE Project Consortium (*36*) in K562 cells corresponding to DNase I hypersensitivity sequencing (DHS-seq); H3K27ac, GATA1, and CTCF chromatin immunoprecipitation sequencing (ChIP-seq); the chromatin state hidden Markov model (ChromHMM); and RNA Pol II ChIA-PET (*37*). To examine transcription factor occupancy at various enhancers, we downloaded the genome-wide binding sites of 100 transcription factors based on ChIP-Seq in K562 cells (wgEncodeRegTfbsClustered track from UCSC Genome Browser). We obtained sequence conservation from the UCSC Genome Browser corresponding to the phastCons 100-mammal multiple alignment (*38*). CTCF motifs were identified using FIMO (*39*) to search for the "V_CTCF_01" and "V_CTCF_02" position weight matrices from TRANSFAC (*36*). We obtained *in situ* Hi-C data for multiple cell types and used 5-Kb resolution KL-normalized observed matrix for all plots and analyses (*25*).

Cloning individual sgRNAs

For each of the selected enhancers (e-GATA1, e-HDAC6, e1-e7), and promoters (*GATA1* and *MYC*) that scored in the screen, we selected 2 non-overlapping sgRNAs with a preference for sgRNAs with high specificity and CRISPRi scores and sgRNAs that overlap the peak of DNase hypersensitivity. For regions that did not score (NS1, *HDAC6* promoter), we selected sgRNAs based on the same criteria, although because these sgRNAs were not high scoring, we also preferred guides predicted to have high efficacy (*40*). As negative controls, we selected 5 sgRNAs from the set without genomic targets. We cloned these sgRNAs as previously described (*41*) into sgOpti.

Generating sgRNA-expressing stable cell lines

We generated stable cell lines expressing single sgRNAs by lentiviral transduction in 8 μg/ml polybrene by centrifugation at 1400 x *g* for 45 minutes with one million 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 0.2 μg/ml puromycin. For each sgRNA, we generated three independent polyclonal cell populations through triplicate infections.

Single sgRNA knockdown

We plated sgRNA-expressing stable cell lines at 200,000 cells/ml in 0.5 μg/ml doxycycline and harvested cells 24 hours later by lysing in Buffer RLT (Qiagen).

RNA extraction and quantitative RT-PCR

We extracted RNA from 20,000-50,000 cells per experiment in Buffer RLT (Qiagen) using Dynabeads MyOne Silane beads (Thermo Fisher), treated samples with TURBO DNase (Thermo

Fisher), and cleaned again with Dynabeads MyOne Silane beads. We used AffinityScript reverse transcriptase (Agilent Technologies, Lexington, MA) and random nonamer primers to convert RNA to cDNA. We performed qPCR using SYBR Green I Master Mix (Roche) and calculated differences using the ΔΔCT method versus *GAPDH*.

To achieve power to detect small effects in gene expression, we performed 3 technical qPCR replicates (from the same cDNA) and took the median value for further analysis. We also included many biological replicates. Specifically, we derived 3 independent lines for each sgRNA and assayed each once as a biological replicate in *GATA1* locus experiments (for a total of 3 replicates) and 4 times for experiments in the *MYC* locus (for a total of 12 biological replicates).

RNA sequencing and analysis

To examine the transcriptional changes resulting from inhibition of a GATA1 enhancer, we performed RNA-sequencing on cell lines expressing individual sgRNAs targeting the GATA1 TSS (2 different sgRNAs), e-HDAC6 (2 different sgRNAs), and non-targeting, negative controls (4 different sgRNAs). We generated RNA sequencing libraries from 3 biological replicates for each sgRNA and processed the data as previously described (*42*). We identified differentially expressed genes ($q < 0.05$, fold-change > 2) with DESeq2 (version 1.6.3) (*43*) and found a significant overlap in the sets of differentially expressed genes between *GATA1* TSS and e-HDAC6 targeting sgRNAs (Figure A-4B), suggesting that e-HDAC6 leads to downstream transcriptional changes consistent with direct regulation of *GATA1*.

Single sgRNA competitive growth assays

For competition experiments we pooled the indicated K562 cells expressing an individual sgRNA and KRAB-dCas9-IRES-BFP with K562s expressing either GFP or RFP (control cells) in 0.5 μg/mL doxycycline. We measured the fractions of CRISPRi and control cells by flow cytometry after 24 hours and again after 7 additional days. We performed each experiment in six replicates including competitions against both the GFP- and RFP-expressing control lines. We quantified the growth phenotype gamma as previously described (*11*).

Luciferase reporter assays for enhancer activity on a plasmid

To test the functions of each putative regulatory element in a classic reporter-based enhancer assay, we created a reporter plasmid derived from pGL4.23 (Promega, Madison, WI) where firefly luciferase is expressed from a 180-bp fragment of the *MYC* promoter (hg19 coordinates: chr8:128748316-128748495). We designed an insertion site ~2 kb upstream of the *MYC* promoter for inserting each candidate enhancer sequence, and we flanked this region with polyadenylation signals in either direction to avoid measuring luciferase activity driven from transcripts initiating from the enhancer elements themselves. The negative control sequence corresponded to a kanamycin resistance cassette.

For each construct, we transfected 500,000 K562 cells using the Lonza (Cologne, Germany) Amaxa 96-well Shuttle according to the manufacturer's instructions for this cell type (except transfecting all 500,000 cells in a single well) with 250 ng of reporter plasmid plus 250 ng of a plasmid expressing *Renilla* luciferase. We harvested cells 48 hours after transfection by spinning once, washing with PBS, and resuspending in 40 μl Passive Lysis Buffer (Promega). We performed the Dual-Luciferase Reporter Assay according to the manufacturer's protocol (Promega). Barplots report firefly luciferase activity normalized to *Renilla* luciferase activity and to the negative control construct for 3 replicate transfections.

Chromatin immunoprecipitation for H3K27ac

We performed ChIP for H3K27ac as previously described, with modifications (*44*). We grew K562 cells expressing individual sgRNAs targeting *MYC* enhancers or negative controls in the presence of doxycycline for 48 hours. We harvested cells, washed once in cold PBS, and crosslinked with 1% formaldehyde in PBS for 10 minutes at 37 °C followed by quenching with glycine for 5 minutes at 37 °C. We washed cells twice in ice cold PBS with 1× protease inhibitor (Roche). We flash froze the pellets and stored at -80°C until sonication, at which time we thawed the pellets on ice and lysed cells in ChIP Lysis Buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl pH 8.0) on ice for 10 minutes. We sonicated batches of 3 million cells in 100 µL using a Q800R2 Sonicator (QSonica, Newtown, CT) at 50% amplitude, 30 s on / 30 s off, for 7.5 minutes to obtain fragment sizes between 150 and 700 bp.

We diluted 100 µL lysate from 1 millions cells in 660 µL ChIP Dilution Buffer (0.01% SDS, 1.1% Triton X-100, 1.12 mM EDTA, 16.7 mM Tris-HCl pH 8.0), and saved an aliquot for whole-cell extract. For immunoprecipitation of H3K27ac (using antibody 39685 from Active Motif, Carlsbad, CA), we incubated 5 µl of antibody with Protein A/G beads (Thermo Fisher) in Blocking Buffer (500 mM Tween-20, 500 mM BSA in 1x PBS) for 2 hours at 4 °C. We then washed the beads once in Blocking Buffer, resuspended the beads in 55 µL Blocking Buffer, and added it to the DNA samples. We incubated the antibody-bead-lysate mixture overnight at 4°C rotating end over end. Next day, we washed the samples as follows: four times with 200 µL of RIPA Buffer (0.1% Na-deoxycholate, 0.1% SDS, 1% Triton X-100, 100 mM NaCl, 1 mM EDTA, 10 mM Tris-HCl pH 8.0), twice with 100 uL RIPA High Salt Buffer (0.1% Na-deoxycholate, 0.1% SDS, 1% Triton X-100, 500 mM NaCl, 1 mM EDTA, 10 mM Tris-HCl pH 8.0), twice with LiCl Wash Buffer (250 mM LiCl, 0.5% NP-40, 0.5% Na-deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.0), and twice with 1× TE. Following the washes, we resuspended beads in Elution Buffer (10 mM Tris-HCl pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.1% SDS) and incubated the resuspended beads at 65 °C for 10 minutes. Following this first brief reverse crosslinking step, we added 5 μL RNase Cocktail (Thermo Fisher) and incubated at 37 °C for 30 minutes, and then added 5 μl Proteinase K (NEB) and incubated at 65 °C for 2 hours. Samples were cooled on ice. DNA was extracted using Agencourt XP (SPRI) beads (Beckman Coulter) at 2× sample volume, followed by elution in 10 mM Tris-HCl pH 8.0. We performed quantitative PCR using Roche 2× SYBR Green Master Mix on a Roche LightCycler 480. We calculated enrichment compared to 5 positive control primers designed against H3K27ac peaks outside of the *MYC* region.

siRNA-mediated knockdown of MYC, GATA1, and PVT1

We transfected 200,000 cells with 10 nM siRNAs obtained from GE Dharmacon (Lafayette, CO) in quadruplicate using the Neon transfection system (Thermo Fisher, settings: 1,450 V, 10 ms width, 3 pulses). We harvested cells in Buffer RLT (Qiagen) 24 hours after knockdown and estimated target gene expression relative to cells transfected with non-targeting siRNAs by quantitative PCR as described above. For competition experiments we transfected fluorescently labeled cells (GFP or RFP) with indicated siRNAs at 10 nM following the described procedure. We pooled cells such that cells transfected with siRNAs targeting PVT1, MYC or GATA1 were matched with differently labeled cells transfected with non-targeting control siRNAs. We measured the GFP and RFP fractions immediately following transfection and again after 4 days by flow cytometry. Each experiment was carried out in quadruplicates and included a label-swap experiment.

Strategy for genetic deletions of enhancers in the MYC locus

To test the effects of enhancers on *MYC* expression through genetic manipulations, one straightforward experiment would be to use CRISPR/Cas9 to generate clonal cell lines containing homozygous knockouts of each putative enhancer and measure the effects on *MYC* using the qPCR assays described above. However, there are several reasons why this experiment is not ideal in our system. First, we observe significant biological variation in MYC expression between clonal cell lines. Second, MYC affects cellular proliferation and thus cells lacking one of these enhancers may be outcompeted. Finally, K562 cells are triploid, making it difficult to obtain cell lines where an enhancer is removed on all 3 alleles.

Accordingly, we developed an alternative strategy (Figure A-7). We used CRISPR/Cas9 to generate clonal cell lines carrying *heterozygous* genetic deletions (on 1 or 2 of the 3 homologous chromosomes) and compared the expression of *MYC* on the modified and unmodified homologous chromosomes in the same cells. We expect that if the enhancer in fact regulates *MYC*, *MYC* expression from the modified allele should be reduced compared to the wild-type allele. This approach is identical in concept to classical *cis-trans* tests. This allele-specific approach can demonstrate that regulation of *MYC* is a direct, *cis* effect of the enhancer rather than an indirect effect (for example, due to the enhancer regulating another gene that in turn regulates *MYC*).

To implement this strategy, we first generated a cell line containing polymorphic sites on each allele of *MYC*. Because K562 cells do not contain polymorphisms in the *MYC* transcript, we knocked in polymorphic tags using CRISPR/Cas9 and homologous recombination. We first chose a targeting site in a *MYC* intron in a region that did not show sequence conservation across mammals. We reasoned that editing such a site would not likely affect the regulation of *MYC*. We designed an sgRNA targeting this site as well as a ssDNA oligo to use as a donor for homologous recombination (Figure A-7A). This oligo contained four random nucleotides (NNNN), allowing us to generate cell lines containing unique polymorphic on each of the 3 alleles. We co-transfected these sgRNAs, Cas9, and the donor oligo in K562 cells, isolated clonal cell lines through serial dilution, and genotyped this intronic site by PCR and sequencing. We identified a clonal cell line containing 3 distinct variants (CTAA, CCCG, and ATCG) in the targeted location. We expanded this cell line (K562-MYC-Tag) and used it for the second round of transfections.

To delete *MYC* enhancers, we designed sets of 4 sgRNAs flanking each element, with 2 sgRNAs on each side. These sgRNAs were designed to delete ~1 kb regions containing the DHS site in the middle of the element. For e3 and e4, we designed the sgRNAs to cut outside of the exons and splice sites of PVT1. We co-transfected the K562-MYC-Tag cell line with Cas9 and sets of 4 sgRNAs, generated clonal cell lines through serial dilution, and genotyped each clone (Figure A-7B). We expanded clones containing deletions on 1 or 2 of the 3 alleles.

For each deletion clone and for 26 wild-type control clones, we use a droplet digital PCR (ddPCR) hydrolysis assay to measure the allele-specific expression of *MYC* and *PVT1*. We used this data, in combination with the genotyping amplicon sequencing, to infer partial phasing of the alleles relative to the polymorphic tags in the *MYC* intron (Figure A-7C). We performed these experiments for e2, e3, and e4 because these loci had SNPs that allowed us to determine which allele was deleted (see below). We compared the allele-specific expression between wild-type and deletion clones to determine how deleting *MYC* enhancers affected MYC expression (Figure A-7D,E).

Additional technical details for each of these steps are included below.

CRISPR/Cas9 transfections and clonal cell line selection

To delete specific sequences, we co-transfected 600 ng of Cas9-expressing plasmids ("PX330-NoGuide"), 300 ng of a pool of sgRNA-expressing plasmids ("pZB-Sg3"), and 600 ng of a plasmid expressing EGFP and a puromycin selectable marker from a CAG promoter (pS-pp7-GFPiP). To create PX330-NoGuide, we modified PX330 (gift from Feng Zhang, Addgene plasmid #44230) (*45*) to remove the sgRNA expression cassette. To generate pZB-Sg3, we cloned a human U6 promoter and optimized sgRNA scaffold sequence (*33*) into a minimal vector with an ampicillin-selectable marker and a ColE1 replication origin. We transfected batches of 250,000 human cancer cells using the Neon Transfection System (Invitrogen), using 3 pulses of 10 milliseconds at 1450 V and plated them into a 96-well plate in 200 μl media. As an internal control for each set of transfections, we performed a transfection using a pool of 4 sgRNAs with no predicted target sites in the human genome. To knock in polymorphic tags into the *MYC* locus, we included 200 ng of ssDNA oligo in the transfection.

We verified efficient transfection by examining GFP expression after 24 hours. To select for transfected cells, we replaced the media 24 hours after transfection with 200 μl media + 4 μg/ml puromycin. One day later, we split the cells into a 6-well plate with 2 ml of 4 μg/ml puromycin. One day later, we replaced the media with 2 ml of media with no puromycin. We allowed cells to grow for 7-8 days, replacing the media every 2-3 days. Once the cells could be reliably counted, we plated 8 96-well round-bottom plates at a dilution of 0.4 cells/well. We grew these plates in 200 ul of 20% FBS media, doing partial media changes every 3-4 days, for 12-16 days. Clonal cell lines were split into multiple copies and grown for 2-14 days before harvesting for biological replicates. We harvested cells for DNA and RNA extraction by removing most of the media and adding 3.5× volume Buffer RLT (Qiagen).

Genotyping deletion clones by PCR and sequencing

To genotype K562 clones, we isolated genomic DNA using Silane beads. For genotyping MYC-Tag insertion clones (Figure A-7A), we performed PCR using primers surrounding the site followed by a second round of PCR to add a different barcode to each sample and sequenced the amplicons on an Illumina MiSeq (Illumina, San Diego, CA).

For genotyping deletion clones, we performed a first round of PCR using primers spanning the deleted region (Figure A-7B) and examined this PCR product using gel electrophoresis. Both wild-type and deletion-sized bands were visible and were used to prioritize clones for further analysis. We next performed a second nested PCR on this product to add sequencing tags and clone-specific barcodes for high-throughput sequencing. We sequenced these products to span the deletion junction; the number of unique amplicons in each clone was used to determine the number of deleted alleles. (This number is technically a lower bound, because in rare cases multiple alleles could be deleted and repaired in the same fashion). Finally, we counter-screened deletion clones for inversions, which can occur when Cas9-mediated cuts occur on both sides of the region, but the cuts are repaired with an inversion of the intervening sequence. We sought to eliminate clones that showed evidence of inversions, which could confound later analysis. For e2, we used primers spanning one side of the intended junction and eliminated clones that showed evidence of an amplicon corresponding to an inverted sequence. For e3 and e4, we were unable to obtain satisfactory PCR primers and so used a restriction digest approach that could distinguish whether the internal sequence was inverted or not. For e3, we digested PCR amplicons with AvrII and PsiI; for e4, we digested with NdeI and BglII (all enzymes from NEB).

Measuring allele-specific MYC and PVT1 expression in deletion clones

We designed and validated ddPCR assays to measure the allele-specific expression of MYC and PVT1. We first cloned the polymorphic regions of MYC and PVT1 from K562-MYC-Tag using the ddPCR-MYCIntron Fwd/Rev and ddPCR-PVT1 Fwd/Rev PCR primers to generate separate plasmid vectors containing each allele of each amplicon. We generated synthetic standard curves by mixing these vectors in specified ratios: 100:0, 90:10, 50:50, 10:90, and 0:100. Each standard curve was generated and quantified in duplicate to confirm that the assays were specific and quantitative.

To perform the ddPCR assay, each 20μl reaction contained 1X ddPCR Supermix for Probes - no dUTP (BioRad, Hercules, CA), 450 nM each of forward and reverse primer, and 500 nM probe. To measure the relative expression of the 3 MYC alleles (Figure A-7C), we used MYCIntron Fwd and Rev along with a FAM-conjugated CTAA or ATCG probe and a HEX-conjugated CCCG probe in two separate assays, then merged the results by comparing to the constant CCCG probe. To measure the relative expression of the 2 PVT1 polymorphisms (Figure A-7C), we used PVT1 Fwd and Rev and probes against T and C alleles in a single assay. Probes were purchased as Custom ZEN Double-Quenched Probes (IDT). Following droplet generation on a QX200 droplet generator (BioRad), we performed 40 cycles of PCR with a 10 minute 55°C combined and melting extension step. We counted droplets using the QX200 Droplet Reader (BioRad) and determined allele specific expression by the ratio of FAM and HEX positive droplets.

To measure the allele-specific expression of each deletion clone, we generated cDNA from cells as described above and performed ddPCR using 1000 cell-equivalents of cDNA for *MYC* and 100 for *PVT1*. We measured each clone using 2 or 3 technical replicates and averaged the ratios between these measurements for further analysis.

