Efficiently searching for enhancers and their target genes in the human genome

Molly Gasperini

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Jay Shendure, Chair

Maitreya Dunham

Stan Fields

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Efficiently searching for enhancers and their target genes in the human genome

Molly Jeanette Gasperini

Chair of the Supervisory Committee:
Professor Jay Shendure
Genome Sciences

A single 3 billion letter genome contains the instructions for the 37 trillion diverse cells that make up one human. To accomplish this, the ~21,000 human genes are expressed and perform function in highly specific combinations per cell. Yet, only 2% of the genome codes for genes. The remaining 98% is made up of highly complicated, loosely patterned DNA referred to as "noncoding sequence". Functional noncoding DNA elements (first termed "enhancers" in 1981) regulate cell-type specific gene expression. Like genes, enhancers disruption is known to cause genetic disease. How can we efficiently search for enhancers within the expansive noncoding genome? The new genome engineering technology CRISPR/Cas9 enables parallelized pooled perturbations to efficiently screen enhancers and the genes they target. In this dissertation, I will cover my development of new pooled methods to screen the noncoding genome.

In the first chapter, I introduce the motivation for these methods, the history of enhancers, their current definitions, and emerging technologies for enhancers' at-scale characterization. In

Chapter 2, I describe a method we devised to scan thousands of CRISPR-induced kilobase-sized deletions ("ScanDel") across a desired noncoding region, programming one unique deletion per cell in a pool and phenotyping them in multiplex by pooled functional selection. However, ScanDel and its contemporaries are limited to evaluating enhancers for their effect upon a single gene. In Chapter 3, I describe a second method designed to overcome this limitation, in which large numbers of CRISPR perturbations are introduced to each cell, followed by single-cell transcriptome sequencing to read out their effect upon any transcript. With this method, we effectively evaluated >70,000 potential enhancer-target gene relationships in one experiment. In Chapter 4, I describe a potential path forward to cataloguing all enhancers in the human genome, and how we might do the same for noncoding variants in human disease.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

When editing the word-count of acknowledgements for a paper, my advisor Jay Shendure noted “If this were the Oscars, the music would’ve played you off the stage by now.” But as I have so many people to thank who have brought me to the end of my 21st grade in school (and dissertations have unlimited word count), I am going to do my best to include most of them here.

Growing up, I was planning on being a novel writer and have only stumbled upon this incredible career by the efforts of many early transformational public-school teachers. In 10th grade biology, my teacher Ms. Tracy Stoops first taught me how the human body responds to a cold and I was hooked. She not only excited me about science but empowered me to envision a future in which I exceled in it. By college, I was learning from most brilliant professors – Linda Martin-Morris, Mary-Pat Wenderoth, Scott Freeman, M.K. Raghuraman – and was inspired by this cutting edge science to persevere through the intimidating, gigantic introductory biology classes of UW.

However, early on in college, I was still too intimidated to apply to labs as an undergraduate researcher, so took a job as a secretary. Yet the kind people in this office quickly saw my heart (and skills!) weren’t destined for administrative greatness, so encouraged me to apply to biology research labs. At the recommendation of a generous seminar teacher Chris Tachibana, I applied to Mary-Claire King’s lab and was lucky enough to be accepted. In Mary-Claire’s lab, I discovered the existence of a completely exciting, challenging, deeply interesting, and dynamic career path that was previously unknown to me. The women of the King lab (Mary-Claire, Caitlin Rippey, Cailyn Spurrell, Sarah Pierce) taught me not only how utterly cool this job was, but were the first in my life to paint a portrait of a scientist as a fierce, passionate, hilarious, kind, brilliant woman.

With the help of the Undergraduate Research Program, I learned you could even be paid to do this job! So after undergrad, I sought a position as a research technician in Louis Kunkel's research lab in Boston. There, I was able to pursue science full time, and learn from the expansive Boston research community. From Genri Kawahara, I learned the joy of scientific pursuit. From Jamie Marshall, I learned how to work hard, set a rigorous bar for your science, and never compromise on it. From Dan Yuan and Michelle Graff, I learned how friends can make an exciting work environment even more enjoyable. From Martha Zepeda, I learned to true to yourself despite the high-pressure intensity of an elite scientific community.

Though I almost stayed in Boston for graduate school, I returned home to UW to join the Genome Sciences PhD program. I was excited but wary of grad school, and didn't at all expect to have the terrific adventures and happiness of the past five years. I have garnered countless scientific mentors here, and in particular want to thank: my dissertation committee of Maitreya Dunham, Josh Akey, Jennifer Nemhauser, and Stan Fields for their many hours of scientific and career advice; Bob Waterston for modeling the epitome of a gracious senior scientist and sharing that graciousness with each student; the GS class of 2014; the entire Trapnell lab; and Kenny Matreyek for teaching me to never believe anything unless you see the data yourself and providing reflective, honest conversations on a career in science. The UWGS departmental staff (GSIT, purchasing, custodial staff, Sandra Pennington, the autoclaving staff) has additionally generously enabled me and the rest of us nerds to pursue research with no extra hassle. In particular, I want to thank Brian Giebel, who received his first of hundreds of emails from me back in 2011 as an undergrad. I have also been lucky to be a part of a wider science communication community (ComSciCon National and PNW; Pacific Science Center Science Communication Fellows) that has always reminded me of my broader passion for science when research was extremely stressful.

I have additionally made many incredible friends who have helped me both process and distract from the pressures of school. Thank you, thank you to Max Dougherty, Rocío Acuña-Hidalgo, José McFaline-Figueroa, Nellie Cruz, Sofía McFaline-Figueroa, Kirsten Cooper, Matt Pantoja, Hubert, Herman, Hannah Pliner, Bennett Ng, Robert Frankel, Ron Hause, Bobeya Krishnek, and The Peeps. In particular; Melissa Chiasson for her jokes, her intelligence, her dog, and her dancing; Kiana Mohajeri for empowering me to let my intensity flag fly and achieve with confidence; Stella for being the best therapy cat I could've wanted; and David Bergsman, who edited my first emails from applying to the King lab and with whom I still debrief every major career event.

The Shendure lab has been a better fit than I ever expected for my graduate school lab. Overall, I couldn't have asked for a better advisor for me than Jay. I am extremely grateful for his patience, optimism, kindness, focus, and brilliance. I never thought I could have done so much, learned so much, and produced so much as a graduate student, and I owe most of that to the resources, mentorship, and push Jay has given. Thanks for putting up with me, appreciating Prince/science joke mashups, and proving that high powered academic science can indeed be conducted in a positive, fun, happy environments.

However, Jay is just one busy guy, and the Shendure lab as a whole has provided me with the remaining ~75% of my graduate education. In particular, I need to thank Greg Findlay, who throughout grad school taught me fundamentals of the lab's science, that taking intelligent risks often leads to success, and how to elevate your scientific thinking. I will always think "what would Greg criticize about this" when I am developing a new idea and the answer will always improve it. Riza Daza and Beth Martin have taught me science and life skills I use every day. Lea Starita has taught me how to slay and given me an extensive introductory education in high-throughput science. I feel so fortunate that Sam Regalado, Silvia Domcke and Jacob Tome joined the lab and

were up for working on projects with me! Darren Cusanovich taught me fundamentals of human genetics and how to maintain kindness and rigor no matter the pressures. I also need to thank Anna Minkina and Seungsoo Kim for their support and putting up with me as a deskmates for years!

I want to thank Charlie Lee for his professional support. When a sequencing run fails, there's nothing more helpful than to hear a therapeutic soliloquy on how each Miseq run is just a grain of sand in the beach of time. I was so lucky to sit by Charlie early in graduate school, where he offered me countless hours of support by being himself - hilarious, calm, passionate, and smart. However, I most want to thank Charlie for his friendship. You are a part of the family now (sucker!).

Through grad school and before, my mom, dad, sister, Aunt Peggy and my brother-in-law have served as my own personal "Fab Five". The extended Gasperini/Erion/O'Bryan horde make up an unbelievably supportive fan club. Their high expectations for striving for personal happiness first (and career success within that) have kept me balanced in a way that I've realized is extremely rare in science. The early skills they taught me – how to always uphold kindness and compassion, how to get a word in edgewise in a loud group, how to tell a good story or it's unlikely people will listen, how to dress for a fancy presentation, how to cook a meal in 15 minutes after a long day at work, how to be curious, how to think deeply, how to work hard – have enabled any academic success I've had. My family-in-law makes up another set of extremely generous and supportive cheerleaders. In addition to support, my mother- and father-in-law Drs. Roger and Leslie Wilson-Hill have been models for how to live a great life while pursuing a great scientific career. My brother-in-law Ian has also been a wonderful friend, and president of our restaurant club.

And last, Andrew – you're pretty great. I don't know how people go through school without you on their team. I don't know how people go through life without you on their team! I love you.

DEDICATION

This dissertation is dedicated to every public-school educator who contributed to the stellar public education I received across Echo Lake Elementary, Einstein Middle, Shorewood High School, and the University of Washington. In particular, I thank my 10th grade biology teacher Ms. Tracy Stoops, who changed my life.

Chapter 1. INTRODUCTION

The human genome encodes a human life. It holds instructions for a person's development from a single-celled zygote into an adult, functions like the repair of a scraped knee or the digestion of a Thanksgiving meal, and how to make the cells that make a smile, a brain, or a heart. The ability of a single sequence of DNA to encode so much life for every human on the earth is nothing short of miraculous. Understanding our genome is critical not only for understanding disease, but also what it means to be human.

We understand many major tenets of how the genome encodes a human. Function is largely encoded in the ~21,000 human genes. Yet, not all genes are active all the time in every cell. They are turned on or off in highly complicated circuits that control development, cell type, and response to the environment. How are these genes regulated? Additionally, only 2% of the genome codes for genes. What is the function of the remaining "noncoding" 98%?

35 years before the work of this dissertation began, small regions of noncoding DNA - called 'enhancers' - were shown to be capable of controlling when genes were turned on or off (Banerji, Rusconi, and Schaffner 1981; Moreau et al. 1981). In the ensuing decades (see Chapter 1.2), enhancers have been found to be only a single class of a diverse suite of noncoding gene regulatory elements (e.g. silencers, insulators, promoters, locus control regions). This suite functions in a highly specific cell-type manner to determine when and where genes are active. For simplicity in this dissertation, I focus on enhancers, though many of the assays and concepts described could be applied to all regulatory elements. However, unlike genes, enhancers carry no strict structure,

which makes them hard to identify by sequence alone. How then can we find all cell-type specific enhancers within the 2.9 billion letters of the noncoding genome?

To understand the function of DNA, classical genetics relies on the perturbation of a sequence followed by the documentation of what function is disrupted (Sullivan 1993). Over the past 40 years, deletion of noncoding sequence has gleaned many insights into where enhancers lie and what genes they control (see Chapter 1.2). Testing enhancers in their native cellular and genomic context is critical for capturing the timepoint at which they act and the gene they target. This creates a vast search space within which enhancers need to be studied, as we need perturbations of all possible enhancer-gene pairs multiplied by all possible cell types and timepoints. The throughput of earlier methods for genomic sequence perturbation prevent efficient search of this space, meaning only a handful of enhancer-gene pairs have been previously validated.

Two years before the initiation of this dissertation work, publication of a new genome engineering technology enabled significant headway into this noncoding search space. Repurposing of the nuclease of a bacterial anti-viral immune system ("Clustered regularly interspaced short palindromic repeats" or CRISPR) enabled the targeted cutting of any chosen genomic sequence (Gasiunas et al. 2012; Jinek et al. 2012). For all implementations described in this dissertation, the CRISPR system works as follows: a programmed synthetic 'guide-RNA' (gRNA) carries the perturbation-inducing Cas9 endonuclease to DNA that matches the gRNA's nucleic acid sequence (Fig. 1.1A). The system was quickly demonstrated to be active in human cells as well (Cong et al. 2013; Mali et al. 2013), priming the technology's use in assaying the human genome.

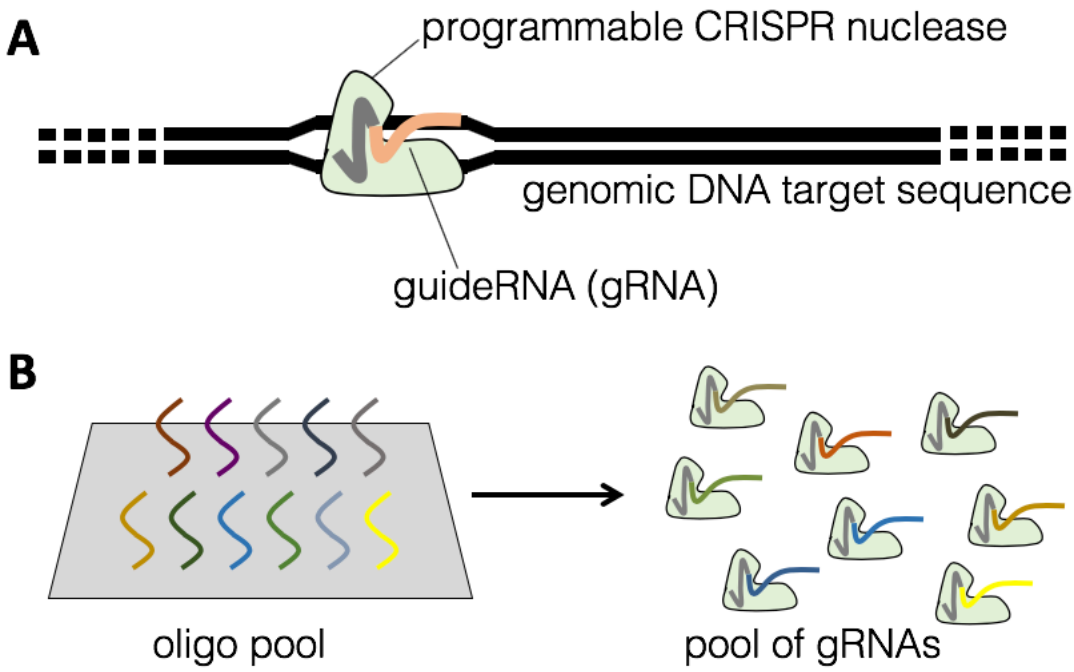


Figure 1.1 The CRISPR/Cas9 genome engineering system is easily parallelizable.

(A) The synthetic CRISPR system consists of a guideRNA (gRNA) that directs a Cas9 nuclease enzyme to a specified genomic target. (B) Hundreds of thousands of gRNAs are easily generated in multiplex by ordering pools of synthesized oligonucleotides that encode the gRNA-programming sequence. This enables cloning and delivery in one pot.

Though it is not the first programmable nuclease available, CRISPR/Cas9's transformative utility is derived from its gRNA. Previous programmable nucleases were only usable in low throughput. The gRNA can be easily reprogrammed to target across the genome by a straightforward modern genomics toolbox: a synthesized 20-base pair oligonucleotide to program the gRNA sequence, recombinant molecular cloning to express it, and delivery to tissue cultured cells by widely available transfection or transduction reagents. By inputting hundreds of thousands of gRNA sequences into the first step of this protocol, it is relatively simple to produce large pools of tissue

cultured cells that collectively hold up to hundreds of thousands of diverse perturbations (Fig. 1.1B). The work of this dissertation began just when pooled CRISPR methods came online. As CRISPR enables disruption of sequence in its native cellular and genomic context, the field was primed for new methods to disrupt enhancers in high-throughput.

In this dissertation, I will cover two advances for pooled CRISPR screens of enhancer-gene pairs. In Chapter 1, I introduce this work's motivation, the history of enhancers, current definitions, and emerging technologies for their discovery, validation, and at-scale characterization. In Chapter 2, I describe a proof-of-concept study in which we used ScanDel to program 4,342 overlapping 1- and 2- kilobase (Kb) deletions that covered 206 Kb centered on *HPRT1*, the gene underlying the Mendelian monogenic disorder Lesch-Nyhan syndrome. Yet, "monogenic" screens like ScanDel can only test enhancers for their control of one gene. In Chapter 3, I describe our method designed to overcome this limitation by testing enhancer perturbations against the whole transcriptome. We adapted the CRISPR enhancer screen framework to a single-cell RNA-seq readout and implemented more efficient use of these transcriptomes by multiplexing the gRNA per cell. We programmed CRISPR perturbations for 5,920 candidate enhancers in the K562 cell line, tested for differential expression of all expressed genes within 1 megabase of each candidate enhancer, and identified 664 enhancer-gene pairs. In Chapter 4, I describe how current methods present a path forward to cataloguing all enhancers in the human genome, and how we might make progress on doing the same for noncoding variants in human disease. This dissertation work aims to contribute to the comprehensive elucidation of the gene-regulatory landscape of the human genome. I hope to convince you that pooled perturbation methods are poised to facilitate this goal.

1.1 The History of Defining an Enhancer

What is an "enhancer"? The term 'enhancer' first appeared in 1981. By this point in time, gene expression was already thought to be controlled by regulatory proteins (Jacob and Monod 1961) that had been demonstrated to be DNA-binding (Mark Ptashne 1967). Where and how do these proteins bind to control expression? Open chromatin regions were suspected to have a role (Axel, Cedar, and Felsenfeld 1973; Weintraub and Groudine 1976), and distal, cell-type specific open chromatin regions had already been identified far from genes' promoters (Stalder et al. 1980). Yet, these distal sites had not yet been shown to control gene expression.

In 1981, these concepts culminated in the first demonstration of a noncoding DNA sequence that distally 'enhanced' a gene's expression (Moreau et al. 1981; Banerji, Rusconi, and Schaffner 1981). On an episomal reporter vector (Fig. 1.2B), a noncoding region of the Simian Virus 40 viral genome could increase expression at a distance remote from the reporter gene's promoter and independent of its orientation. From this came an original definition of an enhancer that is still quoted today: "the transcriptional enhancer element could act in either orientation at many positions… even downstream from the transcription initiation site."(Banerji, Rusconi, and Schaffner 1981). Using a similar reporter vector method, the first mammalian and cell-type specific enhancer was soon identified within the IgH locus (Mercola et al. 1983; Banerji, Olson, and Schaffner 1983; Gillies et al. 1983). Cell-type specific regulatory sequences were next transgenically shown to enhance a reporter *in vivo* (Hanahan 1985).

A comprehensive second-generation definition of an enhancer emerged over the next 10 years: enhancers are free of nucleosomes (measured by unprotected DNA's hypersensitivity to DNaseI)

(Gross and Garrard 1988; Tuan et al. 1985) but proximal to nucleosomes with transcriptional-activity associated histone modifications (Hebbes, Thorne, and Crane-Robinson 1988; D. Y. Lee et al. 1993; Hebbes et al. 1994); composed of protein binding sequence motifs (Serfling, Jasin, and Schaffner 1985); bound by DNA binding proteins that encourage transcription (Ikuta and Kan 1991; Forsberg and Westin 1991); and likely to loop in 3D space to access their target promoters (M. Ptashne 1988; Schleif 1992). However, this definition of an enhancer was generated in low throughput from a small number of individual examples in a handful of cell types. By the end of the century, it was not yet validated for the majority of enhancers genome-wide.

Genome-wide enhancer profiling was finally enabled by the technological advances of the human genome sequence, DNA synthesis, and high throughput sequencing. With these advances, open chromatin was measured en masse by pairing DNAse I hypersensitivity with array-based or sequencing-based measurements (Sabo et al. 2006; A. P. Boyle et al. 2008; Hesselberth et al. 2009; Thurman et al. 2012). Chromatin immunoprecipitation followed by sequencing (ChIP-seq) enabled mapping of histone modifications (Barski et al. 2007; Mikkelsen et al. 2007; Heintzman et al. 2007) and regulatory protein binding (Robertson et al. 2007; Johnson et al. 2007; Visel et al. 2009) profiles across the genome. Whole transcriptome RNA-sequencing identified noncoding transcription of enhancers ("e-RNAs"; De Santa et al. 2010; Kim et al. 2010; Andersson et al. 2014). These functional genomics techniques have been applied at consortia-scale to characterize the regulatory profiles of hundreds of mammalian cell types and conditions (ENCODE Project Consortium 2012; Roadmap Epigenomics Consortium et al. 2015; Stunnenberg, International Human Epigenome Consortium, and Hirst 2016). In turn, this era of technology has given rise to a current consensus definition of enhancers: regions of open chromatin, looping in 3D space to

reach promoters, flanked by histones carrying H3K27ac and/or H3K4me1 modifications, bound by transcription factors (TFs:, e.g. p300), that can be latent, primed, or active (Fig. 1.2A; Spitz and Furlong 2012; Shlyueva, Stampfel, and Stark 2014).

However, the definition outlined above requires an update. First, it is built from biochemically descriptive data rather than true genetic perturbations. Second, this definition does not inform what gene(s) a given candidate enhancer regulates, nor the degree of activation conferred upon that gene. Third, they largely do not inform the impact of human variants, though thousands of noncoding variants have been associated with human disease (Maurano et al. 2012; Finucane et al. 2015) and gene expression (Consortium and GTEx Consortium 2017; Stranger et al. 2012). Fourth, this does not confirm the role of 3D chromatin architecture in gene regulation. And last, use of this definition relies on a bulk, 'one size fits all' model for gene regulation, but contrasting examples can be found for each piece of enhancer-characterization evidence. These points highlight the limitations of the technologies currently used to define enhancers.

In the last decade, multiple new methods have emerged to overcome these limitations: single-cell ("sc") methods to annotate enhancers in individualized cell types (Cusanovich et al. 2015; Buenrostro et al. 2015); higher resolution chromosome conformation capture methods to finely map 3D enhancer-promoter loops (Eagen 2018); massively parallel reporter assays (MPRAs, Fig. 1.2B) to dissect or trap enhancer activity on a reporter vector (Inoue and Ahituv 2015); and high-throughput CRISPR screens to directly perturb enhancers in their genomic context and tie them to their target genes (Fig. 1.2C-D; Klein, Chen, et al. 2018). These represent a new era of human enhancer characterization.

1.2 WHAT IS AN ENHANCER?

How do enhancers control gene expression within the cell? By coordinating a set of cell-type and condition specific transcription factors, they support assembly of the transcription initiation complex at a promoter (Thanos and Maniatis 1995) and encourage RNA polymerase to begin transcription. This activity is thought to be controlled in a cell type specific manner by expression level of the enhancer's interacting TFs. Many current models exist as to how enhancers are activated and reach their target promoter. These include: enhancer tracking, linking, short or long range looping, transcription factories or hubs, transcriptional bursting, and/or phase separation (reviewed in Furlong and Levine 2018).

Currently, annotation or perturbation evidence is available to classify a sequence as an enhancer. Annotation features include: sequence level annotations (conservation and TF binding site motifs); 1D biochemical annotations (open chromatin; H3K27Ac and H3K4me1 modifications on flanking histones for active enhancers; H3K4me1 and HK27me3 for poised enhancers; closed chromatin that has been pre-marked by H3Kme1 for primed enhancers; direct binding of TFs or secondary binding of cofactors such as p300); or 3D biochemical annotations (nuclear spatial proximity as read by 3C, 4C, 5C, or Hi-C) (Fig. 1.2A; Shlyueva, Stampfel, and Stark 2014). Perturbative data includes: episomal validation (Fig. 1.2B; reporter vector or MPRA based demonstration of activation of a minimal promoter and reporter gene); in-genome validation (Fig. 1.2C-D; cellular, genomic epigenetic or genetic perturbation [e.g. CRISPR] resulting in change in target gene expression); or *in vivo* validation (transgenic reporters or direct epigenetic or genetic sequence perturbation, usually in mouse models).

Though enhancers are enriched for these characteristics, counterexamples can be found for each. Not all distal conserved elements are enhancers (Pennacchio et al. 2006) and far more of the gene-distal noncoding genome is annotated as a regulatory element (up to 80% [ENCODE Project Consortium 2012]) than is conserved (1-3.5% [Lindblad-Toh et al. 2011]). TF sequence motifs alone are not perfectly informative as only a portion of the available TF binding sites in the genome are occupied in any one cell type (Spitz and Furlong 2012). Though enriched, histone modification and cofactor binding is not completely predictive of enhancer activity (Visel et al. 2009). Many enhancers are spatially proximal to their target promoter in 3D (Rao et al. 2014), but exceptions exist (Williamson et al. 2012; Ghavi-Helm et al. 2014). Genes can be controlled by a single enhancer or by many in concert (Osterwalder et al. 2018) and vice versa, individual enhancers can target one or multiple gene(s). Some enhancers sit in clusters of a handful (Levings and Bungert 2002) or several hundred enhancers ('super enhancers' [Hnisz et al. 2013; Whyte et al. 2013]) but others are alone. Biochemical annotations do not always correlate with activity on an MPRA.

As enhancers carry contrasting characteristics, it is likely inappropriate to use a "one size fits all" binary definition to classify sequence as an enhancer or not. Sub-classes of gene regulatory circuits may utilize multiple sub-types of enhancers that are potentially defined by biological role (e.g. housekeeping versus developmental genes) or mechanism (e.g. redundantly versus singularly acting enhancers). In order to encompass this diversity, a modern definition of an enhancer could be simple: an enhancer is a noncoding sequence that acts to directly increase the transcription of a distal target gene. I propose that to truly classify a sequence as an enhancer, it must be deleted in the relevant cell-type (either alone or in combination) with an observed decrease in an allelic target gene's expression. Any sequence without this piece of evidence is only a candidate enhancer.

Yet how might we generate a catalogue of all the enhancers of the human genome given the available data and the technological difficulty of deleting every candidate? Layers of enhancer characterization could be used to weakly, moderately, or strongly support a sequence as an enhancer. This will require reconciling contradictory evidence (e.g. an element supported by biochemical annotations but is negative in an MPRA or CRISPR-based assay). Given that most assays are applicable either to large numbers of cells and/or in cell culture only, we will have to contend with cell type heterogeneity of *in vivo* tissues. However, the emerging suite of enhancer characterization assays address many of these challenges.

1.3 Technologies for Discovering and Validating Enhancers

1.3.1 *Current Technology and Limitations*

Sequence features: Annotation of sequence level features is informative for distinguishing where enhancers might lie but is neither necessary nor sufficient. Conservation hotspots can further support the functional candidacy of a region (Pennacchio et al. 2006), but given the cell-specificity of gene regulation, not all enhancers are conserved (Blow et al. 2010; Schmidt et al. 2010). Overlaying these regions with transcription factor binding motifs can add further support (Kulakovskiy et al. 2013), but not all TF's motifs are known (Lambert et al. 2018), and presence of a motif does not confirm binding in a specific cell type (Spitz and Furlong 2012). Automatic sequence-based enhancer annotation (Van Loo and Marynen 2009) is prevented from perfectly predicting activity, as sequence alone is removed from its defining nuclear and cellular context.

Biochemical annotation: Cell type-specific enhancer annotations can be derived from biochemical assays that mark enhancer-flanking histone modifications or where TFs may be bound (via histone

modification/TF ChIP-seq), open chromatin (DNase-seq, MNase-seq, ATAC-seq), DNA methylation (bisulfite sequencing), and the initiation and abundance of transcription (RNA-seq, PRO-seq) (Fig. 1.2A). These data have been collected for diverse cell types for cell-specific profiling of enhancer activity. However, these 'bulk' assays return a median truth in a pool of diverse cells states (Trapnell 2015), rather than single-cell information. Additionally, it remains unknown what percent of biochemically annotated enhancers are unintended technical false positives (Teytelman et al. 2013; Worsley Hunt and Wasserman 2014; Jain et al. 2015) or a product of consistent biochemistry but transient function (Diao et al. 2016). Lastly, these technologies only map candidate enhancers in "1D" linear space rather than with 3D DNA nuclear organization information.

Limitations: Sequence alone and biochemical annotation has nominated over a million candidate human enhancers (Coppola, C Ramaker, and Mendenhall 2016), narrowing down the search space from the entire expansive noncoding genome. However, as they do not pair enhancers to their target genes, these annotations cannot truly validate enhancer function. A critical need remains for assays that validate enhancer-gene pairs in high-throughput.

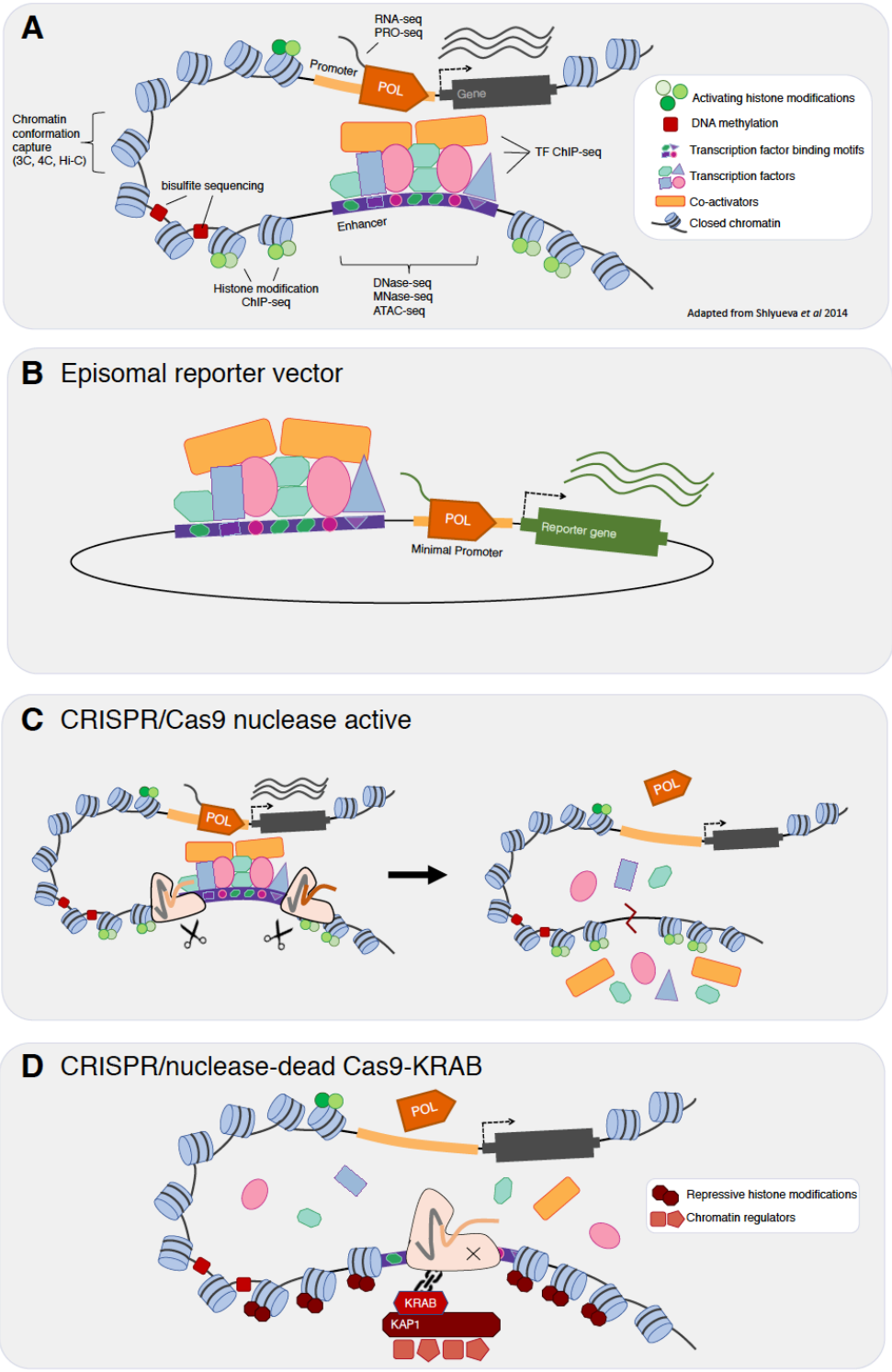


Figure 1.2 Diagram of an active enhancer and assays for enhancer study.

(A) Diagram of an enhancer labeled with biochemical, descriptive assays (adapted from Shlyueva, Stampfel, and Stark 2014). (B) Diagram of measuring enhancer activity on an

episomal reporter vector (as in MPRAs or Banerji, Rusconi, and Schaffner 1981). (C) In-genome CRISPR deletion of an enhancer via long deletion by pairs of gRNAs guiding nuclease active Cas9. (D) In-genome perturbation of an enhancer by nuclease-dead Cas9 tethered to the Kruppel-associated Box repressor domain (Lupo et al. 2013; Thakore et al. 2015).

1.3.2 *Emerging Technology – Annotation*

3D conformation mapping: A long hypothesized model of gene regulation dynamics supports looping of the enhancer in 3D space to access its target promoter (M. Ptashne 1988; Schleif 1992). Recently, scaling of chromosome conformation capture ("3C") methods has provided high-resolution 3D spatial profiling of the human genome. For this, genomic DNA fragments are molecularly linked in the nucleus to their neighboring fragments, and links are deep sequenced to identify proximal fragments (Lieberman-Aiden et al. 2009). These studies identified large-scale compartments of genome organization (including "topologically associated domains" [Nora et al. 2012; de Laat and Duboule 2013; Dixon et al. 2012; Sexton et al. 2012], AB compartments [Lieberman-Aiden et al. 2009]) and enrichment for proximity in enhancer-promoter pairs (Rao et al. 2014). The "3C" methods have further been paired with biochemical assays to enrich for putatively functional pairs (e.g. ChiA-PET [M. J. Fullwood and Ruan 2009; Melissa J. Fullwood et al. 2009], HiChIP [Mumbach et al. 2016], PLAC-seq [Fang et al. 2016], DNase Hi-C [Ma et al. 2015]). If proximity is used as a metric for gene regulatory activity, these provide the first genome-scale datasets of functionally-supported candidate enhancer-gene pairs. Yet, proximity is sometimes maintained even when the gene or enhancer is inactive (Ghavi-Helm et al. 2014; Williamson et al. 2016), or active enhancers gain mobility (Gu et al. 2018). It remains to be determined whether proximity is a causal, residual or required mechanism of gene regulation.