Analysis of allele-specific expression data for deletion clones

To analyze the allele-specific ddPCR data for the deletion clones, we first inferred the phasing of the deletions relative to the polymorphic tags in *MYC*. We identified known polymorphisms near the deleted enhancers that would allow us to phase the deletions by examining DNA sequencing experiments from multiple types of ENCODE experiments (*e.g.*, ChIP-Seq, DHS sequencing). We identified rs67423398 (C/T/T in triploid K562 cells) just outside of the sgRNAs designed at e2 (Figure A-7B), allowing us to directly genotype the deletion bands by amplicon sequencing. For e3 and e4, there were no SNPs in the vicinity of the deletions themselves, but, because each acts as a promoter for *PVT1*, we were able to use a SNP in a downstream *PVT1* exon (rs11604, T/C/C in K562 cells) that allowed us to determine the allele of the deletions by examining which allele of PVT1 RNA was decreased (Figure A-7C). Accordingly, for each e2 clone we performed amplicon sequencing as described in the previous section and determined on which allele(s) the deletion occurred, and for each e3 and e4 clone we performed ddPCR to read out the allele-specific RNA expression of *PVT1*. This allowed us to determine whether the deletion occurred on the unique allele (C for rs67423398 or T for rs11604, C-T) or the ambiguous allele (T for rs67423398 or C for rs11604).

We next phased these polymorphisms based on the unique allele to the polymorphic tags in *MYC*. To do so, we first examined clones that carried deletions on the unique allele and examined their allele-specific expression of *MYC*. For e2, for example, we had 6 independent clones carrying such deletions, and these showed a consistent decrease in *MYC* expression on the CTAA allele (*e.g.*, Figure A-7D). We similarly linked the PVT1 unique allele to CTAA (Figure A-7C). By this strategy, we were able to phase some of the deletions to a unique *MYC* polymorphism (CTAA-C-T allele, Figure A-7C), and the remaining deletions to one of the other two alleles.

For each clone, we then calculated the change in expression of each *MYC* allele relative to 26 wild-type control clones. We first calculated the average expression of each allele in the control clones, which was approximately balanced (31% CTAA, 39% ATCG, 30% CCCG, Figure A-7D). For each clone, we compared the allelic expression fraction to the control clones to determine a fold-change for each allele. We then normalized these fold-changes to maximum of the 3 alleles, assuming that this represents a wild-type allele (*e.g.*, Figure A-7D, right), and termed this the "normalized allele expression". We performed a similar computation on each wild-type clone. Finally, we compared the normalized allele expression between wild-type and deletion clones. For the unique allele (CTAA-C-T), we directly used the *MYC* normalized allele expression. For the remaining alleles (ATCG-T-C and CCCG-T-C), we chose the one of the two alleles with the lowest normalized allele expression, assuming that this was the deletion allele, and similarly generated a distribution of control values by performing a similar procedure on wild-type clones. We combined these comparisons across alleles and compared deletion to control clones using a Wilcoxon rank sum test (Figure A-7E).

Comparison to previous enhancer-promoter predictions

Given our functional mapping of enhancers that regulate *MYC*, we compared our list of true *MYC* enhancers to existing methods for predicting or inferring enhancer-promoter connections. We found that none of these strategies specifically identified more than 2 of the 7 *MYC* enhancers and correctly distinguished the 2 *GATA1* enhancers from neighboring elements that do not affect *GATA1* expression. We describe each of these approaches below.

1. One commonly used strategy for connecting enhancers with target promoters is to assign an enhancer to its nearest gene. It is clear that this does not accurately capture the complexity of enhancer-promoter connections (*5*), but lacking clear alternatives this approach is frequently used to assess which gene an enhancer might regulate. For *GATA1*, this approach does not accurately capture how both e-GATA1 and e-HDAC6, which are closest to *GATA1* and *HDAC6*, respectively, in fact regulate both genes. For *MYC*, e1-e4 would be assigned as regulators of *PVT1*, while e5-e7 would be assigned to the *CCDC26* pseudogene.

Several methods for predicting enhancer-promoter connections are based on correlations in chromatin state across cell types.

2. One such method is based on correlation in histone modification profiles between candidate enhancer-promoter pairs within 125 kb across nine cell types, including K562 cells (*46*). Because of this distance restriction, this method does not make any predictions for *MYC*. For *GATA1*, this strategy misses both e-GATA1 and e-HDAC6, and makes dozens of incorrect predictions.
3. A second method based solely on correlation predicts enhancer-promoter pairs using correlation in DHS for all candidate pairs within 500 kb of one another across 125 cell types, including K562 cells (*19*). For *GATA1*, this method correctly identifies both e-GATA1 and e-HDAC6 but also incorrectly assigns two additional distal enhancers in the regions tested in our screen. For *MYC*, this approach correctly identifies only one of the K562 enhancers (e4) and makes dozens of other predictions that do not overlap e1-e7. (The published catalog from this study does not report which cell type each prediction refers to, and thus some of

these additional predicted enhancers may represent regions that regulate one of the target genes in another cell type.)

4. A third correlation-based method (PreSTIGE) predicts enhancer-promoter pairs by pairing cell-type-specific H3K4me1 signals with cell-type specific gene expression across 12 cell types, using a 100 kb distance plus a subset of CTCF sites to set domain boundaries (*47*). In the *GATA1* locus, PreSTIGE reports that 29 kb of the 74 kb covered by our screen is an enhancer for *GATA1*, including both e-GATA1 and e-HDAC6 but incorrectly reporting many kilobases of additional sequence. In the *MYC* locus, PreSTIGE predicts a single region to be an enhancer; this region does not correspond to any of the enhancers we identify.

In addition to methods based on correlations in chromatin state across cell types, a second category of approaches for inferring enhancer-promoter functional connections is based on measuring their physical interactions with methods based on chromosome conformation capture. Physical contacts between enhancers and promoters correlate with gene activation (*6, 17, 18, 25*)(*1*, *6*, *46*, *47*), and in a few cases increasing the frequency of enhancer-promoter contact has been shown to activate gene expression (*26, 27*). However, long-distance chromatin loops can form without regulatory effects on gene expression (*e.g.*, when a promoter forms a loop with a region that is not an enhancer), and the abilities of various features of chromosome conformation data to predict functional interactions remains unclear (*6*). Accordingly, we examined several features previously noted to correlate with enhancer-promoter connections to determine if they might correctly identify enhancers in the *MYC* locus.

5. We first examined loops as defined by *in situ* Hi-C (*25*). In a Hi-C map of K562 cells at 5 kb resolution, five focal loops involving the *MYC* promoter were reported. Of the five, one corresponds to the long-range loop with e6/e7, one corresponds to NS1, and the other three correspond to CTCF-bound sites that do not overlap *MYC* enhancers. Thus, at the reported

significance thresholds and with the available resolution, these calls do not correspond with the enhancers that regulate *MYC*. Nonetheless, Hi-C data shows that these sites frequently contact *MYC* (**Figure 2-2A**), and higher resolution maps may allow identification of focal loops to these sites. Regardless of the specific loop calls, we find that incorporating this information into our heuristic helps to rank enhancers likely to regulate *MYC* (see Chapter 2).

6. RNA Pol II ChIA-PET has been proposed as a proximity interaction method that enriches for enhancer-promoter interactions (*37*). ChIA-PET in K562 cells (wgEncodeGisChiaPetK562Pol2InteractionsRep1) identifies many interactions between *MYC* and sites throughout the adjacent contact domain (Figure 2-2A). Notably, these do include all 7 of the *MYC* enhancers in K562, but also include dozens of other sites with equal or higher interaction frequencies (Figure 2-2A). Furthermore, ChIA-PET in K562 cells does not detect interactions between *GATA1*/*HDAC6* and either of their enhancers.
7. Various methods developed to predict enhancer-promoter interactions have been developed and trained based on interactions identified in chromosome conformation capture experiments. Consistent with the poor positive predictive value of chromosome conformation capture data as described above, methods trained on this data (*e.g.*, (*48, 49*)) also do not correctly identify *MYC* or *GATA1* enhancers.

Together, these observations highlight the importance of direct functional mapping of regulatory elements. Furthermore, they underscore the opportunity for new models that integrate these two classes of approaches based on chromatin state and proximity interactions in the context of appropriate training data generated through CRISPRi-based mapping of regulatory elements.

Calculating predicted impact of MYC enhancers in K562 cells

To rank the relative importance of putative activating elements near *MYC* in K562 cells, we first created a list of putative regulatory elements in the locus. We downloaded DHS peak calls from ENCODE (narrowPeak files corresponded to both replicates in K562 cells), expanded these peaks by 500 bp, and merged overlapping peaks. For each of these merged peaks, we calculated normalized read count (reads per million, RPM; *not* normalized to length of the element) from H3K27ac and DHS measurements in K562 cells, and retained windows in the top 50% percentile with respect to H3K27ac signal, yielding 93 putative regulatory elements. For each element, we calculated the normalized contact frequency to the *MYC* promoter by consulting KL-normalized observed contact matrices at 5-kb resolution generated by *in situ* Hi-C (*25*). We calculated relative impact by the following formula: Predicted impact = log_2(H3K27ac RPM × DHS RPM × Hi-C contact × Hi-C contact), thereby weighting "activity" and "proximity" approximately equally. Each element was ranked according to this score. In Figure 2-2E, peaks overlapping the MYC enhancers were colored red and plotted versus their CRISPRi score, defined by the maximum CRISPRi score in a window overlapping the element.

To compare the performance of this heuristic with simpler models, we calculated rankings based on H3K27ac ChIP-Seq RPM only, DHS RPM only, and Hi-C contacts only for the same set of 93 putative regulatory elements (Figure A-8A). We note that because these 93 elements were selected based on DHS and H3K27ac signal as described above, this may be an optimistic estimate of the value of each dataset alone.

Additional experimental data will be required to further refine this model and determine whether it is applicable to different gene loci.

Calculating enhancer ranks across cell types

To expand this approach across additional cell types, we downloaded DHS and H3K27ac ChIP-seq data for diverse cell lines and primary tissues from the Roadmap Epigenomics Project (*50*), ENCODE (*36*), and others (*51, 52*). While these data are available across a wide range of cell types (235 samples total), proximity interactions maps are available in a very limited number of cell types. Accordingly, we explored to what extent the topological architecture of the *MYC* locus changes across 7 human cell types previously mapped using *in situ* Hi-C (*25, 53*). We found that key features of the proximity contacts of the *MYC* promoter appeared consistent across cell types, including the long-range contacts to the edges of the topological domain as well as several distinct peaks within these domains (Figure A-8C). These cell-type invariant long-range loops typically corresponded to sites bound by CTCF across multiple cell types, consistent with previous reports (*25*). Beyond these long-range loops, the quantitative interactions of the *MYC* promoter did change somewhat across different cell types, with elevated contact frequency coinciding with the presence of strong H3K27ac occupancy in a given cell type. To capture the features consistent across cell types, we generated a generic proximity profile for the *MYC* locus by averaging the proximity interactions across these 7 cell types, normalizing the absolute magnitude of interactions in each cell type by the signal at the *MYC* promoter itself. This generic profile accurately captured the cell-invariant long-range interactions (Figure A-8C), providing a reasonable template for weighting the contributions of different enhancers in the *MYC* locus across cell types.

To rank elements across the entire domain, we calculated the predicted impact score as described above in 400-bp windows tiled every 100-bp across chr8:127000000-131500000. DHS and H3K27ac were not always available for each of the 235 different samples — accordingly, we used both datasets where available, or calculated an alternative ranking using one or the other dataset (*e.g.*, DHS or H3K27ac normalized read count × normalized Hi-C signal). Given the varying patterns of DHS and H3K27ac signal around a regulatory element (DHS is strong at the center of the element while H3K27ac is depleted in the nucleosome-free region but strong just outside), we smoothed these scores at 2-kb resolution to better compare models generated from DHS or H3K27ac alone. To collapse neighboring windows with strong scores yet retain resolution for the strongest local maximum (*e.g.*, corresponding to the center of the regulatory element), we removed windows that had an overlapping window with a higher score. Finally, we assigned a rank to these remaining windows and focused on the top 10 elements in each cell type.

Analysis of enhancers known to regulate MYC

We curated a list of enhancers that have been shown to regulate *MYC* in their endogenous genomic contexts. (i) An enhancer implicated in *MYC* regulation in the context of colorectal cancer ("Myc-335") was identified based on an association rs6983267 and risk for colorectal cancer (*54, 55*). Genetic knockout of this enhancer in mice leads to an ~40% reduction in Myc RNA expression in the colon, and confers resistance to intestinal tumorigenesis in an APC-/- background (*56*). (ii) An enhancer implicated in *MYC* regulation in the context of lung adenocarcinoma (LUAD) was identified based on a focal amplification of a noncoding region in multiple primary LUAD tumors (*57*). Genetic knockout of this enhancer in a LUAD cell line led to a ~30% reduction in *MYC* expression (*57*) and defects in cellular proliferation. (iii) An enhancer implicated in T-ALL was identified based on focal amplifications of a noncoding region ~1.47 Mb downstream of *MYC* (*58*). This enhancer contacts the *MYC* promoter as assayed by chromosome conformation capture, and a mouse knockout of this element leads to defects in thymocyte development and improved survival in the context of NOTCH1-induced leukemogenesis (*58, 59*). (iv) An enhancer implicated in AML was identified on the basis of strong occupancy by Brg1 in a murine leukemia cell line, and is focally amplified in ~3% of human AMLs. This enhancer (E3) was shown to loop to the *MYC* promoter, and knockdown of Brg1 led to dramatic loss of *MYC* expression (*60*). We extracted coordinates from these previous studies and overlapped these coordinates with highly ranked enhancers in relevant cell types (Figure 2-3B).

Analysis of GWAS variants near MYC

We downloaded a list of variants associated with human phenotypes from the GWAS Catalog at EBI (https://www.ebi.ac.uk/gwas/, accessed May 11, 2016). 121 associations are reported in chr8:127900000-131000000. We used HaploReg v4.1 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php, accessed May 11, 2016) (*61*) to identify SNPs linked to the GWAS index SNP with r^2 >= 0.8 in the European population. The black boxes in Figure 2-3C represent the span of all such SNPs for each variant, collapsed by phenotype to yield 66 unique associations between a human disease or trait and a genetic haplotype. We highlight three examples where these SNPs overlap elements predicted to regulate *MYC*. (i) A SNP linked to increased risk of Hodgkin's lymphoma, which has previously been noted to overlap with B-cell specific H3K27ac signals (*51*), overlaps an element that our heuristic predicts to be quantitatively among the most important for regulating *MYC* in B cell lymphoma cells (Figure 2-3D). (ii) A SNP associated with bladder cancer risk is located in a conserved DHS element active in multiple gastrointestinal tissues, and thus may regulate *MYC* in bladder epithelial cells, for which chromatin data is not available (Figure 2-3D). (iii) A SNP associated with height overlaps a glucocorticoid receptor motif in a conserved H3K27ac-marked element active only in chondrocytes (Figure 2-3D). (DHS data from chondrocytes was not available). Although this SNP is located >1.9 Mb from *MYC*, it resides at the anchor of the long-range chromatin loop near e7 (Figure 2-2A), suggesting that this SNP may affect height by altering the regulation of *MYC* in a chondrocyterelated cell type. Dozens of other predicted regulatory elements overlap disease-associated genetic variants near *MYC*.

Software for data analysis and graphical plots

We used the following software for data analysis and graphical plots: R Bioconductor (version 3.0) (*62*), Gviz (version 1.10.11), gplots (version 2.17.0), GenomicRanges (version 1.18.4) (*63*), rtracklayer (version 1.26.3) (*64*), BEDTools (*65*), Integrative Genomics Viewer (version 2.3.26) (*66*), Pandas (version 0.12.0), Matplotlib (version 1.3.0), Biopython (version 1.61) (*67*), and SciPy (version 0.12.0).

Genome build

All coordinates are reported in human genome build hg19.

Supplemental Figures

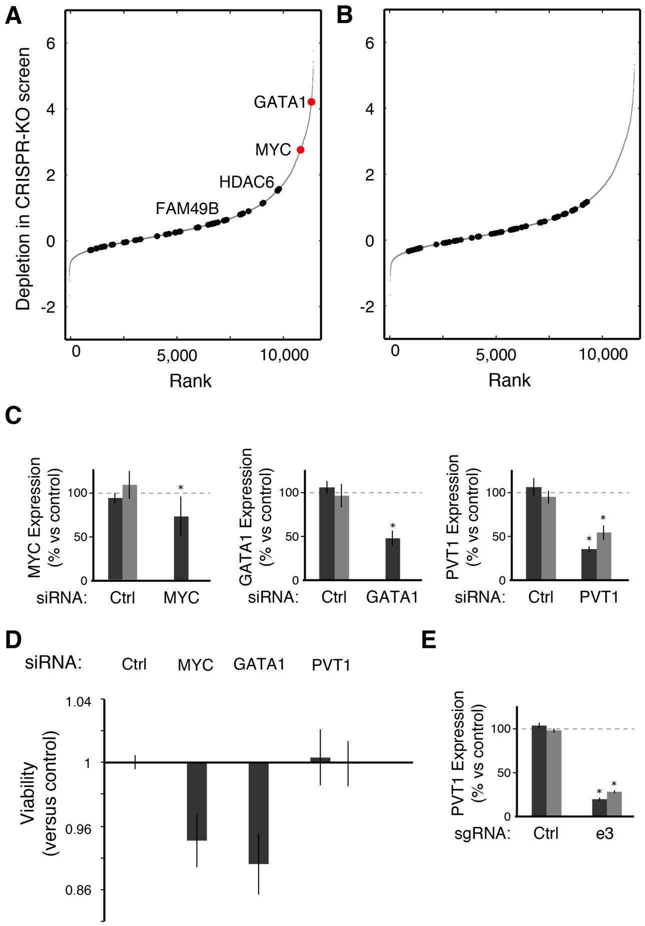


Figure A-1. GATA1 and MYC are encoded far from other genes that strongly affect proliferation in K562 cells. (A) Gray: Depletion (–$\log_2$ fold-change after 14 population doublings) in a previous genome-wide CRISPR knockout screen of all genes expressed in K562 cells (*28*). Higher scores denote stronger effect on proliferation. Black: genes within 500 Kb or in the same topological domain as *MYC* or *GATA1* (highlighted in red). **(B)** Same for the three tiled negative-control regions. **(C)** Knockdown efficiency for siRNAs targeting MYC, GATA1, and PVT1, as assayed by qPCR compared to siRNAs without an RNA target (Ctrl). Gray bars: two different siRNAs for Ctrl and PVT1. Error bars: 95% confidence intervals (CI) for the mean of four independent transfections. *: $p < 0.05$ in T-test versus negative controls. **(D)** Relative viability of cells in a competitive growth assay (gamma). GFP-expressing cells were transfected with siRNAs against GATA1, MYC, PVT1, or siRNAs without a genomic target (Ctrl) and were mixed with RFP-expressing cells transfected with a Ctrl siRNA and grown for four days before counting. Error bars: 95% confidence intervals (CI) for the mean of 4 independent transfections. We tested two different sgRNAs for PVT1. *: $p < 0.05$ in T-test versus negative controls. **(E)** qPCR for PVT1 RNA in cells expressing sgRNAs targeting a TSS of PVT1 (e3) or sgRNAs without a genomic target (Ctrl). KRAB-dCas9 expression was activated with doxycycline for 24 hours before measurement. Gray bars: two different sgRNAs per target. Error bars: 95% confidence intervals (CI) for the mean of 3 independent infections. *: $p < 0.05$ in T-test versus negative controls.

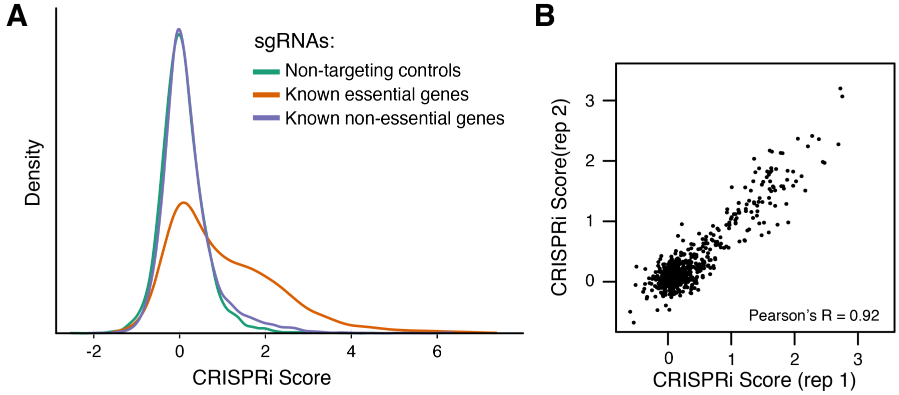


Figure A-2. CRISPRi screen reproducibly depletes sgRNAs targeting promoters of essential genes.

(A) Distributions of CRISPRi scores for sgRNAs targeting the promoters of genes previously identified as essential or non-essential based on a genome-wide CRISPR knockout screen (*28*) and for sgRNAs with no genomic target (control sequences). A higher CRISPRi score indicates stronger depletion over the course of the screen.

(B) Average CRISPRi scores for 600 protein-coding gene promoters in replicate screens.

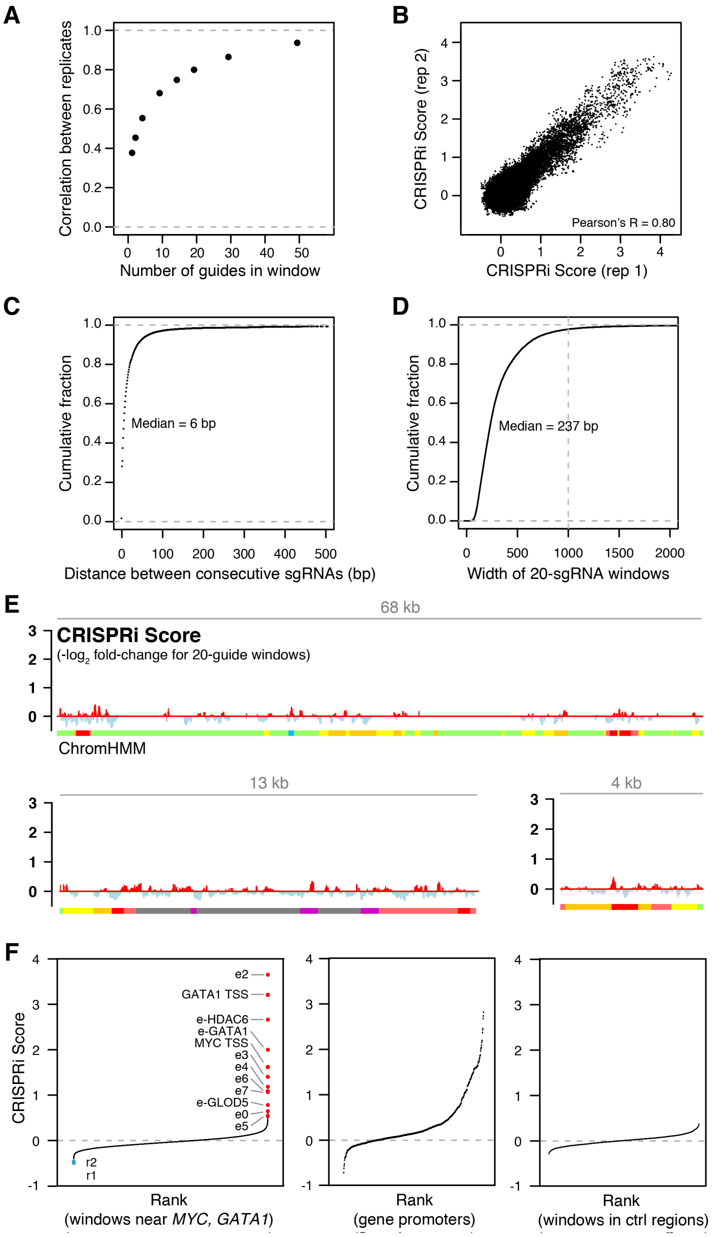


Figure A-3. Sliding window approach for analyzing CRISPRi screens.

(A) Pearson correlation between the two replicate screens for CRISPRi scores averaged across windows of different sizes (2, 3, 5, 10, 15, 20, 30, or 50 consecutive sgRNAs).

(B) CRISPRi scores for all windows of 20 consecutive guides in the replicate screens.

(C) Cumulative density plot of the distance between consecutive sgRNAs. Distribution extends beyond the *x*-axis limits.

(D) Cumulative density plot for the span of 20-sgRNA windows. Windows spanning greater than 1 kb were not considered. Distribution extends beyond the *x*-axis limits.

(E) CRISPRi scores in 20-sgRNA windows for three negative-control regions that are located far from known essential genes (see Methods). These regions show a lack of strong signal as compared with the *GATA1* and *MYC* loci and were used to calculate an empirical false discovery rate for the CRISPRi score.

(F) Gray: CRISPRi score in 20-sgRNA windows for tiled *MYC* and *GATA1* regions (left, ~60,000 windows), the TSSs of protein coding genes from across a range of essentiality (middle, ~600 genes), or tiling regions far from any essential gene (right, ~5,000 windows). Red dots: Most strongly depleted window within identified enhancers and TSSs (other windows nearby, which are also strongly depleted, are not shown for visual clarity). Blue: Most strongly enriched window within putative repressive elements.

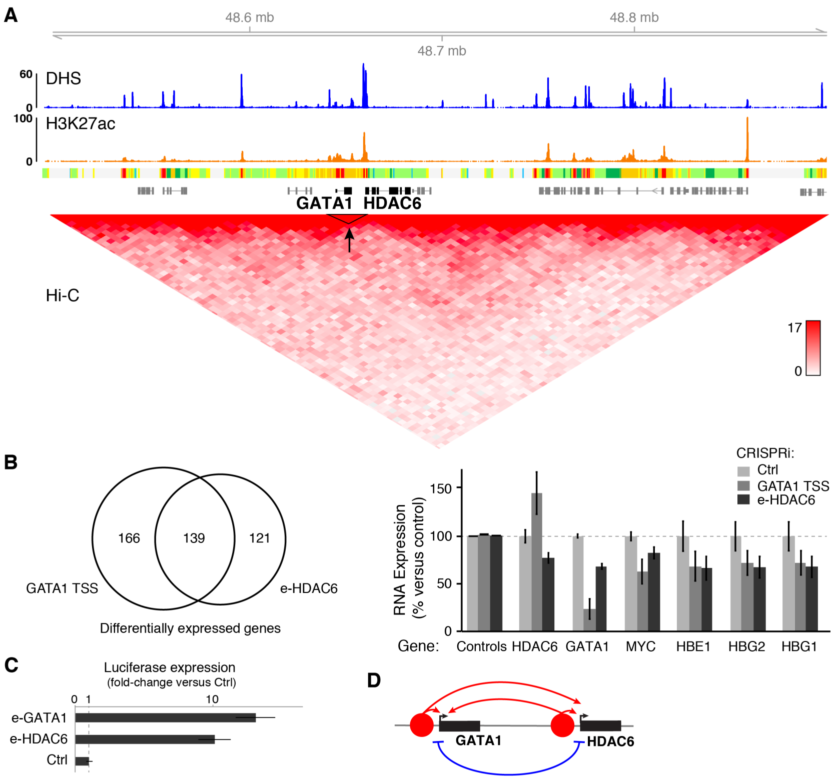


Figure A-4. Characterization of enhancers at the GATA1 locus.

(A) Chromatin state and chromosome conformation in the ~400-Kb topological domain containing *GATA1* and *HDAC6*. K562 DHS, ChIP-Seq data, and chromatin state classifications (ChromHMM) are from ENCODE (*36*) (see Methods). Contact frequency matrix is derived from *in situ* Hi-C maps at 5-kb resolution in K562 cells (KL-normalized observed matrix) (*25*). Black triangle and arrow mark the region of interactions between enhancers (e-GATA1 and e-HDAC6) and the promoters of GATA1 and HDAC6. **(B)** Effects of inhibiting *GATA1* TSS or e-HDAC6 on gene expression of downstream GATA1 target genes. Venn diagram represents differentially expressed genes from RNA sequencing of stable lines expressing the listed sgRNA relative to cells containing negative control sgRNAs (Ctrl). Hypergeometric *p*-value of overlap $<10^{-163}$. Bar plot shows that known target genes of the GATA1 transcription factor (MYC, HBE1, HBG1, and HBG2) (*68-70*) are differentially expressed upon inhibition of e-HDAC6. KRAB-dCas9 expression was activated for 24 hours before measurement. Error bars: 95% CI for the mean of 2 sgRNAs with 3 independently derived stable lines each. Controls: all other expressed genes. **(C)** Expression of firefly luciferase from plasmids containing each enhancer located 2 kb upstream of a *MYC* promoter fragment. Data is normalized to a random sequence of similar size (Ctrl) and to the internal *Renilla* luciferase control (see Methods). Error bars: 95% CI for the mean of 3 independent transfections. **(D)** Regulatory connections in the *GATA1*/*HDAC6* locus: two enhancers (red) regulate both genes, and the promoters appear to repress one another (blue), perhaps by competing for activating signals from the enhancers.

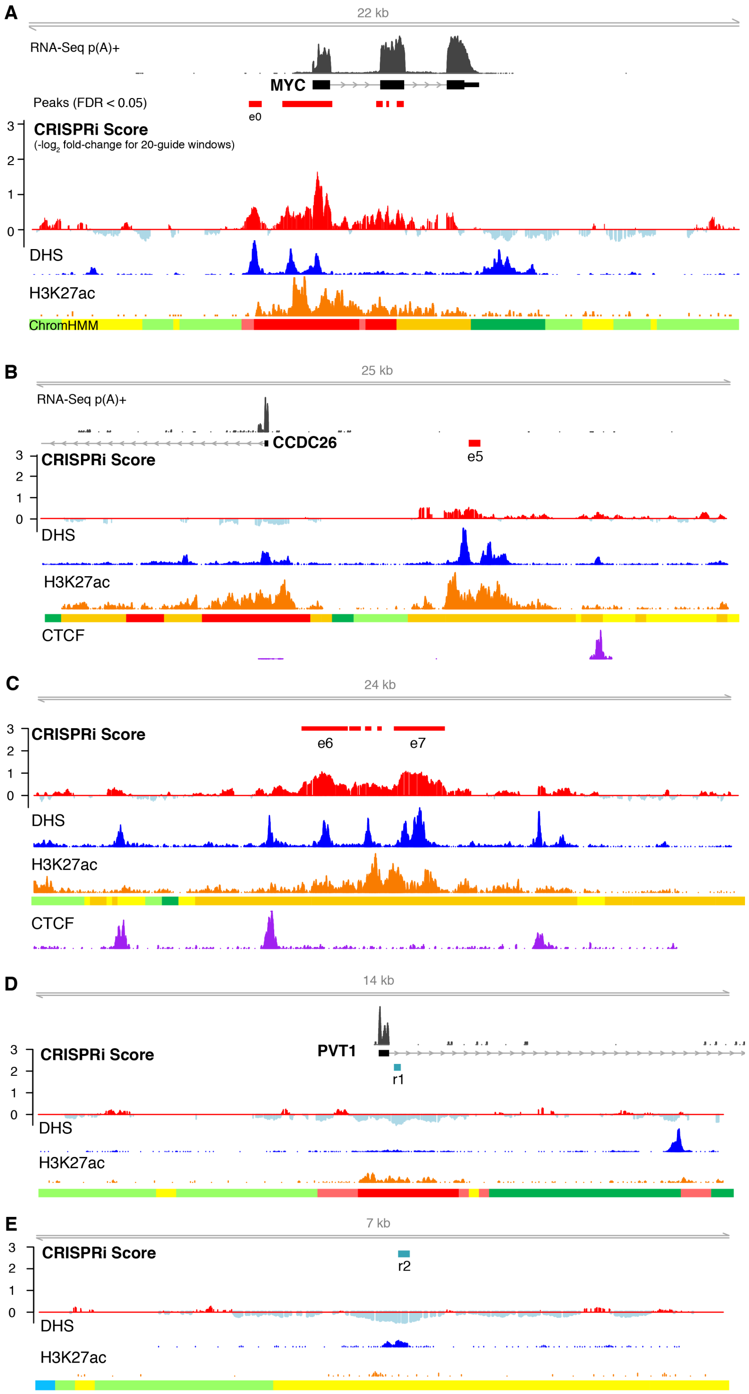


Figure A-5. Regulatory elements at MYC and downstream enhancers. **(A)** CRISPRi screen results in *MYC* gene locus, showing significant peaks at the *MYC* TSS, at several locations in the gene body, and at a known promoter-proximal regulatory element (e0) (*21*). K562 DHS, RNA-Seq, ChIP-Seq data, and chromatin state classifications (ChromHMM) are from ENCODE (*36*). **(B)** Expanded region around e5 and CCDC26 and **(C)** e6/e7 showing strong CTCF occupancy at DHS sites close to the elements. Each CTCF peak has a motif oriented in the reverse direction (toward *MYC*, not pictured). Note that the promoter of CCDC26 does not score as essential, indicating that its expression is not responsible for the proliferative defects observed upon inhibiting e5 or other enhancers. **(D)** Expanded region around the putative repressive elements r1 and **(E)** r2. r1 corresponds to the promoter of an alternative isoform of PVT1.

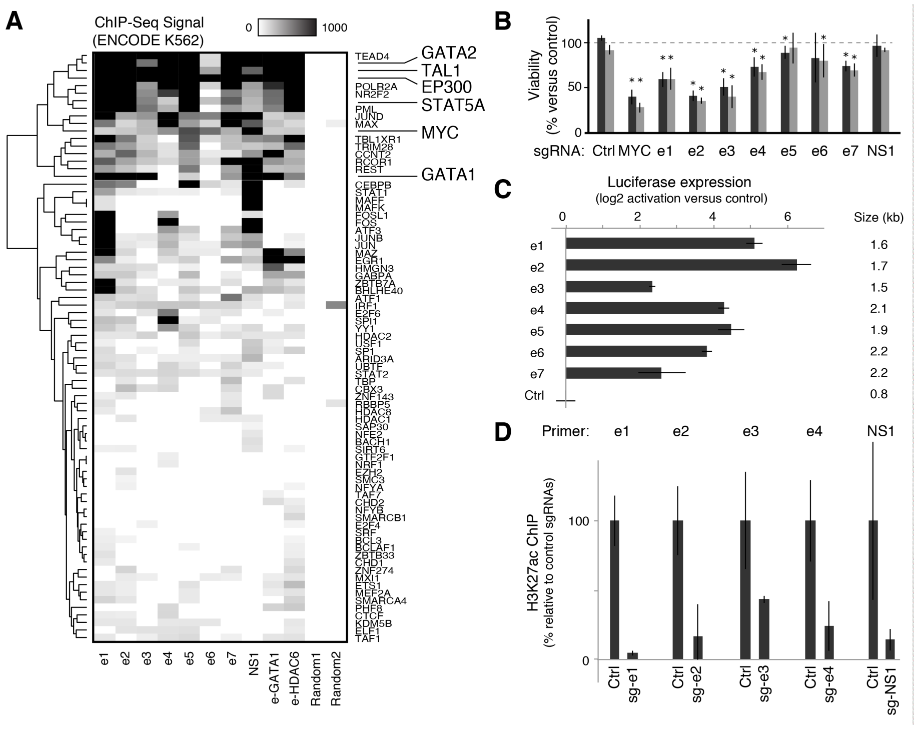


Figure A-6. Characterization of enhancers at the MYC locus.

(A) *GATA1* and *MYC* enhancers bind many activating transcription factors. Transcription factor binding in a 1-kb window centered on each enhancer are shown with their ChIP-Seq signal reported by ENCODE (*36*), which assigns scores to peaks by multiplying the ChIP-seq signal values by a normalization factor calculated as the ratio of the maximum score value (1000) to the ChIP-seq signal value at one standard deviation from the mean, with values exceeding 1000 capped at 1000. For comparison, two random sites near *MYC* are shown. **(B)** Relative viability of cells in a competitive growth assay. Cells expressing the indicated sgRNAs were competed against K562 cells expressing GFP or RFP and grown in doxycycline for 7 days before counting. Gray bars: two different sgRNAs per target. Error bars: 95% CI for the mean of 6 total replicate competition assays using cells from 3 independent infections. *: $p < 0.05$ in T-test versus negative controls. **(C)** Each *MYC* enhancer can activate a reporter gene driven by a *MYC* promoter fragment in a plasmid-based luciferase assay. The size of each enhancer sequence is reported on the right. Ctrl: negative control sequence corresponding to a bacterial kanamycin resistance gene. Error bars: 95% CI for the mean based on three replicate transfections. **(D)** To determine if sgRNAs targeting NS1 successfully affected chromatin state, we performed ChIP for H3K27ac in cells expressing individual sgRNAs targeting e1, e2, e3, e4, or NS1, as well as two non-targeting control sgRNAs (see Methods). We measured ChIP enrichment by qPCR for 5 positive control loci, 3 negative control loci, and the locus targeted by the sgRNA (see Methods). Bars represent enrichment of the indicated locus normalized to the non-targeting control sgRNAs. Error bars: 95% CI for the mean for 5 (Ctrl) or 3 (others) biological replicates.