Single-cell: Current 'bulk' biochemical assays have so far returned a median enhancer profile of input cells. Heterogeneous cell populations or tissues have either been ignored or painstakingly physically dissected (Roadmap Epigenomics Consortium et al. 2015). Single-cell enhancer profiling technologies promise higher resolution (Trapnell 2015). scATAC-seq (Cusanovich et al. 2015; Buenrostro et al. 2015) has enabled the *in vivo* profiling of open and closed chromatin across cell types of entire organs and organisms (Cusanovich, Hill, et al. 2018; Cusanovich, Reddington, et al. 2018). scMNase-seq (Lai et al. 2018) can study single-cell nucleosome positioning. Single-cell '3C' assays have been restricted by resolution (Ramani et al. 2017; Flyamer et al. 2017; Nagano et al. 2017, 2013) or scale (Tan et al. 2018), but co-assays may overcome these limitations (D. S. Lee et al. 2018). Single-cell assays for TF binding (e.g. scChIP-seq [Rotem et al. 2015] or scCut&Run [Hainer et al., n.d.]) are still emerging. These methods have the potential to replace conventional 1D biochemical annotation assays and can be used to design better informed enhancer-gene pairs prediction algorithms (e.g. Cicero [Pliner et al. 2018]). Yet, they still fall short of validating enhancer-gene pairs.

1.3.3 *Emerging Technology – Synthetic Dissection*

Massively parallel reporter assays (MPRAs - Fig. 1.2B): In a single experiment, an MPRA tests activity of thousands of enhancer sequences against a reporter gene. This requires two components: the enhancer library must be cloned into a reporter vector and barcoded in some way such that the frequency of any one enhancer allele is present in both the DNA and RNA. The frequency of a barcode in the RNA over the DNA can then define an enhancer sequence's activity (Gasperini, Starita, and Shendure 2016). Since first being demonstrated for dissection of a promoter in 2009 (Patwardhan et al. 2009), many MPRA technologies have been applied to study enhancers

(reviewed in Inoue and Ahituv 2015). Enhancer activity in an MPRA provides episomal evidence in-line with the original 1981 definition of the SV40 enhancer (Banerji, Rusconi, and Schaffner 1981).

The massive advantage of MPRAs is their ability to simultaneously test high numbers of sequences while requiring only a straightforward genomics toolkit (access to deep sequencing, molecular cloning, and oligo synthesis [Gasperini, Starita, and Shendure 2016]). This scalability and utility allows the study massive numbers of candidate enhancers, including assessment of biochemically annotated enhancer sequences (X. Wang et al. 2018; Vockley et al. 2016; Vanhille et al. 2015; Klein, Keith, et al. 2018), enhancer alleles carrying noncoding human variants associated with transcriptional effects (Ulirsch et al. 2016; Tewhey et al. 2018; Vockley et al. 2015), and even the entire human genome (Y. Liu et al. 2017). A major advantage of MPRAs is their synthetic design, which allows for easy testing of programmed sequences (such as a single nucleotide variant allelic series [Melnikov et al. 2012; Patwardhan et al. 2012] or enhancers with rearranged TF motif composition [Grossman et al. 2017]). Current technologies to deliver allelic series in-genome (Findlay et al. 2014) are still more technologically difficult than an MPRA.

However, current MPRAs are limited by length of sequence tested and confounding choice of the reporter's minimal promoter. Beyond technical limits, testing the enhancer on a reporter vector fundamentally misses critical genomic context features (e.g. 3D chromatin conformation, potential redundancy between enhancers within a locus). Though some MPRAs integrate transgenically into the genome (Inoue et al. 2017; Akhtar et al. 2013; Maricque, Chaudhari, and Cohen 2018), they are fundamentally incapable of pairing enhancers to their target genes.

1.4 Emerging Genome Engineering Methods for Enhancer Screening

The previous methods' limitations are overcome by the emerging suite of pooled CRISPR enhancer screens. These have springboarded off of CRISPR genic screens (Shalem et al. 2014; Y. Zhou et al. 2014; T. Wang et al. 2014) to induce and evaluate massive numbers of enhancer perturbations in their native genomic context. At their core, these screens entail delivery of a gRNA library to a pool of cells, followed by pooled phenotyping of the putative target transcript(s) and pooled measurement of the associated gRNA. These usually deliver Cas9-induced perturbations, including: active Cas9 for sequence disruption (Canver et al. 2015), nuclease dead-Cas9 (dCas9) tethered to an epigenetic repressor domain (Charles P. Fulco et al. 2016) or activator domain (Simeonov et al. 2017). As these perturbations are phenotyped by the expression of the candidate target gene(s), they provide functional evidence for enhancer-gene pairs -- the holy grail of understanding the noncoding genome.

The first generation of CRISPR screens delivered an individual gRNA per cell, usually resulting in 1-10 bp-long deletions or ~1 bp insertions ("indels") (van Overbeek et al. 2016; W. Chen et al. 2018) in 90% of cells (T. Wang et al. 2014; Shalem et al. 2014). A single gene's expression is usually used to phenotype these indel-bearing cells ("monogenic CRISPR screens"). The first monogenic screen tiled indels across a known enhancer of *BCL11A* (Canver et al. 2015). Canver et al. sorted the edited cell pool on the *BCL11A*-dependent switch to fetal hemoglobin, and successfully identified the GATA1 motif critical for the enhancer's function. This first screen was soon followed by at-scale implementations (Wright and Sanjana 2016) that scanned thousands of candidate enhancers per experiment (Korkmaz et al. 2016; Rajagopal et al. 2016; Diao et al. 2016;

Sanjana et al. 2016). But though an indel is likely long enough to disrupt a TF motif, it at most only partially disrupts the entire enhancer. This indel length limits the scalability of these methods, as surveying any substantial amount of sequence requires many indels, many gRNAs, many cells, and many sequencing reads.

The second generation of sequence scans attempts to deliver longer deletions to each cell to create larger effect sizes and scan sequence more efficiently (Fig. 1.2C; Diao et al. 2017; Gasperini et al. 2017; Aparicio-Prat et al. 2015). These "long-deletion scans" deliver pairs of gRNAs relatively near to each other, which can result in drop-out of the intervening sequence between the two cuts. Yet, the further apart the pair, the less often the full deletion occurs (Canver et al. 2014), and this inefficiency (~20% for a 365 bp deletion [Gasperini et al. 2017]) limits the utility of long-deletion scans. However, though inducing a "gold standard" enhancer perturbation, sequence scans are overall limited by either effect size (short indels) or efficiency (long-deletion). Furthermore, the fickle inaccuracy of DSB repair plagues these screens with unprogrammed edit outcomes (Gasperini et al. 2017; Kosicki, Tomberg, and Bradley 2018) and multiple alleles of the targeted locus can receive different edits within the same cell.

Epigenetic perturbations overcome many of these limitations, as enhancers are modified consistently across all target alleles. The dCas9-KRAB repressor domain (Fig. 1.2D; CRISPR-inhibition or "CRISPRi") was the first construct shown to synthetically silence a targeted enhancer by delivering ~1-2 kb of heterochromatin (Thakore et al. 2015), and has since been used in multiple monogenic screens (Charles P. Fulco et al. 2016; Klann et al. 2017; C. P. Fulco et al. 2019). Activating domains (dCas9-VPR or dCas9-p300) have also been used in monogenic screens to

scan for poised enhancers (Klann et al. 2017; Simeonov et al. 2017). Multiple alternative dCas9-tethered domains have been shown to disrupt enhancer activity (Kearns et al. 2015; Kwon et al. 2017; Lei et al. 2017; Vojta et al. 2016; X. S. Liu et al. 2016; Huang et al. 2017) and could next be adapted to the enhancer screen framework. However, though epigenetic scans are technologically advantageous, the synthetic nature of the perturbation requires any hits to be validated by alternative methods (e.g. sequence deletion).

Yet in both types of "monogenic screens", each enhancer is only tested against a limited number of target genes. Perturbation-bearing cells have most often been phenotyped by a single target gene (e.g. mRNA products labeled by FISH [C. P. Fulco et al. 2019], laboriously fluorescently-labeled target genes [Rajagopal et al. 2016; Klann et al. 2017; Diao et al. 2016, 2017], drug responsive protein products [Gasperini et al. 2017; Sanjana et al. 2016], antibody labeled protein products [Simeonov et al. 2017]) and at most a set of genes within a cell proliferation pathway (Korkmaz et al. 2016; Charles P. Fulco et al. 2016). Each "monogenic" selection method requires a specific technical set-up and limits the enhancer-gene pairs tested in any one screen.

To increase throughput, "whole transcriptome CRISPR screens" have been developed to test each enhancer perturbation against any target gene. An enhancer-targeting gRNA-library is still integrated into a pool of cells, but instead of monogenically phenotyping the cells, *all* transcripts per cell are measured via scRNA-seq. The first implementation of this delivered single CRISPRi perturbations per cell and targeted 71 candidate enhancers across 7 genomic loci (Xie et al. 2017). Yet as scRNA-seq is costly and any one enhancer likely affects only a small portion of the whole transcriptome, next-generation methods multiplexed perturbations per cell to increase throughput

to perturb 5,779 enhancers in one pooled experiment (Gasperini et al. 2019). But even multiplexed scRNA-seq experiments are still expensive, and further multiplexing per cell and reduction of scRNA-seq reagent costs are needed. Additionally, whole transcriptome screens are currently limited to epigenetic perturbation, as inconsistent sequence editing efficiency and potentially toxic DSB response prevents the use of multiplexed DSBs in each cell.

dCas9 tethered to base editor variants has emerged as a non-cutting technology that creates sequence disruptions at high efficiency. Base editors do not rely on DSB repair to create edits, and create stereotypic profiles of single nucleotide variants at targeted loci with up to 90% efficiency (Hess et al. 2017). Preliminary data has shown these can potentially be multiplexed across thousands of sites per cell (Smith et al. 2019). Unfortunately, base editors would create effect sizes likely too small for efficient use in enhancer-gene pair screens but are potentially the ideal tool with which to multiplex SNVs' impact upon enhancer activity.

Altogether, these screens have enabled unprecedented genomic testing and pairing of enhancer-gene pairs. Yet, they still need improvement. First, enhancer-pair hit validation outside of screens is only recently becoming field standard, though it's clearly required (Gasperini et al. 2017; Kosicki, Tomberg, and Bradley 2018). Second, as validated enhancer-gene pairs are so rare, these scans all carry unknown false negative rates. How is the field to interpret a biochemically-nominated or MPRA-active candidate enhancer with no signal in a CRISPR screen? Insight will come from further replication of perturbational screens across platforms. Last, these screens have so far produced a relatively small numbers of hits. Is this ~10% rate derived from true biology or limitations of methodology? Potential explanations include: epigenetic perturbations have high

false negatives and sequence deletions have not yet been tested against the whole transcriptome; redundancy remains largely unstudied as few enhancers have been perturbed in combination; and most screens are in terminally differentiated, stable cell lines with no dynamic expression changes. Further innovations in CRISPR screen technology must address these challenges.

Though a major advancement, pooled CRISPR screens study enhancers in *in vitro* pools of cells cultured outside of an organism. Mouse models provide highly biological tissue-specific enhancer validation in an *in vivo* context. Transgenic assays provide valuable tissue-specificity of candidate enhancers tested on reporter plasmids (Visel et al., 2007), but suffer the same limitations as MPRAs (e.g. failure to pair to target gene(s), synthetic locus, unclear what boundaries to use in design of enhancer sequences). CRISPR has made generating genomic sequence deletions in mice much easier and has already enabled new insights into topological chromatin domain disruption (Lupiáñez et al., 2015) and enhancer redundancy (Osterwalder et al., 2018). However, both of these studies rely on limb dysmorphology to phenotype enhancer disruption, and more subtle phenotypes may be challenging to detect. Both methods require advances in throughput. Imperfect synteny prevents the study of some candidate enhancers in mice. Tissue culture innovations such as organoids (Lancaster and Knoblich, 2014) promise closer biological modeling with cell culture scale to complement these animal model limitations.

In Chapters 2 and 3, I will review two enhancer-gene pair CRISPR screens methods for which I led development: the monogenic screen ScanDel and the first multiplexed whole transcriptome screen.

Chapter 2. CRISPR/CAS9-MEDIATED SCANNING FOR REGULATORY ELEMENTS REQUIRED FOR *HPRT1* EXPRESSION VIA THOUSANDS OF LARGE, PROGRAMMED GENOMIC DELETIONS

Chapter 2 has been adapted with minimal modification from:

Gasperini, Molly, Gregory M. Findlay, Aaron McKenna, Jennifer H. Milbank, Choli Lee, Melissa D. Zhang, Darren A. Cusanovich, and Jay Shendure. 2017. "CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for *HPRT1* Expression via Thousands of Large, Programmed Genomic Deletions." *American Journal of Human Genetics* 101 (2): 192–205.

2.1 ABSTRACT

The extent to which non-coding mutations contribute to Mendelian disease is a major unknown in human genetics. Relatedly, the vast majority of candidate regulatory elements have yet to be functionally validated. Here we describe a CRISPR-based system that uses pairs of guide-RNAs (gRNAs) to program thousands of kilobase-scale deletions that deeply scan across a targeted region in a tiling fashion ("ScanDel"). We applied ScanDel to *HPRT1*, the housekeeping gene underlying Lesch-Nyhan syndrome, an X-linked recessive disorder. Altogether, we programmed 4,342 overlapping 1- and 2- kilobase (Kb) deletions that tiled 206 Kb centered on *HPRT1* (including 87 Kb upstream and 79 Kb downstream), with median 27-fold redundancy per base. Programmed deletions were functionally assayed in parallel by selecting for loss of HPRT function with 6-thioguanine. As expected, sequencing gRNA pairs before and after selection confirmed all *HPRT1* exons are needed. However, *HPRT1* function was robust to deletion of any intergenic or

deeply intronic non-coding region, indicating proximal regulatory sequences are sufficient for *HPRT1* expression. Although our screen did identify the disruption of exon-proximal non-coding sequences (e.g. the promoter) as functionally consequential, long-read sequencing revealed this signal was driven by rare, imprecise deletions that extended into exons. Our results suggest no singular distal regulatory element is required for *HPRT1* expression, and that distal mutations are unlikely to contribute substantially to Lesch-Nyhan syndrome burden. Further application of ScanDel may shed light on the role of regulatory mutations in disease at other loci, while also facilitating a deeper understanding of endogenous gene regulation.

2.2 CONTRIBUTIONS

For this manuscript, the initial idea and plan to perform a paired-gRNA noncoding CRISPR screen was conceived by Greg Findlay and Jay Shendure before I joined the lab. During her rotation, Greg and myself developed the idea further and together completed the initial testing (**Fig. 2.2**) and one iteration of the screen that was not included in this publication. During this rotation, the experimental work was led by myself in side-by-side teaching and collaboration with Greg, whereas I conducted all the computational analysis with some help from Aaron McKenna (specifically in using his package Flashfry to design gRNAs for the screen). After the rotation, I performed all of the experimental and computational work for all further experiments (actual screens, validations, PacBio, custom scripts for analyzing the screen data) with the following exceptions: undergraduate research assistant Jennifer Milbank helped me with benchwork in the replicates of the screen; research scientist Choli Lee helped me create a custom sequencing protocol for the paired-gRNAs; undergraduate research assistant Melissa D. Zhang helped me with the tissue culture in the small-pool validations; graduate student Aaron assisted with visualizing the PacBio data using a package he generated for his own thesis work. I used post-doc Darren

Cusanovich's HAP1 ATAC-seq data. The manuscript was written by me and Jay. Both me and Greg were credited as co-first authors, and both myself and Jay were credited as co-corresponding.

2.3 INTRODUCTION

The success of human genetics in identifying the genes and mutations underlying Mendelian diseases has been facilitated by the incontrovertible reality that the majority of causal mutations lie in protein-coding sequences or splice junctions. Indeed, this assumption is explicit in both classic and contemporary practices in genetics (*e.g.* exome sequencing). However, it is clear that distal non-coding mutations make *some* contribution to Mendelian disease. Understanding how often non-coding mutations play a causal role, as well as developing best practices for pinpointing those that do, are critical challenges for the field. For example, in the clinic, even if a person is diagnosed with a monogenic Mendelian disorder on the basis of phenotype, clinical sequencing mainly of coding regions fails to identify a causal mutation ~10% of the time (Chong et al. 2015). However, possible explanations include not only distal regulatory mutations, but also misdiagnosis, somatic mutation, technical false negatives, and others. Furthermore, non-coding loci could contribute to the estimated ~25-50% of undiagnosed but apparently Mendelian cases in which the underlying gene is unknown (Chong et al. 2015; Yang et al. 2013).

The picture is very different for the genetics of common disease, where over 90% of disease-associated SNPs fall in non-coding regions (Maurano et al. 2012). Many resources have been developed to predict the location of putative regulatory elements and the effects of regulatory mutations (Ernst and Kellis 2012; Hoffman et al. 2012; Kircher et al. 2014), with ~88% of all protein-coding genes tied to a *cis*-expression quantitative locus (eQTL) (Consortium and GTEx

Consortium 2017), ~80% of the genome annotated with biochemical function (ENCODE Project Consortium 2012), and numerous tools to link regulatory elements to their target genes (Coetzee et al. 2012; Li et al. 2013; Ward and Kellis 2012; A. P. Boyle et al. 2012). However, the vast majority of these predictions are either confounded (*e.g.* for *cis*-eQTLs, by linkage disequilibrium) or lack functional validation. Indeed, there are few distal non-coding regulatory elements that we can confidently assign to a target gene, or for which we understand the consequences of disruption.

Large-scale functional experiments are clearly an important next step for both common disease genetics (to facilitate the identification of causal regulatory variants and their target genes) and rare disease genetics (to identify distal regulatory elements for Mendelian disease genes where causal non-coding mutations might be found). A number of important studies have undertaken functional work to identify and characterize causal or risk contributory non-coding variants for specific rare and common diseases (Wakabayashi et al. 2016; Weedon et al. 2014; Claussnitzer et al. 2015) but by approaches that are not easily scalable. Within the last year, several studies have used CRISPR/Cas9 genome editing in cell-based screens to introduce and functionally assay large numbers of non-coding mutations at an unprecedented scale (Canver et al. 2015; S. Chen et al. 2015; Diao et al. 2016; Korkmaz et al. 2016; Rajagopal et al. 2016; Sanjana et al. 2016). The common approach of these studies is to introduce complex libraries of guide RNAs (gRNAs) via lentiviral infection to a population of cells at a low multiplicity of infection (MOI), followed by an assay that queries the function or expression of a gene of interest. CRISPR/Cas9 mediates double-stranded breaks at sites specified by the gRNA in each cell, eventually resulting in a mutation at each targeted site via imperfect non-homologous end joining (NHEJ).

A fundamental limitation of these singleton gRNA screens is that because of design constraints (*e.g.* the uneven distribution of protospacer adjacent motif (PAM) sequences, the variable efficiency of gRNAs, and others), the resulting coverage of regions of interest is incomplete and uneven. As the majority of bases will be perturbed by zero or only one gRNA, these studies rely on the aggregate behavior of clusters of target sites within potential regulatory elements (Canver et al. 2015) or arbitrarily sized windows (e.g., 500 base-pairs) (Sanjana et al. 2016), rather than redundant targeting of each base-pair (bp) by independent gRNAs. Furthermore, it is possible that the mutations introduced by NHEJ at single sites (highly heterogeneous but mainly dominated by small 1-10 bp deletions (Tsai et al. 2015)) are insufficient to fully disrupt many regulatory elements. Several recent studies have employed an inhibitory domain guided by nuclease-inactive Cas9 to screen non-coding regulatory regions, *i.e.* CRISPRi (Charles P. Fulco et al. 2016; Klann et al. 2017). Epigenetic modifications mediated by these domains can spread to regions on the order of ~200 bp to 4.5 Kb (Thakore et al. 2015; Horlbeck et al. 2016), and thus mitigate the challenges related to redundancy and coverage of individual gRNA screens. However, CRISPRi screens may be less precise because of this spreading effect and furthermore, do not directly test the consequences of alterations in primary sequence.

Here we sought to overcome these weaknesses by introducing *pairs* of gRNAs to each cell, with the goal of inducing a kilobase-scale deletion of the intervening DNA between two programmed cuts. A principal advantage of this method is that by tiling deletions across a region, each targeted base-pair can be covered with high redundancy (scanning deletion or “ScanDel”). Furthermore, kilobase-scale deletions are much more likely to eliminate the function of an overlapping or fully contained regulatory element, relative to small indels resulting from NHEJ at a single target site.

Our approach is analogous to classic deletion scanning experiments (Reid et al. 1990; Rincón-Limas, Krueger, and Patel 1991), but with advantages in throughput and of targeting much larger regions in the endogenous genome rather than sequences cloned to a plasmid. Similar strategies have recently been described for the interrogation of lncRNA genes (Zhu et al. 2016) and non-coding sequences (Diao et al. 2017). Critically, these implementations (and indeed, all CRISPR genetic screens) rely on indirectly genotyping the lentivirally inserted gRNA sequences, instead of using direct sequencing of edited loci to confirm exactly which CRISPR-induced genotypes are driving effects.

Here we applied ScanDel to survey the genomic locus encompassing *HPRT1* [MIM: 308000], which encodes the enzyme hypoxanthine(-guanine) phosphoribosyltransferase (HPRT). *HPRT1* is a housekeeping gene, a class of genes primarily defined by their broad expression and for which the underlying regulatory architecture remains unclear (Zabidi et al. 2014). Loss-of-function mutations in *HPRT1* result in the X-linked Lesch-Nyhan syndrome (Lesch and Nyhan 1964) [MIM: 300322], in which a minority of individuals present with reduced HPRT enzymatic activity despite the absence of identifiable coding mutations (Fu et al. 2014), Such individuals could carry non-coding mutations that result in reduced *HPRT1* expression. Reduced HPRT activity also causes resistance to the drug 6-thioguanine (6TG), a purine analog and chemotherapeutic agent. Thus, it is straightforward to assay cell populations for loss of *HPRT1* function, as only cells with highly reduced expression of functional HPRT will survive selection by 6TG (Fig. 2.1C). Although there are no known distal regulatory elements of *HPRT1*, its nine exons serve as internal controls.

Adopting the framework of genome-wide CRISPR/Cas9 screens, we synthesized, cloned, and

lentivirally delivered thousands of programmed gRNA pairs to cells at a low MOI. Each gRNA pair targets nearby sites, effectively leveraging CRISPR/Cas9's ability to generate kilobase-scale deletions when NHEJ-mediated repair of two double-stranded breaks results in excision of the intervening DNA segment. In total, we designed and introduced gRNA pairs programming 4,342 overlapping ~1- and ~2- kilobase (Kb) deletions that tiled a 206 Kb region centered on *HPRT1*. 6TG was used to select for cells that had lost *HPRT1* function. By quantifying gRNA pairs both before and after 6TG selection and then directly genotyping putatively important deletions by long-read sequencing, we were able to identify programmed deletions that significantly compromised *HPRT1* expression and function.

2.4 Development of ScanDel

In genome-wide CRISPR/Cas9 screens, a gRNA library is lentivirally delivered to a large pool of cells at a low MOI, such that each infected cell is likely to receive only one gRNA (Shalem et al. 2014; T. Wang et al. 2014; Y. Zhou et al. 2014). Each gRNA induces NHEJ-mediated indels centered at the Cas9-mediated cleavage position within the target sequence, with the goal of perturbing the function of the targeted locus. However, given the small and variable length of indels, the robustness of perturbation is inherently limited, particularly when targeting non-coding sequences in which frameshifts are irrelevant.

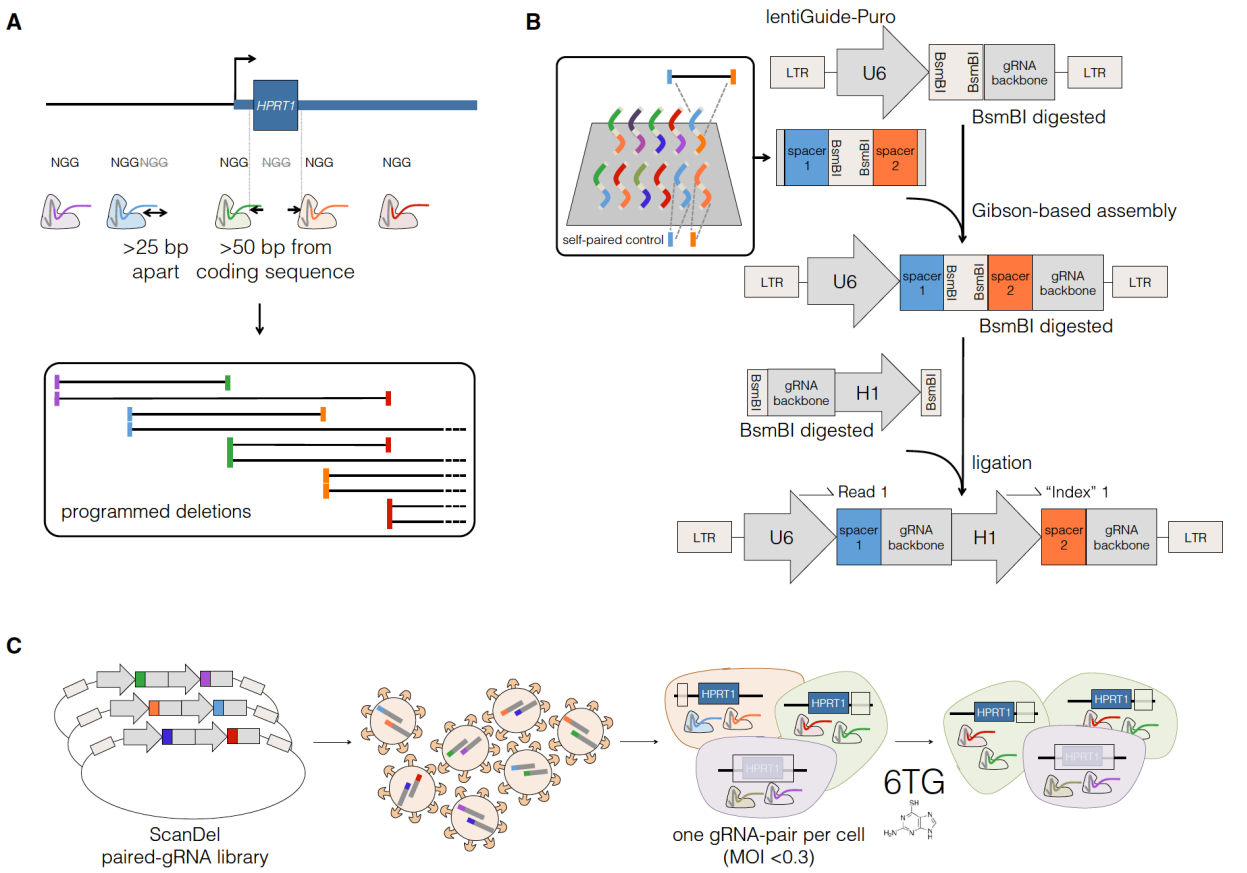


Figure 2.1. Design, delivery, and selection of ScanDel library of CRISPR/Cas9-Programmed deletions for identification of non-coding regulatory elements.

(A) gRNA pairs were designed from a filtered set of protospacers from all Cas9 PAM sequences (50-NGGs) in the HPRT1 locus (see also Figure 2.3A). Sites that were >25 bp apart or >50 bp away from exons were kept. For tile design, each remaining spacer was paired to two downstream spacers targeting sequence ~1 and ~2 kb away. This resulted in high redundancy of independently programmed, overlapping deletions across the locus (see also Figure 2.3B). (B) All spacer pairs corresponding to programmed deletions were synthesized on a microarray (inset). Each spacer was also synthesized as a self-pair as a control for its independent effects. If a self-paired spacer scored positively in the screen, any pairs that used that spacer were removed from the analysis (Fig. 3.3). U6 and gRNA backbone sequence flanked the spacer pairs for

Gibson-mediated cloning into lentiGuide-Puro,35 and mirrored BsmBI cut sites separated the spacer pairs to facilitate insertion of a second gRNA backbone and the H1 promoter. In the final library, each gRNA was expressed from its own PolIII promoter. This design facilitates PCR and direct sequencing based quantification of gRNA-pair abundances. (C) The lentiviral library of gRNA pairs was cloned at a minimum of 203 coverage (in relation to library complexity) and transduced into HAP1 cells stably expressing Cas9 (via lentiCas9-Blast35) at low MOI. After a week of puromycin selection, the cells were sampled for measurement of the baseline abundance of each gRNA pair. The final cell population was harvested after a week of 6-thioguanine (6TG) treatment, which selected for cells that had lost HPRT enzymatic function. The phenotypic prevalence of each programmed deletion was quantified by PCR and deep sequencing of the gRNA pairs before and after selection.

To instead program a kilobase-scale deletion in each cell, we devised the following approach (Fig. 2.1). First, gRNA pairs are designed to program specific deletions (with each gRNA specifying one of the deletion's boundaries, Fig. 2.1A), and the corresponding pairs of 20 bp spacers are synthesized *in cis* on a microarray (Fig. 2.1B). Second, the paired spacers are inserted into the lentiGuide-Puro plasmid between the U6 promoter and the gRNA backbone. Third, a second gRNA backbone and a second RNA Polymerase (Pol) III promoter (H1 or U6) are inserted between the paired spacers. Fourth, libraries of "gRNA pairs" are lentivirally delivered to a large pool of cells at a low MOI, such that each cell receives a pair of gRNAs that programs a single deletion (Fig. 2.1C). Finally, analogous to conventional genome-wide CRISPR/Cas9 screens, deep sequencing of the integrated gRNA pairs is used as a surrogate measure of the prevalence of each

programmed deletion in a population of cells (*e.g.* before and after the cells have been subjected to functional selection) thus capturing the phenotypic consequences of individual deletions.

As an initial test of our paired guide system, we compared the efficacy of using two different promoters for the two guides (a 'U6-H1' system) versus using two copies of the same promoter ('U6-U6'). We tested these lentiviral gRNA pair expression constructs by targeting the same genomic site for deletion with each system (Fig. 2.2). PCR amplification of the site was performed with UMIs in order to minimize biases related to amplicon size (see Methods). The U6-H1 system induced more programmed deletions than the U6-U6 system (20% vs. 10% of reads from cells one week after transduction). The U6-H1 system has several advantages (*e.g.* avoiding recombination between the two U6 promoters during cloning; unique primer design for deep sequencing of each gRNA), and we therefore proceeded with it.

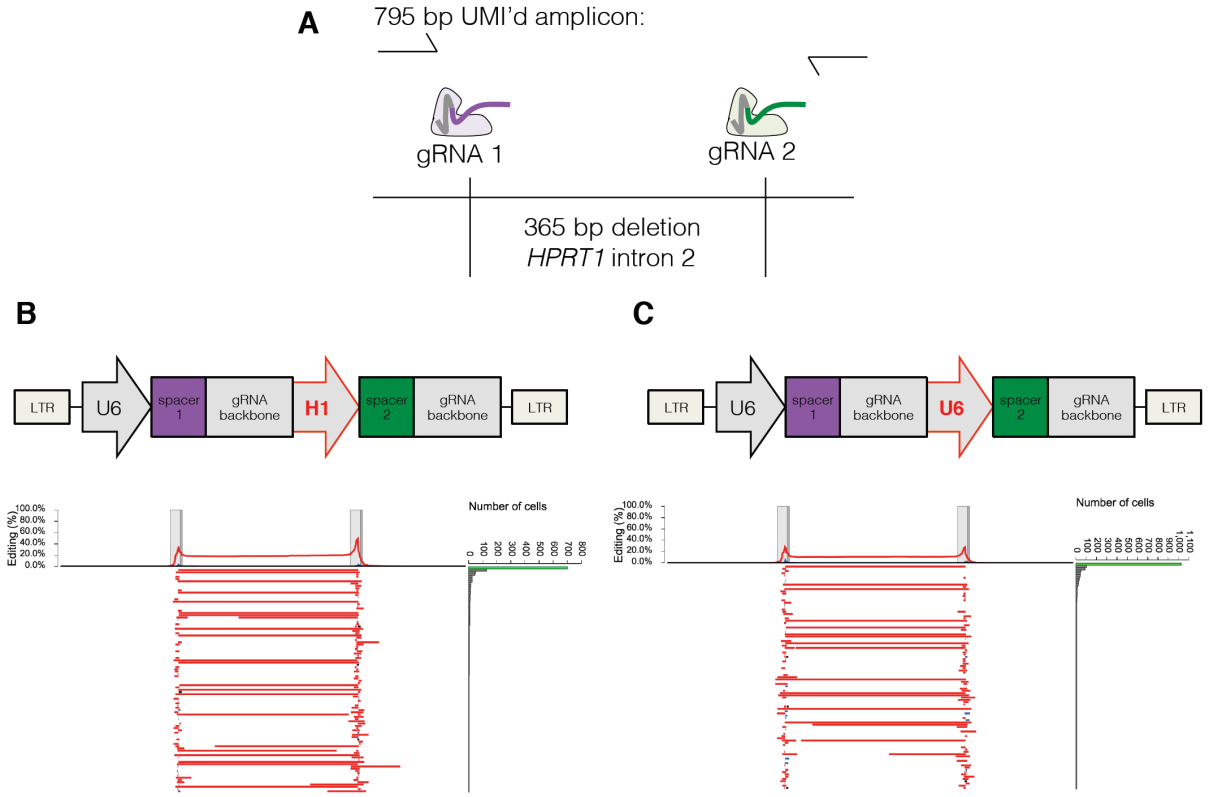


Figure 2.2. The U6-H1 gRNA pair expression construct induces a higher deletion rate.

A) Two spacers were chosen to program a 365 bp deletion within the second intron of HPRT1. To test deletion efficiency of the method as described in Fig. 2.1C, virus was made from the constructs depicted in B and C, and separately transduced into HAP1 at MOI < 0.3. Following 1 week of puromycin selection, gDNA was extracted and the targeted region amplified. The first 3 cycles of this PCR contained a forward primer with a unique molecular tag (UMI) to track reads from the same original cell. Sequencing was performed on a MiSeq. Of note, PCR bias for smaller deletion-holding amplicons was reduced by collapsing reads with the same UMI, but the potential remains for higher clustering efficiency of the shorter amplicons. B) The spacers for the deletion in A were placed behind either a U6 or H1 PolIII promoter. 20% of sampled haplotypes contained the programmed deletion, but 36% of sampled haplotypes remained unedited, implying longer editing time could result in a higher deletion rate. Reads were generated as described in A, and aligned as

described in Methods and Fig. 2.10. The per base-pair editing rate summed across all sampled haplotypes is charted as a percentage at top, and the top 100 most prevalent haplotypes are displayed below it. Red indicates deletions and blue insertions. C) The spacers for the deletion in A were each placed behind a U6 PolIII promoter, and delivered, sampled, and visualized as above. With this expression construct, 10% of sampled haplotypes contained the programmed deletion.

An important caveat for ScanDel, relative to conventional gRNA cell-based screens, is that deletions programmed by gRNA pairs only occur in a minority of cells (Canver et al. 2014; Byrne et al. 2015), with the other major outcomes being small NHEJ-mediated indels at one or both gRNA-targeted sites. For example, in our test of the U6-H1 system, the programmed deletion was found in 32% of cells that had any edit, while the remaining edited cells were mutated at one or both gRNA-targeted sites but retained the intervening sequence. While this complicates interpretation, the problem can be overcome by using a robust functional assay in conjunction with multiple, independent gRNA pairs that query the same genomic region, as well as by including unpaired gRNA controls to ensure that observed effects do not occur with the individual gRNAs that comprise each pair (but rather are dependent on the presence of both gRNAs).

2.5 Application of ScanDel to Survey the 206 Kb Region Surrounding *HPRT1*

With the goal of investigating the potential of non-coding mutations to compromise its function, we applied ScanDel to a 206 Kb region on the X chromosome centered on the *HPRT1* gene (Figs. 2.1A & 2.3A). We designed pairs of gRNAs that programmed deletions tiling across the 206 Kb region, including tiles that overlapped *HPRT1* exons in order to allow coding regions to serve as

positive controls. As deletion length has been shown to affect deletion rate(Canver et al. 2014), deletions were programmed to be consistently either ~1 or ~2 Kb in length (Fig. 2.1A). This design resulted in 4,342-programmed deletions that tiled across the region, collectively covering each base-pair a median of 27 times (Fig. 2.3B). Testing each base-pair with numerous independently programmed, tiling deletions is expected to reduce noise and also increase resolution (as all successfully made deletions tiling a critical regulatory element should exhibit positive selection). However, to guard against the possibility that individual gRNAs' effects could confound analysis (*e.g.* via off-target mutations, or on-target small ~10 bp indels), we also included all spacers in the library as pairs with themselves ('self-pairs'; Fig. 2.1B inset, Fig. 2.4). Additionally, we included 330 negative control gRNA pairs not expected to survive 6TG selection, as they program deletions in non-genic regions far from *HPRT1* or use spacers made of random sequence not present in the reference genome (hg19).

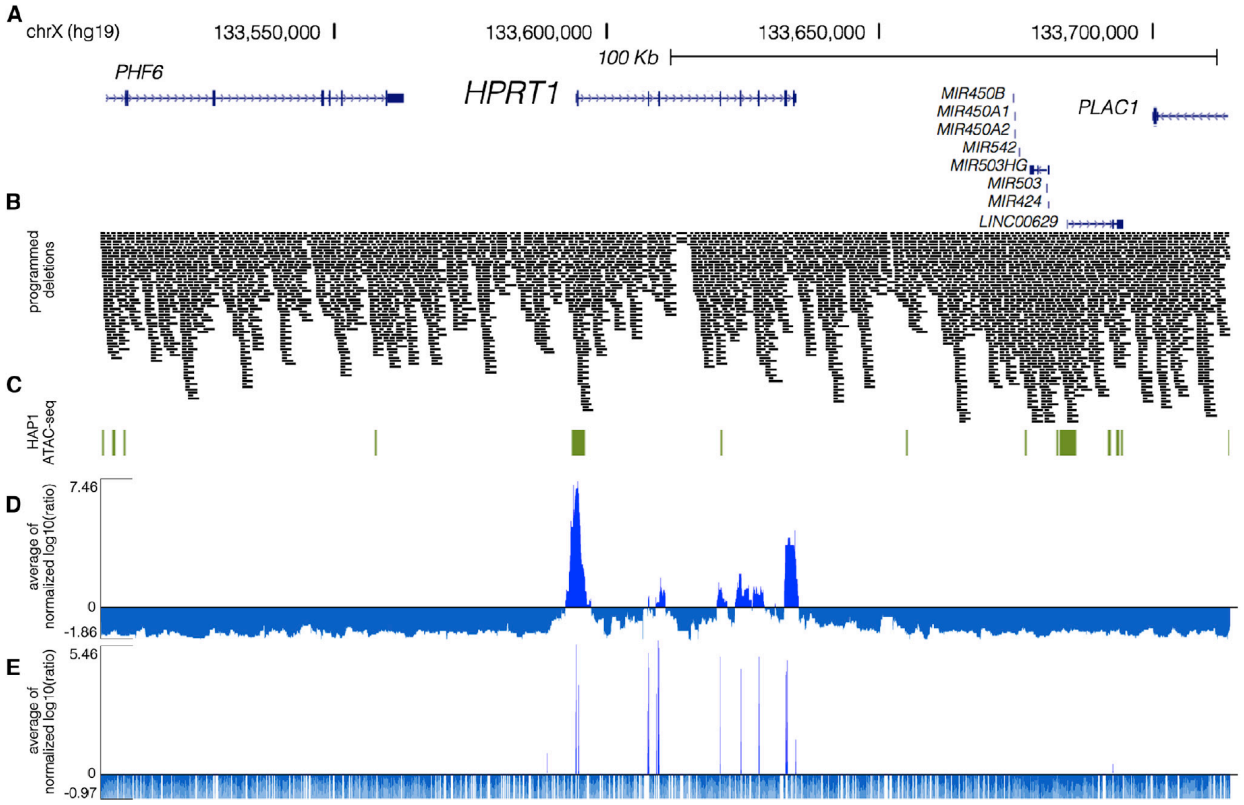


Figure 2.3. High-coverage ScanDel library across the *HPRT1* locus reveals a paucity of critical distal regulatory elements

(A) Deletions were programmed across 206.1 kb of the HPRT1 locus and its surrounding sequence (chrX: 133,507,694–133,713,798, hg19; UCSC Genes track in blue). (B) A total of 4,342 overlapping 1 or 2 kb deletions were programmed (see Fig. 2.1A) to tile across the locus such that each base pair was interrogated by a median of 27 independently programmed deletions. A high density of repeat elements resulted in reduced coverage of a region within intron 3 of HPRT1. Deletions are represented by black bars spanning the gRNA pair's programmed cut sites. (C) HAP1 ATAC-seq hotspots (green) indicate regions of open chromatin in the cell line. Of note, a hotspot extends 600 bp upstream and 1.6 kb downstream of exon 1. (D) ScanDel scores were assigned to each base pair as the average of all selection scores (log10(after/before)) for gRNA pairs that programmed deletions to span that base pair (Material and Methods). If a gRNA pair

used a spacer that was positively selected on its own as a self-pair, that gRNA pair was removed from the analysis. Given that depleted gRNAs are usually completely absent after 6TG, their negative scores are of arbitrary negative magnitude. To avoid over-weighting negative values, we determined a minimum score from each replicate's gRNA-pair score distribution (Fig. 2.5), and scores below it were set at this minimum. For each biological replicate, the base pair's score was normalized to the replicate's median of positive scores. The average of the two biological replicates' normalized scores for that base pair is displayed (positive scores in royal blue and negative scores in blue-gray). (E) An individual-gRNA mutagenesis screen of the same region was also performed and covered only ~70% of bases in the region as a result of the sparsity of high-quality designable spacers. Individual base pairs were scored on the basis of nearby cut sites under the assumption that each gRNA queries a ~10 bp region. The plotted scores were calculated as in (D) (positive scores in royal blue and negative scores in blue-gray).

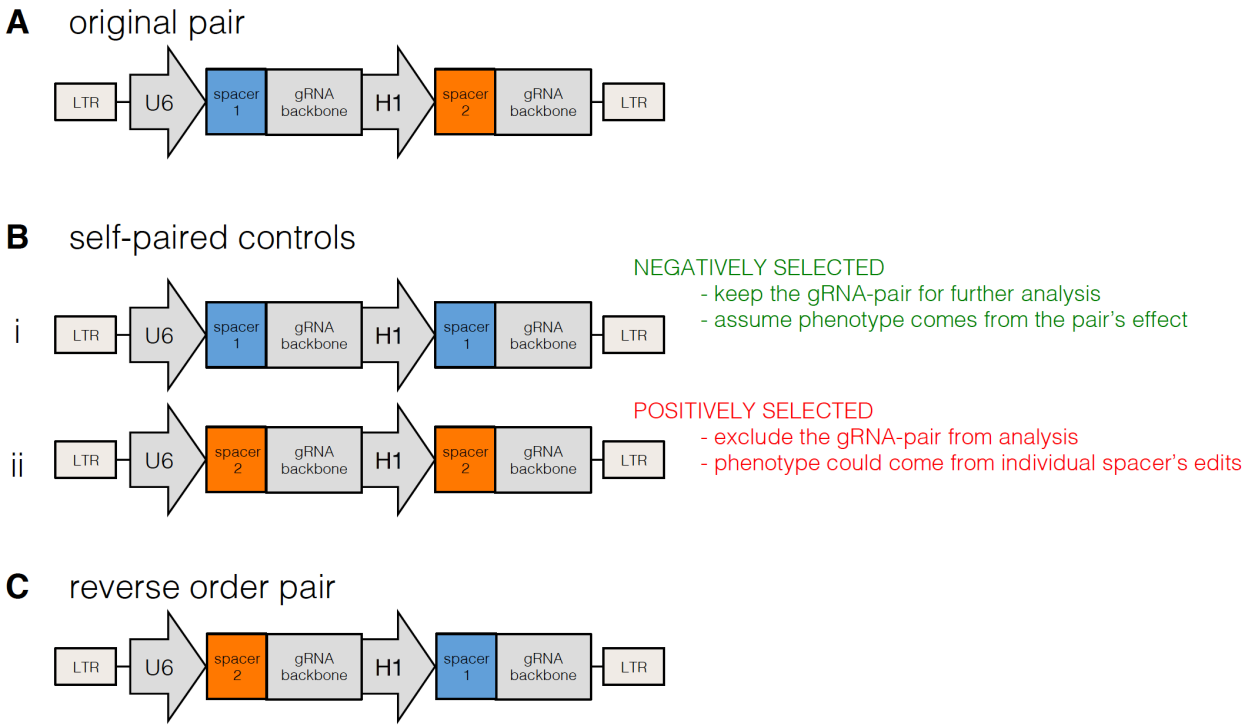


Figure 2.4. Self-paired spacers in the ScanDel library reveal phenotypes independently created by individual spacers.

A) The spacers used in every designed gRNA pair had their own self-paired control included in the programmed gRNA pair library. B) The self-paired controls consisted of the exact same spacer included behind each promoter in the expression construct (two for each pair; (i) and (ii)). If a self-paired spacer was positively selected, any gRNA pairs that included that spacer were excluded from further analysis. This avoided any confounding effects of alternative repair outcomes that result from an individual gRNA's edit that could cause 6TG resistance (e.g. a ~10 bp indel disrupting a transcription factor binding site, or disrupting an off-target locus that affects 6TG resistance, or an individual gRNA inducing translocations of HPRT1 at a high rate). By excluding these gRNAs, we can more confidently attribute observed phenotypes to programmed deletion induced by the gRNA pairs. C) Each gRNA pair was included in both possible orderings on the microarray. This was intended to minimize the impact of differences between

the promoters, as well as to increase the chance that each deletion will be represented in the library, as synthesizing each pair twice reduces loss due to synthesis errors and cloning bottlenecks.

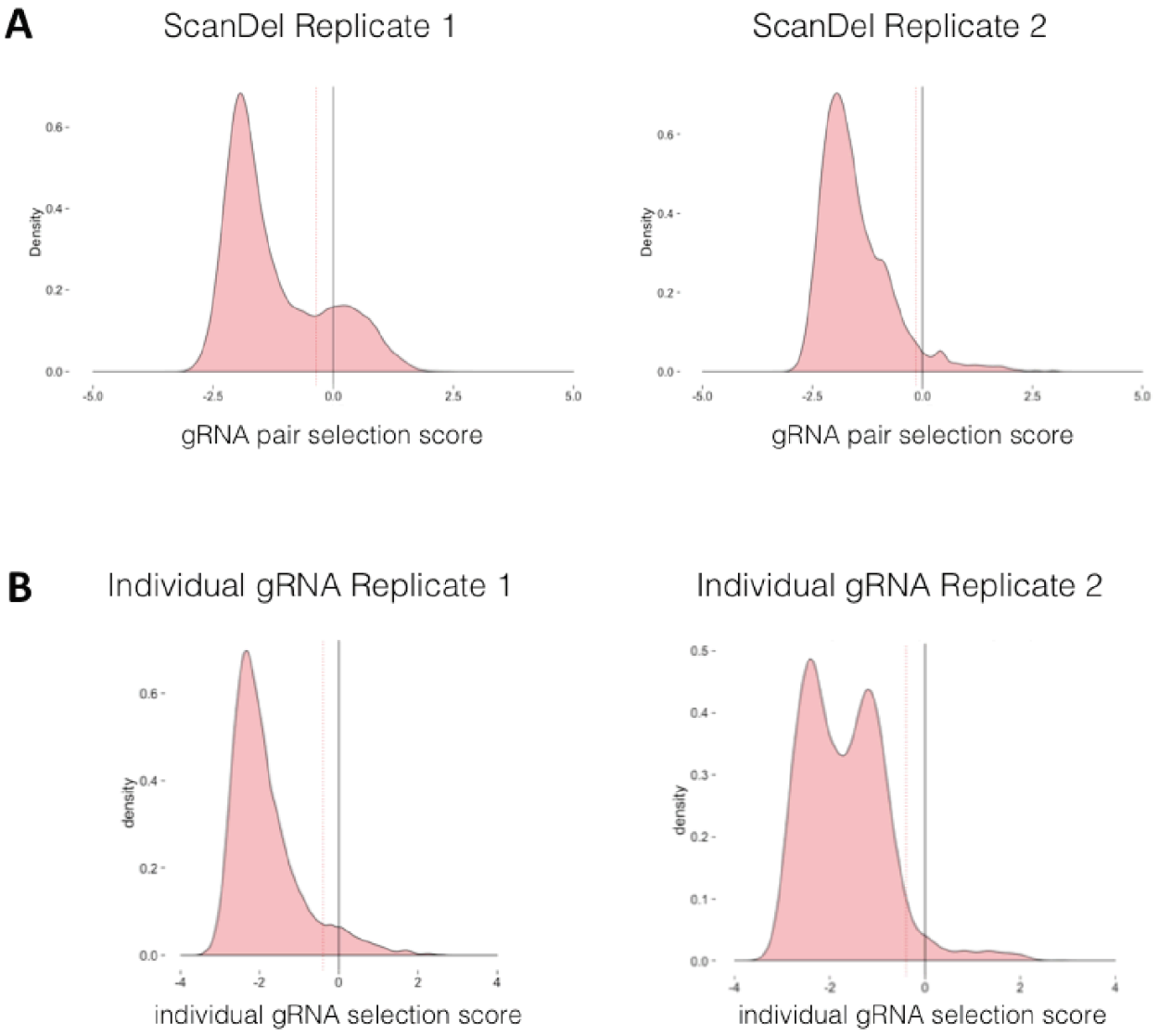


Figure 2.5. Distribution of selection scores across biological replicates for ScanDel gRNA pairs or individual gRNAs.

A) Each gRNA pair in the ScanDel screens was assigned a selection score (log10(after/before 6TG)). The minimum selection score threshold described in Methods (-0.35 for replicate 1, -0.15

for replicate 2) is drawn with a dotted red line. B) Each gRNA in the individual gRNA screen was assigned a selection score as in A, for each replicate. The minimum negative selection score threshold (-0.4 for both replicates) is drawn with a dotted red line (explanation in Methods).

The gRNA pair library was array-synthesized, cloned, and delivered via lentiviral infection to HAP1 cells in replicate (Fig. 2.1 B, C). Cell populations were sampled before and after one week of the 5 μM 6TG selection, with PCR amplification and deep sequencing of gRNA pairs to quantify abundance at each time-point. The functional selection score was calculated as the log10 ratio of normalized read counts after selection relative to before 6TG treatment ("selection score" as log10(after/before 6TG)). Positively scoring self-paired spacers were flagged, and gRNA pairs that used these flagged spacers were excluded from further analysis (11% of pairs in replicate 1 and 3% of pairs in replicate 2). To integrate signal from overlapping programmed deletions, we calculated a "per base-pair" metric as the mean of selection scores of all deletions overlapping a given base (Fig. 2.3D, Methods). This per base-pair score across the *HPRT1* locus was well-correlated between biological replicates (Pearson: 0.708; Fig. 2.6). Importantly, none of the negative-control gRNA pairs that were sampled in each of the two replicates were positively selected in both experiments (Fig. 2.7).

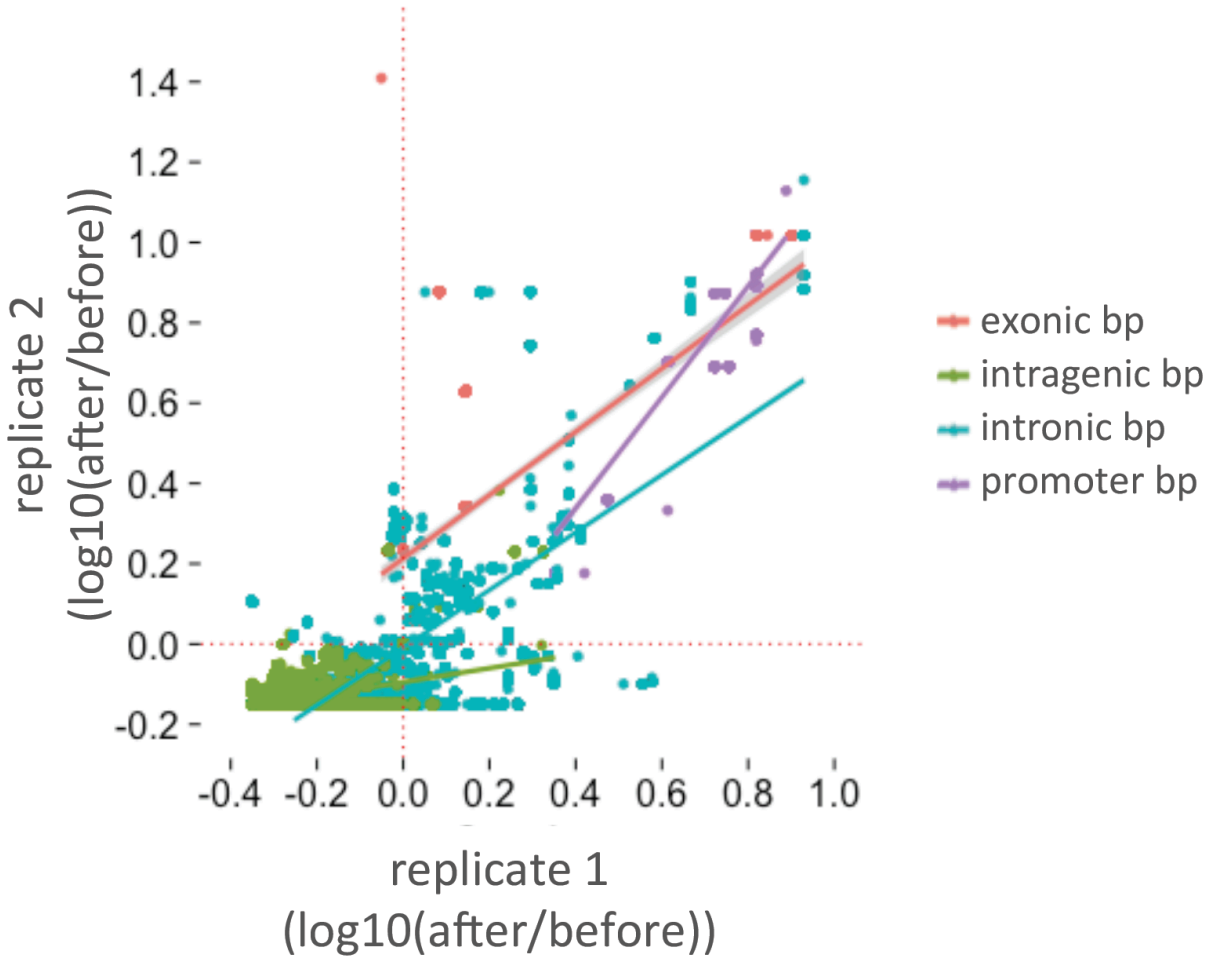
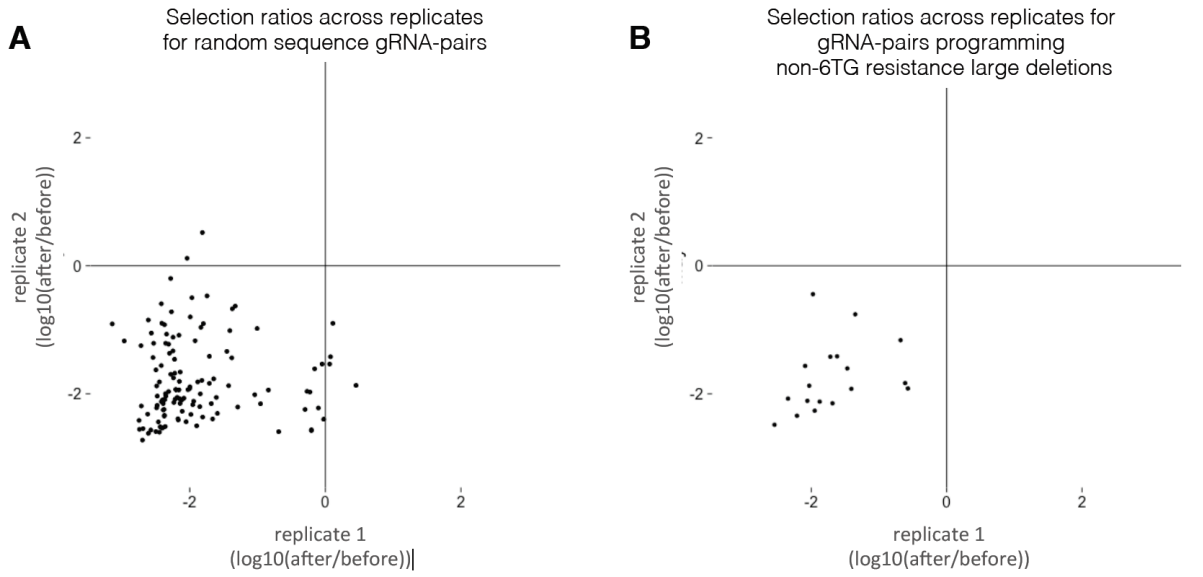


Figure 2.6. ScanDel scores correlate across two biological replicates.

The ScanDel selection scores for each biological replicate were calculated per base-pair by averaging the log10(after/before 6TG) for every programmed deletion that covers that base-pair. Least squares lines and points are colored by sequence content category. The stronger correlation for the 'intronic' category is driven by sequences proximal to the exons as seen in Fig. 2.10. Red corresponds to exons (Pearson: 0.736); green to intragenic regions (Pearson: 0.417); blue to intronic regions (within 2 Kb of an exon, Pearson: 0.628; deeply intronic, Pearson: -0.0194); and purple is the promoter (1 Kb upstream of the TSS, Pearson: 0.905).



C	positively selected in both	negatively selected in both	positively selected in rep 1	negatively selected in rep 1	positively selected in rep 2	negatively selected in rep 2
random sequence not found in hg19 (**A**)	0	236	8	256	4	244
deletions targeting non-6TG resistance regions (**B**)	0	18	1	31	0	18

Figure 2.7. None of the negative control gRNA pairs were positively selected by 6TG in both ScanDel replicates.

A) Negative control gRNA pairs targeting random sequences not found in hg19 were given a selection score of log10(after/before 6TG). Only gRNA pairs sampled in both replicates are plotted. B) Additional negative control gRNA pairs were programmed to create 1 and 2 Kb deletions in regions not expected to cause 6TG resistance. Selection scores were calculated for each gRNA pair as in A, and plotted for gRNA pairs found in both replicates. These region's coordinates were randomly generated from poorly conserved sequence 1 not within 10 Kb of any gene and far from HPRT1 (chr8:23768553-23771053, chr4:25697737-25700237, chr9:41022164-41024664, chr5:12539119-12541619, chr6:23837183-23839683, chr8:11072736-11075236). C)

Table showing counts of positively and negatively selected negative control gRNA pairs across experiments.

Crucially, all nine *HPRT1* exons exhibited strong functional scores, confirming the sensitivity of ScanDel as applied here to detect sequences essential to *HPRT1* function (Fig. 2.8). However, all of the reproducibly positive non-coding signal across the 206 Kb region was immediately proximal to an *HPRT1* exon. This result suggests that there is no distal regulatory element in the 206 Kb region that is essential to *HPRT1* expression in HAP1 cells.

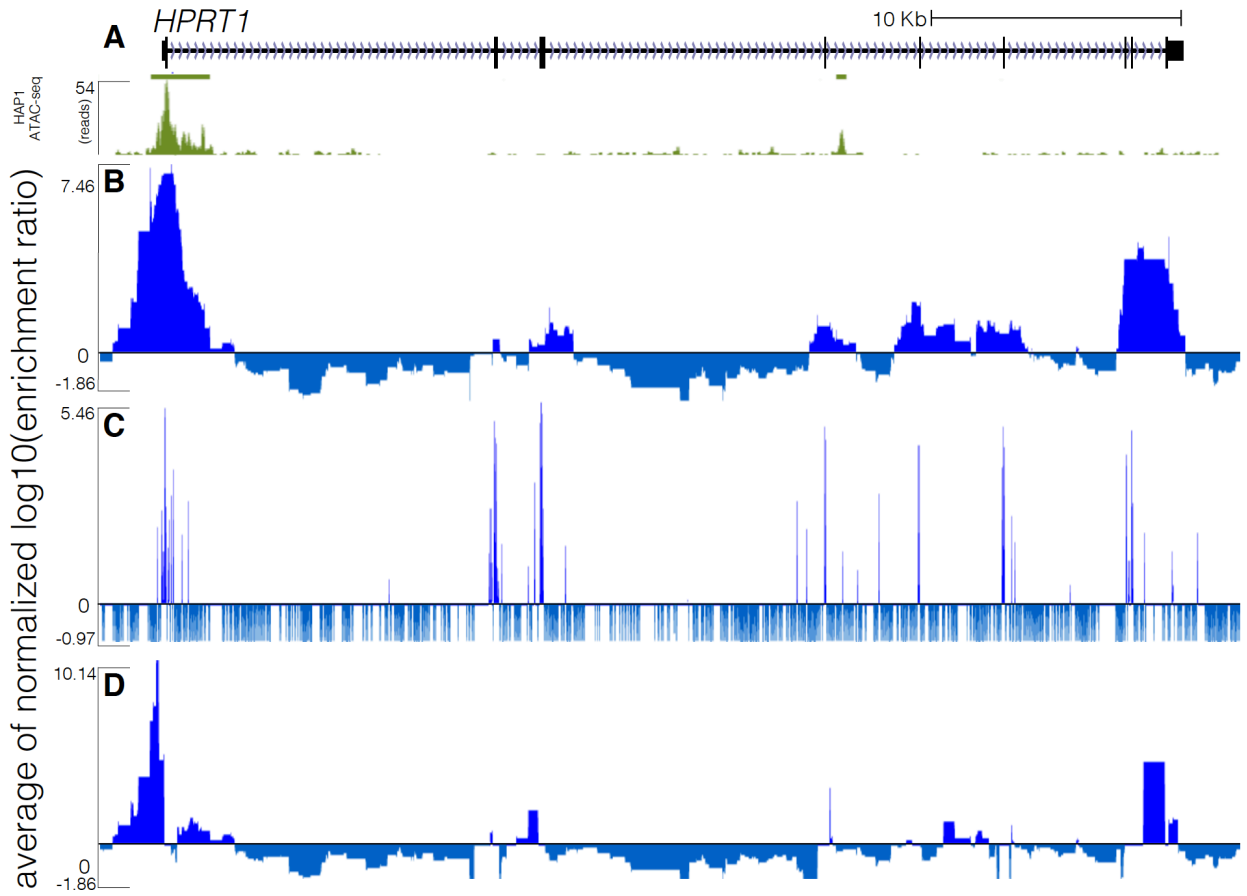


Figure 2.8. All exons and some exon-proximal non-coding regions score strongly in both the ScanDel gRNA pair screen and the individual gRNA screen.

A) ATAC-seq data (green) from the HAP1 cell line displayed for the HPRT1 locus (chrX:133,591,675-133,637,198, hg19). Bars depict hotspots 2 and beneath is the pile-up representation of ATAC-seq reads. B) The same ScanDel data is displayed as in Fig. 2C but zoomed-in on the HPRT1 locus. Each basepair's score is the mean of the log10(after/before 6TG) values for all the programmed deletions that cover that base-pair. These scores are normalized to the median positive score from the replicate. The average of the two replicates' scores for each base-pair is displayed. C) The same individual gRNA data is displayed as in Fig. 2D but zoomed in on HPRT1. Each base-pair score is the mean of the log10(after/before 6TG) values for all the inferred ~10 bp deletions that remove that base-pair. The normalized average of the two replicates' scores for that base-pair is displayed. D) The same ScanDel track as in B but with per base-pair scores calculated after excluding any deletions programmed to disrupt an exon.

Near exons, non-coding regions exhibiting positive signal did so even when deletions that also overlapped the exons themselves were excluded from the analysis (Fig. 2.8D). This suggested the presence of essential, proximal regulatory sequences. We noted that the positively scoring regions immediately upstream and downstream of the first exon overlapped with a region of open chromatin identified by performing ATAC-seq in HAP1 cells, supporting the region's role in gene regulation (Fig. 2.3C, Fig. 2.8A and 2.9). Together, these observations motivated us to attempt validation experiments for this region, with the goal of directly confirming which deletions of putative regulatory elements were impairing *HPRT1* function (Fig. 2.10A, E).

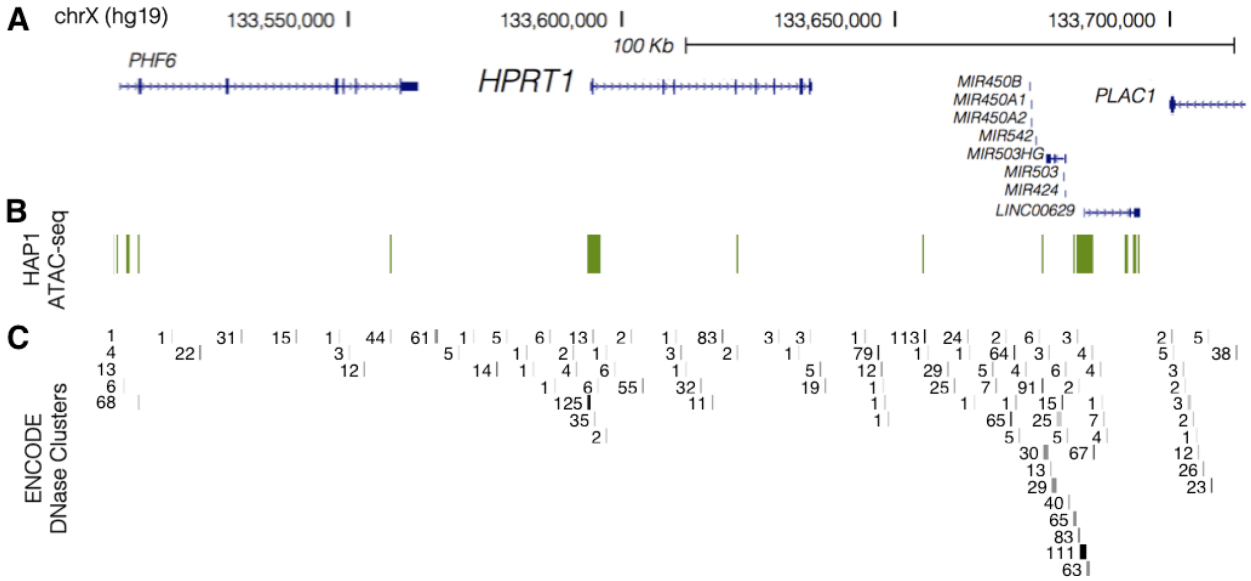


Figure 2.9. Regions of accessibility compared across HAP1 and 125 ENCODE cell types.

A) The 206.1 Kb encompassing HPRT1 and its surrounding sequence interrogated by this screen (chrX:133,507,694-133,713,798, hg19, UCSC Genes track in blue). B) Regions of open chromatin in HAP1 cells (green) as profiled by ATAC-seq. C) Clusters of DNase accessibility peaks across 125 cell lines assayed by the ENCODE project. Each accessible region is labeled with the number of cell lines in which it is detected. Though there are many cell-type specific peaks, the HAP1 open chromatin regions match sites commonly accessible across many cell lines.