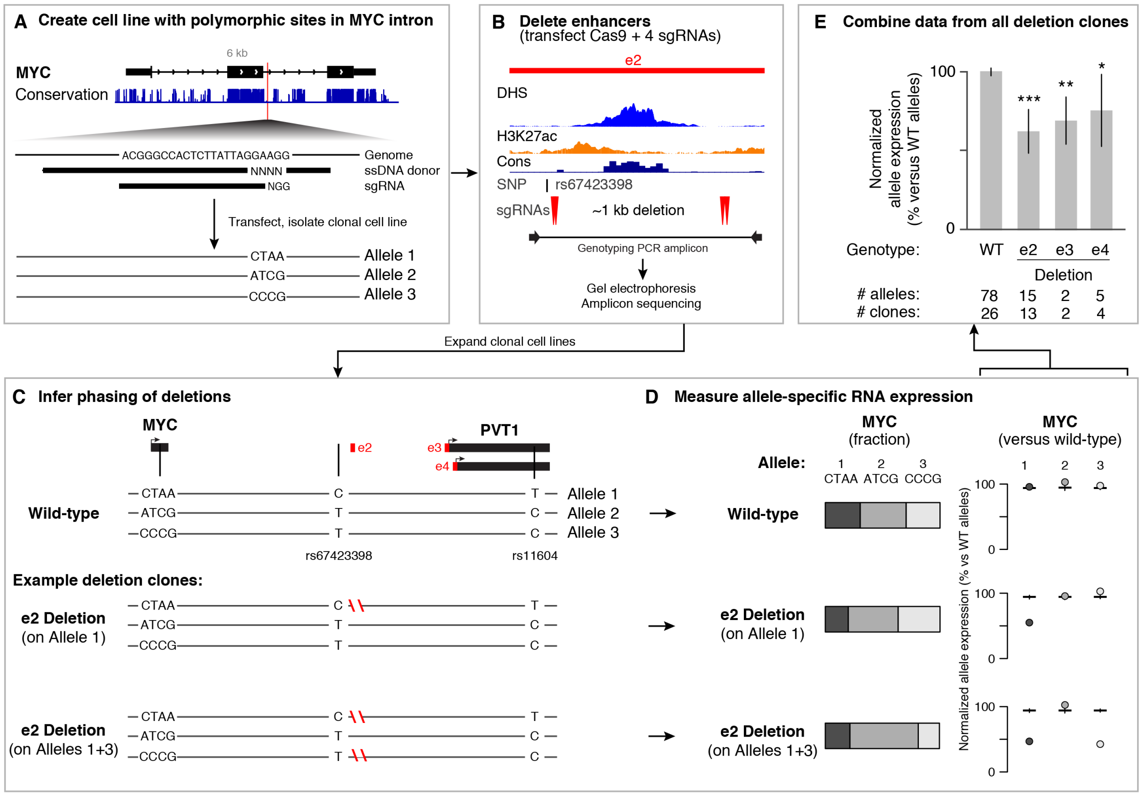


Figure A-7. Genetic deletions of enhancers in the MYC locus.

(A) Strategy for generating a cell line containing polymorphic sites on each allele of *MYC*. We used CRISPR/Cas9 to knock in a random 4-mer sequence into an intronic site in the *MYC* locus that was not conserved across mammals (red line). We co-transfected a plasmid expressing Cas9, a ssDNA oligo donor, and an sgRNA, picked clonal cell lines, genotyped by amplicon sequencing, and isolated a clone with three unique alleles. **(B)** Strategy for deleting enhancers, showing e2 as an example. To delete each enhancer, we designed 4 sgRNAs flanking the DHS peak in the center of each element, two on each side. We co-transfected these 4 sgRNAs and isolated clones containing deletions on 1 or 2 of the 3 alleles. The rs67423398 SNP was contained in the genotyping PCR amplicon and was used to determine which allele of e2 was deleted. **(C)** Overview of sites relevant to enhancer deletions in the *MYC* locus, including inferred phasing of polymorphic sites. Bottom: Genotypes for example deletion clones. **(D)** Allele-specific RNA measurements for representative clones. For each clone, we determined the fraction of RNA molecules carrying each of the *MYC* alleles using ddPCR (bar plots). We calculated a fold-change for each allele in deletions versus controls and normalized this to the highest of these three values within each clone (see Methods). This yielded the "normalized allele expression" (right). Dots: values for one clone. Horizontal bars: mean with 95% confidence interval for 26 wild-type clones. **(E)** Deletions of e2, e3, and e4 led to a 30-40% decrease in the expression of *MYC* on the corresponding allele compared to wild-type alleles in the same cells. We compared normalized allele expression values between wild-type and deletion alleles using a Wilcoxon rank-sum test. *: $P < 0.05$. **: $P < 0.01$. ***: $P < 10^{-4}$.

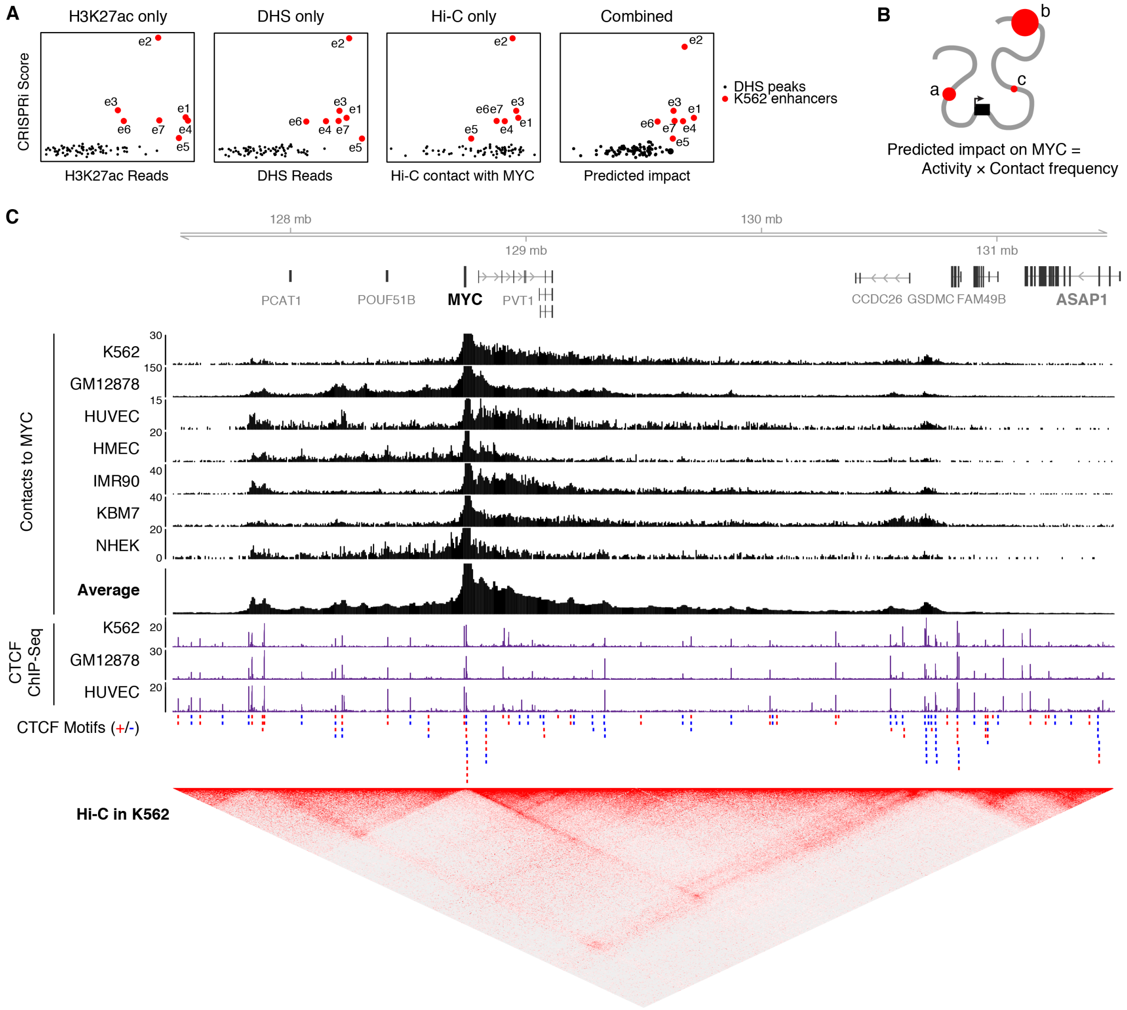


Figure A-8. Heuristic model for predicting enhancer function in the MYC locus.

(A) Comparison of models using H3K27ac only, DHS only, Hi-C only, or a combination of all three (Predicted Impact, same as Figure 2-2E). This ranking is applied to 93 elements selected based on DHS and H3K27ac signal (see Methods), and thus provides an optimistic estimate of the power of each individual source of information for predicting MYC enhancers. **(B)** Heuristic framework for predicting the relative impact of regulatory elements on *MYC* expression. Impact depends on activity (estimated by quantitative H3K27ac and DHS signal, represented by size of red dot) and the frequency with which it contacts the *MYC* promoter (estimated based on Hi-C, represented by distance from gene). For the three example enhancers, their relative impact would be a = b > c. **(C)** Comparison of Hi-C and CTCF ChIP-Seq signal in the *MYC* locus across cell types. Contact frequency with the *MYC* promoter is derived from *in situ* KL-normalized Hi-C maps at 5-kb resolution across 7 cell types (*25*). *Y*-axis differs between cell types according to the depth of sequencing. The average contact profile used in our enhancer ranking calculations across cell types was created by averaging the normalized contact frequencies from these 7 cell types. CTCF motifs are colored according to their orientation: red = positive strand, blue = negative strand.

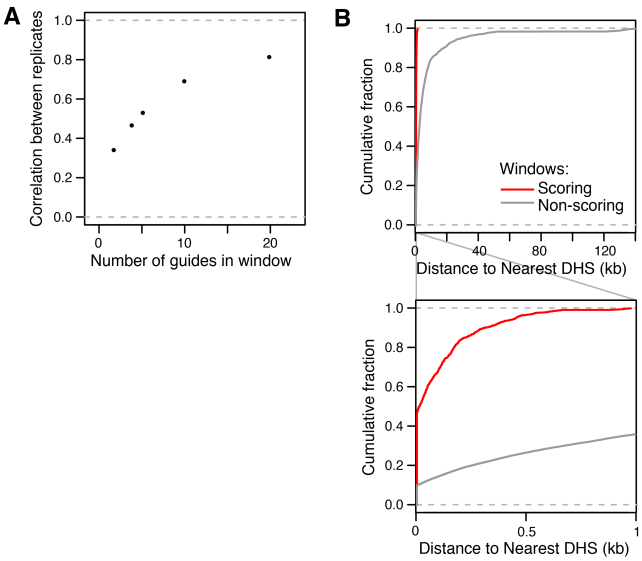


Figure A-9. Design of new CRISPRi libraries

(A) Pearson correlation between the two replicate screens for CRISPRi scores from windows of different sizes – 2, 4, 5, 10, 20 sgRNAs – downsampled by taking every 10th, 5th, 4th, 2nd, or every sgRNA, respectively. Reducing the density of coverage reduces reproducibility. **(B)** Cumulative density plot of the distance between 20-sgRNA windows and the nearest DHS peak, with the first kb highlighted below. All significantly enriched or depleted windows (Scoring) are less than 1 kb from a DHS peak, compared to <35% of all other windows (Non-scoring).

References

1. N. Rajagopal *et al.*, High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**, 167-174 (2016).

2. M. C. Canver *et al.*, BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192-197 (2015).

3. G. Korkmaz *et al.*, Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192-198 (2016).

4. D. Shlyueva, G. Stampfel, A. Stark, Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-286 (2014).

5. J. van Arensbergen, B. van Steensel, H. J. Bussemaker, In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol* **24**, 695-702 (2014).

6. J. Dekker, T. Misteli, Long-Range Chromatin Interactions. *Cold Spring Harb Perspect Biol* **7**, a019356 (2015).

7. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).

8. O. Shalem *et al.*, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).

9. P. I. Thakore *et al.*, Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature methods* **12**, 1143-1149 (2015).

10. O. Parnas *et al.*, A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**, 675-686 (2015).

11. L. A. Gilbert *et al.*, Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647-661 (2014).

12. Y. Guan *et al.*, Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin Cancer Res* **13**, 5745-5755 (2007).

13. Y. Y. Tseng *et al.*, PVT1 dependence in cancer with MYC copy-number increase. *Nature* **512**, 82-86 (2014).

14. J. M. Engreitz *et al.*, Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452-455 (2016).

15. V. R. Paralkar *et al.*, Unlinking an lncRNA from Its Associated cis Element. *Molecular cell* **62**, 104-110 (2016).

16. S. L. Ameres, P. D. Zamore, Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol* **14**, 475-488 (2013).

17. M. Bulger, M. Groudine, Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339 (2011).

18. B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, W. de Laat, Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell* **10**, 1453-1465 (2002).

19. R. E. Thurman *et al.*, The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

20. M. J. Guertin, A. L. Martins, A. Siepel, J. T. Lis, Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS Genet* **8**, e1002610 (2012).

21. C. D. Arnold *et al.*, Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077 (2013).

22. M. P. Creyghton *et al.*, Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010).

23. S. Bonn *et al.*, Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat Genet* **44**, 148-156 (2012).

24. E. Lieberman-Aiden *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

25. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

26. W. Deng *et al.*, Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244 (2012).

27. W. Deng *et al.*, Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849-860 (2014).

28. T. Wang *et al.*, Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101 (2015).

29. D. Levens, You Don't Muck with MYC. *Genes Cancer* **1**, 547-554 (2010).

30. P. D. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**, 827-832 (2013).

31. M. A. Horlbeck *et al.*, Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife* **5**, (2016).

32. S. Djebali *et al.*, Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

33. B. Chen *et al.*, Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479-1491 (2013).

34. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).

35. Q. B. Li, J.B.; Huang, H; Bickel, P.J., Measuring reproducibility of high-throughput experiments. *arXiv*, (2011).

36. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

37. G. Li *et al.*, Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98 (2012).

38. W. J. Kent *et al.*, The human genome browser at UCSC. *Genome research* **12**, 996-1006 (2002).

39. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).

40. H. Xu *et al.*, Sequence determinants of improved CRISPR sgRNA design. *Genome research* **25**, 1147-1157 (2015).

41. N. E. Sanjana, O. Shalem, F. Zhang, Improved vectors and genome-wide libraries for CRISPR screening. *Nature methods* **11**, 783-784 (2014).

42. J. M. Engreitz *et al.*, RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* **159**, 188-199 (2014).

43. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).

44. M. Garber *et al.*, A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Molecular cell* **47**, 810-822 (2012).

45. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823 (2013).

46. J. Ernst *et al.*, Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49 (2011).

47. O. Corradin *et al.*, Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome research* **24**, 1-13 (2014).

48. B. He, C. Chen, L. Teng, K. Tan, Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci U S A* **111**, E2191-2199 (2014).

49. S. Whalen, R. M. Truty, K. S. Pollard, Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48**, 488-496 (2016).

50. C. Roadmap Epigenomics *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).

51. R. J. Ryan *et al.*, Detection of Enhancer-Associated Rearrangements Reveals Mechanisms of Oncogene Dysregulation in B-cell Lymphoma. *Cancer Discov* **5**, 1058-1071 (2015).

52. J. Huang *et al.*, Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Dev Cell* **36**, 9-23 (2016).

53. J. R. Dixon *et al.*, Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).

54. S. Tuupanen *et al.*, The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41**, 885-890 (2009).

55. J. B. Wright, S. J. Brown, M. D. Cole, Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol* **30**, 1411-1420 (2010).

56. I. K. Sur *et al.*, Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360-1363 (2012).

57. X. Zhang *et al.*, Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet* **48**, 176-182 (2016).

58. D. Herranz *et al.*, A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nat Med* **20**, 1130-1137 (2014).

59. Y. Yashiro-Ohtani *et al.*, Long-range enhancer activity determines Myc sensitivity to Notch inhibitors in T cell leukemia. *Proc Natl Acad Sci U S A* **111**, E4946-4953 (2014).

60. J. Shi *et al.*, Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes & development* **27**, 2648-2662 (2013).

61. L. D. Ward, M. Kellis, HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-934 (2012).

62. R. C. Gentleman *et al.*, Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).

63. M. Lawrence *et al.*, Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118 (2013).

64. M. Lawrence, R. Gentleman, V. Carey, rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841-1842 (2009).

65. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

66. J. T. Robinson *et al.*, Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).

67. P. J. Cock *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).

68. M. Rylski *et al.*, GATA-1-mediated proliferation arrest during erythroid maturation. *Mol Cell Biol* **23**, 5031-5042 (2003).

69. Y. Woon Kim, S. Kim, C. Geun Kim, A. Kim, 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**, 6944-6955 (2011).

70. Q. Gong, A. Dean, Enhancer-dependent transcription of the epsilon-globin promoter requires promoter-bound GATA-1 and enhancer-bound AP-1/NF-E2. *Mol Cell Biol* **13**, 911-917 (1993).

Appendix B. Supplemental Material for Chapter 2

Supplemental Notes

Note B1. 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 (10 out of 98), these two situations also account for most of the DE-G pairs with repressive effects (20 out of 24).

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. We noted that 24 of the 98 significant DE-G pairs (20%) in our data involve elements that, upon CRISPRi inhibition, led to *increased* expression (average +15%) of a nearby gene. However, these effects do not appear to result from *cis*-acting "repressors"; 6 of these 19 unique elements have activating effects on at least one other nearby gene (Figure B-12A). 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 (Figure B-12B-D).

CTCF Sites

The second situtation involves elements bound by CTCF, a protein that affects gene regulation by shaping 3D genomic architecture (*1*) (34% of tested DHS elements bind CTCF). 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} vs. H3K27ac^{low} elements, we found clear differences between the two classes (Figure B-11). H3K27ac^{high} CTCF elements had larger effects on gene expression (average 32%) and were far more often activating rather than repressive (23 vs. 3), consistent with these elements primarily affecting gene expression as enhancers. The ABC model accurately predicts the effects of the perturbation of these elements (AUPRC = 0.53, Figure B-11B). In contrast, H3K27ac^{low} CTCF elements had smaller effects (average 10% vs. 32% for H3K27ac^{high} CTCF elements, rank-sum test $p = 0.002$), had balanced effects on gene expression (5 activating and 8 repressive vs. 23 and 3 for H3K27ac^{high} CTCF elements, Fisher's exact $p = 0.002$), and the ABC model performed less well (AUPRC = 0.11, Figure B-11C).

Note B2. Regulatory effects of promoters on nearby genes

In addition to DE-G pairs, our CRISPR dataset in K562 cells included 1114 distal promoter-gene (DP-G) pairs (where the CRISPR-targeted element is located <500 bp from a TSS).

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 53 significant DP-G pairs (out of 1114 total tested), and the ABC score was only moderately predictive of these effects (AUPRC=0.16, Figure B-14). Importantly, the DP-G pairs in our dataset behaved qualitatively differently from the DE-G pairs: promoters more frequently had repressive effects (27 of 53 DP-G pairs, 51%, versus 20% 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 (*2, 3*); 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 (*4*); and transcriptional interference, in which transcription of one gene physically blocks transcription of another (*5*). 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 (*2, 3, 6*). 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 $p = 0.001$).

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 5 promoters whose inhibition affected 2 or more tested genes in our FlowFISH dataset (*GATA1*, *KLF1*, *LYL1*, *PPP1R15A*, and *SEC61A1*). Of these, 3 encode transcription factors and 2 encode regulators 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. These 5 promoters also more often had repressive effects than other promoters we found to affect the expression of nearby genes (median 2 repressed genes vs 0, rank-sum test $p = 0.004$). Based on this evidence, we expect that the effects of these 5 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 (Figure B-12D).

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 18 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 (*7, 8*).

Transcriptional interference

We identified 4 promoters (2 alternative promoters for each of 2 genes) where CRISPR perturbation caused an increase (6-36%) 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 (Figure B-14C), suggesting that these promoters might repress the neighboring gene via transcriptional interference (*5*).