2.6 Direct Genotyping of Deletions that Survive Functional Selection

With the goal of validating the positive signal upstream of the first exon, we repeated the experiment with a small pool of 4 gRNA pairs targeting the putative *HPRT1* promoter (Fig. 2.10B). We then amplified 3 Kb of this region by PCR and performed long-read sequencing of the amplicons (Pacific Biosciences). As expected, before 6TG selection, the programmed deletions were all well-represented in the population, although deletions with boundaries deviating from

Cas9 cut sites (*i.e.* 'unprogrammed') were also detected (Fig. 2.10C). However, after selection with 6TG, deletions with unprogrammed boundaries predominated, including those unseen before 6TG, and those that extend beyond the transcriptional start site (TSS) (Fig. 2.10D). The fact that these initially rare deletions were strongly selected (while 2 Kb promoter deletions that did not cross the TSS were not) suggests that even relatively proximal sequences upstream of the *HPRT1* TSS are not strictly essential for expression. Based on the results of these validation experiments, we conclude that only a narrow window of non-coding sequence immediately upstream of the TSS and 5'UTR is required for *HPRT1* expression.

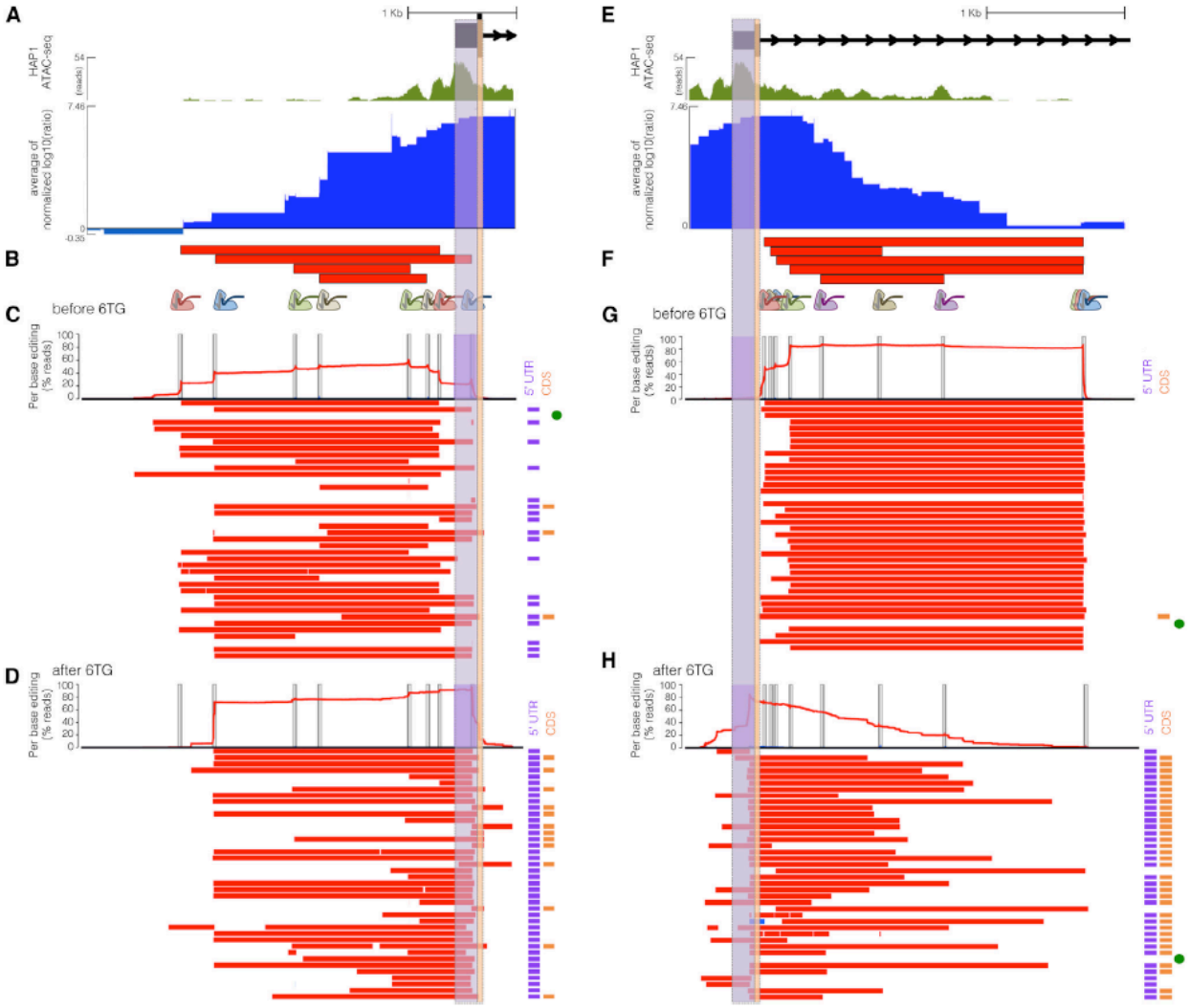


Figure 2.10. Long-read sequencing of edits derived from exon-proximal ScanDel gRNA pairs reveals rare, unprogrammed, exon-interrupting deletions that drive selective effects.

(A) A putative promoter is implicated by open chromatin (HAP1 ATAC-seq broad peaks, green) surrounding exon 1 of HPRT1 (UCSC Genes, black). ScanDel signal in the 2 kb upstream of HPRT1 also suggests the possibility of critical regulatory sequences in this region (chrX: 133,591,603–133,594,626, hg19; blue as in Fig. 2.3D). The 5' UTR and coding regions of exon 1 are highlighted in purple and orange, respectively. (B) Four gRNA pairs targeting the promoter

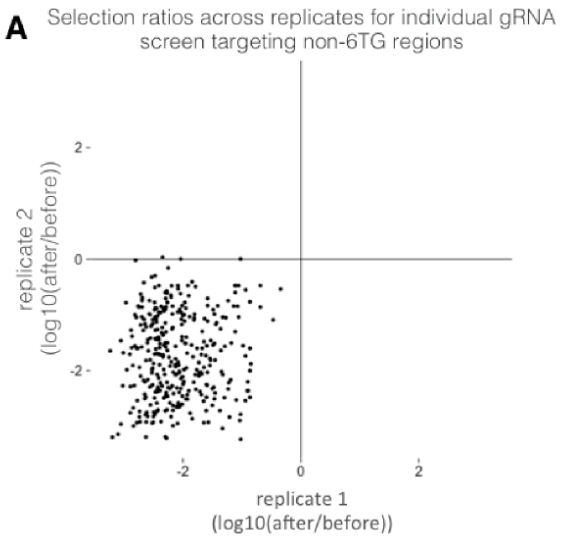
were cloned as a small pool, delivered, and selected with 6TG to enable sequencing of the edited locus (programmed deletions are displayed as red bars). A 3 kb region was amplified and sequenced with long reads (Pacific Biosciences). (C) The chart at the top displays the per-base percentages for deletions (red) and insertions (blue), and target sites are indicated by vertical gray bars. Horizontal bars show the edits found on each haplotype (red, deletions; blue, insertions; ranked by decreasing prevalence). All programmed deletions, in addition to rare, unexpected deletions, were abundant before 6TG treatment. The notations to the right indicate whether the edits interrupt the TSS or 50 UTR (purple bar) and/or coding sequence (orange bar). The unedited haplotype is marked with a green dot. Of note, PCR and sequencing on the PacBio RSII were biased toward smaller fragments, limiting accurate quantitative comparison of read counts from differently sized edits. (D) Haplotypes from 6TG-selected cells are plotted as in (C), revealing that only edits interrupting the TSS or 50 UTR survive selection and that no programmed or ‘‘promoter only’’ deletions survive selection. (E) Open chromatin (green as in A) and ScanDel signal suggest the presence of critical non-coding regulatory sequences in the first _2.7 kb of intron 1 (chrX: 133,593,871–133,596,998, hg19). (F) Five gRNA pairs that drove the signal in this intronic region were cloned and selected with 6TG as a small pool, as in (C). (G) A 3.1 kb region spanning the most-50 part of intron 1 was amplified and sequenced from cells sampled before 6TG selection. Haplotypes and per-base editing rates are diagrammed as in (C). (H) Post-6TG selection haplotypes from the intron-1-targeted cells are plotted as in (G), revealing that the vast majority of surviving edits disrupt the exon. Two edited haplotypes do not interfere with the exon, but these are present at approximately the level of unedited haplotypes, suggesting that 6TG resistance in these cells is caused by mutations elsewhere.

We next sought to validate the positive signal downstream of the first exon. To do so, we again repeated the experiment with a small pool of just 5 gRNA pairs targeting the first ~2.7 Kb of intron 1 (Fig. 2.10F). We then amplified the region and again performed long-read sequencing of the amplicons (Pacific Biosciences). As with the promoter, the programmed deletions were all well-represented before 6TG selection, although deletions with unprogrammed boundaries are also detected at a low rate (Fig. 2.10G). After selection, deletions with unprogrammed boundaries predominated again, particularly those that extended into the first exon, thereby disrupting coding sequences (Fig. 2.10H). A low rate of non-exonic deletions survived post-6TG, but these were present at the same level as unedited reads, implying that there may be some other explanation for 6TG resistance in these cells. Thus, as with the promoter, the positive signals that we originally observed for deletions in the first intron were likely consequent to the positive selection of rare 'on-target-but-with-incorrect-boundaries' deletions that extend into the first *HPRT1* exon.

2.7 An Individual guideRNA screen of the Same Region for Comparison to ScanDel

We next compared our ScanDel results against a more conventional screen relying on only individual gRNAs (Sanjana et al. 2016; Canver et al. 2015; S. Chen et al. 2015; Diao et al. 2016; Korkmaz et al. 2016) (Fig. 2.3E). For this, we cloned a second lentiviral library consisting of 12,151 individual gRNAs targeting the same 206 Kb region and assayed HPRT function in HAP1 cells as previously. Under the assumption that each individual gRNA potentially disrupts a ~10 bp region, this experiment at best interrogates ~70% of bases within the 206 Kb region due to the sparsity of PAM sites (as compared to our coverage of the entire locus at median ~27-fold redundancy per base-pair with ScanDel). 86% of exon-targeting gRNAs were positively selected and exonic selection scores were well correlated between biological replicates (Pearson: 0.781).

Of 612 negative control gRNAs, none that were sampled in each replicate were positively selected in both experiments (Fig. 2.11). In non-coding sequence, scores were poorly correlated between biological replicates, with a paucity of reproducible, positively selected signal (Pearson: 0.156, Fig. 2.12).



B	positively selected in both	negatively selected in both	positively selected in rep 1	negatively selected in rep 1	positively selected in rep 2	negatively selected in rep 2
gRNA targeting non-6TG resistance regions (**A**)	0	336	2	520	3	344
random sequence not found in hg19	0	9	0	12	0	9

Figure 2.11. None of the negative control random-sequence gRNAs were positively selected in both individual gRNA screen replicates.

A) Selection scores across replicates for individual gRNAs that target regions not expected to induce 6TG resistance (as described in Fig. 2.8). Only gRNAs sampled in both replicates are plotted. B) Table of the negative control gRNAs selected in both, either, or neither biological replicate.

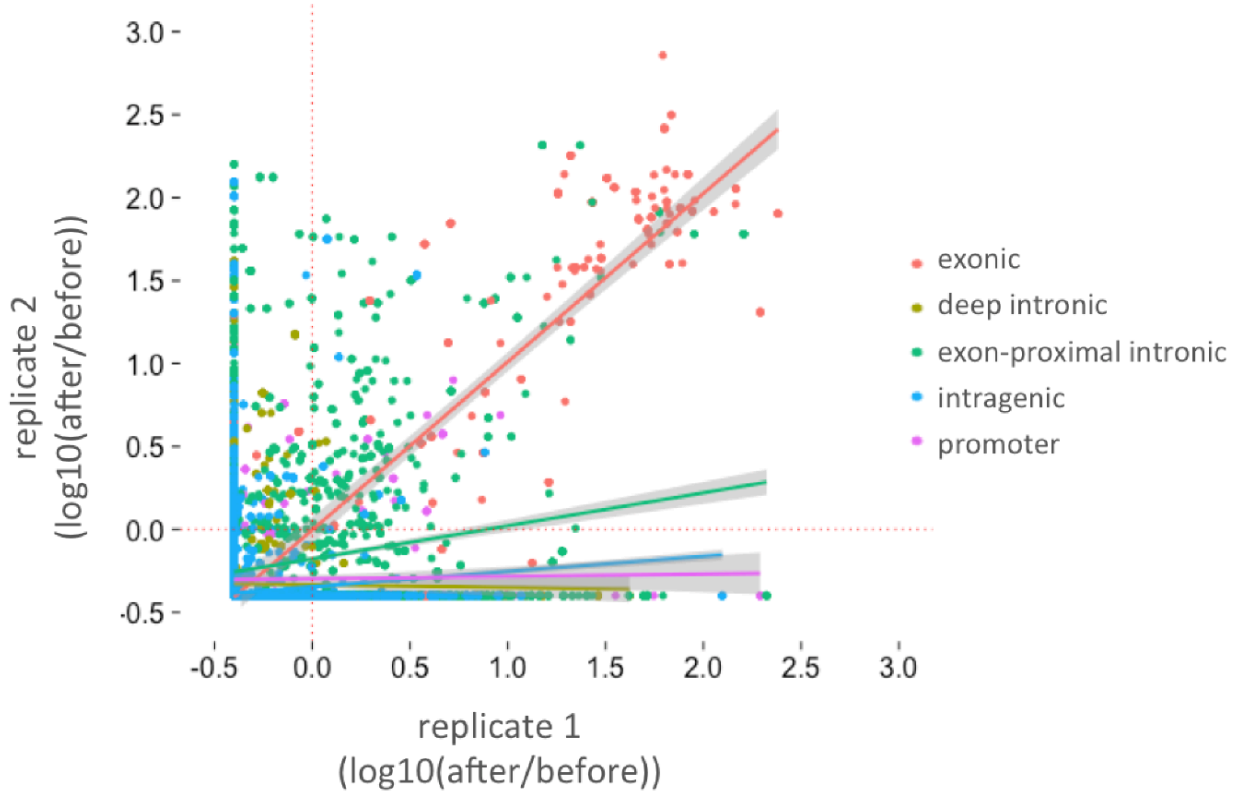


Figure 2.12. Correlation of the individual gRNA screen scores across two biological replicates. The individual gRNA scores for each biological replicate were calculated per base-pair and presented as mean of log10(after/before 6TG) between replicates. Least squares lines and points are colored by sequence content category. Specifically, intronic sequence within 2 Kb of an exon is colored in green (Pearson: 0.176); exons are red (Pearson: 0.818); deep intronic is yellow (Pearson: -0.14); intragenic sequences are blue (Pearson: 0.070; and promoter sequence (2 Kb upstream of the TSS) is purple (Pearson: 0.022).

Notably, we did observe a greater proportion of positively scoring gRNAs in the vicinity of exons – *i.e.* whereas only 2% of intergenic gRNAs were positively selected, 7.5% of deep intronic (>2 Kb away from an exon boundary) and 20.5% of proximal intronic (<2 Kb from an exon boundary) gRNAs were positively selected (Fig. 2.13A). Given our earlier observation with ScanDel of rare,

'on-target-but-with-incorrect-boundaries' that were confounding when targeting near exon boundaries, we next performed similar validation experiments on individual gRNAs that targeted non-coding sequences nearby exons (Fig. 2.13B). We chose 10 gRNAs in the *HPRT1* promoter region (Fig. 2.13C), and repeated the individual gRNA experiment with a small pool of just these 10 gRNAs, again using long reads (Pacific Biosciences) to sequence the locus before (Fig. 2.13D) and after 6TG selection (Fig. 2.13E). Similar to our results with ScanDel in this region, the only mutations that survived 6TG selection were initially rare deletions whose boundaries extended past the TSS and into the 5' UTR and/or coding sequence (Fig. 2.13D). This result strongly underscores that caution should be exercised in the interpretation of results from CRISPR-based screens of non-coding regions, whether performed with individual gRNAs or gRNA pairs, and the importance of sequencing-based validation of edited regions in the context of such screens.

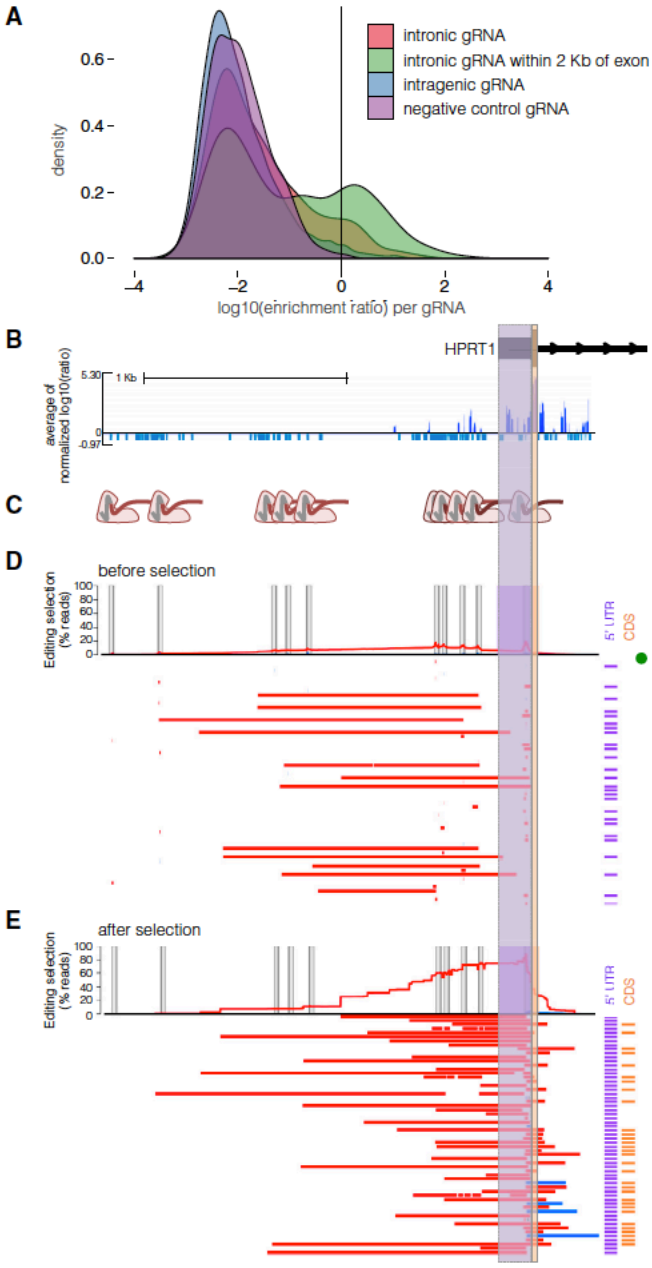


Figure 2.13. Direct genotyping of edits from an individual-gRNA mutagenesis screen also reveals rare, unexpected edits disrupting exon 1 of *HPRT1*.

(A) A greater proportion of gRNAs targeting non-coding sequence within 2 kb of exons were positively selected in an individualgRNA screen across the HPRT1 locus (data shown from replicate 1; Fig. 2.3E). Each gRNA was assigned a score equal to the log10 (after/before 6TG). (B) gRNAs that target upstream of the transcriptional start site were positively selected. The 2.4

kb region sequenced for genotype validation (chrX: 133,592,240–133,594,646, hg19), i.e., a zoom-in of data from the whole region in Fig 2.3E, is also shown. (C) For validation, ten gRNAs in this 2.4 kb promoter region were cloned into a low-complexity library, delivered to HAP1 cells expressing Cas9, and selected with 6TG. After selection, the 2.4 kb promoter region was amplified for long-read sequencing. (D) Reads from before 6TG selection are plotted as in Fig. 2.10C. In brief, the per-base percentage of haplotypes that carried a deletion (red) or insertion (blue) is charted. The edits of the most-prevalent haplotypes from long-read sequencing are drawn as colored bars, and the notations to the right indicate whether the edits interrupt the TSS or 50 UTR (purple) or coding sequence (orange) of exon 1. A green dot signifies the unedited haplotype. Target-site programmed edits are observed and are mainly composed of the expected small indels, in addition to rarely occurring larger deletions. PCR and sequencing on the PacBio RSII were biased toward smaller fragments, limiting accuracy of quantitative comparison of the read-count prevalence of different sized edits. (E) The most abundant haplotypes from cells after 6TG selection are visualized as in (D). Only mutations that interrupt exon 1 survived 6TG selection.

2.8 DISCUSSION

We developed a method that uses CRISPR/Cas9 and pairs of gRNAs to experimentally test the functional consequences of thousands of programmed, kilobase-scale genomic deletions in a single experiment. We applied this method to perform the systematic investigation of the regulatory architecture of a housekeeping gene via editing of the endogenous genome. Upon introducing a set of densely tiling deletions spanning a 206 Kb region centered on the gene *HPRT1*, we found no evidence for any distal regulatory element that is critical for its activity, as measured by 6TG sensitivity in HAP1 cells. A screen of this same region with individual gRNAs supported this finding. The dearth of positive selection from disruption of non-coding regions contrasts with the

strong positive selection observed from disruption of any exon of *HPRT1*, either by programmed deletions or individual guides.

HPRT1 is a widely expressed housekeeping gene (The GTEx Consortium 2015) with no eQTLs identified by the Genotype-Tissue Expression Project (Consortium and GTEx Consortium 2017), and thus may not require multiple (or any) distal regulatory regions for its expression. The simplest explanation of our results is that sequences immediately proximal to the *HPRT1* transcriptional start site may be sufficient to confer the level of expression that provides sensitivity to 6TG, such that even if we disrupt distal regulatory elements that subtly modulate expression, they would go undetected by our strong selection. For future applications of ScanDel, implementing more quantitative readouts will be critical. For example, ScanDel is compatible with any functional selection that reliably separates cells on the basis of gene expression (*e.g.* knocking in GFP to a locus of interest, and then using FACS to stratify ScanDel-edited cells on the basis of expression). Such quantitative readouts may facilitate validation of the many candidate regulatory elements (and cognate target gene assignments) nominated by eQTL and functional genomics studies (Won et al. 2016; Kumasaka, Knights, and Gaffney 2016). We anticipate that the application of ScanDel to non-housekeeping genes coupled to a more quantitative readout will likely identify more regulatory elements than found for *HPRT1*, especially for genes that play key roles in development and cell fate determination.

Another possibility, albeit an unlikely one, is that critical regulatory elements for *HPRT1* lie outside of the 206 Kb window that we surveyed. For example, the gene resides at the terminus of a ~300 Kb topologically associated domain identified in HAP1 cells that spans ~185 Kb beyond

our interrogated region (Sanborn et al. 2015) (Fig. 2.14). This could potentially be addressed by increasing the complexity of the library of programmed deletions in order to densely tile a larger region, or by simply increasing the size of each programmed deletion to interrogate more sequence per gRNA pair.

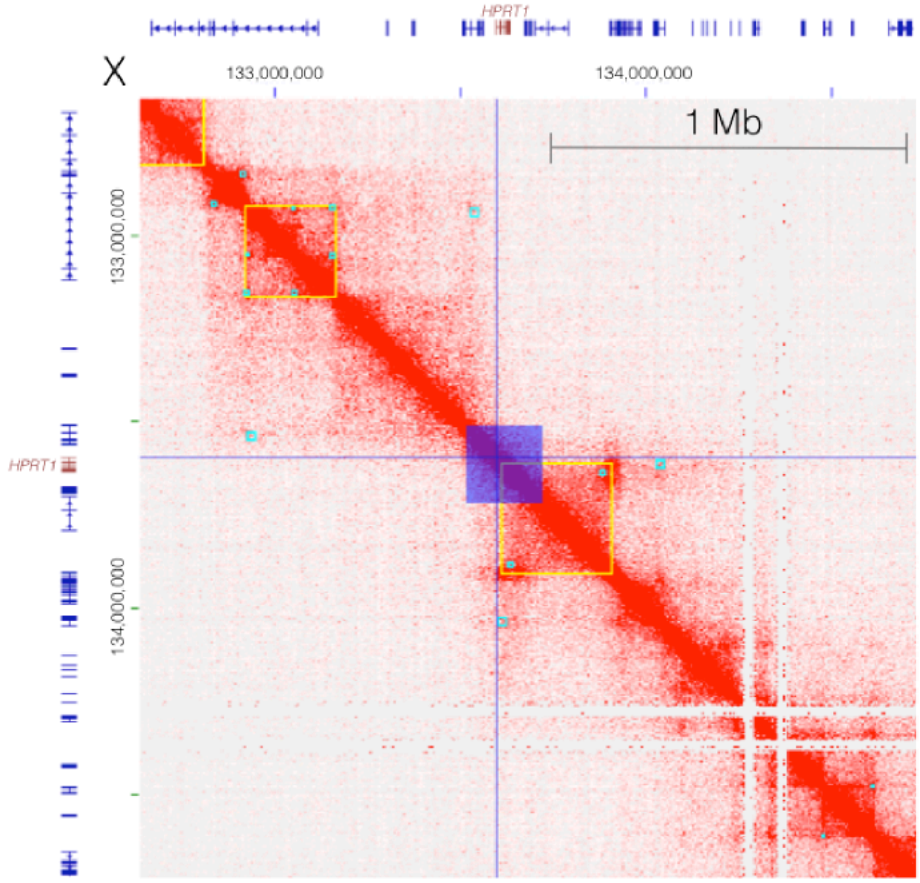


Figure 2.14. Region interrogated with ScanDel only partially surveys a 300 Kb HAP1 topologically associated domain.

A heatmap of interactions between 5 Kb bins along chrX:132,669,000-134,716,000 (hg19) in HAP1 cells (Juicebox 1.4, balanced normalization). RefSeq gene annotations are drawn across the axes, with the *HPRT1* gene model drawn in red. Blue lines mark its TSS and the 206 Kb surveyed by ScanDel is highlighted as a dark blue box. Light blue boxes mark peaks and yellow boxes mark domains as called by Sanborn et al.

We note that the paucity of regulatory sequences discovered by CRISPR/Cas9-based screening is not exclusive to this study. Collectively, individual gRNA CRISPR/Cas9 screens have surveyed over a megabase of prioritized non-coding sequences, but only a handful of gRNAs tested have robust phenotypic effects that validate (Canver et al. 2015; Diao et al. 2016; Korkmaz et al. 2016; Rajagopal et al. 2016; Sanjana et al. 2016). One explanation is that the assays being used are insufficiently sensitive and fail to detect modest regulatory effects. This could be addressed through the implementation of more quantitative assays.

A second explanation is that as implemented, genome editing has poor sensitivity due to redundancy in mammalian gene regulation. Redundancy of transcription factor binding sites within enhancers could prevent ~1-10 bp indels introduced by individual gRNAs from sufficiently disrupting function. Indeed, this was part of the motivation for developing ScanDel, whose programmable kilobase-scale deletions exceed the size of enhancers. Although we did not identify distal enhancers, the essentiality of the TSS and portions of the 5'UTR in our assay was detected primarily by deletions substantially larger than 1-10 bp (Fig. 2.13D, E), suggesting paired gRNA libraries will be effective for enhancing sensitivity. However, there may also be redundancy amongst sets of distal regulatory elements, a question which can only be fully addressed by combinatorial perturbations.

A third explanation is that gene expression levels depend in part on historical events, such that disruption of an enhancer in a differentiated cell line would not result in the same outcome as disrupting the same enhancer prior to differentiation. This could be potentially addressed by

performing lentivirally-mediated genome editing steps in stem cells, followed by differentiation to a cell type of interest. Any differences in functional consequences that are dependent on the timing of mutation would be of great interest.

Our results also provide a cautionary example of the importance of validation by direct genotyping in the context of CRISPR/Cas9-based screens of non-coding sequences. NHEJ generates a wide assortment of mutations, and strong selections may recover rare editing outcomes. For example, whereas targeting regions adjacent to exons might have been interpreted to reflect the presence of critical proximal regulatory elements, validation experiments using a long-read sequencer showed that this signal was caused by rare deletions that extended into exonic sequence. Many of these unexpected events would have been difficult to detect had we been relying solely on a short-read sequencing platform to genotype editing outcomes. Additionally, validating CRISPR/Cas9-based screens by assessing selection for specific edited haplotypes adds biological information. Here, with long-read genotyping we were able to identify a set of variable deletions that either did or did not drive selection, thus enabling greater resolution (Fig. 2.10C, D).

We also note that in experiments relying on pairs (or more) of gRNAs to program deletions, it is critical to include controls that quantify the effects of the individual gRNAs comprising these pairs, as these can have direct effects or off-target effects that might be misinterpreted as being consequent to the programmed deletion. While this manuscript was in preparation, a study was published that similarly used gRNA pairs to program deletion of a large number of lncRNAs, followed by phenotyping for cellular growth (Zhu et al. 2016). Although the results are of great interest, these important controls were not included for the vast majority of spacers used. It will

also be important to confirm the validity of each of this screen's findings through direct genotyping.

2.9 CONCLUSIONS

Even with the aforementioned open questions and remaining technical hurdles, it is critical that we continue to advance and apply methods for multiplex perturbation of the regulatory landscape with genome editing. The importance of experimental perturbation is highlighted by our results. The non-coding region surrounding *HPRT1's* first exon resides in open chromatin in this cell line (Fig. 2.3, Fig. 2.8), yet our results with ScanDel and subsequent validation experiments indicate the essential regulatory region is only a small part of the broader ATAC-seq peak. Perturbing the endogenous genome represents a highly complementary approach to the more classic strategy of reporter assays (Patwardhan et al. 2009; Banerji, Rusconi, and Schaffner 1981), in which short sequences are tested for their regulatory potential on an episomal vector. Of note, the results of early reporter assay-based tests of potential regulatory sequences flanking *HPRT1* are largely consistent with our findings but also identify three sequences immediately proximal to the first or second exons that are critical for episomal *HPRT1* expression). Though this discrepancy could be due to cell type or species differences (as two of these elements were required only in mouse embryonic stem cells but not human cells (Reid et al. 1990), and the remaining one was only tested in Chinese hamster fibroblasts (Rincón-Limas, Krueger, and Patel 1991)), it could also be due to differences in regulatory element activity when assayed via episomes versus genome editing. For example, elements necessary to drive expression of a gene on a plasmid may not be required in the genome, where redundancy is more likely. This underscores the ongoing challenge that genome editing can address: understanding how short sequences with regulatory potential coordinate with one another across endogenous loci to give rise to specific levels of expression.

In summary, ScanDel enables the multiplex characterization of the functional consequences of thousands of programmed, kilobase-scale deletions to the endogenous genome in a single experiment. We applied ScanDel to *HPRT1,* a housekeeping gene in which disruptive mutations cause Lesch-Nyhan syndrome, introducing densely tiled 1-2 Kb deletions across a 206 Kb region encompassing the gene, covering each base-pair with median ~27-fold redundancy. Our results demonstrate the absence of distal *cis*-regulatory elements in this region that are critical for *HPRT1* expression. In the future, we anticipate that large-scale perturbation of putative regulatory elements in their endogenous context with methods such as ScanDel will provide further insights into gene regulation and the contribution of non-coding mutations to human disease.

2.10 METHODS

Tissue culture: HAP1 cells were purchased from Horizon Discovery and cultured in Iscove's Modified Dulbecco's Medium with L-glutamine and 25 mM HEPES (Gibco). The HAP1 cell line was derived from the near-haploid KBM7 line (male cells of chronic myelogenous leukemia origin) by introduction of induced pluripotent stem cell factors. Despite the cell line's male origin, HAP1 cells no longer hold a Y chromosome (Essletzbichler et al. 2014). HEK293T cells were purchased from ATCC and cultured in Dulbecco's Modified Eagle's Medium with high glucose and sodium pyruvate (LifeTechnologies). Both media were supplemented with 10% Fetal Bovine Serum (Rocky Mountain Biologicals) and 1% Penicillin-Streptomycin (Gibco), and grown with 5% CO_2 at 37° C.

gRNA library design : To generate a list of gRNAs, we identified all 20 bp protospacers followed by a 5'-NGG PAM sequence from chrX:133,507,694-133,713,798 (hg19). We then excluded protospacers that had a perfect sequence match elsewhere in the genome, and scored the remaining gRNAs for both on-target and off-target activity. We considered off-target sequences that had five or fewer mismatches to the putative gRNA, and calculated an aggregate off-target score using the method of (Hsu et al. 2013). In addition we scored each site for on-target efficiency (Doench et al. 2014). Final deletion pairs were matched using spacers that did not contain BsmBI restriction sites, were not predicted to have off-target hits in other 6TG resistance genes or in KBM7 essential genes (the HAP1 parental cell line), were greater than 25 bp apart, further than 50 bp from an exon, and passed on-target (above 10) and off-target (above 25) thresholds. Contrastingly, the individual gRNA library included all of the spacers targeting the same region, excluding those predicted to have 2,000 or more off-targets or to have off-targets with 4 or fewer mismatches within the targeted *HPRT1* region.

Building the gRNA pair library: This library cloning method was developed in parallel to similar recently published methods(Aparicio-Prat et al. 2015) and is modified from the GeCKO single gRNA cloning scheme(Sanjana, Shalem, and Zhang 2014; Shalem et al. 2014). First, the lentiGuide-Puro backbone (Addgene #52963) is digested with BsmBI (FastDigest Esp3I, Thermo) and gel purified. The paired spacers (flanked with lentiGuide-Puro overlap sequences) are synthesized twice on a microarray (CustomArray, Inc.) such that each pairing is represented in both possible orders (Fig. 2.4).

To ensure quality of array synthesis, 1 ng of the oligo pool was amplified with Kapa HiFi Hotstart ReadyMix (KHF, Kapa Biosystems) and run on a gel to confirm oligos are of the expected 108 bp length. After PCR purification with Agencourt AMPure XP beads (Beckman Coulter), the amplicon is cloned into lentiGuide-Puro using In-Fusion HD Cloning Plus (Clontech) and transformed into Stable Competent *E. coli* (NEB C3040H) to minimize repeat-based recombination of the lentivirus. This ensuing library (lentiGuide-Puro-2xSpacers) now contains each pair of spacers, but is still missing the additional gRNA backbone and PolIII promoter.

We next cloned in the additional gRNA backbone and H1 promoter between each spacer pairing to enable expression of the two independent gRNAs. The gRNA backbone-H1 promoter fragment was ordered as a gBlock (IDT) with flanking BsmBI sites to allow ligation into the BsmBI-digested lentiGuide-Puro-2xSpacers library. The gBlock and the lentiGuide-Puro-2xSpacers are each digested with BsmBI, purified, ligated together with Quick Ligase (NEB M2200S), and transformed into Stable Competent *E. coli* to create a final lentiGuide-Puro-2xgRNA library.

To prevent bottlenecking of the library, these cloning steps are performed with enough replicates at high efficiency to maintain a minimum of 20x average library coverage (relative to the expected library complexity). Sequencing of the lentiGuide-Puro-2xgRNA library revealed 97.8% retention of diversity from the designed paired spacers. However, 16% of library reads held unprogrammed, interswapped pairs. 88.5% of these swaps are only seen in a single read, implying a more likely cause is template switching during either PCR or cluster generation. For all experimental analysis, only reads of gRNA pairs that perfect matched programmed pairs were considered.

Building the individual gRNA library: The spacers of this library were similarly synthesized on an array, amplified, and purified as above. The lentiGuide-Puro backbone was linearized as above, and the library cloned into it using the NEBuilder HiFi DNA Assembly Master Mix (NEB). This plasmid was transformed into Stable Competent *E. coli*, generating enough transformants for 30x average coverage. This method produced 98.5% retention of complexity from the designed array.

Lentiviral library production, delivery, and 6-thioguanine selection: Lentivirus was produced using Lipofectamine 3000 (Life Technologies) to transfect HEK293T with the lentiviral vector libraries made above and 3rd generation packaging plasmids (pMDLg/pRRE Addgene 12251, pRSV-Rev Addgene 12253, pMD2.G Addgene 12259). Supernatant was collected 72 hours after transfection, centrifuged at 300 rcf for 5 minutes to remove cell debris, and passed through a 0.45 μm syringe filter.

To create a monoclonal HAP1 cell line stably expressing Cas9, HAP1 cells were transduced with lentivirus produced using lentiCas9-Blast (Addgene 52962), selected with 5 μg/mL Blasticidin (Thermo Fisher Scientific), and single-cell sorted via FACS.

HAP1-Cas9-Blast monoclonal cells were plated to be at 30% confluency on the day of lentiviral gRNA/pair transduction. To transduce, 5% of the recipient cells' media was replaced with filtered virus, limiting the MOI to < 0.3. Media was changed after 24 hours, and selection for transduced cells began 48 hours post-transduction. Puromycin was added at 2 μg/mL for two days to assess the percentage of cells transduced, and then cells were maintained in 1 μg/mL for 5 more days.

After puromycin treatment, an initial population of cells was collected. Selection for loss of HPRT function was performed by applying 5 uM 6TG to the remaining cells at <50% confluency for 7 days. An additional concern is that minor changes in gene expression caused by ScanDel-mediated mutations in regulatory elements will not be strong enough to confer resistance. To mitigate this, we used the lowest dosage of 6TG that completed HAP1 selection after seven days. 6TG concentrations of 6-60 μM are reported in the literature to achieve effective selection in this timeframe, depending on cell type (Jacobs and DeMars 1984; Monnat 2009). We tested our monoclonal HAP1-lenti-Cas9-Blast line at concentrations just below this range (1 μM, 2.5 μM, and 5 μM 6TG). After 7 days, the 5 μM treatment had no readily identifiable surviving cells, whereas the 2.5 μM treatment retained a sparse population, and the 1 μM treatment produced appreciably more outgrowing colonies. Based on these results, we proceeded with selections using 6TG at 5 μM for seven days. Enough cells were transduced and sampled at each timepoint to maintain minimum 2,000x average coverage of the library in each population.

Sequencing of the baseline (*i.e.* pre-6TG) population revealed 98.4% of diversity of the lentiGuide-Puro-2xgRNA library was preserved from replicate 1, and replicate 2 retained 78.8%. As our deletions are highly overlapping, we proceeded with replicate 2 as all base-pairs are interrogated despite the lower diversity. We observed 95.6% retention of programmed library diversity in replicate 1 of single gRNA plasmid library and 71.2% of replicate 2.

Interswapped gRNA pairs were observed in 35.5% of reads from the baseline pre-6TG sample. This is an increase from the 16% observed in reads from the lentiGuide-Puro-2xgRNA plasmid library. This suggests additional template switching during the library's amplification from gDNA,

which requires more cycles of PCR. However, since we are directly sequencing each gRNA spacer as a read out opposed to using barcoded libraries (Zhu et al. 2016) and only taking exact sequence matches, this does not pose a problem.

gRNA library amplification and sequencing from HAP1 cells: gDNA was extracted from the cells sampled before and after 6TG selection using the DNeasy Blood & Tissue kit (QIAGEN). KHF was used for all amplification steps. The libraries were initially amplified from a minimum of 6 ug of gDNA divided across thirty 50 μL reactions, ensuring sampling of ~2 million haploid genome equivalents at each timepoint. Two additional PCRs were performed to add sequencing adapters and sample indices to the amplicon, with AMPure bead purification between each reaction. Amplification conditions were optimized using qPCR to minimize overamplification of the construct.

Sequencing was performed on an Illumina Miseq using a 50-cycle kit. Read 1 and the Illumina Index read were used to sequence the two gRNAs in the paired gRNA construct prior to paired-end turnaround, and Read 2 was used to sequence the 9 bp sample index.

Calculation of a selection score assignment per base-pair: Custom Python scripts counted tallies of gRNAs (for individual gRNA library experiments) or gRNA pairs before and after selection. These counts were normalized to the total number of reads per sample. An enrichment ratio was calculated for each gRNA/pair by dividing its normalized read count after selection by its before selection read count. A selection score is the log10 of the enrichment ratio (log10(after/before)). If a gRNA or gRNA pair was absent before selection, it was excluded from further analysis. Any

gRNA pairs that used a self-paired gRNA with an independent selection ratio > 0 were also excluded from further analysis.

If a gRNA/pair is absent after 6TG selection, its selection score as calculated will be a negative number relatively large in magnitude that is somewhat arbitrarily determined by the number of pre-selection reads. Thus, to limit the contribution of these scores to average measurements derived from many independent deletions, we set a minimum selection score equal to the middle of the bimodal distribution between the positively and negatively selected deletions of each replicate (Fig. 2.5). For example, in ScanDel replicate 1, if the log10-value of a selection score was less than -0.35, that gRNA pair's score was set to -0.35. Each individual base-pair was assigned a per base-pair selection score by taking the mean of all deletions programmed to cover that base-pair. The per base-pair score was normalized to the median score for all positive scores in that replicate. The per base-pair selection score of each replicate was averaged to get the final selection score per base-pair. Per base-pair scores were uploaded as a bedgraph for visualization on the UCSC Genome Browser.

For the individual gRNA mutagenesis screen, we calculated selection scores per base-pair similarly, assuming a 10 bp deletion was made by each gRNA queried. If a base-pair was scored at the minimum negative threshold in one screen, it was given that value for the consensus selection score of the two replicates.

Bulk ATAC-seq of HAP1 cells: Two biological replicates were separately maintained (on 10cm dishes, split 1:10 three times per week) and processed separately. Chromatin accessibility in the

HAP1 cell line was profiled with the ATAC-seq protocol (Buenrostro et al. 2013) with slight modifications. The media for 10cm plates of confluent HAP1 cells was aspirated and replaced with 2 mL of ice cold lysis buffer ('CLB+'; made as described in the original paper, but supplemented with protease inhibitors (Sigma cat. no. P8340)). Cells were incubated on ice for 10 minutes in CLB+ and then were dislodged with a cell scraper and transferred to a 15 mL conical tube and pelleted at 500 rcf for 5 min at 4° C. Nuclei were re-suspended in 1 mL of CLB+ and counted on a hemocytometer. 50,000 nuclei in 22.5ul of CLB+ were combined with 2.5 µL of TDE1 enzyme and 25ul of TD buffer (Illumina). Tagmentation conditions were as described in the original paper (37° C for 30 min). After MinElute purification into 10 µL EB buffer (Qiagen), 5 µL of tagmented DNA was amplified in 25 µL reactions for 12 cycles using the NEBNext Master Mix (NEB). Reactions were monitored with SYBR Green to ensure that samples were not overamplified. PCR products were cleaned once with a QiaQuick PCR Cleanup Kit (Qiagen) and once with 1x AMPure beads (Agencourt). The quality of the library was assessed on a 6% TBE gel and the yield was measured by Qubit (1.0) fluorometer (Invitrogen).

Samples were sequenced on two paired-end Illumina NextSeq 500 runs. Read lengths were 2x75 bp for the first run and 2x151 bp for the second run, so the second run was truncated to 75 bp. Sequencing reads were also trimmed for read-through of adapter sequences and quality with Trimmomatic (Bolger, Lohse, and Usadel 2014) ('NexteraPE-PE.fa:2:30:10:1:true TRAILING:3 SLIDINGWINDOW:4:10 MINLEN:20' parameters) and then mapped to the 1000 genomes integrated reference genome 'hs37d5' with bowtie2 (Langmead and Salzberg 2012), using the '-X 2000 -3 1' parameters. Only properly paired and uniquely mapped reads with a mapping quality above 10 were retained ('samtools -f3 -F12 -q10'). Reads mapping to the mitochondrial genome

and non-chromosomal contigs were also filtered out. In addition, duplicate reads were removed with Picard. After checking QC metrics on the individual replicates, reads from the two libraries were combined for downstream analysis. Hypersensitive sites were called (at a 1% false discovery rate) with the Hotspot algorithm (John et al. 2011).

Validation and direct genotyping of positive signal from the screens: gRNA pairs that drove the ScanDel signal surrounding *HPRT1*'s first exon were cloned into simple lentiGuide-Puro-2xgRNA libraries. The TSS ScanDel validation library contained four pairs and the intron 1 library contained five. For the individual gRNA screen TSS library, ten gRNAs were cloned into lentiGuide-Puro. These constructs were lentivirally delivered to HAP1-Cas9-Blast cells, selected with 6TG, and gDNA extracted as described above.

As the expected deletions could remove up to two kilobases, the loci were sequenced with a Pacific Biosciences RSII (University of Washington PacBio Sequencing Services, P6C4 chemistry, RSII platform). To prepare libraries for PacBio sequencing, the TSS- or intron 1-targeted regions were amplified from 800 ng of gDNA each, using four 50 µL KHF reactions with primers adding sample indices and SbfI or NotI cut sites. The purified amplicons (Zymo Research DNA Clean & Concentrator-5) were digested with SbfI-HF (NEB) and NotI-HF (NEB), leaving sticky ends. 5'-phosphorylated SMRT-bell hairpin oligos (IDT) containing the PacBio priming site, hairpin-forming sequence, and resulting sticky ends for either SbfI or NotI were annealed by heating to 85° C and snap frozen in 10mM Tris 8.5, 0.1mM EDTA, 100 mM NaCl. These were ligated at 10x molar excess to the digested amplicons, destroying the restriction site once attached. To remove

undigested amplicons and primers, this ligation was performed in the presence of further SbfI and NotI, and followed by treatment with Exo7 (Affymetrix) and Exo3 (Enzymatics).

Only reads with over five circular consensus sequence passes and containing the expected first twelve 5' and 3' base-pairs of the amplicon were used for further analysis. Reads positive for complex inversions (>= 100 basepairs) were removed from the library using the Waterman-Eggert algorithm with match, mismatch, gap open, and gap extend scores of 2, 10, 10, and 5, respectively (Döring et al. 2008). The resulting reads were then were aligned to the amplicon reference using the NEEDLEALL (Rice, Longden, and Bleasby 2000) aligner with a gap open penalty of 10 and a gap extension penalty of 0.5. Insertions were required to start within a window of five bases up or downstream of the putative cut site. Deletions were required to either start or end within the same 10 bp window or span the window. Reads that carried the same edit pattern were collapsed into haplotypes, and figures were generated using a custom D3 script.

Comparing deletion rate of U6-H1 versus U6-U6: Two protospacers were chosen to program a 365 bp deletion within the second intron of *HPRT1* and their spacers were cloned into a U6-H1 construct and U6-U6 construct (Fig 2.2). Virus was produced and delivered to cells, which were selected with puromycin, and gDNA extracted as described above. The locus was amplified in four successive rounds of nested PCR. The first reaction was only 3 cycles and included a forward primer with a 10-bp unique molecular index (UMI). The second reaction amplified any UMI-tagged fragments. The third and fourth reactions added sample indices and Illumina flow cell adapters. The products were AMPure cleaned between each reaction at a concentration that would lose primer dimer but retain the smaller deletion-holding fragments, and sequenced on a MiSeq.

Any reads that contained the same UMI or edit pattern were collapsed using custom scripts and their alignments were visualized with the same D3 script as above.

Chapter 3. A GENOME-WIDE FRAMEWORK FOR MAPPING GENE REGULATION VIA CELLULAR GENETIC SCREENS

Chapter 3 was adapted with minimal modifications from:

Gasperini, Molly, Andrew J. Hill, José L. McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D. Zhang, Dana Jackson, et al. 2019. "A Genome-Wide Framework for Mapping Gene Regulation via Cellular Genetic Screens." *Cell*, Volume 176, Issue 1, p377-390.e19, January 10.

3.1 ABSTRACT

Over one million candidate regulatory elements have been identified across the human genome, but nearly all are unvalidated and their target genes uncertain. Approaches based on human genetics are limited in scope to common variants, and in resolution by linkage disequilibrium. We present a multiplex, expression quantitative trait loci-inspired framework for mapping enhancer-gene pairs by introducing random combinations of CRISPR/Cas9-mediated perturbations to each of many cells, followed by single-cell RNA-seq. Across two experiments, we used dCas9-KRAB to perturb 5,920 candidate enhancers with no strong *a priori* hypothesis as to their target gene(s), measuring effects by profiling 254,974 single-cell transcriptomes. We identified 664 (470 high-confidence) *cis* enhancer-gene pairs, which were enriched for specific transcription factors, non-housekeeping status, and genomic and 3D conformational proximity to their target genes. This framework will facilitate the large-scale mapping of enhancer-gene regulatory interactions, a critical yet largely uncharted component of the *cis*-regulatory landscape of the human genome.

3.2 CONTRIBUTIONS

For this manuscript, the initial idea for multiplexing CRISPR perturbations per cell like a true human eQTL study was generated by me and Jay. However, the initial conception was to implement this by generating hundreds of monoclonal cell lines. After struggling with this for two years and inspired by recent work by Andrew Hill and Jose McFaline (Hill et al. 2018), I conceived of the switch to a single-cell CRISPRi perturbation framework (3 months prior to the publication of Xie et al. 2017).

I performed all primary experimental work for this manuscript (sgOPTI-CROP-seq synthesis, gRNA cloning, transductions, single-cell RNA-seq for all 254,974 cells, bulk RNA-seq, Nextseq sequencing). Secondary experimental work was provided by post-doc Jose McFaline (digesting sgOPTI-CROP-seq vector with BsmBI, helping run the 10X machine for the 30 lanes requires for the 207,324 cells scRNA-seq experiment), research scientist Dana Jackson (brief assistance with the 30 lanes requires for the 207,324 cells scRNA-seq 10X experiment), research scientist Beth Martin (assistance with tissue culture and genotyping the monoclonal validation lines), undergraduate research Melissa Zhang (assistance with tissue culture and bulkRNAseq of the half of the singleton validations); lentivirus was made by the Fred Hutch core, Pacbio sequencing by the UW Pacbio core, and the Novaseq run performed by Northwest Genomics Center. All flowFISH and post-sort gDNA extraction was developed and performed by research scientist Anh Leith on cells generated by me.

All secondary analysis and figures were generated by me, with the exception for Figure 3.10E/3.11B-C (generated by Hi-C expert graduate student Seungsoo Kim), and Figure 3.12D,

3.13, and 3.14 by graduate student Andrew Hill. Jacob Schreiber identified 1/6th of the candidate enhancers tested in the at-scale screen by using a model he built to unbiasedly survey the genome for candidate enhancers. Andrew (with assistance from me) performed the primary scRNA-seq processing, gRNA assignment, and differential expression testing, building off his Hill et al. 2018 scRNAseq and gRNA assignment pipelines. He also provided significant data analysis advice and computational mentorship to me throughout the entire project. The manuscript was written by me and Jay, and we were both cited as co-corresponding authors.

3.3 INTRODUCTION

Consequent to an era of biochemical surveys of the human genome (*e.g.* ENCODE) and 'common variant' human genetics (*i.e.* GWAS & eQTL studies), we are awash in candidate regulatory elements and phenotype-linked haplotypes, respectively (ENCODE Project Consortium 2012; MacArthur et al. 2017). Determining whether and how each candidate regulatory element is truly functional, as well as pinpointing which noncoding variant(s) are causal for each genetic association, will require functional characterization of vast numbers of sequences.

We and others have recently adapted cell-based CRISPR/Cas9 genetic screens to evaluate candidate regulatory sequences in their native genomic context (Sanjana et al. 2016; Gasperini et al. 2017; Canver et al. 2015; Korkmaz et al. 2016; Diao et al. 2016; Rajagopal et al. 2016; Klann et al. 2017; Charles P. Fulco et al. 2016; Diao et al. 2017). However, two aspects of these studies limit their scalability. First, they focus on the regulation of a single gene per experiment, typically entailing the development of a gene-specific assay. Second, each cell is a vehicle for one CRISPR-mediated perturbation, with the specificity-conferring guide-RNAs (gRNAs) usually introduced

via lentivirus at a low multiplicity of infection (MOI). With millions of candidate regulatory elements and ~20,000 regulated genes in the human genome, these limitations preclude the comprehensive dissection of the *cis*-regulatory architecture of even a single cell line.

Here we introduce a framework (Fig. 3.1A) designed to overcome both limitations. First, by using single-cell RNA-seq (scRNA-seq) instead of gene-specific assays, one experiment can globally capture perturbations to gene expression (Adamson et al. 2016; Dixit et al. 2016; Xie et al. 2017; Hill et al. 2018; Jaitin et al. 2016), with no strong *a priori* hypothesis as to the target gene of each regulatory element tested. Second, by introducing gRNAs at a high MOI, each individual cell acquires a unique combination of perturbations against the isogenic background of a cell line. Introducing multiple perturbations per cell markedly increases power (Fig. 3.1B). An association framework inspired by eQTL studies (Morley et al. 2004; Stranger et al. 2012) is used to map *cis*- and *trans*- effects by comparing gene expression in the subset of cells that contain a given gRNA to those that lack that guide. This strategy is analogous to conventional eQTL studies, but with individuals replaced by cells; variants replaced by unique combinations of gRNAs per cell to induce multiplex CRISPR-interference (CRISPRi); and tissue-level RNA-seq replaced by scRNA-seq. However, unlike eQTL studies, the resolution of our screen is not constrained by linkage disequilibrium, nor is it limited to studying sites in which common genetic variants happen to exist. Although we recognize the imperfection of the analogy given that a reverse genetic screen using CRISPRi is far from equivalent to mapping the natural genetic variation that underlies QTLs, the fact that we were directly inspired by the eQTL framework led us to originally term this method ‘crisprQTL mapping’.

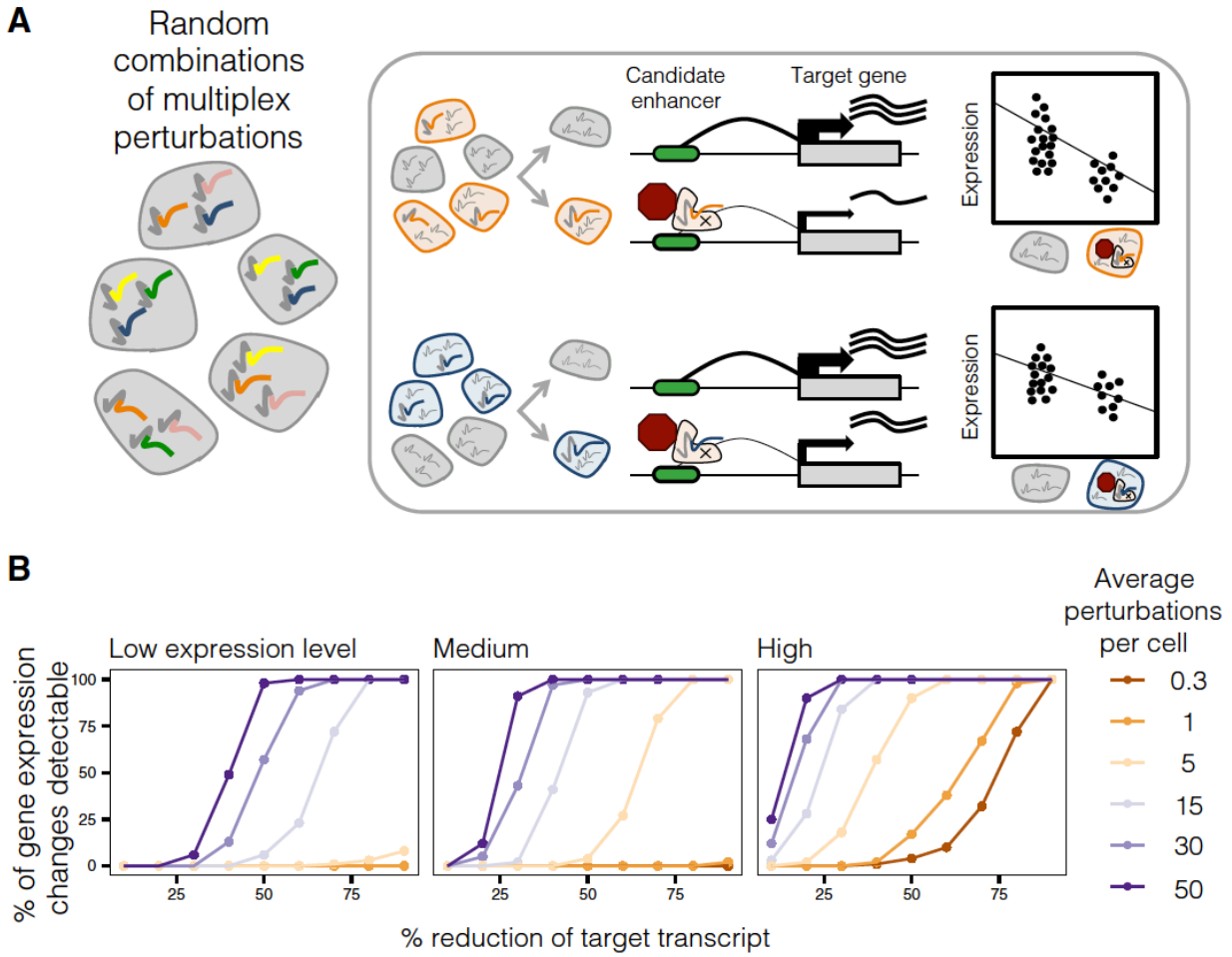


Figure 3.1. Multiplex enhancer-gene pair screening.

(A) Enhancer-gene pairs are screened by introducing random combinations of CRISPR/Cas9 candidate enhancer perturbations to each of many cells, followed by scRNA-seq to capture expression levels of all transcripts. Then, all candidate enhancers are tested against any gene by correlating presence of any perturbation with reduction of any transcript. (B) Multiplex perturbations increase power to detect changes in expression in single cell genetic screens while greatly reducing the number of cells that need to be profiled. Power calculations on simulated data show that increasing the number of perturbations per cell increases power to detect changes in expression, including for genes with low (0.10 mean UMIs per cell), medium (0.32) or high (1.00) mean expression. X-axis corresponds to the simulated % repression of target transcript.

3.4 A Proof-of-Concept Multiplex Enhancer-Gene Pair Screen

To establish the feasibility of the assay formerly known as crisprQTL mapping, we targeted 1,119 candidate enhancers in the chronic myelogenous leukemia cell line K562, with CRISPRi as our mode of perturbation. For CRISPRi, we used a nuclease-inactive Cas9 tethered to the KRAB repressor domain to induce heterochromatin across a ~1-2 kilobase (Kb) window around a gRNA's target site (Thakore et al. 2015). The 1,119 candidate enhancers were all intergenic DNase I hypersensitive sites (DHSs) representing various combinations of H3K27 acetylation, p300, GATA1, and RNA Pol II binding (Fig. 3.2A). Candidate enhancers were required to fall within the same TAD as at least one gene from the top decile of K562 expression, and were collectively distributed across 510 TADs on every chromosome (Rao et al. 2014). 5,611 of the 12,984 genes expressed in K562 cells fell within 1 megabase (Mb) of at least one candidate enhancer (K562-expressed genes defined as those observed in at least 0.525% of profiled cells).

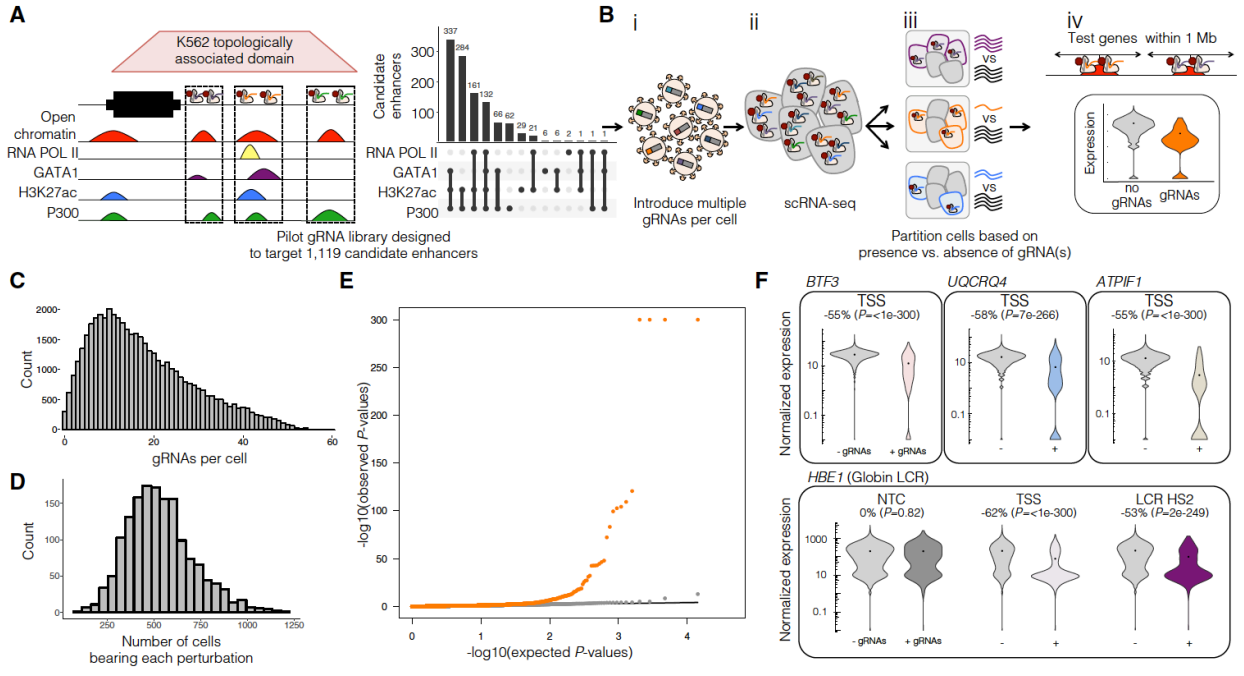


Figure 3.2. Pilot multiplex enhancer-gene pair screen testing 1,119 candidate enhancers in K562 cells.

(A) 1,119 candidate enhancers were chosen based on intersection of enhancer-associated features, and each targeted by two gRNAs. (B) Schematic of this multiplex enhancer-gene pair screening method: i) gRNAs were cloned into a lentiviral vector, and delivered to K562 cells at a high MOI; ii) scRNA-seq was performed on these cells, with concurrent capture of the multiple gRNAs present in each cell; iii) For each candidate enhancer, cells were partitioned based on whether or not they contained a gRNA targeting it; iv) For each such partition, we tested for differential expression between the two populations for any gene within 1 Mb of the candidate enhancer. (C) gRNAs were delivered to K562 cells at a high MOI, with median of 15 +/- 11.3 gRNAs identified per cell. (D) A total of 47,650 single cell transcriptional profiles were generated. Each perturbation was identified in a median of 516 +/- 177 cells. (E) Quantile-quantile plot of the differential expression tests. Distributions of observed vs. expected *P*-values for candidate enhancer-targeting gRNAs (orange) and NTC gRNAs (gray; downsampled) are shown. (F) Expression of selected TSS (top row) and β-globin LCR positive controls (bottom row). Nearly all targeted TSSs, and all positive controls, showed significant differential expression of the expected target genes between cells *with* (+) vs. *without* (-) targeting gRNAs, in contrast with NTCs. Percent changes and *P*-values show the effect size and significance of differential expression of the denoted target gene between these cell groups.

Two gRNAs were designed to target each candidate enhancer. Additional pairs of gRNAs served as positive controls (targeting the transcription start sites (TSSs) of genes sampled from the top decile of K562 expression, or alternatively hypersensitivity sites of the β-globin locus control region (LCR)) and negative controls (50 non-targeting controls or 'NTC' that target nowhere or in a gene desert).

This gRNA library was cloned into the lentiviral CROP-seq vector modified to include a CRISPRi-optimized backbone (Datlinger et al. 2017; B. Chen et al. 2013; Hill et al. 2018) and K562 cells were transduced at a high MOI (Fig. 3.2B). After 10 days to allow for effective CRISPRi, the transcriptomes of 47,650 single cells were profiled. With a targeted amplification protocol (Dixit et al. 2016; Adamson et al. 2016; Hill et al. 2018), we identified a median of 15 +/- 11.3 gRNAs per cell (Fig. 3.2C). Each candidate enhancer or control was targeted in a median of 516 +/- 177 cells (Fig. 3.2). For each targeted element, we partitioned the 47,650 cells based on whether they did or did not contain gRNA(s) targeting it. We then tested for a reduction in the expression of each K562-expressed gene within 1 Mb of that element (Fig. 3.2, (Stranger et al. 2012)). We also tested the 50 NTCs against all K562-expressed genes within 1 Mb of any targeted candidate enhancer. For perspective, with a 'one gRNA per cell' framework, achieving equivalent power would require profiling the transcriptomes of ~715,000 single cells.

A quantile-quantile plot showed an excess of significant associations involving the targeting of candidate enhancers relative to NTC controls (Fig. 3.2). We defined a 3.5% empirical FDR threshold based on the NTC tests as they are subject to the same sources of error as the element-targeting gRNAs. At this threshold, 94% (357 of 381) of TSS-targeting positive controls repressed their associated genes, as did all β-globin LCR controls (examples shown in Fig. 3.2). Additionally, we reidentified a known enhancer 3.6 Kb upstream of *GATA1* (Charles P. Fulco et al. 2016).

At this same threshold, targeting of 11% of candidate enhancers (128 of the 1,119) repressed 1+ gene(s) within 1 Mb. As there were 13 candidate enhancers whose targeting impacted more than one gene (Fig. 3.3A), this analysis yielded a total of 145 enhancer-gene pairs. Of the 105 downregulated target genes (Fig. 3.3B), 26 were impacted by targeting of more than one of the 128 candidate enhancers (Fig. 3.3A).

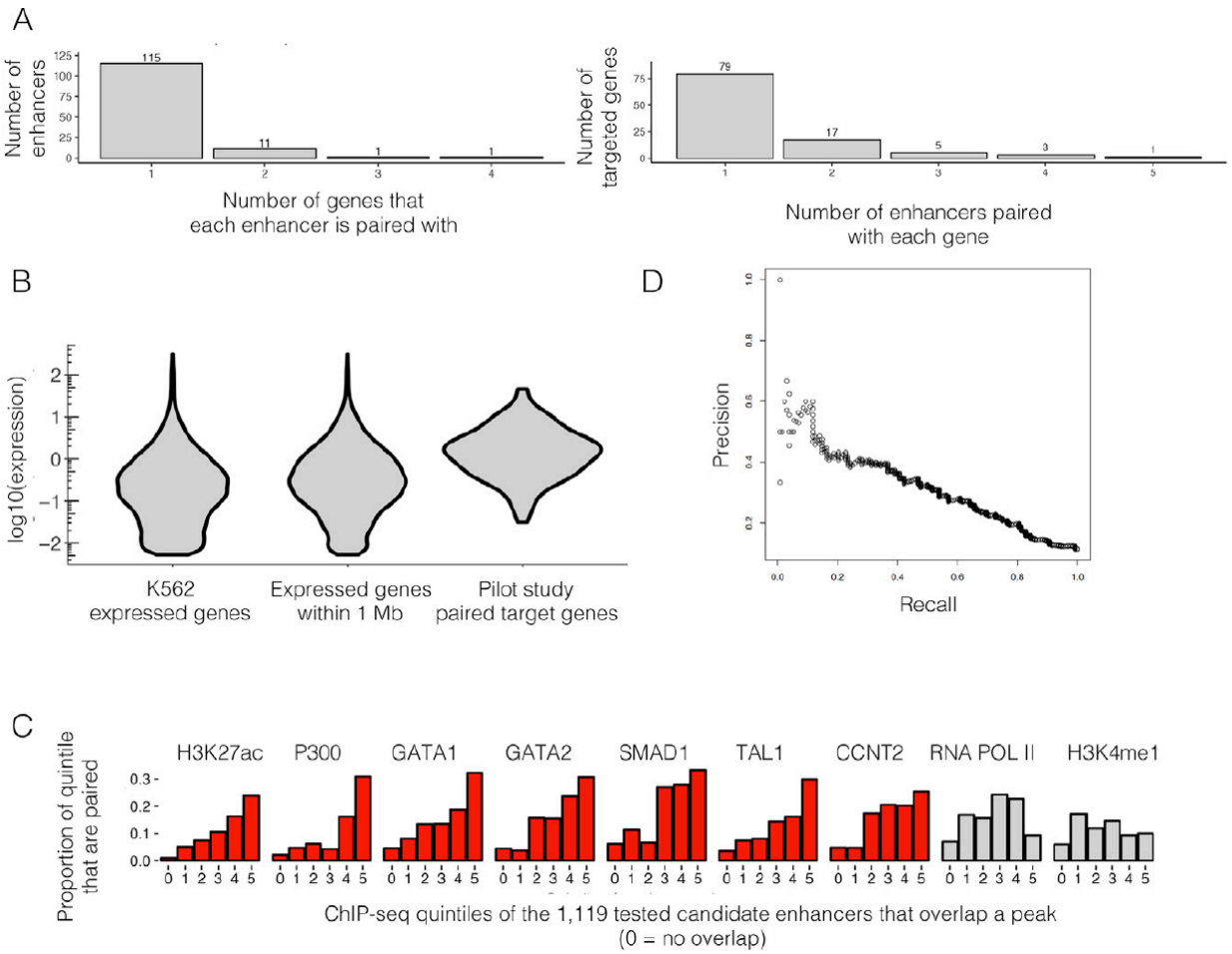


Figure 3.3. Details of 145 enhancer-gene pairs originally identified in the pilot screen.