Note B3. 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 (*9*) (see Methods), we tested simply using the theoretical contacts expected from extrusion globule and fractal globule models (Contact$_{Globule}$ is proportional to Distance$^{-\gamma}$, with $\gamma = 0.7$ and 1, respectively) (*9*). Both scores performed nearly as well as the ABC score based on Hi-C data (AUPRC = 0.64 for both, versus 0.66 for ABC, Figure B-9A,C). In comparison, Activity x Loop, Activity x Domain, Activity x Distance, and Activity x Contact$_{Globule}$ models with more extreme values of γ performed less well (Figure B-9). 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 (Figure B-9G,H). These and other 3D loops are present across many cell types (*10, 11*). Accordingly, we tested estimating Contact for a given pair of loci using the average contact frequency for those loci in Hi-C data from 8 other human cell types. We found that a 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 slightly better performance in the full K562 CRISPR dataset (AUPRC = 0.68 versus 0.66 respectively; Figure B-9A).

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.

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 by lentiviral transduction in 8 μg/ml polybrene by centrifugation at 1200 x g for 45 minutes with 200,000 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 (*10*).

Defining candidate elements

We defined candidate regulatory elements in 6 human cell types (K562, GM12878, NCCIT, LNCaP, primary hepatocytes, and primary erythroid progenitors), and 1 mouse cell type (mESCs).

For K562 and mESC, we concatenated all peaks called by ENCODE in both replicate DNase-seq experiments and merged resulting peaks. This resulted in 174,403 peaks in K562. We then removed any peaks overlapping regions of the genome which have been observed to accumulate anomalous number of reads in epigenetic sequencing experiments ('blacklisted regions' (*12*) downloaded from https://sites.google.com/site/anshulkundaje/projects/blacklists) — with the exception of 5 peaks in mESCs, which were either tested by CRISPR experiments or which were promoters of genes nearby tested elements, and which were not removed. Given that the ENCODE peaks were initially 150bp in length, we extended each of these peaks 175bp to arrive at candidate elements that were 500bp in length.

For GM12878, NCCIT, LNCaP, primary hepatocytes, and primary erythroid progenitors we called peaks using MACS2 based on either DNase-seq or ATAC-seq as a measure of chromatin accessibility. We initially considered all peaks with pvalue < .1 and removed peaks overlapping blacklisted regions. We then resized these peaks to be 500bp in length centered on the peak summit. In order to approximately match the number of candidate elements considered in K562, we then counted DNase-seq (or ATAC-seq) reads overlapping these regions and kept the 175,000 regions with the highest number of read counts. To this peak list, we added 1 kb regions centered on the transcription start site of all genes.

Any overlapping peaks resulting from this extension within a cell type were merged. We define these extended and merged peaks as *candidate elements.*

Guide selection for CRISPRi-FlowFISH screens

We designed gRNAs within K562 candidate elements as previously described (*10*) and used all gRNAs within each candidate element after removing those with specificity scores <50 or with homopolymer stretches of more than 7 As, Gs, or Cs, or 4 Ts.

Gene selection for CRISPRi-FlowFISH screens

We used a series of filters for each probeset and screen to ensure robust, comprehensive, and quantitative discovery of regulatory elements for each gene (Figure B-4). We initially tested PrimeFlow probesets for genes expressed at >20 TPM in K562s in five genomic loci (Figure B-3). We first screened probesets by flow cytometry and selected those with >2-fold signal vs unstained cells. Next we performed a tiling CRISPRi-FlowFISH screen (see below) and focused our analysis on the screens that showed the following characteristics: (i) maximum unscaled knockdown among 20-gRNAs windows within 500 bp of the TSS >50%; (ii) variance in non-targeting, negative-control gRNAs <1; and (iii) >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 28 genes.

CRISPRi-FlowFISH Screens

We cloned gRNA libraries purchased from CustomArray (now GenScript) for each of 5 genomic loci (Figure B-3), transduced into K562s harboring a doxycycline-inducible KRAB-dCas9, and selected for transduced cells as previously described (*10*). We induced KRAB-dCas9 expression with 1 μg/ml doxycycline for 48 hours. We used 30M cells for each screen.

We used the PrimeFlow RNA Assay Kit (Thermo Fisher; Catalog number: 88-18005) according to the manufacturer's instructions with some modifications. Specifically, we split each screen into three

10 million cell reactions and performed five total washes with 35ºC wash buffer after following the staining protocol. We stained each sample for the gene of interest with an Alexa Fluor 647 (AF647, "Type 1") probeset and against a positive control housekeeping gene with Alexa Fluor 488 (AF488, "Type 4"). For most screens we used control gene *RPL13A*, but because *BAX*, *BCAT2, FTL*, *NUCB1*, and *PPP1R15A* are <700 kb from *RPL13A*, we used *ACTB* for these.

Fluorescence activated cell sorting

We diluted the stained cells in PBS with 0.5% BSA to a concentration of $2x10^7$ cells/ml and filtered using a 30µm filter (CellTrics, Catalog number 04-004-2326). We sorted 30 million cells for each screen into six bins based on fluorescence intensity of target genes using the Astrios EQ Sorter (Beckman Coulter B25982). To control for differences in staining efficiency for each cell, we normalized the fluorescence associated with the gene of interest to that of the control gene. Specifically, we used the color compensation tool to subtract a portion of each cell's AF647 signal based on the intensity of its AF488 signal such that the mean AF488 signal in the top and bottom 25% of cells based on AF647 was within 10%. If necessary, we then reduced the level of compensation until the fraction of cells with AF647 signal equal to 0 was no more than 5%. We set the gates for each bin on the compensated signal to capture 10% of the cells according to the percentiles (i) 0-10% (ii) 10-20%, (iii) 35-45%, (iv) 55-65%, (v) 80-90%, and (vi) 90-100%.

Genomic DNA extraction and gRNA sequencing

We collected the sorted cells by centrifugation at 800g for 5 minutes, resuspended cells in 100uL of Lysis buffer (50mM Tris-HCl, pH 8.1, 10mM EDTA, 1% SDS), and incubated at 65ºC for 10 minutes for reverse crosslinking. Once the samples cooled to 37 ºC, we added 2ul of RNase Cocktail (Invitrogen, catalog #AM2286), mixed well, and incubated the mixture at 37ºC for 30 minutes.

Finally we added 10μl Proteinase K (NEB, catalog number P8107S), mixed well, and incubated the mixture at 37ºC for 2 hours followed by incubation at 95ºC for 20 min. We extracted genomic DNA using Agencourt XP (SPRI) beads (Beckman Coulter). We sequenced gRNA integrations as previously described (*10*).

Analysis of CRISPRi-FlowFISH screens

To determine the effects of each gRNA on fluorescence, we used a maximum likelihood estimation (MLE) method. First, we normalized gRNA frequencies in each bin by dividing each gRNA count by the total read count for all gRNAs in that bin and summed normalized counts across PCR replicates. Next, we used the limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm MLE method in the R stats4 package to fit the read counts in each fluorescence bin to the log-normal distribution that would have most likely produced the observed counts in the bins. The effect size is from the mean of the log-normal fit. We assumed the gRNAs targeting the TSS of the assayed gene have a "true" effect size of 85% (based on previous observations that show CRISPRi effects of 80-90% across a panel of genes (*13*)), but that some portion of the FlowFISH signal is due to non-specific binding of the probe. Accordingly, we scaled the effect size of each gRNA within each screen linearly so that the strongest 20-gRNAs window within 500 bp of the target genes TSS gRNAs 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 *t*-test to determine whether the mean effect size of the gRNAs in each candidate element deviated significantly from the mean of scrambled-sequence, negative control gRNAs. We computed the FDR for elements using the Benjamani-Hochberg method applied per gene, and used an FDR threshold of 0.05 to call significant E-G interactions.

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

(i) There was less than 80% power to detect a 25% effect for this E-G pair.

(ii) The element overlapped the gene's promoter.

(iii) The element was within the gene body or extended up to 2 kb downstream of the 3' end of the gene.

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 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 the ABC model. We did not consider element-gene pairs where the element was that gene's own promoter. 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 (*10*). We used RT-qPCR data from this study to represent the effect sizes for the 7 and 2 enhancers that significantly affected *MYC* and *GATA1* expression, respectively. For all other elements, we estimated their effect sizes on gene expression based on the linear relationship between *MYC* expression and proliferation (*10*).

Klann 2017. Klann *et al.* used CRISPRi (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 (*14*). 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 (Supp Figure 3B in Klann *et al.* 2017).

Ulirsch 2016. Ulirsch *et al.* used CRISPR used one gRNA per enhancer to introduce small deletions at each of 3 enhancers in K562 cells (*15*). We obtained the original qPCR data from the authors and assessed expression differences between homozygous knockout and wild-type clones using a *t* test.

Wakabayashi 2016. Wakabayashi *et al.* used one gRNA per enhancer to introduce small deletions at each of 5 enhancers in K562 cells (*16*). We obtained the original qPCR data from the authors and assessed expression differences between homozygous knockout and wild-type clones using a *t* test.

Thakore 2015. Thakore *et al.* used KRAB-dCas9 to inhibit an enhancer (HS2) in the globin locus in K562 and performed RNA-seq (*17*). 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 (*18*). We downloaded the raw data from GSE85011 and quantified transcript abundance with kallisto (v. 0.43.0). 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 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.

Engreitz 2016. We previously generated homozygous and heterozygous knockout clones of 12 lncRNA and 6 mRNA promoters in mES cells on a 129S1/*Castaneus* hybrid genetic background, and measured the effects on gene expression using allele-specific RNA-seq (*2*) 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 (*2*), and identified several additional nearby genes that showed significant allele-specific effects on expression. For this analysis, "nCtrl" and "nKO" refer to the number of wild-type and knockout *chromosomes* for each locus.

mES cell enhancer deletions (this study). We also included data from new experiments in which we deleted two putative enhancers in mES cells via transfection of multiple gRNAs and measured the effects on nearby genes using allele-specific RNA sequencing, as previously described (*2*). These two enhancers were selected on the basis of previous plasmid reporter assays showing enhancer activity for these elements (*19*) 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 (*2*). We assessed statistical significance using DESeq2 as described above.

Moorthy 2017. Moorthy *et al.* generated enhancer knockouts in mES cells on a 129S1/*Castaneus* hybrid genetic background, and measured the effects on gene expression using allele-specific RNA-seq as well as RT-qPCR (*20*). 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 (*2*). 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 (*21*). 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, 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 (*22-32*).

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 (*33*). 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 compenent 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 150bp paired-end reads of insert sizes between 150bp and 400bp (in increments of 10bp), (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 sucessfully 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 (*34*) and computed differential expression with DESeq2 as described in Fuentes *et al.* (*33*). We considered a gene significantly differentially expressed if its Benjamini-Hochberg adjusted pvalue 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.* (*33*): 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 which

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

We only considered genes which

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

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. 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.

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%) (*10*), 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 permited only 23 negative DE-G pairs for analysis of the perturbation data in other cell types (Figure B-15), 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 920.

Power calculations for FlowFISH experiments. For each candidate element, we used a *t*-test (equal variances) to compare the MLE effects of the gRNAs in that element to the MLE effects of 668-3505 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 a single tested E-G connection with insufficient power.

Power calculations for qPCR datasets. We used a *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 (*35*). 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.

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 "distal 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 method, 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 63% precision in our initial K562 dataset could identify DE-G connections in other cell types with 91% recall and 75% precision (Figure B-15).

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 = 0.66 vs 0.75, respectively; Fig 4).

Epigenomic datasets, processing, and analysis.

DNaseI 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 varienty of sources including ENCODE (*33, 36-39*). We generated our own H3K27ac ChIP-seq data in F1 129/Castaneus hybrid mESCs grown in 2i media as previously described (*2*), 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) (*11*). We also generated new *in situ* Hi-C maps of male mouse V6.5 embryonic stem cells grown in 2i conditions as previously described (*11*), 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 (*40*).

NCCIT ATAC

We performed ATAC-seq on 10K cells NCCIT cells in duplicate according to the protocol described by Buenrostro *et al.* (*41*) with some modifications. Specifically, we used Sigma Nuclei EZ lysis buffer for lysis for 10 minutes while spinning 500xG at 4C, resuspended with the lysis buffer, and spun again for 3 minutes. We then resuspended the nuclei pellet with a tagmentation buffer containing 12.5 uL of TD buffer, 1.25 uL of Tn5 transposase, 7.5 uL of PBS and 2.75 uL 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 (*33*). We aligned reads using BWA (v0.7.17) (*42*), removed PCR duplicates using the MarkDuplicates function from Picard (v1.731), and removed reads with mapq < 30. We used MACS2 (v2.1.1) (*43*) to call peaks on Cas9 ChIP-seq using the non-targeting conditions as controls as described in (*33*).

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 multiplied by the strength of the enhancer (i.e., the ability of the enhancer to activate transcription upon contacting a promoter) (*10*). 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.

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

ABC score for effect of element E on gene G = Activity of E × Contact frequency between E and G / Sum of (Activity × Contact Frequency) over all candidate elements within 5 Mb.

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 normalized Hi-C contact frequency between E and the promoter of gene G, and elements are defined as ~500bp 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.

We detail the calculation of the ABC score and discuss these assumptions below.

Calculating enhancer activity from DHS and H3K27ac ChIP-seq signals

We estimated enhancer activity of candidate elements using a combination of quantitative DNase-seq and H3K27ac ChIP-seq signals. For a given element, we counted DHS and H3K27ac reads (per million) in DNase peaks (150 bp from ENCODE), which we extended by 175 bp on either side (to 500 bp total; average length after merging overlapping peaks = 597 bp, Figure B-2B) because H3K27ac ChIP-seq signals are strongest on the nucleosomes flanking the nucleosome-free DHS peak. We computed the geometric mean of DNase-seq and H3K27ac ChIP-seq signals because we expect that strong enhancers should have strong signals for both, and that elements that have only one or the other likely represent other types of elements. (Elements with strong DNase-seq signal but no H3K27ac ChIP-seq signal might be CTCF-bound topological elements. Elements with strong H3K27ac signal but no DNase-seq signal might be sequences that are close by to strong enhancers but do not themselves have enhancer activity, due to the spreading H3K27ac signal over hundreds to thousands of bp.)

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

Calculating contact frequency from cell-type specific Hi-C data.

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

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

i. Rows and columns corresponding to KR normalization factors less than .1 were removed (these typically correspond to 5-kb bins with very few reads).

ii. Each diagonal entry of the Hi-C matrix was replaced by the maximum of its four neighboring entries. Justification: 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:

i. We extract 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.

ii. 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.

iii. We add a small adjustment ("pseudocount") to ensure that the contact frequency for each E-G pair is non-zero. For E-G pairs within 1 Mb, the adjustment is equal to the expected contact frequency at 1Mb (as predicted by the power-law relationship between contact frequency and genomic distance, see below), and for E-G pairs at distance d (d > 1Mb), the adjustment is equal to the expected contact at distance d. In each case the adjustment is scaled to be in the same units as described in (i). Adding the adjustment sometimes results in a quantitative Contact greater than 100; in such cases, the Contact is reduced to 100.

Calculating the contribution of one candidate element relative to others in the region.

To calculate the relative effect of each element to the expression of a gene, we normalize 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 (see below). 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 (*2, 3*). 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.

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 size 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 (Figure B-8). 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.

Alternative methods to estimate Contact in the ABC model

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 8 human Hi-C datasets. These averaged profiles were created as follows:

i. For each gene in the genome, we extract the row corresponding to this gene from each Hi-C matrix (KR normalized, at 5KB resolution)

ii. Each of these profiles is then scaled using the cell-type specific power law parameters relative to the K562 power law parameters (see below)

iii. Finally, the total Hi-C signal in each cell-type specific profile is normalized to sum to one and then averaged across cell types to create the average profile at a given locus

Normalizing Hi-C Profiles Using the Power-Law Fit

We find 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 parameters from the power law relationship in that cell type (see below). The scaling factor at distance *d* is given by $(scale_{\mathrm{ref}}/\, scale_{\mathrm{celltype}}) * d \,\hat{}\, (gamma_{\mathrm{ref}} – gamma_{celltype})$, where $scale_{\mathrm{ref}}$ and $gamma_{\mathrm{ref}}$ are the given reference parameters. For this study we used the power-law parameters in K562 as a reference.

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:

i. We aggregate all entries of the Hi-C matrix located greater than 10kb and less than 1Mb from all gene promoters (KR normalized at 5kb resolution)

ii. We then perform 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 and the intercept is the *scale* 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.

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 enhancer-class to the 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.

Identifying expressed genes for ABC predictions

When using the ABC model to predict enhancer-gene connections genome-wide, 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 transcripts per million (TPM) > 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.

Comparison to other published enhancer-gene prediction methods

We evaluated the performance of the following published enhancer-gene prediction methods in predicting 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.*, DNAase1, 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 (*44*). 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/#SD1 (*45*). 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 (*46*). 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.

Hi-ChIP 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 (*31*). 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%) (*10*). 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. To test their effects in the genome, we designed a pool containing 2-3 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

(Figure B-9H. Because *PVT1* has multiple promoters in K562 cells (*47*), we verified the effects we observed in FlowFISH using RT-qPCR with primers corresponding to a specific *PVT1* isoform (that uses e3 as a promoter) as previously described (*10*).

siRNA-mediated knockdown of GATA1

We transfected 200,000 K562 CRISPRi cells (from the same population of cells that was used in the CRISPRi-FlowFISH screens) with siRNAs (from Ambion, Thermo Fisher Scientific) 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 (*48*) with minor modifications. Specifically, we isolated RNA from 100,000 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) 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 uL Nextera Tagment DNA Tn5 transposase (Illumina, San Diego, CA, 15027916) in a 10 uL 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 chosing 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-targetting controls with the design formula "~perturbation+dose" (to control for the doses of siRNAs). Genes within 1 Mb of *GATA1* with Benjamani-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 (*49*) or uniform expression across many tissues (*47, 50*) for 847 total ubiquitously expressed genes, For mouse, we used the list of 4781 uniformly expressed genes provided in Li *et al. (51).* 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 genes, including 28 genes from this study and 2 genes (*MYC* and *HBE1*) from previous studies (*10, 14*). In this subset of the data, we found 58 regulatory DE-G pairs for the 22 tissue-specific genes and 1 regulatory DE-G pairs for the 8 ubiquitously-expressed genes (Fisher's exact $p < 10^{-4}$), as reported in the main text. We note that the same trends hold 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 (140 significant pairs out of 2304 tested) than for ubiquitously expressed genes (6 significant pairs out of 777 tested, Fisher's exact $p < 10^{-11}$).

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. We downloaded CTCF ChIP-seq peak calls generated by ENCODE 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$^{\text{High}}$ or H3K27ac$^{\text{Low}}$, 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 762 total DE-G pairs that involved (i) CTCF-bound elements unlikely to function as enhancers (H3K27ac$^{\text{Low}}$, 755 DE-G pairs), or (ii) DE-G pairs likely to result from indirect effects (18 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 18 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.