(A) Histogram per enhancer of the number of genes paired with that enhancer (3.5% empirical FDR in pilot screen, *left*). Histogram per gene of the number of enhancers paired with that gene (3.5% empirical FDR in pilot screen, *right*). (B) Expression of target genes paired with candidate

enhancers in the pilot screen. expression = mean transcript UMIs/cell in the entire 47,650 cell pilot dataset for: K562 expressed genes; those that fell within 1 Mb of a targeted candidate enhancer in the pilot experiment; and for the 105 genes targeted in the pilot experiment's pairs. In the pilot screen, tested candidate enhancers were required to fall within TADs that contained genes highly expressed in K562s. As these were then only tested for pairing with genes within 1 Mb, the pilot screen's target genes are potentially biased towards being highly expressed. This enrichment for highly expressed genes is not seen in the at-scale experiment, where tested candidate enhancers were not required to be in the same TAD as a highly expressed gene. (C) Relative to the 1,119 candidate enhancers tested, the 128 paired candidate enhancers from the pilot experiment (3.5% empirical FDR) tend to fall in enhancer-associated ChIP-seq peaks that show stronger signals. All ChIP-seq peaks that overlap the 1,119 candidate enhancers were divided into quintiles of strength, defined as the average enrichment in ChIP-seq peak region (0 = no such peak overlaps the candidate enhancer, 1 = lowest, 5 = highest). Histograms of the proportion of each 1,119-quintile that were called as enhancer-gene pairs are shown. Red = *P*-value < 0.005 for independent logistic regression for predicting a candidate enhancer as paired based on this peak type. (D) Precision-recall curve for a multivariate logistic regression classifier based on ENCODE enhancer-associated biochemical features that differentiates the 128 paired candidate enhancers from the remaining of the 1,119 candidate enhancers. The median AUPR from five-fold cross-validation was 0.31.

We examined the characteristics of paired enhancers whose targeting significantly impacted expression of 1+ genes in *cis.* We found paired candidate enhancers to be enriched for high ChIP-seq peak strength (based on average enrichment in ChIP-seq peak region) for enhancer-associated histone modifications (H3K27ac, logistic regression *P*-value = 4e-5, candidate enhancers in the

top quintile were 1.4-fold more likely to be paired than those in the bottom quintile), certain co-activators (p300, *P*-value = 4e-16, 1.1-fold) and lineage-specific TFs (GATA1 *P*-value = 2e-7, 1.4-fold; GATA2 *P*-value = 3e-10, 1.5-fold; SMAD1 *P*-value = 1e-6, 1.4-fold; TAL1 *P*-value = 6e-6, 1.1-fold; CCNT2 *P*-value = 3e-7, 1.4-fold), whereas RNA Pol II and H3K4me1 were not associated (Fig. 3.3C). Using these features, as well as average enrichment within the DHS and whether each had been previously validated *in vivo (Visel et al. 2007)*, we trained a multivariate logistic regression classifier to distinguish the 128 paired candidate enhancers from the 991 candidate enhancers for which we did not identify a target gene, achieving an AUPR of 0.31 (area under precision-recall curve; median from five-fold cross validation; Fig. 3.3D).

3.5 A Scaled Multiplex Enhancer-Gene Pairs

To demonstrate this approach at a substantially greater scale, we performed a second experiment targeting five times as many candidate enhancers (n = 5,779). First, two-thirds of these (n = 3,853) were new DHSs chosen by the classifier trained on the first experiment (Fig. 3.3D; Fig. 3.4). Second, as this set may be biased towards annotations used to select the initially targeted candidate enhancers (Fig. 3.2A), we also targeted 948 exploratory DHSs chosen independent of the model (see Methods). Third, we re-targeted 978 of the 1,119 initially targeted pilot candidate enhancers, including the aforementioned candidate enhancers paired with target genes in the pilot. Altogether, candidate enhancers targeted in this scaled experiment were within 1 Mb of 10,560 of 13,135 K562-expressed genes. As previously, we designed two gRNAs per candidate enhancer. However, to evaluate whether poorly efficacious gRNAs might contribute to false negatives, we designed an additional two gRNAs for 377 of the 978 re-targeted candidate enhancers (Fig. 3.4B). Finally, in addition to gRNA pairs targeting 5,779 candidate enhancers, we included the same positive and negative control gRNA pairs targeting 381 TSSs, the globin LCR, and 50 NTC pairs.

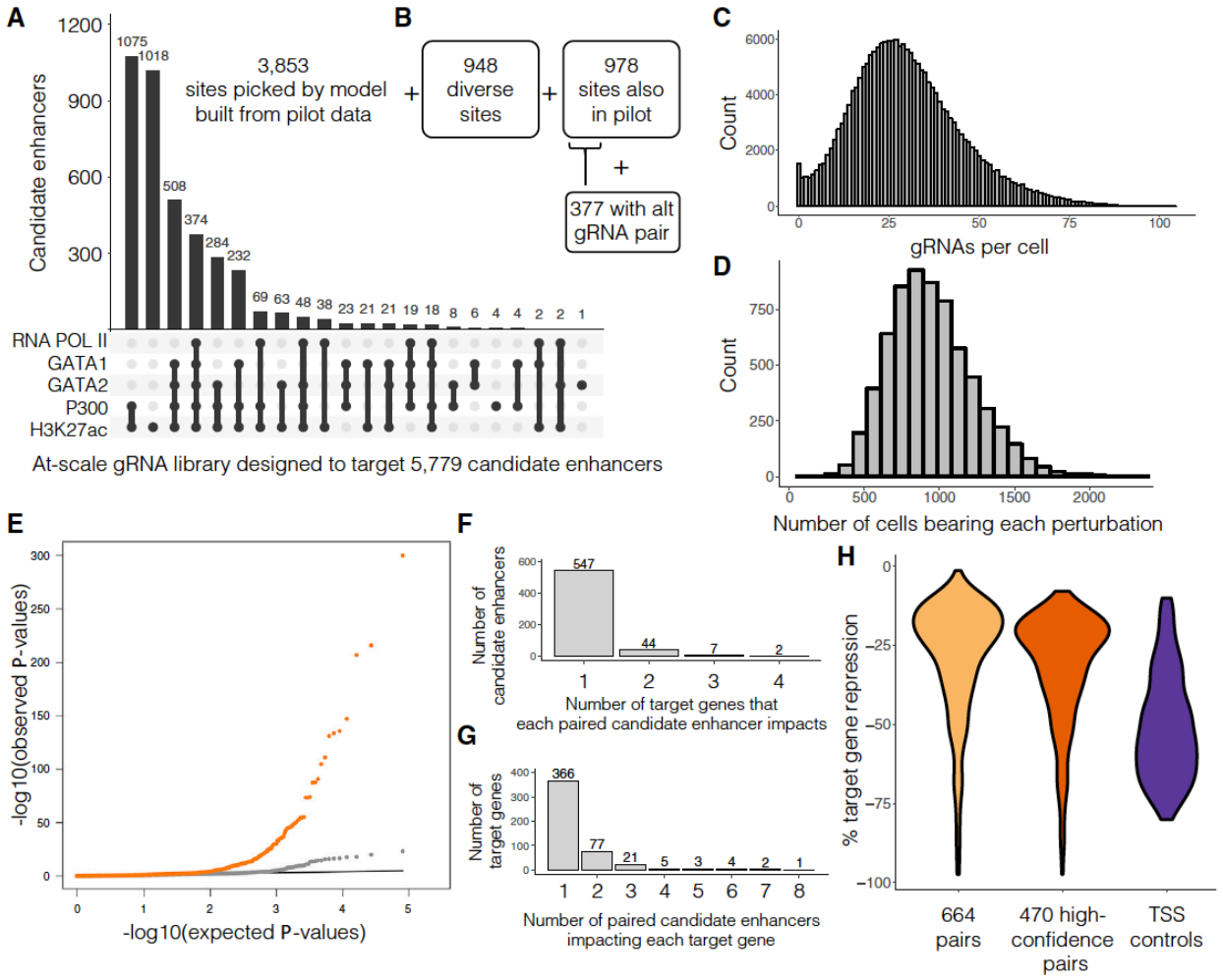


Figure 3.4. Multiplex enhancer-gene pair screening at scale in K562 cells.

(A) For the scaled experiment, gRNAs were designed to target a total of 5,779 candidate enhancers. Characteristics are shown for 3,853 sites chosen by a model informed by the hits identified in the pilot experiment. (B) 948 exploratory candidate enhancers were sampled from all DHSs. 978 candidate enhancers from the pilot were re-targeted with the same gRNA pair, and 377 of these were also targeted with a second, alternative gRNA pair. (C) gRNAs were again delivered to K562 cells, but at a higher MOI than the pilot experiment (median 28 +/- 15.3 gRNAs identified per cell). (D) A total of 207,324 single cell transcriptional profiles were generated. Each perturbation was identified in a median of 915 +/- 280 single cells. (E) Q-Q plot of the differential expression

tests. Distributions of observed vs. expected *P*-values for candidate enhancer-targeting gRNAs that were correlated with decrease in target gene expression (orange) and NTC gRNAs (gray; downsampled) are shown. (F) Histogram of the number of target genes impacted by each candidate enhancer identified as part of a pair (10% empirical FDR). (G) Histogram of the number of paired candidate enhancers detected as regulating each target gene (10% empirical FDR). (H) Effect sizes for the 664 enhancer-gene pairs that pass a <0.1 empirical FDR, the 470 high-confidence enhancer-gene pairs, and the 97% of TSS controls that are detected as repressing their target genes.

K562 cells were transduced at an even higher MOI than in the proof-of-concept experiment. We profiled the transcriptomes of 207,324 single cells and identified a median of 28 +/- 15.3 gRNAs per cell (Fig. 3.4C). Each candidate enhancer was targeted in a median of 915 +/- 280 single cells (Fig. 3.4D). Testing for associations as previously, a quantile-quantile plot again showed an inflation of significant associations involving the targeting of candidate enhancers (Fig. 3.4E). Using the NTCs to set a more inclusive empirical FDR of 10%, 97% (369 of 381) of TSS-targeting positive controls repressed their associated genes, as did all β-globin LCR controls. At this same threshold, of the 5,779 candidate enhancers, we identified 600 as repressing 1+ gene(s) within 1 Mb. These included 397/3,853 model-selected candidate enhancers (10%), 35/948 systematically sampled exploratory DHS (4%), and 168/978 previously targeted candidate enhancers (17%). As targeting of 53/600 candidate enhancers downregulated more than one gene (Fig. 3.4F), we collectively identified a total of 664 enhancer-gene pairs. As 113 genes were downregulated by targeting of more than one candidate enhancer, these pairs involved 479 target genes (Fig. 3.4G). These ranged in effect size from -1.4% to -97.5% target gene repression (Fig. 3.4H).

To evaluate reproducibility, we compared our results for the 978 candidate enhancers targeted in both experiments. Applying the same empirical FDR threshold of 10% to each dataset, 187/978 were identified as paired candidate enhancers in the pilot experiment, and 168/978 as paired candidate enhancers in the scaled experiment. Of these, 105 were identified in both experiments (hypergeometric test of overrepresentation *P*-value 7e-45; 3.3-fold enriched over expectation). The pairs identified in both experiments had stronger effect sizes (median 25% vs. 13% repression), better correlated effect sizes (Spearman's rho for % repression: 0.82 vs. 0.16; Fig. 3.5A), and involved more highly expressed genes (median 0.90 vs. 0.63 UMIs per cell), than pairs identified in only one experiment.

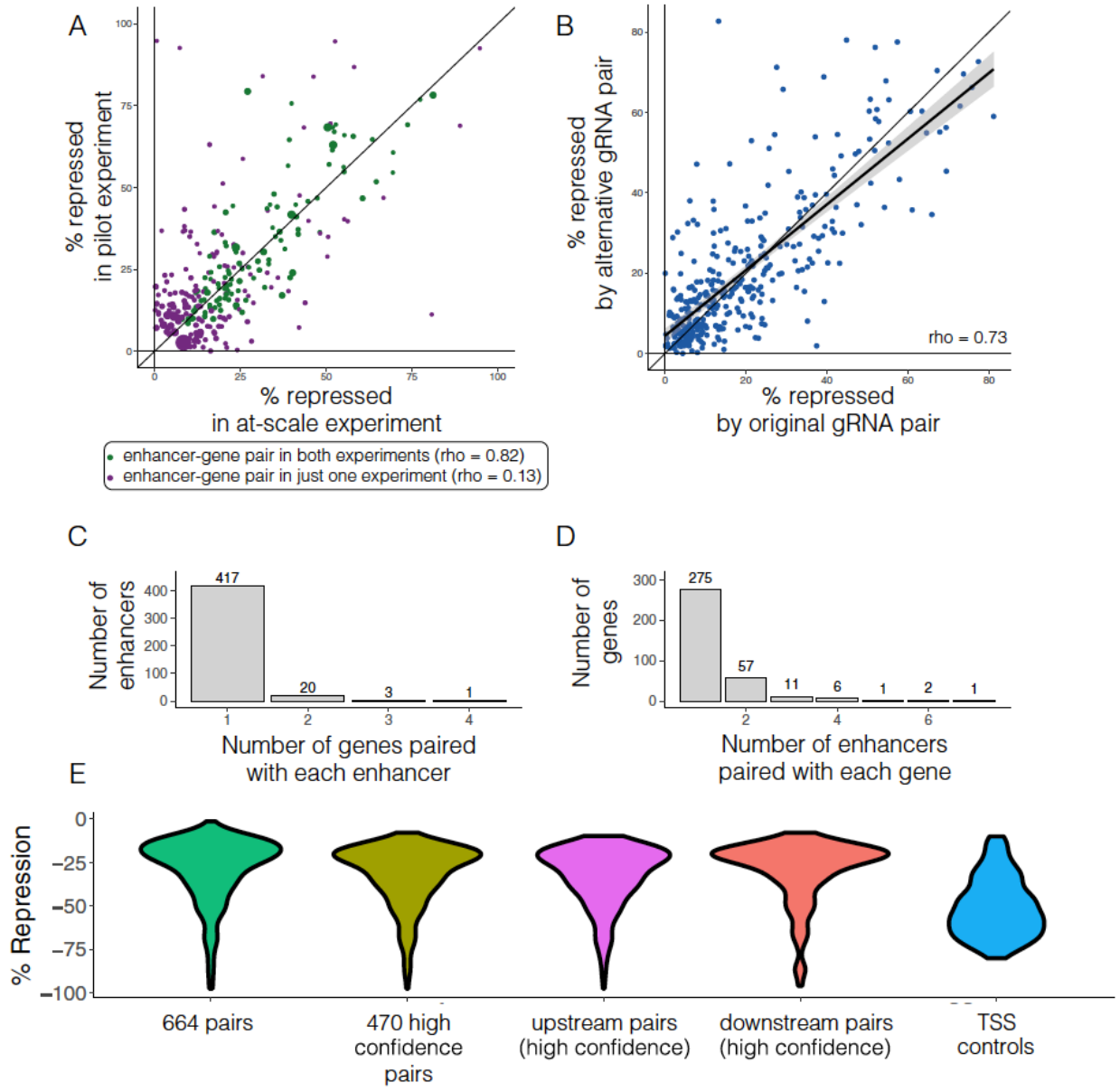


Figure 3.5. Replication of effect across experiments and alternative gRNA pairs.

(A) The percent target gene repression of an enhancer-gene pair in the pilot versus the scaled experiments (green: called as a pair in both experiments; purple: called as pair in only one experiment). (B) The effect sizes on the most highly repressed gene for each pair of gRNA pairs targeting the same candidate enhancer (see Methods). (C) Histogram per enhancer of the number of genes paired with that enhancer (high confidence pairs of the at-scale screen). (D) Histogram per gene of the number of enhancers paired with that gene (high confidence pairs of the at-scale

screen). (E) Effect sizes from enhancer-gene pairs identified in the at-scale screen. % repression of target transcript for the 664 enhancer-gene pairs that pass a <0.1 empirical FDR, the 470 high confidence enhancer-gene pairs, the high confidence pairs in which the enhancer is upstream of the target gene, high confidence pairs in which the enhancer is downstream, and the 97% of 381 TSS controls that are detected as repressing their target genes in the at-scale screen.

As noted above, an additional pair of gRNAs for 377/978 re-targeted candidate enhancers were included in this experiment, to facilitate evaluation of the extent to which poorly efficacious gRNAs might contribute to false negatives. In the scaled experiment at a 10% empirical FDR, 109/377 of the original gRNA pairs and 119/377 of the new gRNA pairs mediated enhancer-gene pairs. Of these, 84 were directed at the same candidate enhancers, a highly significant overlap (hypergeometric test of overrepresentation *P*-value 4e-33; 2.4-fold enriched over expectation). Furthermore, the effect sizes on the most highly repressed genes for gRNA pairs targeting the same candidate enhancer were well-correlated (Spearman's rho for % repression: 0.73; Fig. 3.5B). Overall, this analysis suggests that targeting candidate enhancers with more than two gRNAs could modestly increase our sensitivity.

Due to the noise from variability in expression levels, effect sizes, and gRNA quality, we defined a high-confidence subset of reproducible enhancer-gene pairs as those identified in both experiments at the 10% empirical FDR (112 pairs; 359/381 (94%) of targeted TSSs also met this criteria), as well as those internally reproducible between the 2 independently targeting gRNAs for candidate enhancers only tested in the scaled experiment (358 pairs; 337/381 (88%) of targeted TSSs also met this criteria). Putting these sets together, we annotated 470 enhancer-gene pairs as

high-confidence, involving 441 candidate enhancers (Fig. 3.5C) and impacting expression of 353 target genes. These ranged in effect size from -7.9% to -97.5% (Fig. 3.4H). We use this high-confidence subgroup for all summary analyses described below, unless otherwise noted. Of note, 24 candidate enhancers are paired with multiple target genes (Fig. 3.5D); it is possible that some of these pairings represent indirect effects, *e.g.* if a gene that is the primary target of the enhancer is involved in the regulation of the other gene.

3.6 Replication or Validation of 22 Selected Enhancer-Gene Pairs in Singleton Experiments

We next sought to individually replicate 15 enhancer-gene pairs with a range of effect sizes (-10% to -81%) and 6 "null" candidate enhancers not paired with any target gene. We transduced K562 cells separately with small pools of gRNAs targeting individual candidate enhancers, and investigated the impact on gene expression via bulk RNA-seq. For 12/15 replication experiments targeting candidate enhancers associated with downregulation of a target gene, the effect sizes were similar in magnitude and direction of effect (Fig. 3.6A-D; Fig 3.7-3.8). For all 9 experiments predicted to cause >30% repression, replication effects were also significant in a test of differential expression (*cis* adjusted *P*-value < 0.1). Of the 6 lines targeting a "null" candidate enhancer, none significantly decreased expression of a gene located within 1 Mb of the target (*cis* adjusted *P*-values > 0.1).

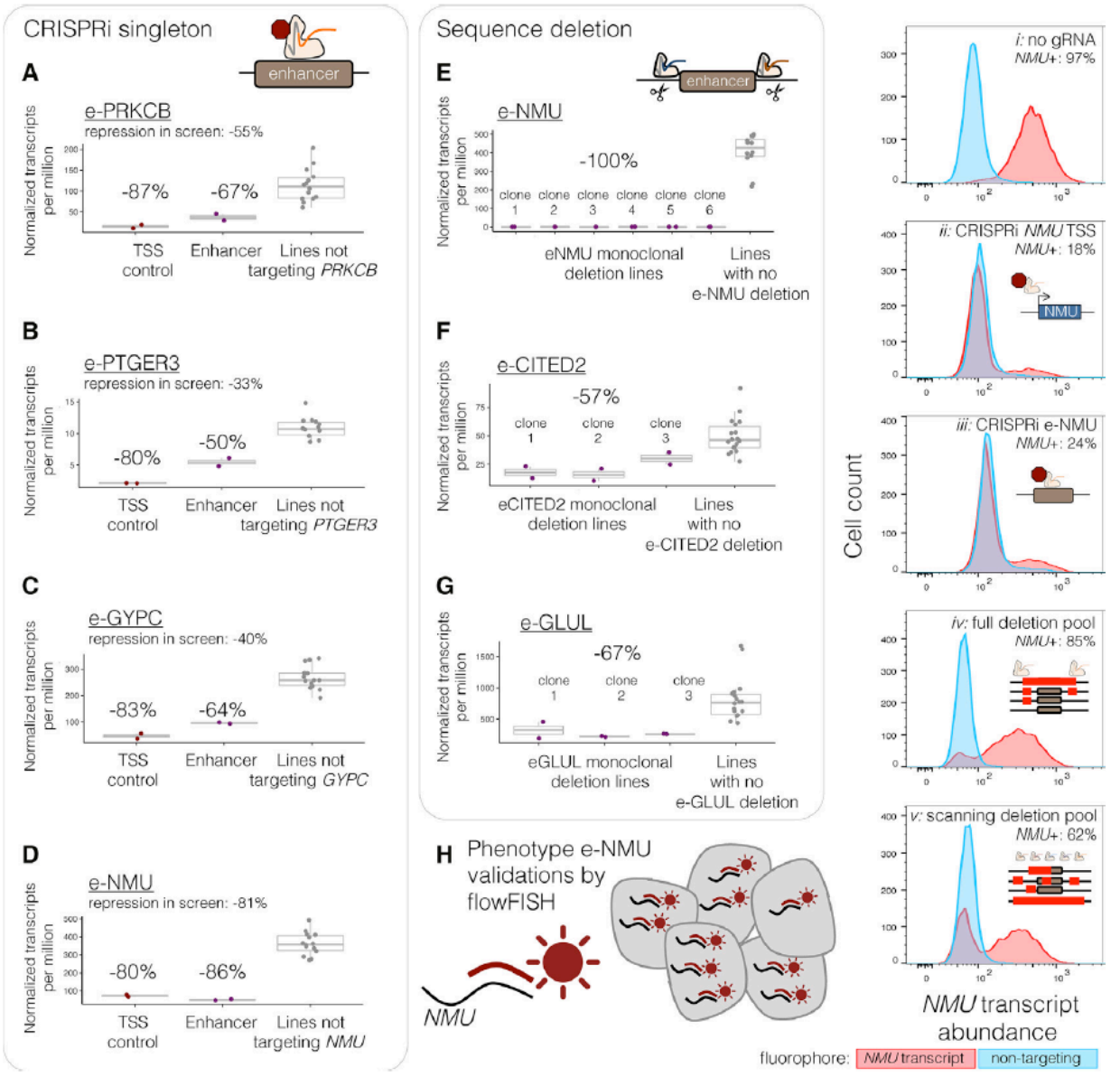


Figure 3.6. Replication and validation of selected enhancer-gene pairs in singleton experiments. For each singleton replication experiments of enhancer-gene pairs (A-D), bulk RNA-seq was performed on CRISPRi+ K562 cells transduced with gRNAs targeting e-NMU, e-PRKCB, e-GYPC, e-PTGER3 (purple) or the TSSs (dark red) of their respective target genes. Target gene expression in the singleton-target cell lines (red/purple) as compared to replication experiments in which the other 4 candidate enhancers or TSSs were targeted (gray). Eleven other singleton

CRISPRi experiments are summarized in Fig. 3.7. To validate three enhancer-gene pairs by sequence deletion (E-G), monoclonal lines were generated with full deletion of the locus's genomic sequence in three to six independent clones (e-NMU, e-CITED2, and e-GLUL), followed by bulk RNA-seq. (H) *NMU*-targeting cells were phenotyped by fluorophore-labelling of intracellular *NMU* transcripts by RNA flowFISH. *ii*-*iii*: singleton CRISPRi targeted cells as in Fig. 3.6A. *iv*-*v*: a heterogeneous pool of cells engineered such that a portion (based on deletion efficiency) harbor full or scanning deletions of e-NMU.

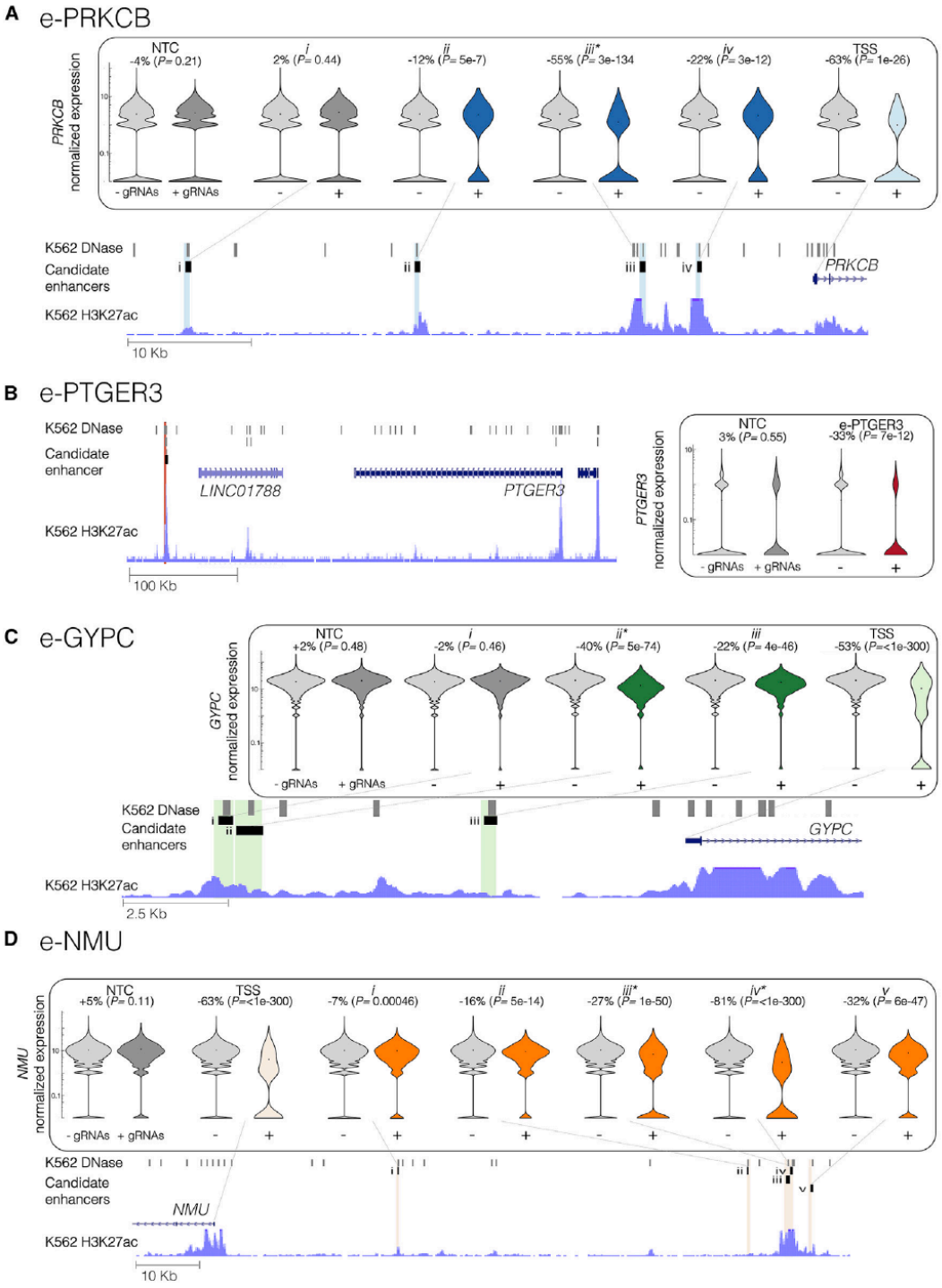


Figure 3.7. Highlighted examples of enhancer-gene pairs.

Asterisks denote the candidate enhancers that were targeted as part of a singleton replication experiments (Fig. 3.6). + and - denote the cells from the at-scale screen with or without gRNAs

targeting that locus. Percent changes and *P*-values denote the size and significance of a differential expression between these cell groups. (A) Three candidate enhancers (labeled *ii*-*iv*) that reside 32, 14, and 9 Kb upstream of *PRKCB* were paired with *PRKCB*, but a fourth (*i*) that lies 50 Kb upstream was not (shown: hg19 chr16:23791225-23851797). (B) A single candidate enhancer located 371 Kb downstream of *PTGER3* was paired with *PTGER3* (shown: chr1:71104684-71582921). (C) Two candidate enhancers paired with *GYPC* (*ii*-*iii*) lie in the 11 Kb region upstream of *GYPC*. However, a third candidate enhancer (*i*) immediately adjacent to *ii* was not paired with *GYPC* (shown: chr1:71104684-71582921). (D) Targeting five candidate enhancers (*i*-*v*) located 30.5, 87, 93.4, 94.1 and 97.6 Kb upstream of *NMU*, significantly reduced expression of *NMU* (shown: chr1:71104684-71582921). Target genes' normalized expression presented on log scale.

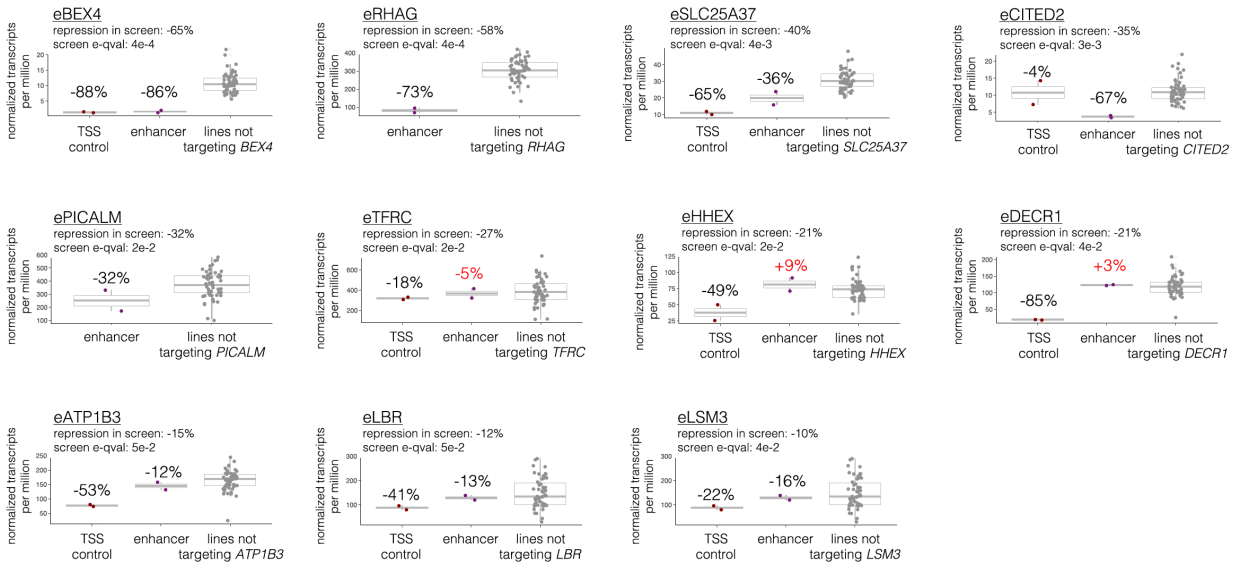


Figure 3.8. Eleven further singleton CRISPRi experiments.

For each singleton replication experiments of enhancer-gene pairs, bulk RNA-seq was performed on CRISPRi-positive K562 cells transduced with CRISPRi-optimized CROP-seq gRNAs targeting the labeled paired-enhancer (purple, denoted with an "e" prefix) or the TSSs (dark red) of their

respective target genes. Target gene transcript expression in the singleton-target cell lines (dark red/purple) as compared to 'non-targeting' lines (gray; singleton experiments in which the other 10 candidate enhancers or TSSs were targeted plus a line transduced with non-targeting gRNAs). Repression in screen = differential expression from at-scale screen. Screen e-qval = Benjamini-Hochberg corrected empirical *P*-value from at-scale screen. Normalized transcripts per million (tpm) from sleuth. %s above boxplots = sample's % repression in bulkRNA-seq calculated from (transcript's mean tpm between the sample's two technical replicates) / (transcript's mean tpm from all the 'non-targeting' lines). % repression labeled light red if in disagreement with the enhancer-gene pair in the at-scale screen.

Although the field often refers to singleton independent re-testing via CRISPRi as 'validation', it is a recapitulation of the modality of perturbation of the screen, and perhaps better classified as another form of replication. Therefore, we also performed a more stringent validation by generating 3+ monoclonal homozygous deletion lines for each of 3 enhancers (effect size in scRNA-seq screen: e-NMU = -81%, e-CITED2 = -35%, e-GLUL = -21%; Fig. 3.9). All three selected enhancers are quite distal from the gene whose expression they regulate (>50 Kb). These homozygously deleted lines all had the expected and magnitude of direction of effect (Fig. 3.6E-G), indeed with stronger effect sizes than seen by CRISPRi perturbation in the scRNA-seq screen (effect size with deletion: e-NMU = -100%, e-CITED2 = -57%, e-GLUL = -67%).

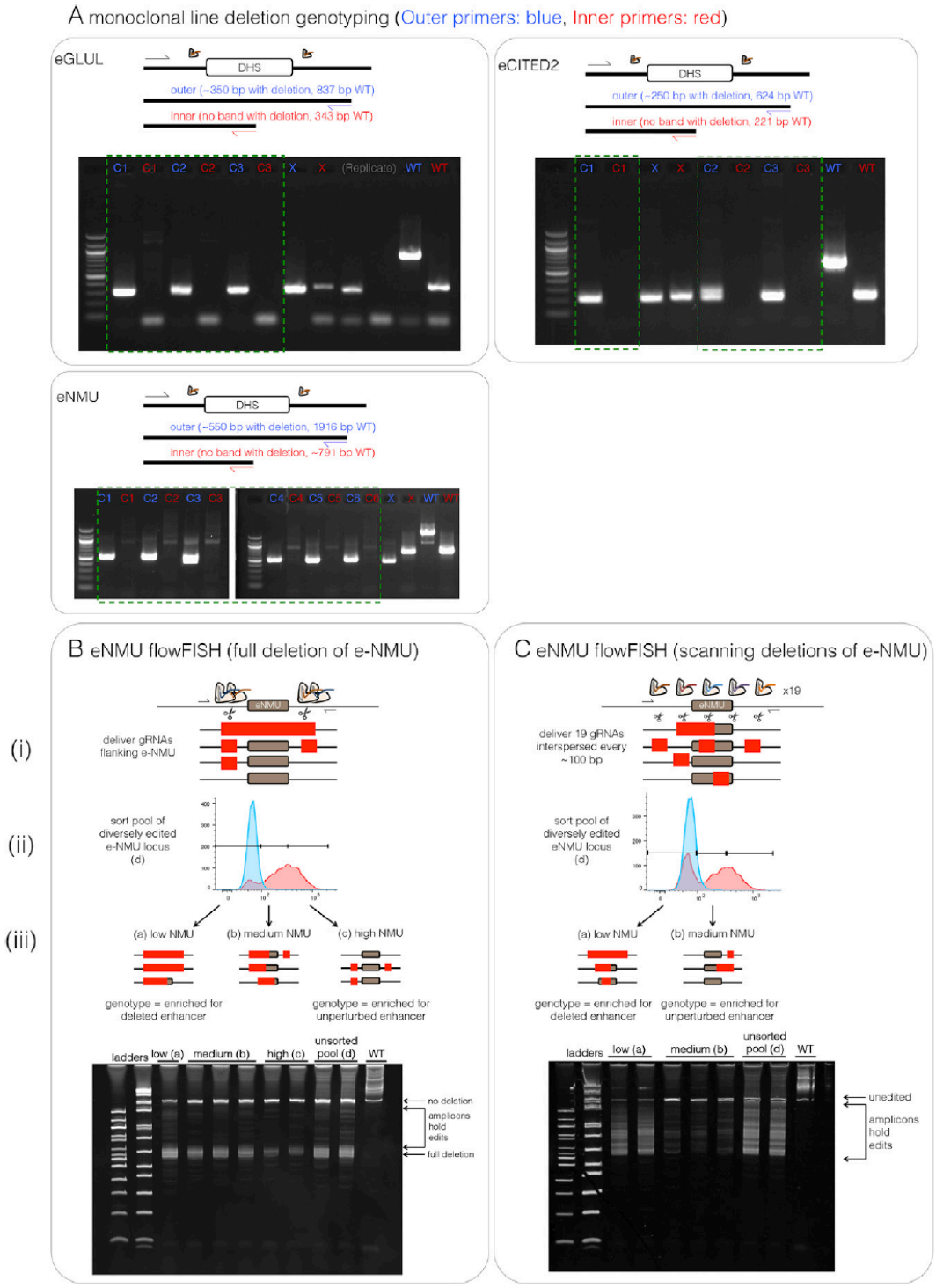


Figure 3.9 Details of sequence deletion validation, Related to Fig. 3.6E-H.

(A) Genotyping PCR design and gels for the homozygous sequence deletion monoclonal lines, as featured in Fig. 3.6E-G. Outer primers were designed to amplify the entire candidate enhancer

locus; shorter band in these 'outer' lanes (blue label, as compared to 'WT' lane) represents presence of a full deletion. Inner primers were designed to amplify only if a wildtype allele remained ("red" labeled lanes); presence of a band indicates a remaining wildtype locus. Primers design is schematized at the top. Clones with a deletion band (in the "outer" PCR lane) *and* no wildtype band (in the "inner" PCR lane) were submitted to bulkRNA-seq. Green dashed outline represents the clones used in Fig. 3.6. Nomenclature of "C1" and "C2" etc correspond to "clone 1" and "clone 2" *et cetera* as labeled in Fig. 3.6. 'WT' lanes = same parental K562 cell line that was transfected with gRNA targeting *HPRT1*. Ladder = NEB 100 bp (N3231L). "X" = cell line did not harbor homozygous deletions. (B-C) e-NMU sequence-disrupted cells were phenotyped by *NMU* RNA flowFISH, as featured in Fig. 3.6H. First, K562 cells were transfected with nuclease-active Cas9 and gRNAs either flanking (B) or scanning (C) the e-NMU locus to create a heterogeneous population of cells (*i*) in which a portion (based on editing efficiency) harbor full or partial deletion of e-NMU. Then, intracellular *NMU* expression was labeled via flowFISH (*ii*) and cells were sorted into bins of low (a), medium (b), or high (c) *NMU* expression (as to sort genotypes based on the effect upon disruption of e-NMU function, *iii*). Last, gDNA was extracted from the cells in each bin, and the e-NMU locus was amplified (primers diagrammed at the top of the figure). Unsorted pool (d) = unsorted but edited cells to demonstrate original distribution of genotypes in the original heterogenous pool. Each lane is a replicate PCR of gDNA (10 ng per reaction) from that same sorted sample. Ladder L = 100 bp (NEB), Ladder R = 1 Kb ext (Invitrogen). WT = untreated parental K562s. Remaining full-length alleles in the 'low' expression bins could correspond to inaccuracy of flowFISH, alleles with very small edits, or (as K562s are pseudotriploid) heterozygous cells that still retained a largely uninterrupted copy of e-NMU on one or two alleles.

In our validations of the *NMU* candidate enhancer ("e-NMU"), we also applied RNA flowFISH (Choi et al. 2018), and again observed decreased *NMU* expression in singleton CRISPRi populations targeting *NMU's* TSS (-79% less *NMU* than untreated cells) and e-NMU (-73% less *NMU*, Fig. 3.6H *ii-iii*). We also used flowFISH to phenotype a heterogeneous pool of cells that harbored a mix of full, partial, or no deletions of e-NMU, generated by transient transfection of flanking pairs of gRNAs. 12% of the cells showed reduced *NMU* expression in comparison to untreated cells (Fig. 3.6H *iv*), which is in-line with expected full deletion efficiency (Gasperini et al. 2017). Cells were sorted into bins of low, medium, or high NMU expression. PCR of the e-NMU locus revealed enrichment of the full deletion in the low and medium *NMU* bins, whereas full deletion was rarer in the high *NMU* bin (Fig. 3.9B). To further dissect e-NMU, we additionally transfected with 19 gRNAs interspersed every ~100 bp across e-NMU to generate deletions of diverse lengths and locations, inducing reduction of *NMU* expression in 35% of cells compared to untreated (Fig. 3.6H *v*). PCR of e-NMU again showed a similar enrichment of longer deletions in the cells with lower *NMU* expression (Fig. 3.9C).

In summary, of the high confidence pairs that we re-tested by singleton CRISPRi and/or singleton CRISPR-mediated deletion, 13/16 matched with respect to both their direction and magnitude of effect size, whereas 3/16 failed to validate. This false positive rate is consistent with the 10% FDR that we used to assign a threshold for calling pairs (*P*-value on whether 3/16 disagrees with 10% FDR = 0.21).

3.7 Selected Examples of Enhancer-Gene Pairs

We highlight four of the enhancer-gene loci in Fig. 3.7. An 'e-' prefix is used to denote candidate enhancers that we targeted in singleton replication experiments. In the scaled experiment, we

targeted four candidate enhancers across the region upstream of *PRKCB*. The furthest of these (*i* in Fig. 3.7A, 50 Kb upstream) did not have an effect, but candidate enhancers 32, 14, and 9 Kb upstream of the TSS were associated with repression of *PRKCB* (*ii*-*iv* in Fig. 3.7A). The strongest of these, located 14 Kb upstream, was also targeted and replicated in both the pilot and singleton experiments (*iii*, "e-PRKCB").

In the pilot, scaled, and singleton replication experiments, we targeted only one candidate enhancer within 1 Mb of *PTGER3* ("e-PTGER3", Fig. 3.7B), located 371 Kb downstream of the *PTGER3* TSS. In each of the three experiments, targeting of e-PTGER3 consistently repressed expression of *PTGER3.*

We targeted three candidate enhancers in the region upstream of *GYPC*, a human erythrocyte membrane protein. Targeting of candidate enhancers 4.5 Kb upstream (*iii*, Fig. 3.7C) and 10 Kb (*ii*) upstream of *GYPC*'s TSS resulted in its repression in the scaled experiment. Interestingly, a candidate enhancer so close to e-GYPC as to likely be unresolvable from it by CRISPRi (*i*, Fig. 3.7C) did *not* result in repression of *GYPC* in the scaled experiment, potentially attributable to poor gRNA quality or another source of false negatives.

Targeting of multiple candidate enhancers decreased expression of the same gene, *NMU*, which encodes neuromedin U, a neuropeptide that plays roles in inflammation as well as erythropoiesis (Gambone et al. 2011). One candidate enhancer was associated with light repression of *NMU* (*i* in Fig. 3.7D, located 30.5 Kb upstream of the *NMU* TSS). An additional four candidate enhancers were located in close proximity to one another, but nearly 100 Kb upstream of the *NMU* TSS (*ii*-

v in Fig. 3.7D, located 87, 93.4, 94.1 and 97.6 Kb upstream). Because of their proximity, these closely located candidate enhancers internally replicate e-NMU within the scaled experiment, in contrast to the neighboring candidate enhancers of e-GYPC.

3.8 Insights into the Properties of Human Enhancers and their Target Genes

3.8.1 *Distance Between Paired Enhancers and Promoters*

We find that of the class of enhancers surveyed here (non-intronic, unbuffered by other enhancers), paired enhancers are separated from the TSS of their target genes by a median distance of 24.1 Kb (Fig. 3.10A, top row; note that this analysis is restricted only to high-confidence pairs that fall upstream of their target genes (n = 354), to avoid bias from the length of the gene body consequent to the fact that we avoided targeting intronic candidate enhancers for which CRISPRi might directly inhibit transcription. Upstream and downstream enhancers do not exhibit large differences in their effect size distributions (Fig. 3.2E)). Given that we tested for associations against all genes within 1 Mb of each candidate enhancer (Fig. 3.10A, fourth row; median distance of 440.2 Kb, similarly restricted to upstream tests), this supports a very strong role for proximity in governing enhancer-promoter choice. Nonetheless, 153/470 (33%) of enhancer-gene pairs involved skipping of at least one closely located TSS of another K562-expressed gene (Fig. 3.10B). Interestingly, low-confidence enhancer-gene pairs (*i.e.* the subset of the 600 that were not high-confidence and also fall upstream; n = 127) were also enriched for proximity to their target genes, suggesting that a substantial proportion of these are bonafide enhancers (Fig. 3.10A, second row; median distance of 45.0 Kb).

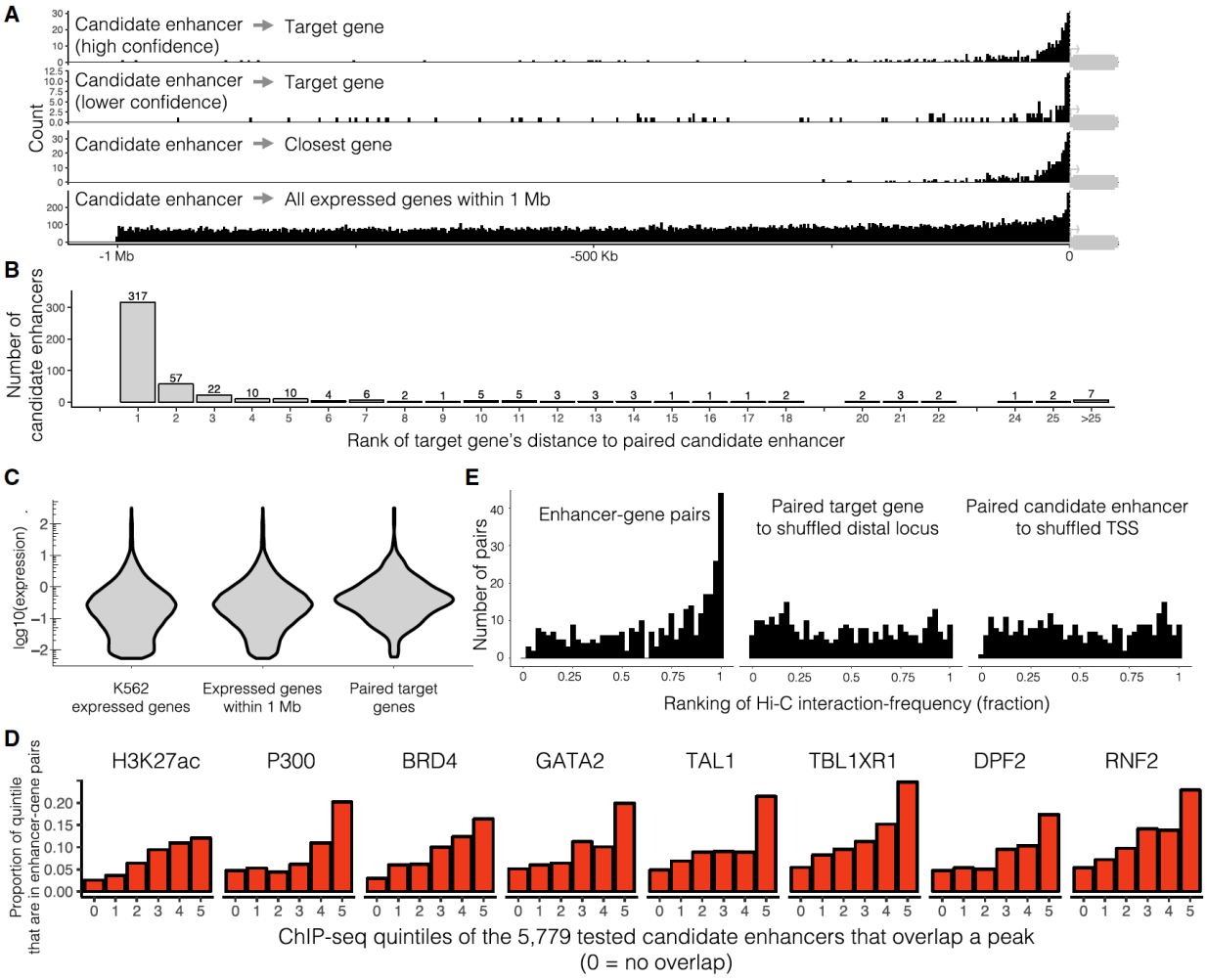


Figure 3.10 Characteristics of K562 enhancer-gene pairs.

(A) Paired candidate enhancers fall close to target genes. Distribution of distances between the paired candidate enhancers and: their target gene's TSS (top row, high confidence pairs, second row, lower confidence pairs), the TSS of whatever K562-expressed gene is closest (third row), or the TSS of every K562-expressed gene within 1 Mb (fourth row). Plotted with respect to gene orientation. Of the 470 high confidence pairs, this plot displays only the 354 that fall upstream of the target genes (as the gRNA library does not include candidate enhancers within 1 Kb of any gene body, downstream enhancers are biased to fall further from the target TSS). A TSS-focused zoom of this plot is included as Fig. 3.11E. (B) 317 of 470 high-confidence pairs target the most

proximal K562-expressed gene. Target genes are ranked by their absolute distance to the paired candidate enhancer (1 = closest, 2 = second closest, etc.). (C) This framework captures regulatory effects on genes from a broad range of expression levels (expression = mean transcript UMIs/cell in the entire 207,324 cell dataset, for 13,135 K562-expressed genes, 10,560 of these within 1 Mb of a targeted candidate enhancer in the scaled experiment, and 470 high-confidence enhancer-gene pairs). See also Fig. 3.11D. (D) Paired candidate enhancers tend to fall in enhancer-associated ChIP-seq peaks that show stronger signals. All ChIP-seq peaks that overlap the scaled experiment's 5,779 candidate enhancers were divided into quintiles defined as the average enrichment in ChIP-seq peak region (0 = no such peak overlaps the candidate enhancer, 1 = lowest, 5 = highest). Histograms of the proportion of which candidate enhancers in each quintile that were paired with a target gene are shown for the eight most-enriched ChIP-seq datasets. (E) Enhancer-gene pairs interact more frequently in K562 Hi-C data (*left*, fractional ranking of enhancer-gene pairs' Hi-C interaction-frequency against all other possible interactions at similar distances within the same TAD, K-S test against a uniform distribution *P*-value <2e-16), as compared to two control distributions: paired target gene TSSs paired with a shuffled genomic locus (*middle*, K-S test vs. actual enhancer-gene pairs distribution = *P*-value 2e-7) or paired candidate enhancers paired with a shuffled genomic locus (*right*, K-S test vs. actual enhancer-gene pairs distribution = *P*-value 1e-9). See also Fig. 3.11B-C.

Of our 359 'positive control' TSSs whose targeting successfully repressed the expected gene in both experiments, 35 reduced expression of 1+ additional genes (45 apparent promoter-promoter relationships in total). 15 of these 45 involved overlapping promoters (TSSs within 1 Kb), such that the observed effect of CRISPRi is likely direct. As for the remaining 30, one possibility is that

these represent examples of promoters acting as enhancers, as recently reported (Diao et al. 2017; Charles P. Fulco et al. 2016). Additionally, as repressive epigenetic effects may spread a few Kb from the target site, it is possible that CRISPRi of promoters may be silencing proximal enhancers as well. However, these 30 are largely *not* enriched for proximity to affected genes (Fig. 3.11A; median distance of 405.3 Kb, similarly restricted to upstream tests), in contrast with enhancer-gene pairs (median distance = 24.1 Kb). We therefore hypothesize that these are more likely consequent to *trans* effects of repressing the primary target of these TSS-targeting gRNAs. In other words, rather than these gRNA-targeted promoters acting as noncoding regulatory elements of other genes, the reduction in protein levels of the targeted gene may secondarily affect the expression of other genes.

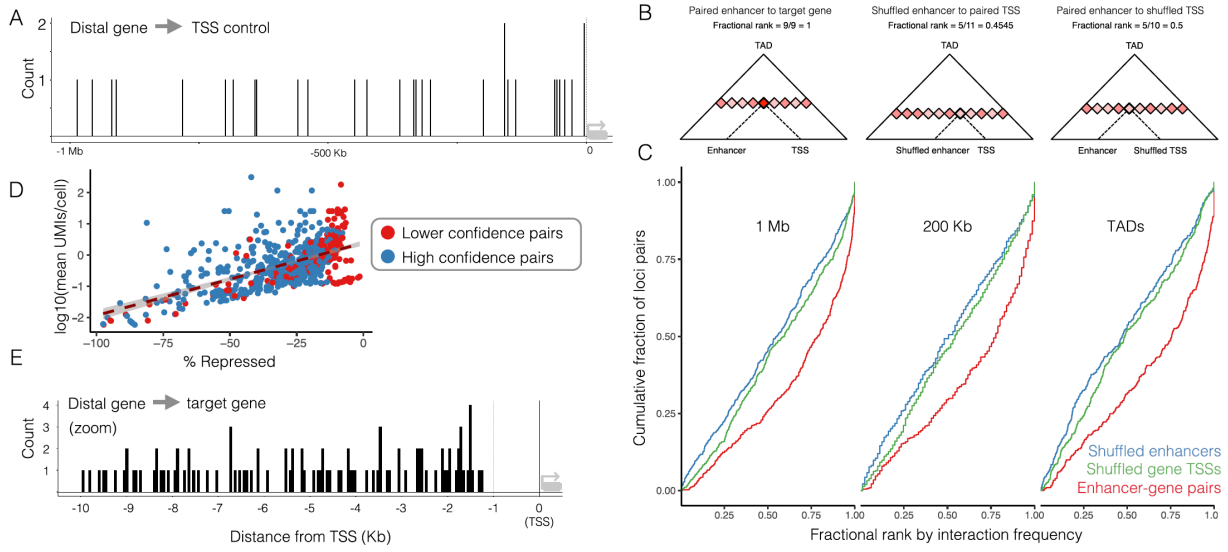


Figure 3.11 Details on characteristics of K562 enhancer-gene pairs.

(A) Distribution of distances between "positive control" TSSs and any secondarily repressed genes. Of our 359 'positive control' TSSs whose targeting successfully repressed the expected gene in both experiments, 35 reduced expression of 1+ additional genes (45 apparent promoter-promoter relationships in total). 15 of these 45 involved overlapping promoters (TSSs within 1 Kb) and are not shown here as the observed effect of CRISPRi is likely direct. The distances that

the remaining 30 secondarily repressed genes fall upstream of the targeted TSS are shown. In contrast with enhancer-gene pairs (Fig. 3.10A), these 30 are largely *not* enriched for proximity to affected genes. Dashed line = target gene TSS. (B-C) Hi-C interaction frequency analysis, (B) Example schematic of fractional ranking by interaction frequency analysis. The interaction frequency of each loci pair (color of pixel) is ranked within the interaction frequencies of all distance-matched genomic-pairs in the same TAD (the stripe of pixels shown in schematic). For the two null distributions in Fig. 3.10E, each pair's target gene's TSS is given a shuffled enhancer (and then ranked again within this new distance distribution), or the pair's candidate enhancer is given a shuffled TSS (and then ranked again within this new distance distribution). Shuffled TSSs and enhancers are drawn from the same distance distribution as the actual enhancer-gene pairs. (C) The same fractional rank by interaction frequency analysis within the same TAD as shown in Fig. 3.10E, but also comparing ranking to all pairs within 1 Mb or 200 Kb of the chosen enhancer-TSS pair. Red = enhancer-gene pairs, blue = hit-gene to shuffled enhancer pair null distribution, green = hit enhancer to shuffled gene TSS pair null distribution. (D) Correlation of effect size of enhancer-gene pair versus expression level of target gene. Effect size (% transcript repressed) was correlated with expression level of targeted gene (Spearman's rho for 664 inclusive pairs: 0.56; Spearman's rho for 470 high confidence pairs: 0.53). This is likely consequent to power, as small effects (less than -25%) are not detected on lowly expressed genes (less than 0.12 UMIs/cell). log10 of the mean UMIs/cell is denoted per target gene transcript. (E) A "zoom-in" of Fig. 3.10A to the 10 Kb upstream of the target gene's TSS (rather than 1 Mb). 101 of 354 upstream, high confidence enhancer-gene pairs fall within 10 Kb of the TSS. Same restrictions to enhancer-gene pairs plotted here as in Fig. 3.10A. Gray line = TSS, red line = 1 Kb upstream of TSS (all protospacers within 1 Kb of a TSS were excluded from any candidate enhancer gRNA library).

3.8.2 *Characteristics of Target Genes*

The 353 genes included in 1+ 470 high-confidence enhancer-gene pairs had several notable characteristics. First, their expression levels are distributed similarly to the full set of 10,560 genes against which we tested (Fig. 3.10C), suggesting we are reasonably well-powered to detect regulatory effects on even modestly expressed genes. Second, housekeeping genes were underrepresented, relative to all tested genes (hypergeometric test *P*-value = 3e-5 and 2.1-fold depleted using the housekeeping gene list of (Eisenberg and Levanon 2013); hypergeometric test *P*-value = 2e-6 and 3.9-fold depleted using the housekeeping gene list of (Lin et al. 2017)). Similar depletions of housekeeping genes are observed when we instead compare paired target genes to the K562-expressed genes most proximal to tested candidate enhancers. Although these analyses support the view that a prevailing characteristic of housekeeping genes may be a dearth of distal regulatory elements (Gasperini et al. 2017; Ganapathi et al. 2005), we cannot fully rule out that the possibility that this result is influenced by our choice of candidate enhancers to target. Finally, paired target genes were enriched for genes with roles in leukocyte migration and differentiation, consistent with distal enhancers shaping the expression of K562-specific genes.

3.8.3 *Characteristics of Paired Enhancers*

We also examined the characteristics of the candidate enhancers for which targeting significantly impacted expression of 1+ genes in *cis.* First, as compared with the full set of 5,779 candidate enhancers targeted in either or both experiments, we tested if the 441 high-confidence candidate enhancers were enriched for strong peaks in 169 K562 ChIP-seq datasets (ENCODE Project Consortium 2012). We identified 87 that were significantly enriched (threshold of an adjusted *P*-value < 0.005), but the eight most significant were co-activators (p300 logistic regression *P*-value = 1e-46, candidate enhancers in the top quintile were 1.8-fold more likely to be paired than those

in the bottom quintile; BRD4 *P*-value = 2e-33, 1.6-fold); an enhancer-associated histone modification H3K27ac (*P*-value = 8e-37, 1.6-fold); the MYC activator TBL1XR1 (*P*-value = 2e-34, 1.5-fold), and lineage-specific TFs (TAL1 *P*-value =2e-33, 1.6-fold; GATA2 *P*-value =1e-31, 1.5-fold; DPF2 *P*-value = 5e-31, 1.5-fold; RNF2 *P*-value = 2e-33, 1.5-fold; Fig. 3.10D). Other expected enhancer-associated marks also exhibited significant enrichment (CCNT2 *P*-value = 4e-21, 1.3-fold; H3K4me1 *P*-value = 1e-19, 1.8-fold; MYC *P*-value = 2e-12, 1.3-fold). However, many of these features are correlated, and BRD4, H3K4me1, TRIM24, p300, H3K27ac, ETS1, and ZNF274 were the only significant predictors in a multivariate logistic regression (*P*-value < 0.01). Of note, high conservation as measured by median phyloP scores (Pollard et al. 2010) was not enriched in these candidate enhancers as compared to all tested candidate enhancers (independent logistic regression *P*-values > 0.5).

Second, we examined whether paired enhancers were more likely to intersect with K562 super-enhancers. Overall, 474 of the 5,779 candidate enhancers that we tested fell within 65 K562 super-enhancers (Cao et al., 2017); however, a much higher proportion of high-confidence paired enhancers belonged to this set (102/441). Several super-enhancers contained multiple targeted enhancers that were paired with the same gene. More specifically, 20 genes were linked with two candidate enhancers, and 6 genes were linked with three or four candidate enhancers, that were located within the same super-enhancer.

Third, we evaluated enrichment of transcription factor (TF) motifs in either our associated enhancers or the promoters of their target genes. Motifs for the known blood TFs KLF-1, -5, -6, -15, leukemogenesis-related SALL4, and the MYC-interacting ZN281 were enriched in the

promoters of the inclusive set of 479 paired-target genes, as compared to the promoters of all genes within 1 Mb of a tested candidate enhancer. Similarly, motifs for a largely distinct set of known blood TFs (TAL1, KLF-1, -3, -4, -5, -8, GATA-1,- 2, -3) and AP2C were enriched in the inclusive set of 600 paired enhancers, as compared to the overall set of 5,779 candidate enhancers tested.

3.8.4 *Pairs of Transcription Factors Act Together Across Enhancer-Gene Pairs*

To investigate whether there was any discernible logic underlying why particular enhancers were associated with particular promoters, we next sought to identify pairs of TFs that are "co-enriched" in the inclusive set of 664 enhancer-promoter pairs, *i.e.* they occur across pairs at a higher frequency than expected by chance given their background frequency in each category. We identified 6 TF pairs whose sequence motifs were co-enriched in this way, suggesting potential interactions. For example, presence of the NR2C2 motif (implicated in regulation of the globins (Tanabe et al. 2007)) in a paired promoter was associated with presence of a KLF1 or RXRA motif in the corresponding paired enhancer. On the other hand, presence of the GATA3 motif in a paired promoter was associated with the *absence* of a KLF1 motif in the corresponding paired enhancer.

We also explored such pairings via ChIP-seq data. Although ChIP-seq peaks often reflect indirect binding, such secondary partners might still play a role in the restriction of enhancer-promoter interactions. We identified 24 TF pairs that are "co-enriched" in enhancer-promoter pairs. Unfortunately, none of the TF pairs identified in either analysis had corresponding ChIP-seq datasets or high quality consensus motifs for *both* TFs involved in the pair, preventing cross-confirmation between the two modalities of analysis.

3.8.5 *Comparison of Enhancer-Gene Pairs to Hi-C Based Measurements of Physical Proximity*

We sought to evaluate whether our enhancer-gene pairs are enriched for physical proximity as measured by the global chromosome conformation mapping technique Hi-C. To control for the dominant effects of genomic distance and TADs in Hi-C datasets, we ranked the Hi-C contact frequencies in K562 cells (Rao et al. 2014) for the 71% of the enhancer-gene pairs that fell in the same TAD (333/470 high-confidence pairs) against all other possible interactions at similar distances within the same TAD (median 66 other genomic-loci pairs, range 6 to 260, Fig. 3.11B-C). Upon plotting the fractional ranks of high-confidence pairs, we found their contact frequencies to be strongly enriched at the highest ranks (K-S test against a uniform distribution *P*-value < 2e-16, Fig. 3.10E). To ensure that this enrichment was not an artifact of paired enhancers or genes interacting more frequently with all neighboring loci (as in FIREs (Schmitt et al. 2016)), we repeated this analysis twice, but shuffling the genomic loci paired to either the enhancers or genes (keeping these shuffled pair sets' overall distance distributions the same as the original enhancer-gene pair set's distance distribution). This did not result in the same enrichment as seen in the high confidence pair distribution (K-S test of high confidence enhancer-gene pair vs. enhancer-pair shuffling *P*-value 1e-9; high confidence enhancer-gene pair versus TSS-pair shuffling *P*-value 2e-7), consistent with more frequent looping specifically between the high confidence enhancer-gene pairs (Fig. 3.10E). Though enriched for proximity, we note that only a minority of our hits are called as proximate to their target genes based on this analysis; as such, many enhancer-gene pairs would not have been identified if we had limited tested candidate enhancers to those physically proximate to a promoter according to Hi-C or related data.

3.9 CRISPRi IS HIGHLY MULTIPLEXABLE WITHIN CELLS

To our knowledge, prior to this study, it was unknown whether extensively multiplexing gRNAs within a single cell would dilute the efficacy of CRISPRi. To evaluate this, we conducted a biological replicate of the pilot experiment, targeting the same 1,119 candidate enhancers but at a low MOI. From this experiment, we profiled the transcriptomes of 41,284 cells and identified a median of 1 +/- 1.6 gRNAs per cell (Fig. 3.12A). Each perturbation was only seen in a median of 43 +/- 16 cells, as compared with 516 +/- 177 cells in the high MOI pilot experiment (Fig. 3.12B). At a 10% empirical FDR, only 316 TSSs and 69 enhancer-gene pairs were identified in the low MOI experiment, as compared with 359 TSSs and 226 enhancer-gene pairs in the high MOI pilot experiment, validating the substantial increase in power resulting from multiplexed perturbation (Fig. 3.1B). As the same 381 TSS controls were targeted in the low MOI, pilot and scaled experiments, we compared the degree of repression conferred by CRISPRi at increasing MOI (median 1 vs. 15 vs. 28 gRNAs per cell), and found them to be well-correlated (Spearman's rho's ranging from 0.73 to 0.87; Fig. 3.12C). On average, the degree of repression conferred by targeting a TSS in both high MOI experiments was only ~6% less than by targeting it in the low MOI experiment (Fig. 3.12D). Similarly, for candidate enhancers paired in the scaled experiment (10% empirical FDR) that were also targeted in the low MOI and pilot experiments, effect sizes were well correlated (Spearman's rho's ranging from 0.54 to 0.70; Fig. 3.12C), and effect sizes ratios clustered around 1 (Fig. 3.12D). Overall, these results suggest that multiplexing gRNAs within individual cells, even to MOIs of ~28, does not dilute the efficacy of CRISPRi.