The performance of the ABC score is markedly higher on this filtered dataset, with the AUPRC rising from 0.77 to 0.82 for tissue-specific genes and 0.66 to 0.72 for all genes (Figure B-13B). 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

R (3.1.1) with Bioconductor (3.0)(*52*), 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).

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.

Supplemental Figures

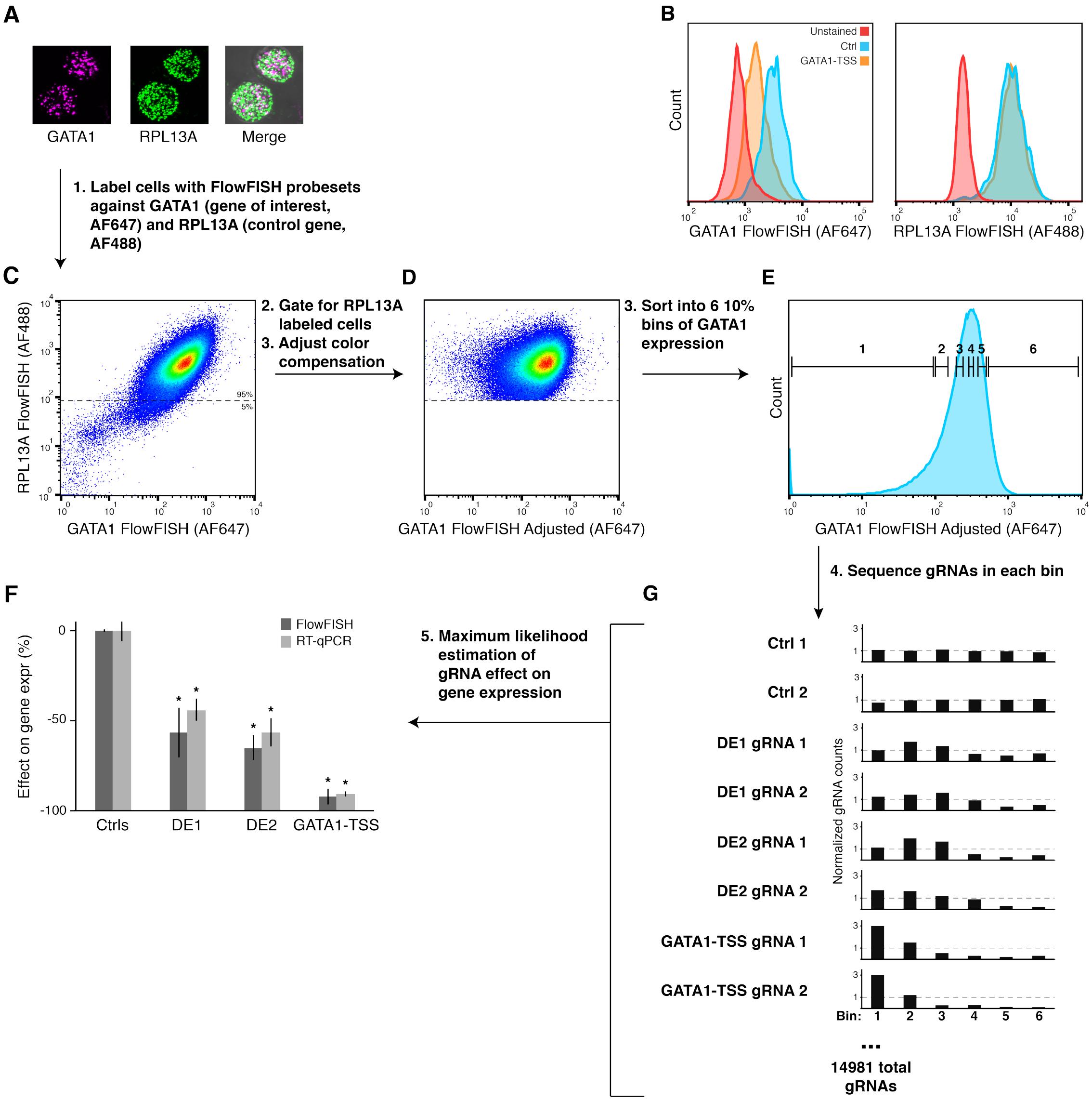


Figure B-1. Sorting and sequencing strategy for CRISPRi-FlowFISH Screens. (A) K562 cells labeled with FlowFISH probesets against *RPL13A* (control gene) and *GATA1* (gene of interest) imaged by fluorescence microscopy. **(B)** Histograms of FlowFISH signal (arbitrary units of fluorescence) for *GATA1* (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). **(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.

Figure B-1 (Continued). (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, and 6 Ctrl gRNAs for RT-qPCR, each measured in biological triplicate. *: $p < 0.05$ in *t*-test versus Ctrl. **(G)** Counts in each of the 6 bins for single gRNAs targeting the *GATA1* TSS, the two *GATA1* enhancers (DE1 and DE2) identified in Fulco *et al. (10)*, and representative negative controls (Ctrl).

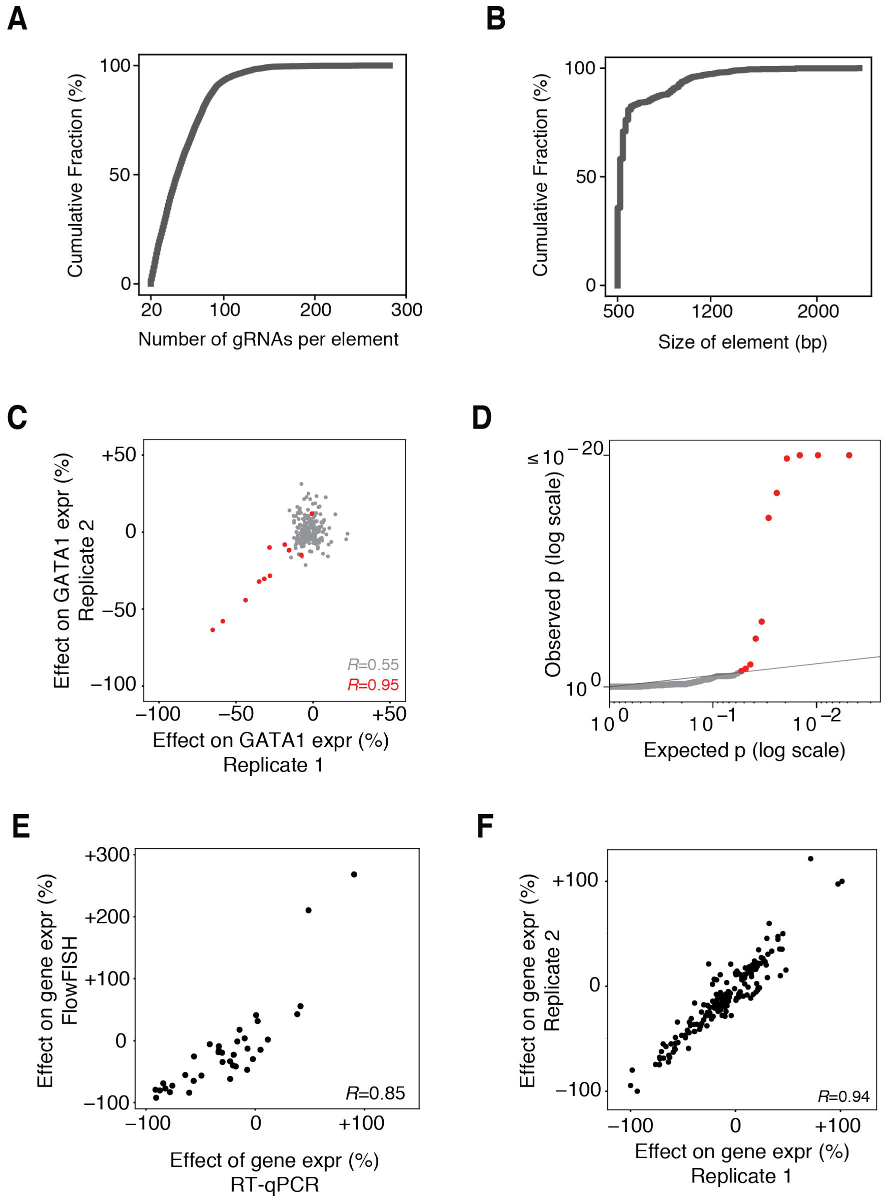


Figure B-2. CRISPRi-FlowFISH reproducibly quantifies effects of regulatory elements. **(A**) Cumulative distribution plot of the number of gRNAs in each tested candidate element. **(B)** Cumulative distribution plot of the width of each tested candidate element. **(C)** Correlation between replicate CRISPRi-FlowFISH screens for *GATA1*. Red points denote elements significantly affecting expression. Pearson R = 0.95 for significant elements, 0.55 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)** Correlation between effect on gene expression as measured by CRISPRi-FlowFISH screening and RT-qPCR for all 36 E-G pairs tested by both methods. Value is the mean effect of the two gRNAs for each element. **(F)** Correlation between effects on gene expression for all significant E-G pairs measured by replicate CRISPRi-FlowFISH screens.

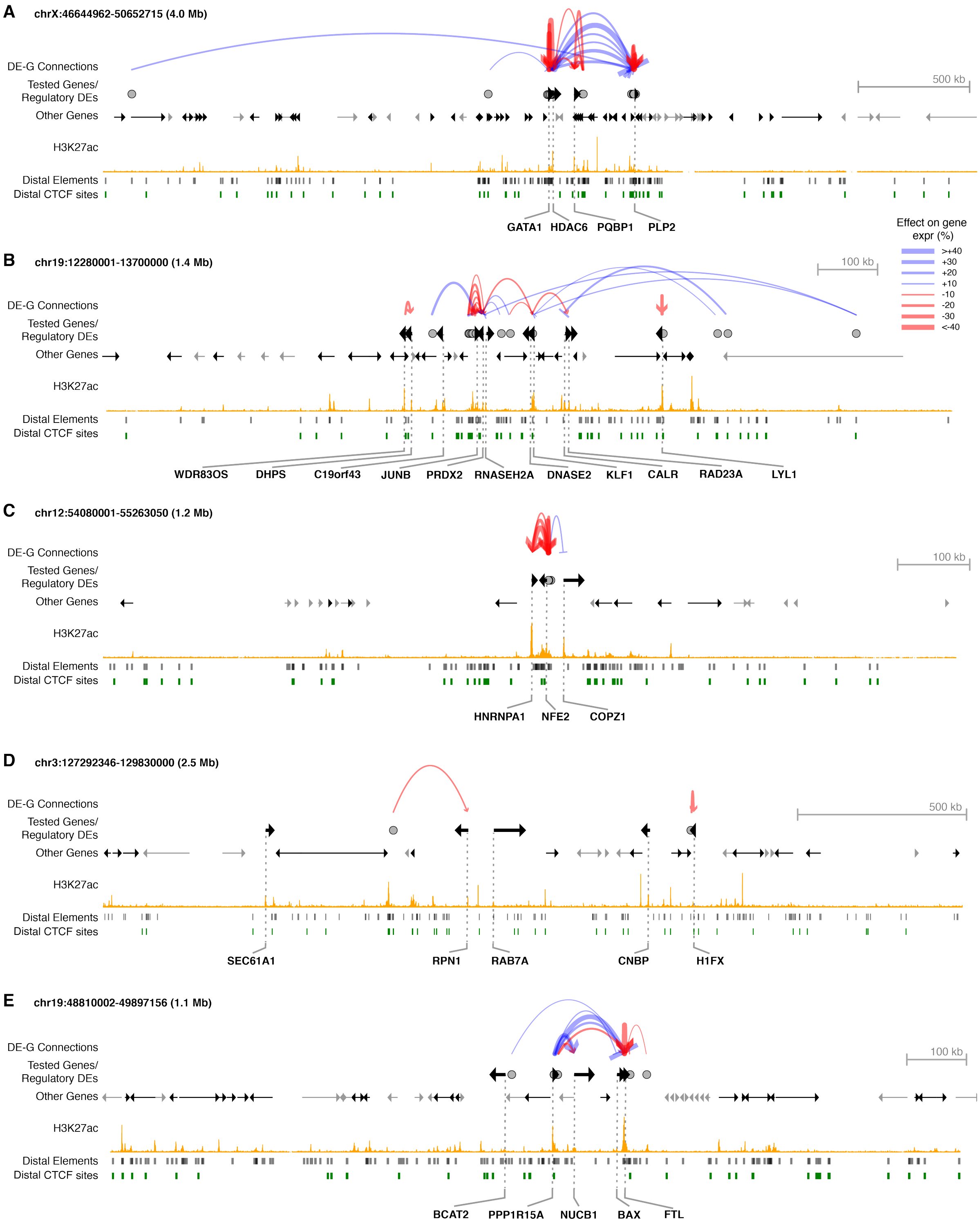


Figure B-3 (Legend on next page).

Figure B-3 (Continued). 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.

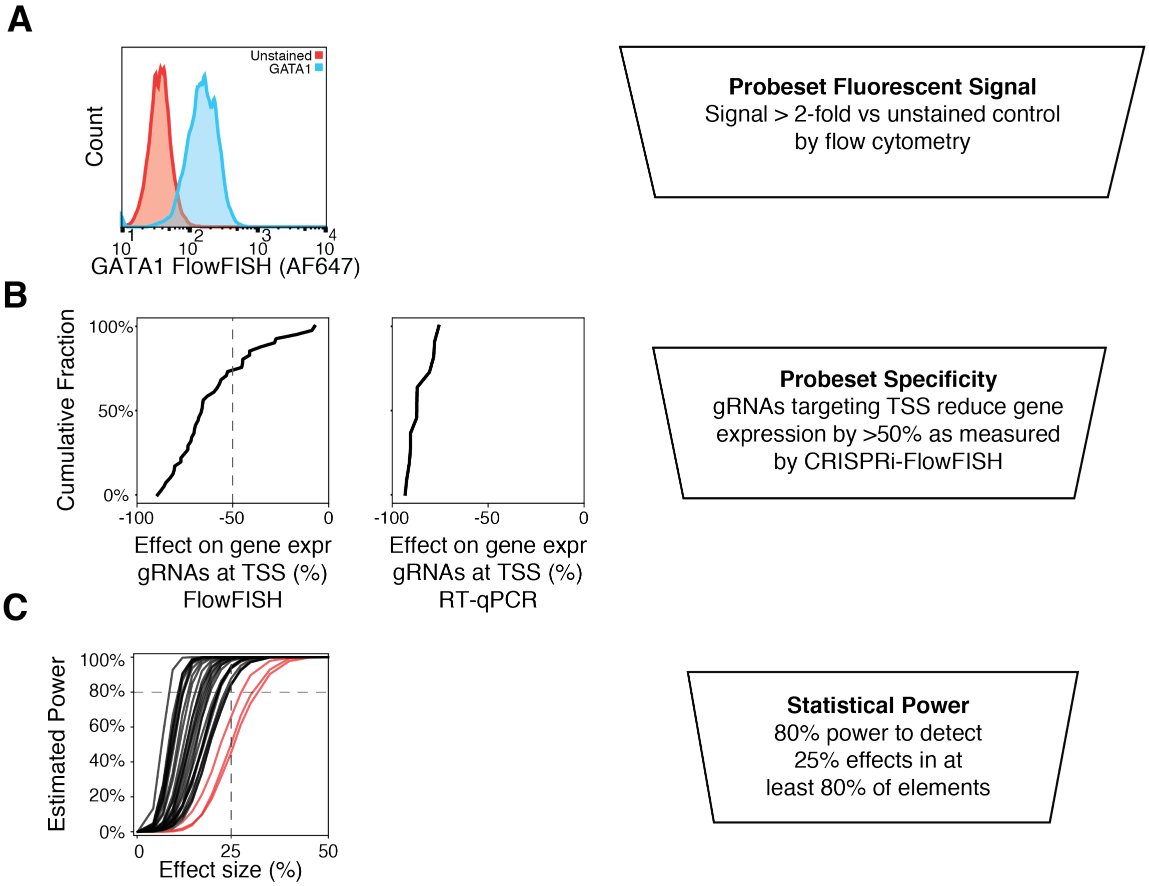


Figure B-4. Quality filters for CRISPRi-FlowFISH probesets and screens. (A) Histogram of FlowFISH signal, as measured by flow cytometry, comparing K562 cells stained with *GATA1* probes compared to unstained, negative-control cells. We required probesets to have >2-fold mean fluorescent signal in stained 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 those screens with at least 80% power to detect a 25% effect for at least 80% of tested elements. Red lines represent screens that did not meet this power threshold.

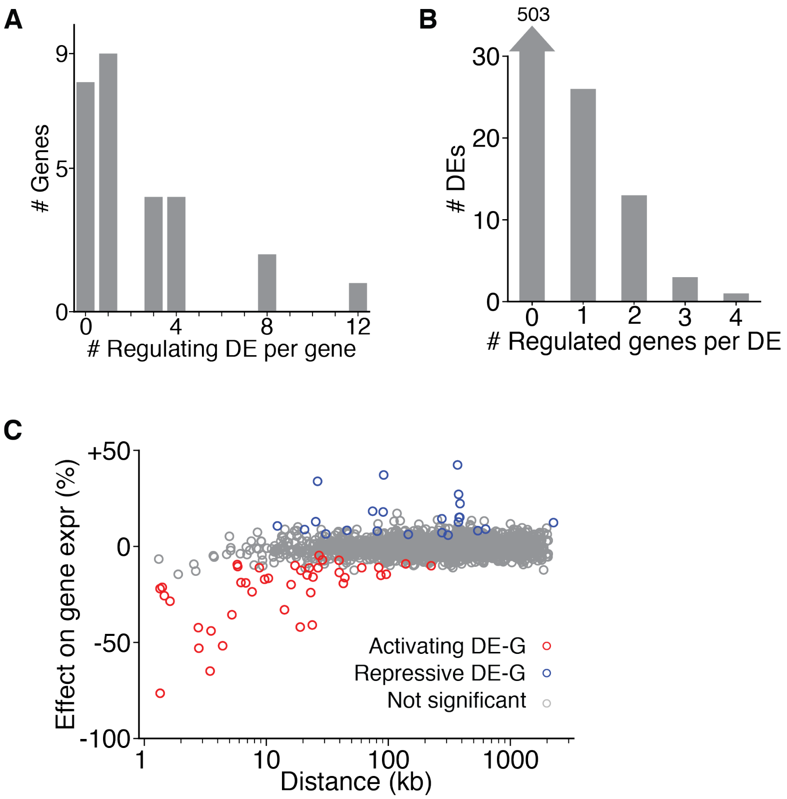


Figure B-5. Properties of the CRISPRi-FlowFISH dataset. (A) Histogram of the number of distal elements affecting each gene in CRISPRi-FlowFISH experiments. **(B)** Histogram of the number of genes affected by each distal element tested in CRISPRi-FlowFISH experiments. **(C)** Comparison of genomic distance with observed changes in gene expression upon CRISPRi perturbation. Each dot represents one tested DE-G. Red dots: connections where perturbation resulted in a decrease in the expression of the tested gene. Blue dots: perturbation resulted in an increase. Grey dots: had no significant effect.

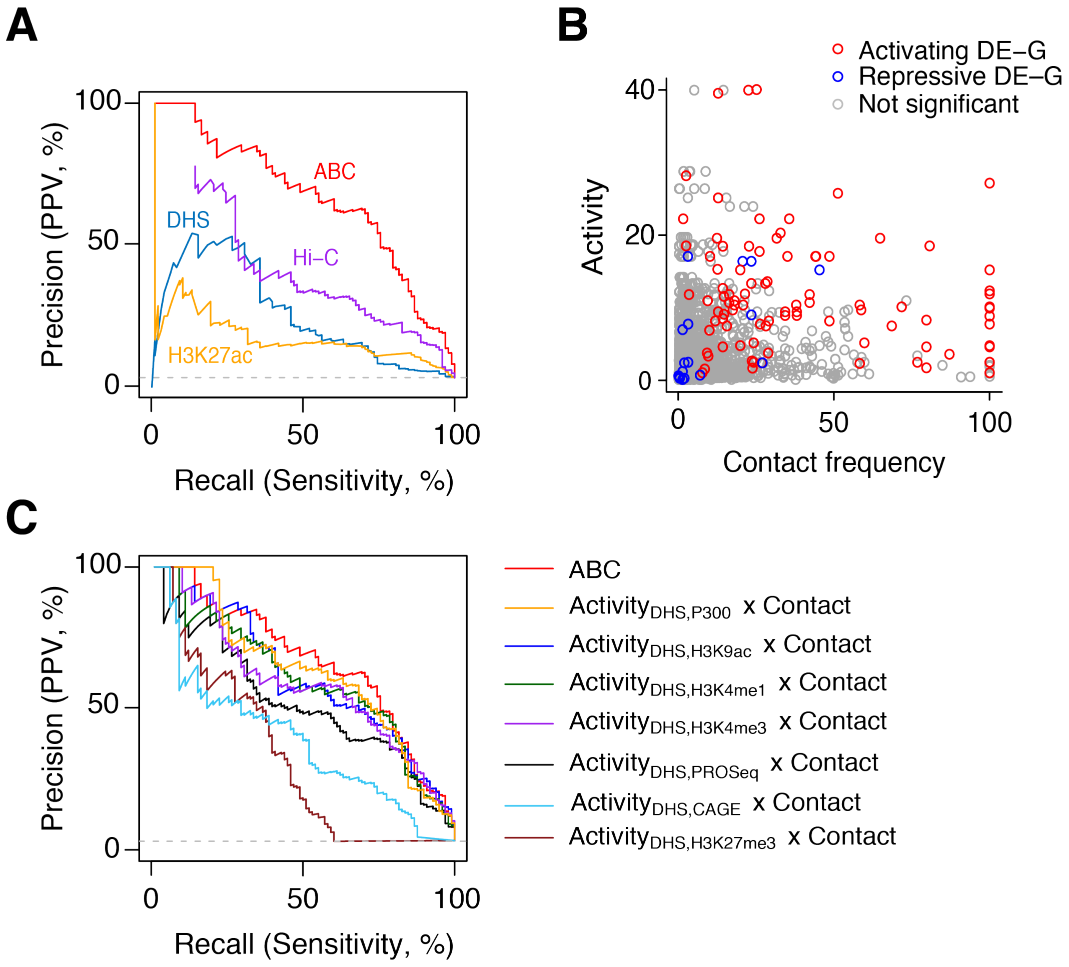


Figure B-6. Comparison of ABC score to other predictors. (A) Precision-recall curves for classifying regulatory DE-G pairs, comparing each of the components of the ABC score. **(B)** Scatterplot of Activity and Contact frequency for each tested DE-G pair. KR-normalized Hi-C contact frequencies are scaled for each gene so that the maximum score of an off-diagonal bin is 100 (see Methods). **(C)** Precision-recall curves comparing different measures of Activity. Activity$_{\text{Feature1,Feature2}}$ = sqrt(Feature1 RPM x Feature2 RPM). (ABC score corresponds to Activity$_{\text{DHS,H3K27ac}}$ x Contact). Figure legends are ordered from best to worst AUPRC.

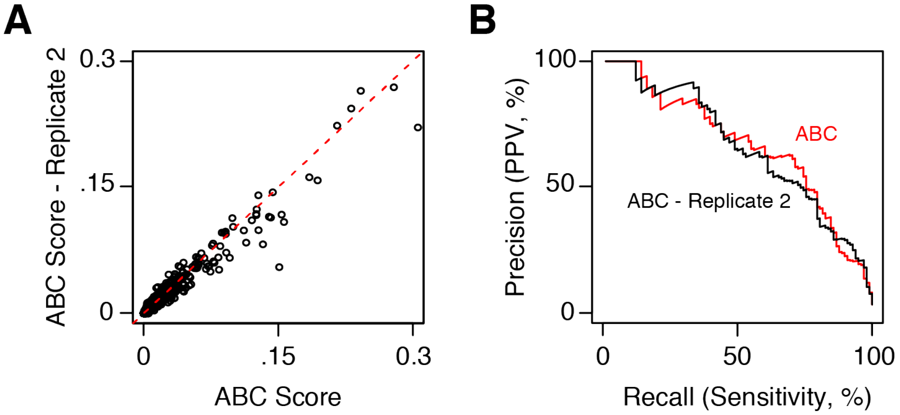


Figure B-7. The ABC Score is reproducible between replicates of the epigenomic datasets. **(A)** Scatter plot of ABC Score computed using biological replicates for DHS and H3K27ac (Pearson *R* = .98). **(B)** Precision-recall curves for classifying regulatory DE-G pairs (Positive DE-G pairs are those 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.

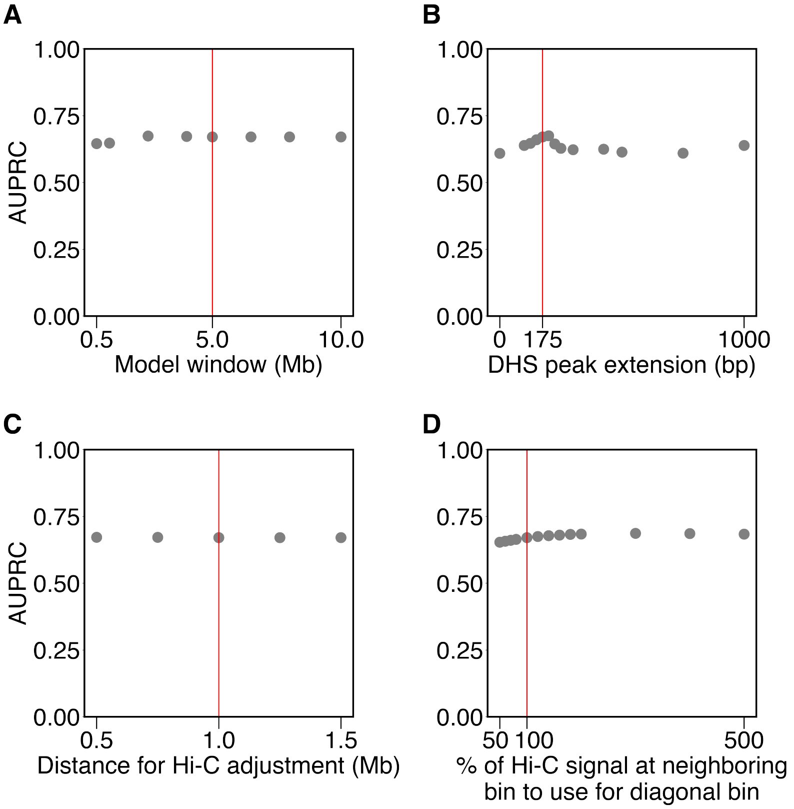
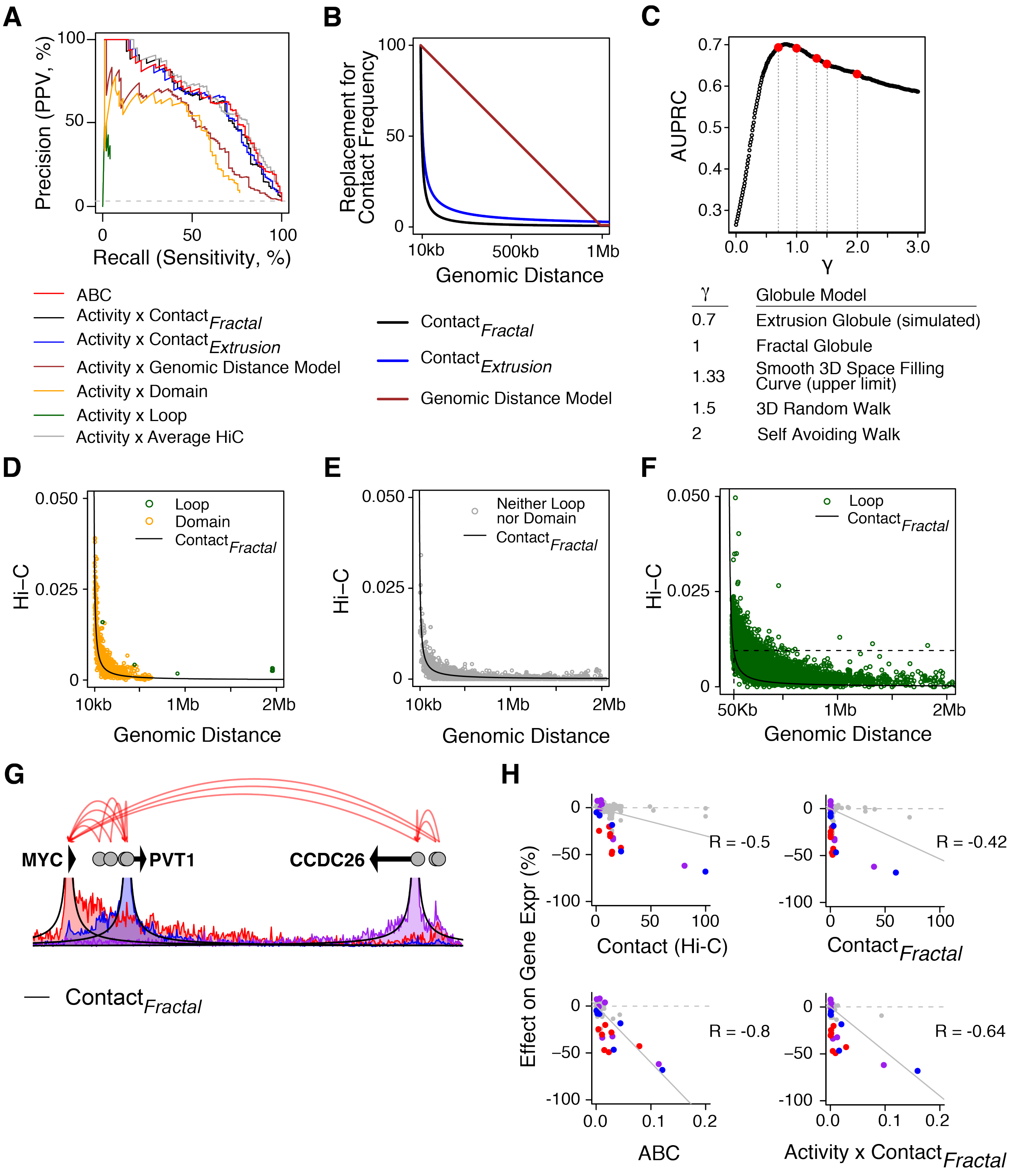


Figure B-8. Sensitivity of ABC score performance to chosen parameters. Changing the parameters of the ABC score does not dramatically affect performance near the default values. Each panel presents the area under the precision recall curve (AUPRC) for the ABC score when changing the specified parameter. 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 of Hi-C data, each diagonal entry of the Hi-C matrix is replaced by the maximum of its four neighboring entries when estimating contact frequency at distances < 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

Figure B-9. (Legend on next page).

Figure B-9. (Continued) 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(.01, (1e6 - *Distance*)/1e6). *Activity x Contact*$_{\mathit{Fractal}}$*, Activity x Contact*$_{\mathit{Extrusion}}$*:* Contact component is proportional to *Distance*$^{-\gamma}$. *Contact*$_{\mathit{Fractal}}$ uses $\gamma = 1$, *Contact*$_{\mathit{Extrusion}}$ 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 (γ = 0.7 and 1) well represent the empirically observed Hi-C contacts at various distances (*53*). **(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) (*53*). **(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*$_{\mathit{Fractal}}$ 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 (*11*). 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*$_{\mathit{Fractal}}$ model (G) Visualization of Hi-C tracks anchored at the MYC, PVT1 and CCDC26 promoters (colored lines), compared to the *Contact*$_{\mathit{Fractal}}$ model (black lines). (H) Computation of the ABC score for DE-G pairs in the MYC locus using Hi-C vs the *Contact*$_{\mathit{Fractal}}$ model. Using Hi-C data better predicts the quantitative effects of enhancers in this locus.

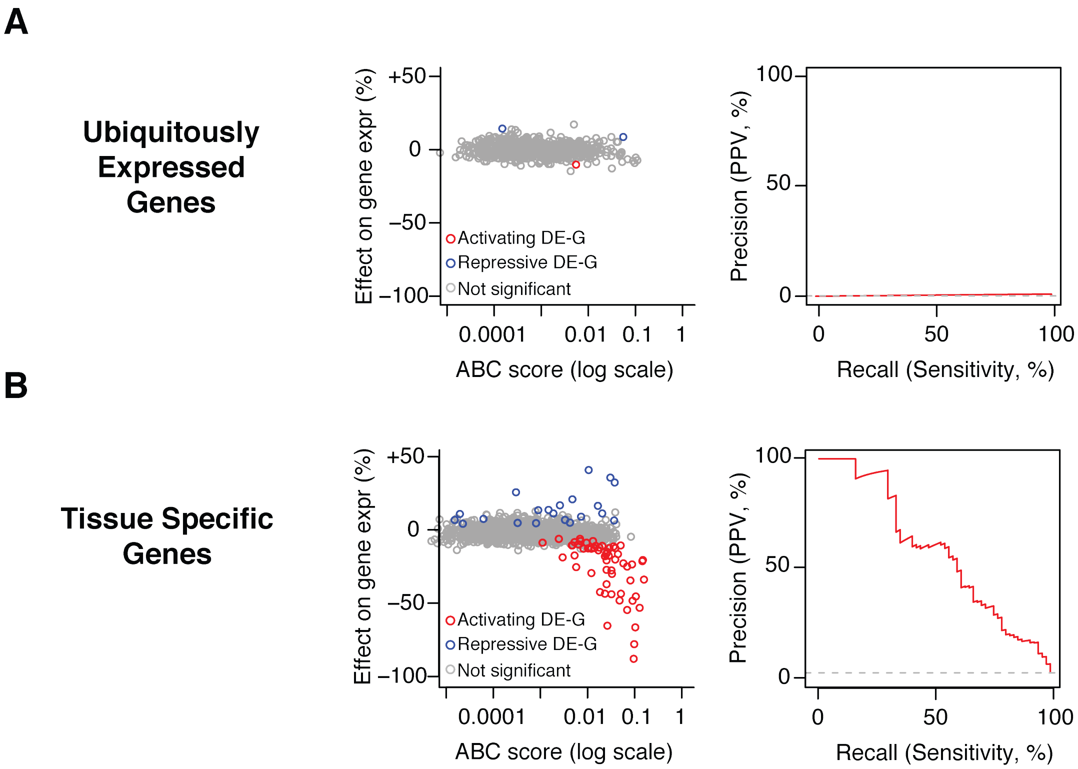


Figure B.10. Tissue-specific genes have more distal enhancers than ubiquitously-expressed genes. (A) Left: Comparison of ABC scores (predicted effect) with observed changes in gene expression upon CRISPR perturbations. Each dot represents one tested DE-G pair where G is a ubiquitously-expressed gene. 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 (28 genes from this study, 2 from previous studies).

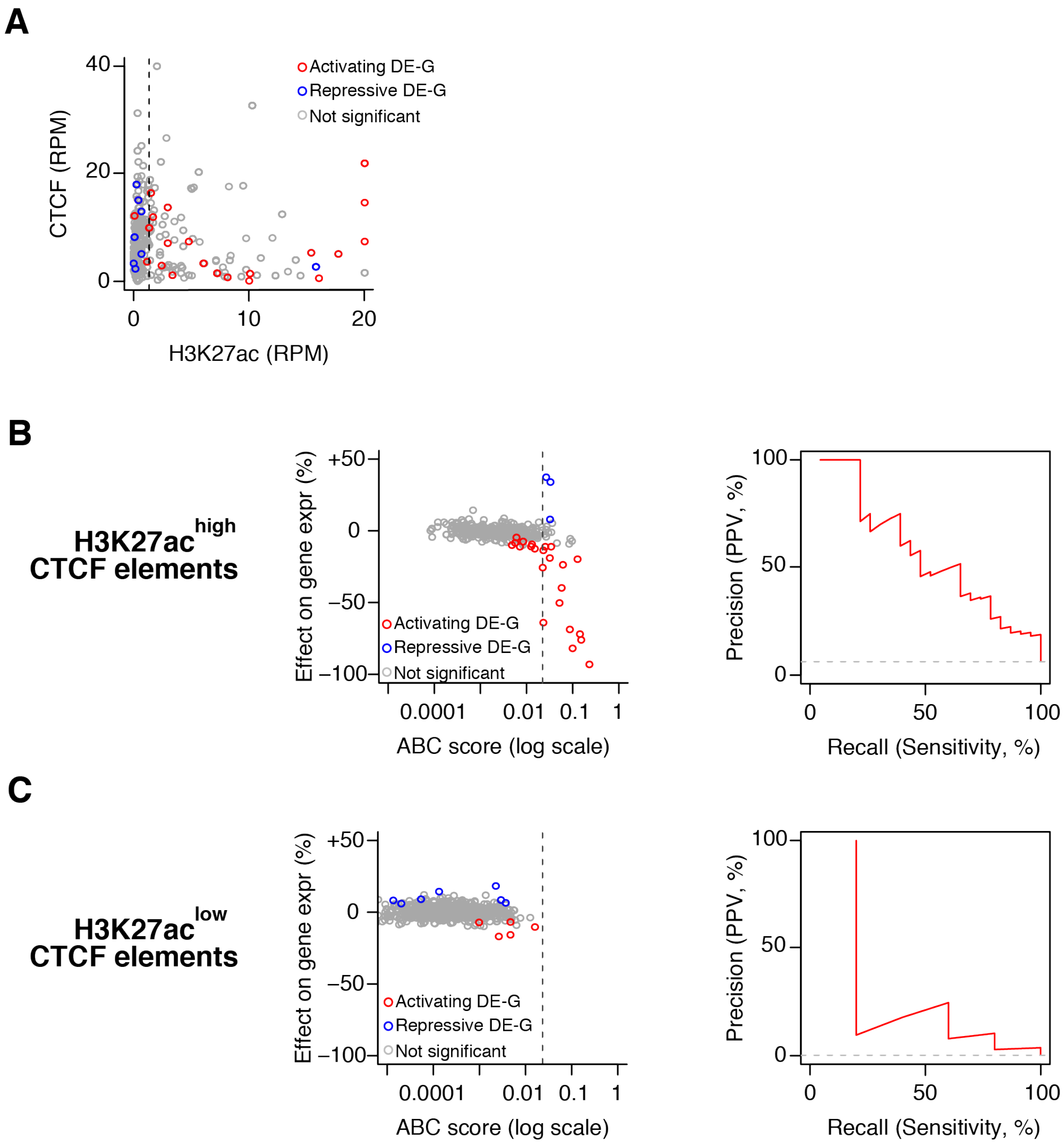


Figure B-11. Analysis of CTCF-bound elements. (A) Scatterplot of CTCF signal (reads per million) vs H3K27ac signal (reads per million) for all DE-G pairs where the DE is bound by CTCF (see Methods). Dotted black line corresponds to the median H3K27ac signal for all distal elements in the dataset. We 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.

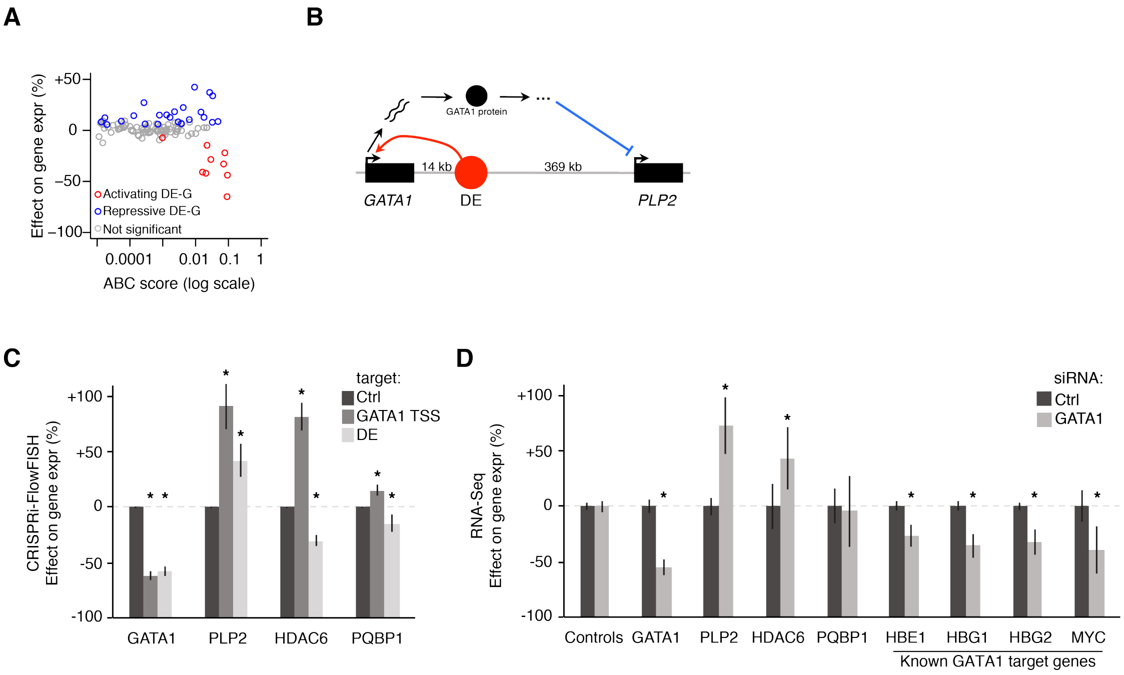


Figure B-12. Elements that repress a distal gene are likely explained by indirect regulatory effects. (A) Comparison of ABC scores (predicted effect) with observed changes in gene expression upon CRISPR perturbations. Each dot represents one tested DE-G pair where the element represses at least one gene. **(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 *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. *: BH-adjusted $p < 0.05$ in t-test versus negative controls (see Methods). **(D)** Effects of inhibiting *GATA1* with siRNAs on gene expression of *GATA1, PLP2, HDAC6, PQBP1*, and known GATA1 transcription factor targets (*54-56*) 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 (*ACTB*, *B2M*, *C1orf43*, *CHMP2A*, *EMC7*, *GAPDH*, *GPI*, *PGK1*, *PPIB*, *PSMB2*, *PSMB4*, *REEP5*, *RPL13A*, *SNRPD3*, *TBP*, *TUBB*, *VCP*, and *VPS29*). Error bars: 95% confidence interval for the mean of two siRNAs with four independent transfections each. *: BH-adjusted $p < 0.05$ from DESeq2 for *GATA1* siRNA versus Ctrl (see Methods).

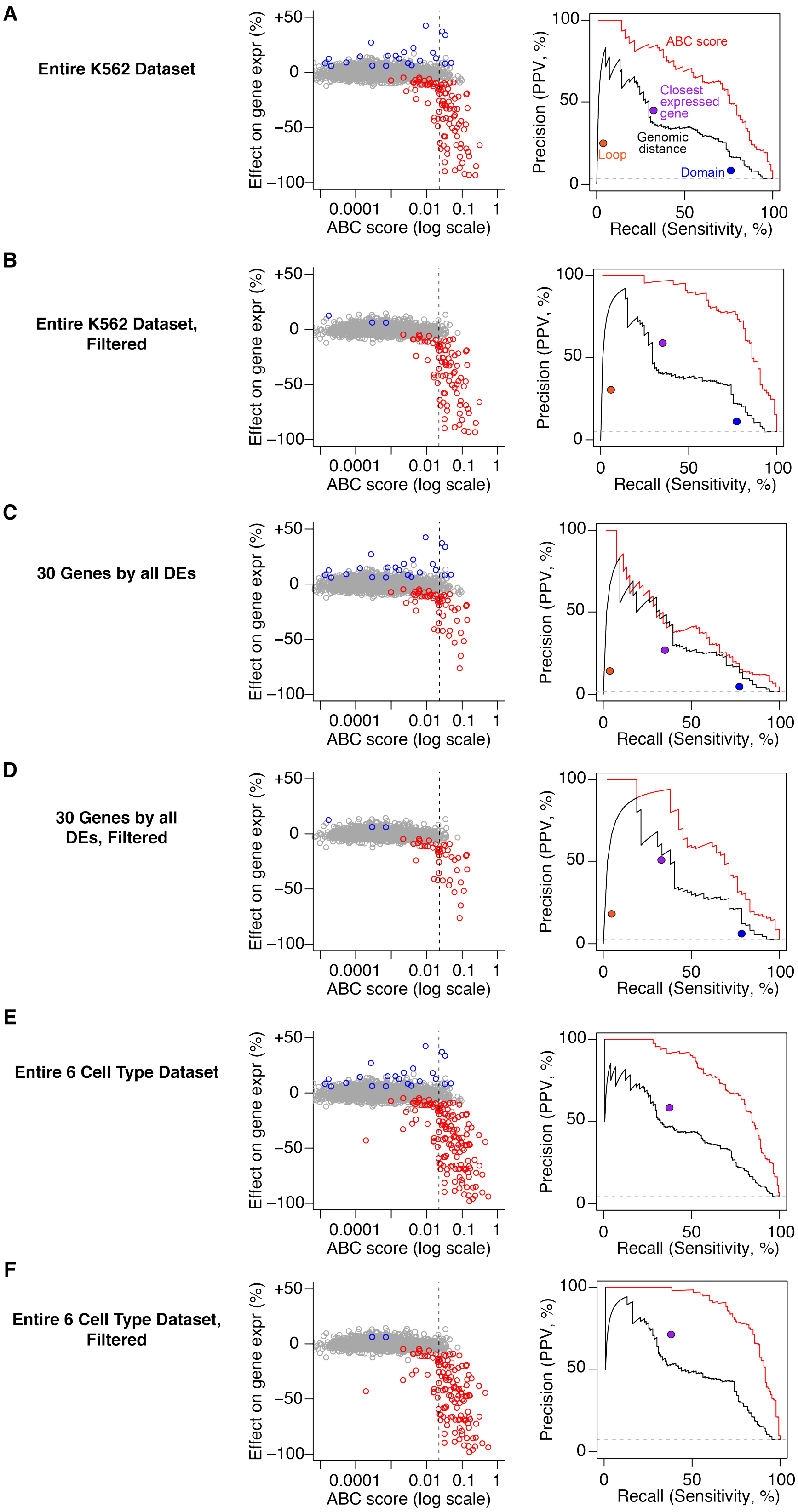


Figure B-13. (Legend on next page).

Figure B-13 (Continued) Performance of the ABC score after filtering ubiquitously expressed genes, 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. **(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).

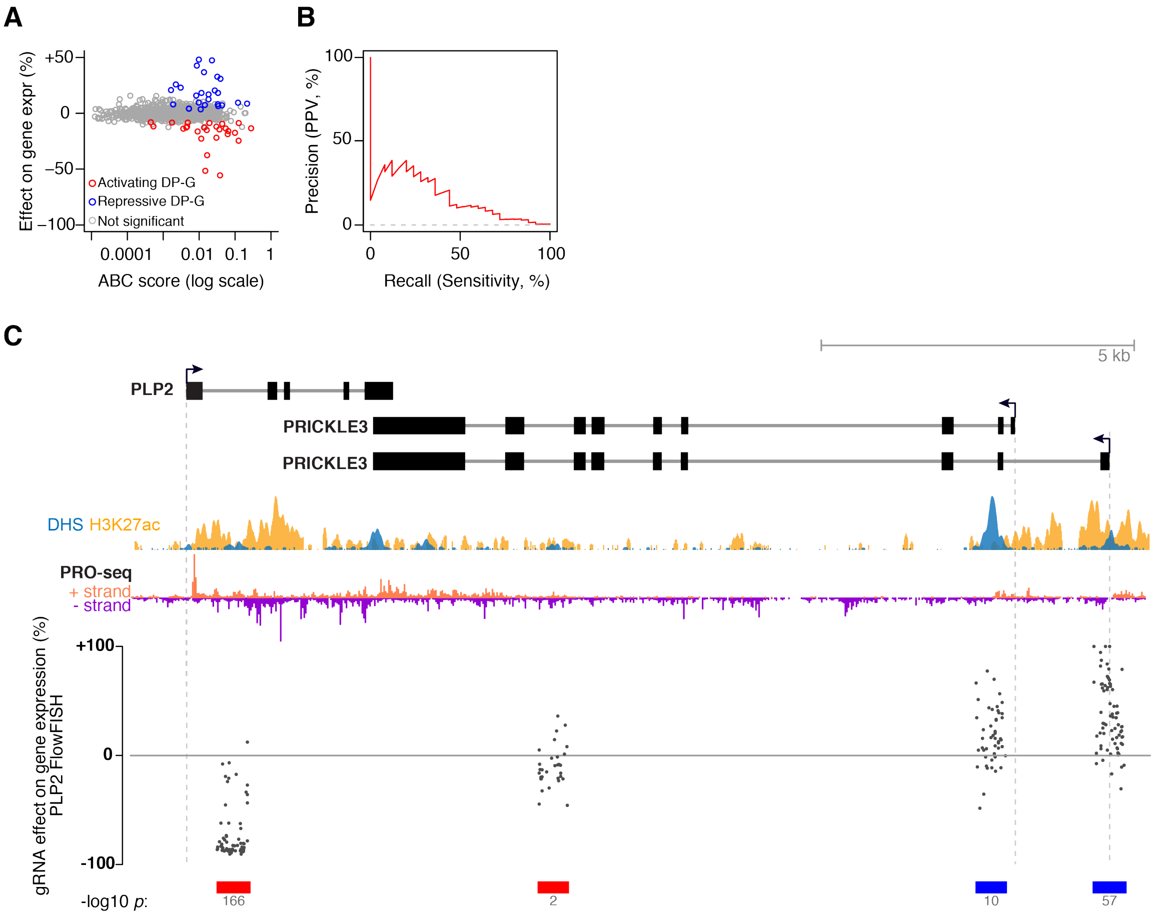


Figure B-14. Effects of promoters on nearby genes. (A) Comparison of ABC scores (predicted effect) with observed changes in gene expression upon CRISPR perturbations in K562 cells. Each dot represents one tested DP-G pair (where the element itself is a promoter). **(B)** Precision-recall curve for classifying 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.

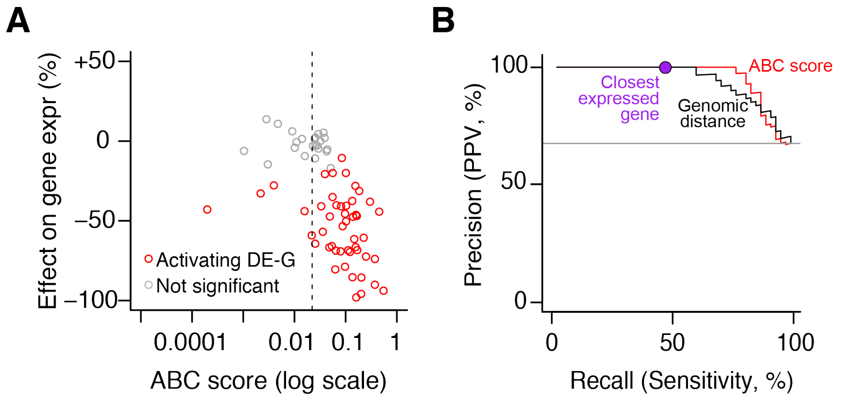


Figure B-15. The ABC model generalizes across cell types.

(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 Methods. See Fig 4 for data in these types using a lenient power filter of 80% power to detect a 50% effect on gene expression).

References

1. J. E. Phillips, V. G. Corces, CTCF: master weaver of the genome. *Cell* **137**, 1194-1211 (2009).

2. J. M. Engreitz *et al.*, Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452-455 (2016).

3. L. T. M. Dao *et al.*, Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet* **49**, 1073-1081 (2017).

4. D. I. Martin, S. Fiering, M. Groudine, Regulation of beta-globin gene expression: straightening out the locus. *Curr Opin Genet Dev* **6**, 488-495 (1996).

5. K. E. Shearwin, B. P. Callen, J. B. Egan, Transcriptional interference--a crash course. *Trends Genet* **21**, 339-345 (2005).

6. V. R. Paralkar *et al.*, Unlinking an lncRNA from Its Associated cis Element. *Molecular cell* **62**, 104-110 (2016).

7. S. W. Cho *et al.*, Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell* **173**, 1398-1412.e1322 (2018).

8. T. M. Townes, R. R. Behringer, Human globin locus activation region (LAR): role in temporal control. *Trends Genet* **6**, 219-223 (1990).

9. E. Lieberman-Aiden *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

10. C. P. Fulco *et al.*, Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773 (2016).

11. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

12. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

13. L. A. Gilbert *et al.*, CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451 (2013).

14. T. S. Klann *et al.*, CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat Biotechnol* **35**, 561-568 (2017).

15. J. C. Ulirsch *et al.*, Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530-1545 (2016).

16. A. Wakabayashi *et al.*, Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *Proc Natl Acad Sci U S A* **113**, 4434-4439 (2016).

17. P. I. Thakore *et al.*, Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature methods* **12**, 1143-1149 (2015).

18. S. J. Liu *et al.*, CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, (2017).

19. X. Chen *et al.*, Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117 (2008).

20. S. D. Moorthy *et al.*, Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome research* **27**, 246-258 (2017).

21. S. Xie, J. Duan, B. Li, P. Zhou, G. C. Hon, Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular cell* **66**, 285-299.e285 (2017).

22. Y. Li *et al.*, CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**, e114485 (2014).

23. H. Y. Zhou *et al.*, A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes & development* **28**, 2699-2711 (2014).

24. J. Xu *et al.*, Developmental control of polycomb subunit composition by GATA factors mediates a switch to non-canonical functions. *Molecular cell* **57**, 304-316 (2015).

25. S. Blinka, M. H. Reimer, Jr., K. Pulakanti, S. Rao, Super-Enhancers at the Nanog Locus Differentially Regulate Neighboring Pluripotency-Associated Genes. *Cell reports* **17**, 19-28 (2016).

26. N. Rajagopal *et al.*, High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**, 167-174 (2016).

27. J. Huang *et al.*, Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat Commun* **9**, 943 (2018).

28. R. Tewhey *et al.*, Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529 (2016).

29. X. Wang *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).

30. S. Spisak *et al.*, CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. *Nat Med* **21**, 1357-1363 (2015).

31. M. R. Mumbach *et al.*, Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* **49**, 1602-1612 (2017).

32. K. Musunuru *et al.*, From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-719 (2010).

33. D. R. Fuentes, T. Swigut, J. Wysocka, Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *Elife* **7**, (2018).

34. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527 (2016).

35. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).

36. L. J. Core *et al.*, Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-1320 (2014).

37. D. J. Hazelett *et al.*, Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet* **10**, e1004102 (2014).

38. M. R. Corces *et al.*, Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193-1203 (2016).

39. M. C. Canver *et al.*, Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nat Genet* **49**, 625-634 (2017).

40. N. C. Durand *et al.*, Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98 (2016).

41. J. D. Buenrostro, B. Wu, H. Y. Chang, W. J. Greenleaf, ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, 21 29 21-29 (2015).

42. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, (2013).

43. J. Feng, T. Liu, B. Qin, Y. Zhang, X. S. Liu, Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728-1740 (2012).

44. Q. Cao *et al.*, Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat Genet* **49**, 1428-1436 (2017).

45. G. Li *et al.*, Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98 (2012).

46. S. Whalen, R. M. Truty, K. S. Pollard, Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48**, 488-496 (2016).

47. A. R. Forrest *et al.*, A promoter-level mammalian expression atlas. *Nature* **507**, 462-470 (2014).

48. D. Alpern, V. Gardeux, J. Russeil, B. Deplancke, Time- and cost-efficient high-throughput transcriptomics enabled by Bulk RNA Barcoding and sequencing. *bioRxiv*, (2018).

49. J. Zhu, F. He, S. Song, J. Wang, J. Yu, How many human genes can be defined as housekeeping with current expression data? *BMC genomics* **9**, 172 (2008).

50. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends Genet* **29**, 569-574 (2013).

51. B. Li *et al.*, A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. *Sci Rep* **7**, 4200 (2017).

52. R. C. Gentleman *et al.*, Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).

53. A. L. Sanborn *et al.*, Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-6465 (2015).

54. Q. Gong, A. Dean, Enhancer-dependent transcription of the epsilon-globin promoter requires promoter-bound GATA-1 and enhancer-bound AP-1/NF-E2. *Mol Cell Biol* **13**, 911-917 (1993).

55. M. Rylski *et al.*, GATA-1-mediated proliferation arrest during erythroid maturation. *Mol Cell Biol* **23**, 5031-5042 (2003).

56. Y. Woon Kim, S. Kim, C. Geun Kim, A. Kim, 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**, 6944-6955 (2011).