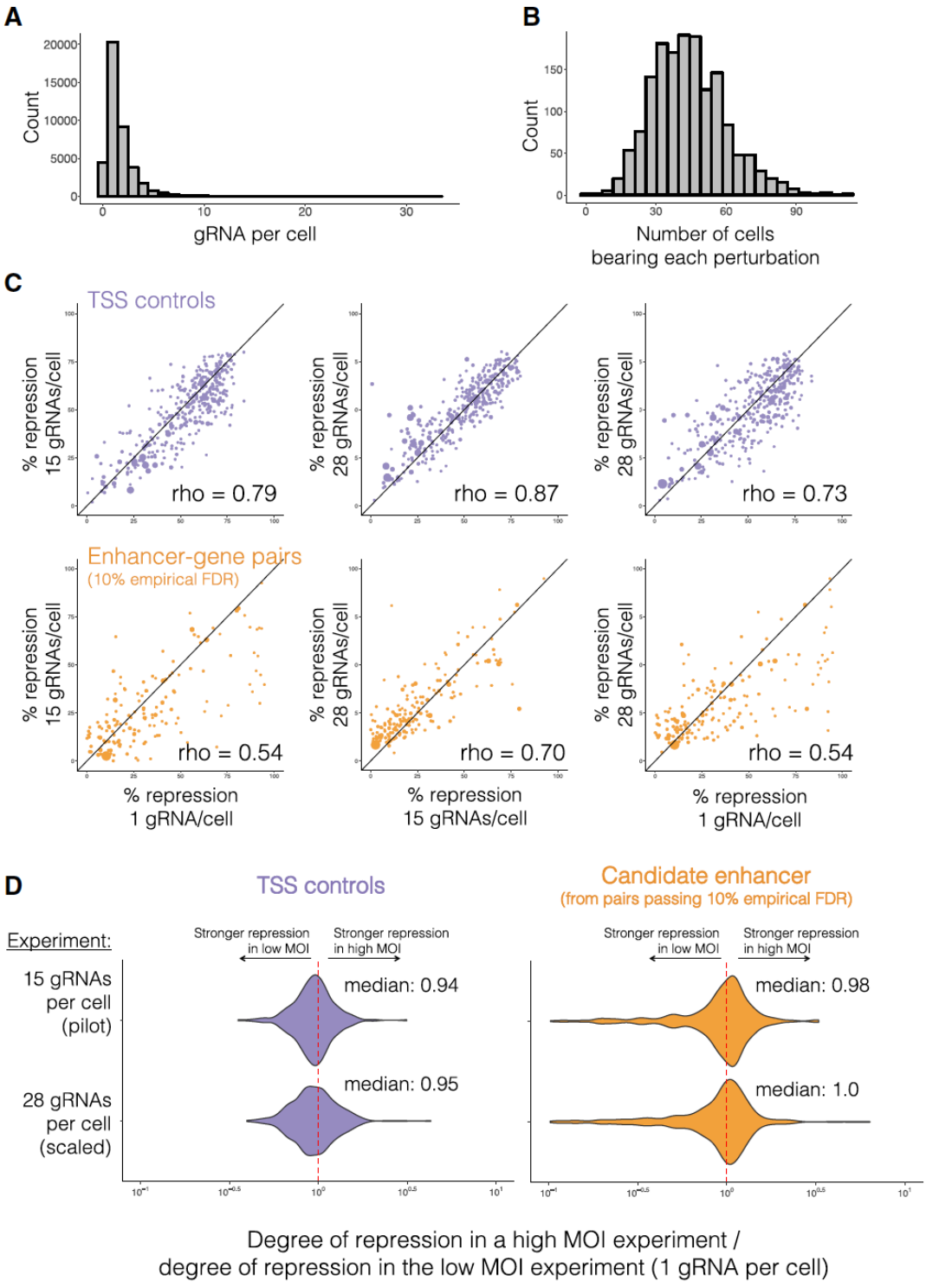


Figure 3.12 CRISPRi is robust to multiplexing within a cell.

(A) A biological replicate of the pilot study, targeting the same 1,119 candidate enhancers and 381 TSSs, was performed at a low MOI (median 1 +/- 1.6 gRNAs identified per cell). (B) A total of

41,284 single cell transcriptional profiles were generated. Each perturbation was identified in a median of 43 +/- 16 single cells. (C) Correlation of effect sizes for TSS controls (*top*, purple) or enhancer-gene pairs identified in the scaled experiment (10% empirical FDR, *bottom*, orange) across increasing rates of gRNA per cell (*left*, 1 vs 15; *middle*, 15 vs 28; *right*, 1 vs 28 gRNAs/cell). Point sizes are proportional to each target gene's expression level. (D) The ratios of repression for each TSS control or paired candidate enhancer (as identified with a 10% empirical FDR in any experiment) in the low MOI experiment vs. a high MOI experiment (top = median 1 gRNA vs. 15 gRNAs; bottom = median 1 gRNA vs. 28 gRNAs). The candidate enhancer outliers with stronger effect sizes in the low MOI experiment (right panel, ratios in left tail) are largely due to stochastic undersampling of lowly expressed target genes in the low MOI experiment (see also Fig. 3.13).

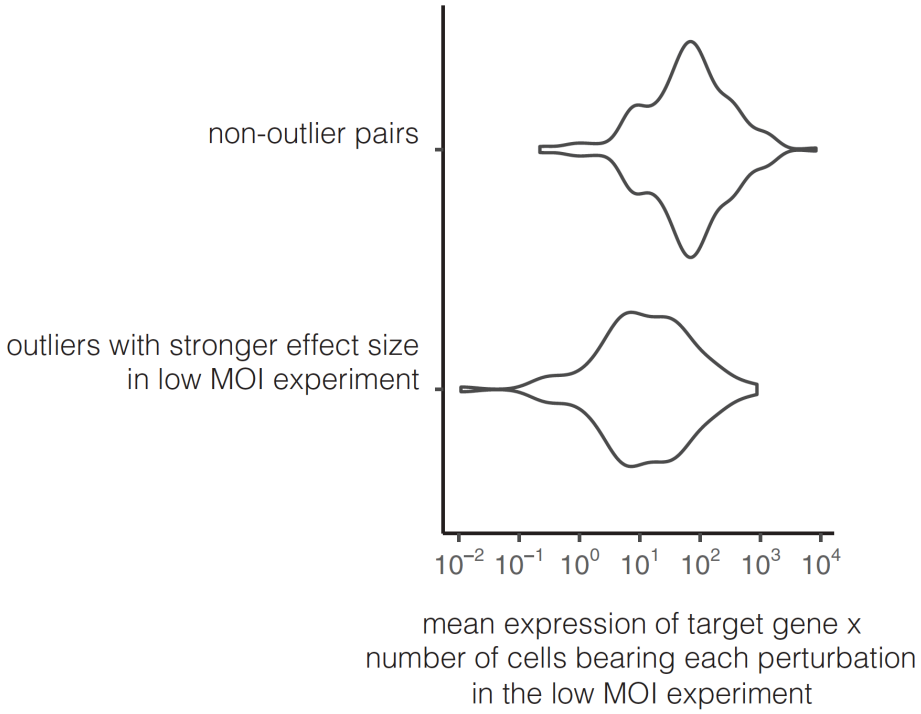


Figure 3.13 Outliers with greater effect size in low MOI replicate are likely due to low expression and low cell count in low MOI replicate.

The mean expression of the target gene in the low MOI 41,284 cell dataset as a function of the number of cells bearing each perturbation in that experiment.

3.10 DISCUSSION

Understanding the regulatory landscape of the human genome requires the validation and identification of target genes for the vast numbers of candidate enhancers that have been nominated by biochemical marks or that reside on haplotypes implicated by GWAS or eQTL studies. Our multiplexed enhancer-gene pair screening method has the potential to help address this challenge. In the scaled experiment, we evaluated 78,776 potential *cis* regulatory relationships involving 5,779 candidate enhancers and 10,560 expressed genes. In contrast, nine recently published CRISPR screens of noncoding sequences cumulatively studied regulatory effects on a total of 17 genes (Korkmaz et al. 2016; Rajagopal et al. 2016; Canver et al. 2015; Sanjana et al. 2016; Charles P. Fulco et al. 2016; Klann et al. 2017; Diao et al. 2016, 2017; Gasperini et al. 2017). By delivering a median of 28 perturbations to each of 207,324 cells, this experiment was powered equivalently to a 'one gRNA per cell' experiment profiling 5.8 million single cell transcriptomes. Of note, one recent study used scRNA-seq as a readout for the effects of CRISPR-based perturbations of 71 candidate regulatory elements on ~100 genes in seven genomic regions (Xie et al. 2017). However, its power and scope was limited by a low MOI (Fig. 3.1B) and a gRNA barcoding strategy that suffers from a ~50% rate of template switching (Hill et al. 2018; Xie et al. 2018).

For future iterations of target prioritization for multiplexed enhancer-gene pair screening, several characteristics of our identified enhancer-gene pairs are important to keep in mind. Foremost, although a wide range of effect sizes (7.9% to 97.5% for the 470 high-confidence pairs, Fig. 3.4H) were observed on genes with a broad range of expression levels (0.0058 to 313 UMIs/cell, Fig. 3.10C), effect sizes were correlated with expression levels (Spearman's rho 0.53; Fig. 3.11D). This is likely consequent to power, as small effects are more challenging to detect on lowly expressed

genes. Additionally, we note that although we identified many genomic features that were significantly correlated with the likelihood of belonging to an identified pair, a pilot-trained classifier informed by biochemical marks did not appreciably increase our hit rate in the at-scale screen. Furthermore: 1) 29% of enhancers did not fall within the same TAD as their target gene; 2) though enriched for proximity in 3D space as measured by Hi-C, the majority of enhancer-gene pairs are not identified as contacts in such datasets; and 3) though enriched for sequence-level proximity, one-third of enhancer-gene pairs involved skipping of at least one closely located TSS of another K562-expressed gene. These observations underscore the difficulty of the prediction task, and we recommend that future screens do not overly bias themselves towards looking under the lamppost, until additional examples accrue and the rules of mammalian gene regulation are better understood.

Although it may be surprising that *cis* changes in gene expression were identified for only ~10% of the candidate enhancers tested here, there are several potential caveats to bear in mind. First, previous studies have identified shadow enhancers acting to mask the effects of perturbing individual enhancers (Hong, Hendrix, and Levine 2008), although a genome-wide survey of such enhancer redundancy has yet to be conducted. To investigate such interactions more thoroughly, future iterations of our method could randomly distribute programmed pairs of multiplexed enhancer perturbations per locus. Second, other technical caveats include: a) Not all enhancers may be susceptible to dCas9-KRAB perturbation; b) gRNAs may be variably effective in targeting enhancers (Fig. 3.5B); c) Some enhancers required for the initial establishment rather than maintenance of gene expression could be missed in a screen in a stable immortalized cell line; d) We did not comprehensively survey the noncoding landscape surrounding each gene, and the

marks we used to define candidate enhancers may be excluding some classes of distal regulatory elements. These caveats are respectively addressable in the future by using other epigenetic modifiers or nuclease-active Cas9, by using more gRNAs per candidate enhancer, by combinatorial perturbation of selected loci (Xie et al. 2017), by using cell models of differentiation, and by densely tiling selected loci with perturbations.

However, the fact that our paired candidate enhancers are predicted by the strength of enhancer-associated marks (*e.g.* H3K27ac, p300) supports the assertion that we are identifying *bona fide* enhancers, and simultaneously weakens the case for elements that were negative. Also, our study provides new insights into key properties of human enhancers, *e.g.* the distribution of distances between at least some types of enhancers (*i.e.* unbuffered, upstream) and their target genes. A full understanding of the precise rules governing enhancer-promoter choice is a topic of great interest, and will be facilitated by the identification of more enhancer-gene pairs.

A limitation of enhancer-gene pair screening as implemented here relates to the resolution of CRISPRi. In the future, this can potentially be improved upon by adapting enhancer-gene pair screening to use single or pairs of gRNAs with nuclease-active Cas9 to disrupt or delete candidate enhancers at the sequence level. A separate concern is whether high MOI transduction is inducing a cellular inflammatory response, and therefore biasing discovery. However, although some genes with roles in inflammation are amongst our paired target genes (e.g. NMU, IL6), we only observed pathway-level enrichment of one immune-system related pathway. Moreover, the effect sizes observed in our high MOI vs. low MOI experiments were well correlated.

To date, ENCODE has catalogued over 1.3 million human candidate regulatory elements based on biochemical marks (http://screen.umassmed.edu/), while GWAS have identified over 75,000 unique haplotype-trait associations (https://www.ebi.ac.uk/gwas/). Validating candidate elements, fine-mapping of causal regulatory variants, and identifying the target genes of both enhancers and regulatory variants, represent paramount challenges for the field. Given the scale of the problem, we anticipate that the multiplex, genome-wide framework presented here for mapping gene regulation may help overcome these challenges.

3.11 METHODS

Cell Lines and Culture: K562s cells are a pseudotriploid ENCODE Tier I erythroleukemia cell line derived from a female (age 53) with chronic myelogenous leukemia (B. Zhou et al. 2018). K562 cells expressing dCas9-BFP-KRAB (Addgene #46911, polyclonal) were a gift of the Bassik lab, grown at 37°C, and cultured in RPMI 1640 + L-Glutamine (Gibco) supplemented with 10% fetal bovine serum (Rocky Mountain Biologicals) and 1% penicillin-streptomycin (Gibco). K562s were authenticated by bulk/single-cell RNA-seq and visual inspection.

HEK293Ts (a human embryonic kidney female cell line) used for housemade virus production were cultured at 37°C in DMEM also supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin. HEK293Ts were authenticated by visual inspection.

gRNA-library design: (Note about terminology used below: A gRNA-group is defined as all the gRNAs that are targeting the same candidate enhancer or positive control site. To note, all novel

TSS and candidate enhancer targeting gRNA-groups are referred to as “perturbative gRNA-groups”, whereas all others are referred to as “control gRNA-groups”.)

Pilot Library - 1,119 candidate enhancers: Picking candidate enhancer regions: K562 DNase-seq narrowPeaks (ENCSR000EKS) < 1 Kb away from any gene (GENCODE March 2017 v26lift37) were bedtools-intersected (Quinlan and Hall 2010) with K562 Hi-C domains (Rao et al. 2014) that contained at least one of the top 10% most highest expressed genes in a previously generated 6,806 single-cell K562 data set. The remaining regions were largely taken from intersections with K562 GATA1 ChIP-seq narrowPeaks (ENCSR000EFT, lifted to hg19), H3K27ac ChIP-seq narrowPeaks (ENCSR000AKP, lifted to hg19), RNA Pol II ChIP-seq narrowPeaks (ENCSR000AKY), and EP300 ChIP-seq narrowPeaks (ENCSR000EHI) (Fig. 3.2A). Ten further sites were handpicked and do not overlap either of these four marks.

Candidate enhancer gRNAs: NGG-protospacers within these candidate enhancers were scored using default parameters of FlashFry (McKenna and Shendure 2018), and the two top-quality-scoring gRNA per region were chosen as spacers to be used in the gRNA library (scores prioritized by Doench2014OnTarget > Hsu2013 > Doench2016CDFScore > otCount).

TSS positive control gRNAs: 381 genes were randomly sampled from the highly-expressed genes within the same Hi-C domains (as described above) and 2 gRNA were chosen per gene from spacers with the best empirical and predicted scores of the hCRISPRiv2 library (Horlbeck et al. 2016). To note - these spacers are designed as 19 bp, rather than the full 20 of the spacers used in the rest of our gRNAs.

NTC gRNAs: 50 scrambled-sequence spacers with no targets in the genome and 11 protospacers targeting 6 gene-devoid regions of the genome (hg19 chr4:25697737-25700237, chr5:12539119-12541619, chr6:23837183-23839683, chr8:11072736-11075236, chr8:23768553-23771053, chr9:41022164-41024664) were chosen as evaluated by Benchling's CRISPR tool. These were randomly paired to create a gRNA group. More were chosen from 6 random regions of the hg19 genome (chr4:25697737-25700237, chr5:12539118-12541619, chr6:23837183-23839683, chr8:11072736-11075236, chr8:23768553-23771053, chr9:41022164-41024664) using FlashFry (McKenna and Shendure 2018) to total 50 targeting these gene-devoid regions of the genome. A further 39 NTCs were sampled from those recommended by (Horlbeck et al. 2016). A gRNA to the CAG promoter was additionally included as an internal control (excluded from analysis for simplicity).

Distal enhancer positive control gRNAs: 15 gRNAs targeting the HBE1 TSS, and HS1-4 of the Globin LCR were chosen as validated from (Xie et al. 2017; Klann et al. 2017). These were manually paired based on their target sites to create gRNA-groups.

Note: Our initial FlashFry quality annotations when designing the pilot experiment did not label a small number of protospacers with perfect repeat off-targets, permitting their inclusion in our library (81 of 2,238 spacers ordered in the pilot library; only 9 gRNA-groups with both spacers affected). gRNA-groups with an impacted spacer were rare in our 145 significant enhancer-gene pairs. We also note that we still expect these guides to target their intended site, but with potentially more off-targets. This error was fixed for evaluating gRNA quality in the scaled experiment.

At-Scale Library - 5,779 candidate enhancers: Choice of new and repeated sites: A logistic regression classifier built using the 145 enhancer-gene pairs originally identified in the pilot experiment (see Aggregate analysis of enhancer-gene pairs: ChIP-seq strength quintile analysis and logistic regression classifier) was used to select the top 5,000 intergenic open chromatin regions in K562s (as defined by DNase-seq narrowPeaks (ENCSR000EKS)). Of these, 3,853 were over 1 Kb away from boundaries (GENCODE March 2017 v26lift37) of any genes expressed in the pilot 47,650 K562 single-cell dataset, were not previously included in the pilot library, and had minimum two gRNAs with high quality as again determined by FlashFry. Of the top 5,000, 120 corresponded to a candidate enhancer in one of the original 145 pilot enhancer-gene pairs, and 851 of these corresponded to candidate enhancers targeted in the pilot library but not originally identified as part of a enhancer-gene pair. We additionally included 7 more candidate enhancers not top-ranked by our model, but identified as part of the original 145 enhancer-gene pairs. The only candidate enhancer that was identified in an original 145 pilot enhancer-gene pair but not included in this library had no high quality gRNAs by this second library's standards (see Pilot library: Note). Only 15 sites did not overlap any of the marks shown in Fig. 3.4A.

Design of 377 alternative gRNA pairs: Two alternative gRNAs were designed for 377 of the sites repeated from the pilot library. NGG-protospacers within these candidate enhancers were again scored using default parameters of FlashFry (McKenna and Shendure 2018), and the third and fourth top scoring spacers were chosen to be used as an alternative gRNAs.

Choice of 948 exploratory candidate enhancers: Because the logistic regression classifier is biased toward the annotations that were used to select the initially targeted candidate enhancers (Fig. 3.2A), we additionally used submodular subset selection to include DHSs optimized for a diversity of epigenomic features (Wei, Iyer, and Bilmes 2015). We first removed from the full set of 29,833 DHSs (ENCSR000EKS) those 1,119 DHSs that were a part of the original screen. Note that we did not remove the 128 DHSs that had been selected again by the logistic regression model, because doing so would bias our remaining DHSs away from the same annotations. Then we calculated the Pearson correlation of overlapping epigenomic marks between the remaining DHSs. Lastly, we applied a facility location function (Mirchandani and Francis 1990) to this similarity matrix and used a greedy submodular selection algorithm to identify 948 additional DHSs as exploratory candidate enhancers. The top two highest quality gRNAs (as scored by FlashFry) were included to target each candidate enhancer.

Note on choice of gRNA design for future screens of CRISPRi candidate enhancers: We used our set of enhancer-gene pairs to assess if there was a specific gRNA-target location within the candidate enhancer that increased CRISPRi efficacy. We correlated enhancer-gene pair effect size with each gRNA's absolute distance to center of either DHS-peak or overlapping p300 ChIP-seq peak. However, neither the absolute-distance-to-center-of-DHS-peak (Pearson's r: 0.02) nor the absolute-distance-to-center-of-overlapping-p300-peak correlated with effect size (Pearson's r: 0.07). Thus, we currently only recommend prioritizing gRNAs that fall within an open chromatin site based on quality and on-target efficiency as assessed by a gRNA quality algorithm like Flashfry (McKenna and Shendure 2018).

gRNA-library cloning: The lentiviral CROP-seq gRNA-expression vector (Datlinger et al. 2017) was modified by Q5-Site Directed Mutagenesis (New England BioLabs, F:5-acagcatagcaagtttAAATAAGGCTAGTCCGTTATC-3 R:5-ttccagcatagctcttAAACAGAGACGTACAAAAAAG-3) to incorporate the previously described gRNA-(F+E)-combined backbone optimized for CRISPRi (B. Chen et al. 2013; Hill et al. 2018), Addgene #106280). Prepared vector was digested with BsmBI (FastDigest Esp3I, Thermo Fisher Scientific), "filler" sequence removed by gel extraction, and cleaned (Zymo Research DNA Clean & Concentrator-5) vector without "filler" was used for all downstream cloning.

Spacer libraries were ordered as single stranded pools (CustomArray, 5-atcttgtggaaaggacgaaacaccGNNNNNNNNNNNNNNNNNNgtttaagagctatgctggaaacagcatagcaagt-3). 1 ng of each pool was amplified (F = 5-atcttGTGGAAAGGACGAAACA-3, R = 5-acttgctaTGCTGTTTCCAGC-3, 64C Tm, Kapa Biosystems HiFi Hotstart ReadyMix (KHF), see Special Note below, as we now recommended a different R primer = 5-CTGTTTCCAGCATAGCTCTTAAAC-3) and purified amplicons (Zymo Research DNA Clean & Concentrator-5) were cloned into CRISPRi-optimized CROP-seq vector prepared as described above (NEBuilder® HiFi DNA Assembly Cloning Kit, NEB, 100 fmol purified vector : 200 fmol cleaned insert). 2 ul of each product was transformed into Stable Competent E. coli (NEB C3040H) in enough replicates to produce >20 transformant clones per gRNA in the library. Plasmid DNA was purified using ZymoPURE Maxiprep kits, following by DNA Clean and Concentrator cleaning (Zymo Research).

Special note: In Sanger sequence of the final gRNA plasmid libraries and in the 8-15 bp immediately downstream of the spacer (7 bp of the gRNA backbone transcript captured in all single-cell RNA-sequencing datasets), we identified that ~80% of gRNAs harbored a small insertion or deletion (vast majority 1 bp deletions, Fig. 3.14A) in between the spacer and the R primer 5-acttgctaTGCTGTTTCCAGC-3 used in the initial amplification of spacer-oligos. We inferred that this is due to slippage of the KHF polymerase as it copies the secondary structure of the first stem extension loop added as part of the more stable sgOPTI backbone. In the scRNA-seq data, ~70% of gRNA carried a 1 bp deletion, ~8% carried a 2 bp deletion, and ~2% carried a 3 bp deletion (Fig. 3.14A).

Fortunately, 1 bp deletions did not correlate with significant disruption of CRISPRi efficacy in the scRNA-seq data. (1 bp deletion % reduction) / (full length gRNA reduction) ratio was 1.01 (high confidence enhancer-gene pair) or 0.958 (TSS control). For 2 bp deletions, this ratio was also not extreme (0.959 (high confidence pair) or 0.806 (TSS control)). However, for 3 bp deletions (very rare), the ratio was 0.908 (high confidence pair) or 0.644 (TSS control). Overall correlation of all these deletion lengths to full length efficacy was very high (Fig. 3.14B).

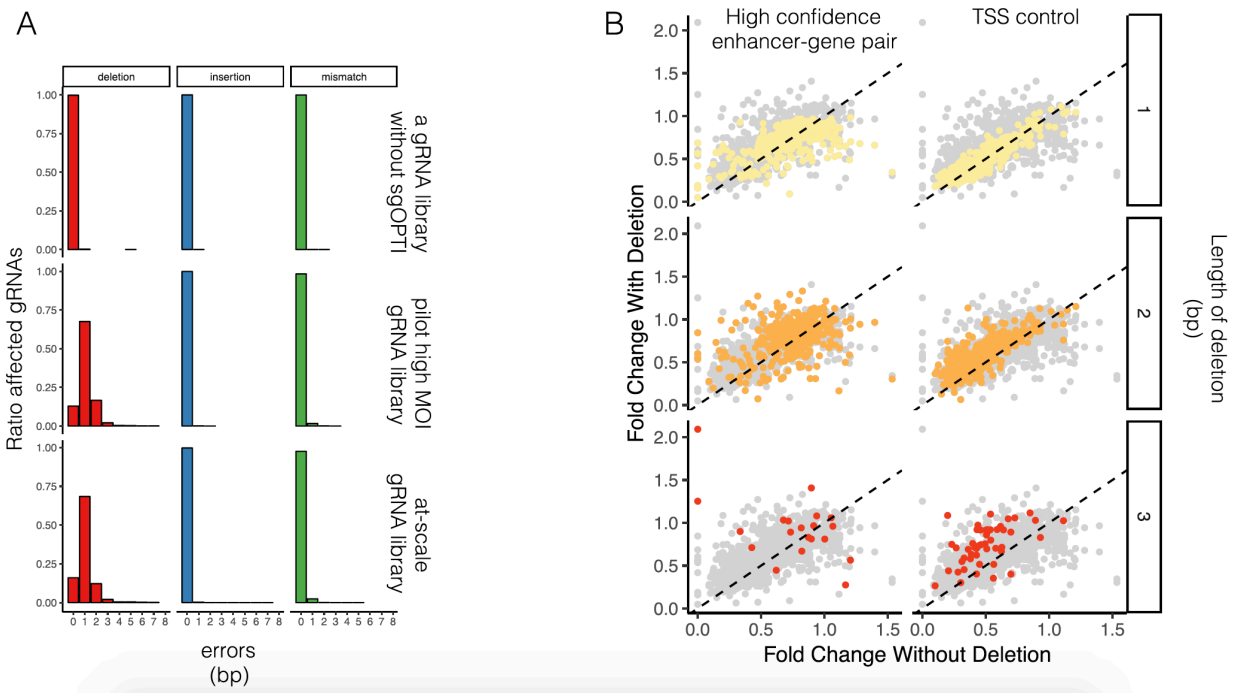


Figure 3.14 Supplementary details related to sgOPTI-backbone error correction.

(A-B) Quantification of errors in synthesis of the sgOPTI gRNA backbone across the three scRNA-seq datasets. (A) Deletion (red), insertion (blue), or mismatch (green) rate in the 8-15 bp downstream of the spacer in the gRNA backbone as captured by gRNA-transcript enrichment from scRNA-sequencing data. Data is shown for scRNA-seq datasets of a gRNA library that does not have sgOPTI added to the backbone (but was cloned, amplified and sequenced in a similar manner), the pilot high MOI gRNA library, and the at-scale gRNA library. (B) The impact of indels on effect sizes for paired-candidate enhancers (high confidence set) and TSS positive controls. The effect size of gRNAs with versus without perfect backbones, stratified by length of deletion. Gray points = a unique dot is plotted for the subgroups of each paired enhancer/TSS gRNA, divided by if they harbor 0, 1, 2, or 3 bp deletions. Colored points = set of gRNAs bearing the specified deletion length. Only points for which there are >=50 cells in a given deletion-length group are plotted to ensure reasonable estimates of fold change.

Thus, the vast majority (~90%) are either wildtype or harbor 1 bp deletions that create zero-to-little effect on CRISPRi efficacy. 8% of the remaining gRNA harbor 2 bp deletions that also largely do not affect CRISPRi efficacy. However, to avoid this problem in cloning future gRNA libraries into the sgOPTI-CROP-seq plasmid, we now recommend amplifying with a reverse primer that is flush with the spacer (5-CTGTTTCCAGCATAGCTCTTAAAC-3), potentially enabling a boost in repression efficacy.

Virus production and transduction: The Fred Hutchinson Co-operative Center for Excellence in Hematology Vector Production core produced all virus for the multiplexed enhancer-gene pair screening experiments. For the singleton CRISPRi recapitulation, virus was made in-house by co-transfecting (Lipofectamine 3000, ThermoFisher, L300015) HEK293Ts with the small pools of CRISPRi-optimized CROP-seq with the ViraPower™ Lentiviral Packaging Mix (ThermoFisher). After 3 days, supernatant was syringe filtered with a 0.45 uM filter (cellulose acetate, VWR) to prepare virus for transduction.

Cells were transduced (8 μg/mL polybrene) with varying titers and amounts of virus to achieve differing MOI. 400,000 and ~2.5 million original cells were transduced for the pilot and at-scale experiments, respectively. At 24 hours post-transduction, cells were spun and resuspended with virus- and polybrene- free media. At a total 48 hours post-transduction, 2 μg/mL puromycin was added to the culture, and changed to 1 μg/mL puromycin at the next passage for maintenance. A total of 10 days post transduction, cells were collected for scRNA-seq or bulkRNA-seq.

Single cell transcriptome capture

~4000-8000 cells were captured per lane of a 10X Chromium device using 10X V2 Single Cell 3' Solution reagents (10X Genomics, Inc). Six lanes were used for both the low and high MOI 1,119-pilot library experiments, and 32 lanes were used for the scaled experiment. All protocols were performed as per the Single Cell 3' Reagent Kits v2 User Guide (Rev B), except prior to the enzymatic shearing step, 10% of full length cDNA was taken for PCR enrichment of gRNA-sequences off the CRISPRi-optimized CROP-seq transcripts as described below. After RT, the 32 lanes of the scaled experiment were split into two batches (16 lanes each) for the remainder of the prep to enable easier handling.

gRNA-transcript enrichment PCR: A three-step hemi-nested PCR reaction was performed to enrich gRNA sequences from the 3' UTR of puromycin resistance gene transcripts produced by the CRISPRi-optimized CROP-seq integrant. PCR was monitored by qPCR to avoid overamplification, and each reaction was stopped immediately before it reached saturation.

PCR 1: 10-13 ng of full-length 10x scRNA-seq cDNA were amplified in each 50 µl KHF reaction (annealing temp 65C), spiked with SYBR Green (Invitrogen) for qPCR monitoring (10% of all unfragmented 10x cDNA).

F: U6_OUTER 5- TTTCCCATGATTCCTTCATATTTGC -3

R: R1_PCR1 5- ACACTCTTTCCCTACACGACG-3

PCR 2: Sample replicates were pooled, cleaned with 1x Agencourt AMPure XP beads (Beckman Coulter), and 1/25th of the cleaned pooled product was amplified in a 50 µl KHF reaction spiked with SYBR Green and monitored as above (annealing temp 65C).

F: U6_INNER_with_P7_adapter 5-
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcTTGTGGAAAGGACGAAACAC -3
R: R1-P5 5-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG-3

PCR 3: The PCR 2 replicate reactions were pooled and 1x AMPure cleaned. 1/25th of the cleaned pooled product was amplified in a 50 µl KHF reaction (spiked with SYBR Green and monitored as above, annealing temp 72C) and products cleaned once again via 1x Ampure.

F: 5-CAAGCAGAAGACGGCATACGAGATIIIIIIIIIIGTCTCGTGGGCTCGG-3 (standard NEXTERA P7 indexing primer)
R: R1-P5 again

Sequencing of scRNA-seq libraries:Pilot library experiments - The final libraries were sequenced on a NextSeq 500 using four 75-cycle high-output kits (R1:26 I1:8, I2:0, R2:57) for each experiment (low and high MOI). Scaled library experiments - The final library was sequenced by the Northwest Genomics Center on a NovaSeq 6000 using an S4 flow cell (R1:26, I1:8, I2:0, R2:91). All libraries were sequenced to ~20% sequencing saturation.

Digital gene expression quantification: Sequencing data from each sample was processed using the Cell Ranger software package as provided by 10x Genomics, Inc., to generate sparse matrices of UMI counts for each gene across all cells in the experiment.

Each lane of cells was processed independently using cellranger count, aggregating data from multiple sequencing runs. The pilot library experiments were each processed with cellranger 2.0.2; the at-scale library experiment was processed with cellranger 2.1.1.

Definition of genes well-expressed or 'detectably expressed' in K562: Unless otherwise notes, genes were defined as well expressed or detectably expressed in K562 if they had at least one read in 0.525% of cells in their respective single cell RNA-seq datasets.

Assigning genotypes to cells: gRNAs were assigned to cells in the following method (Hill et al. 2018): Sequences corresponding to the gRNA-containing CRISPRi-optimized CROP-seq transcripts are extracted from the cellranger position sorted BAM file after running our custom indexed libraries through the cellranger pipeline to tag reads with corrected cell barcodes and UMIs. gRNA sequences are extracted and corrected to the library whitelist within an edit distance of two, and gRNA-cell pairs are tracked when a valid cell barcode and UMI are both assigned to the read. Likely chimeric reads are detected and removed to reduce noise in the assignments as previously described. We utilized thresholds to set minimum acceptable values for the total reads for a gRNA-cell pair and for the proportion of all CROP-seq transcript reads accounted for by each gRNA observed in a cell to distinguish noise from real assignments (Hill et al. 2018). Here, given the larger number of guides contained in each cell, we find that UMI counts provide a much cleaner

distribution than read counts and have used UMI counts in all calculations. For the 1,119 pilot library experiments we used 0.01 read counts and 5 UMI in both our low and high MOI for each of these thresholds. For the scaled library experiment, we used 0.005 read counts and 5 UMI. Only cell barcodes that appear in the set of passing cells output by cellranger, which imposes an automated threshold on the total UMIs observed in cells, are carried forward in downstream analysis.

Differential expression tests: In our cis analyses, we tested each perturbing gRNA-group against genes within 1 Mb of the gRNA. These gRNA-gene pairs were identified by using bedtools to intersect the DHSs targeted by the gRNA library with 1 Mb windows in either direction of TSS annotations from GENCODE March 2017 v26lift37 (total of 2 Mb, centered around the TSS). In our trans analysis, all gRNA-groups were paired with all genes that were defined as expressed in K562. In both cis and trans analyses, NTCs were tested against any genes used to test perturbing-gRNAs.

For each gRNA-group we assigned a label of "1" to cells that contained a gRNA belonging to that group and a label of "0" to all other cells in the dataset. Monocle2 (Qiu et al. 2017) was used to perform a differential expression test, using the negbinomial.size family, over this categorical label to find differentially expressed genes between these two groups. Due to its support of complex model formulas, Monocle2 does not provide model coefficients as part of the differential expression results. We created a modified version of the differentialGeneTest function and associated helper functions that return both the intercept term and the coefficient of the group assignment to facilitate more robust prioritization and characterization of hits from our screen. The

negative binomial family uses log as the link-function, so we can calculate the initial expression level as exp(intercept), and the fold change in expression between the two groups as exp(group_coefficient + intercept) / exp(intercept). We verified data from our power simulations that the appropriate effect sizes can be obtained with this method using the coefficients output by VGAM.

For the scaled experiment, as we collected a much larger number of lanes and observed the highest MOI, we regressed out the number of guide RNAs observed in a cell (as a proxy for the number of integrants), the percentage of total transcripts observed that are mitochondrial, and the prep batch (as following reverse transcription, the 32 lanes were prepared in two batches to make handling easier). In practice, we observe a modest boost in sensitivity when regressing out each of these factors in DE testing. This was done using the full model formula ~gRNA_group+guide_count+percent.mito+prep_batch and the reduced model ~guide_count+percent.mito+prep_batch in Monocle2.

Calling hits from differential expression test results: All differential expression test results were performed for all K562 expressed genes within 1 Mb of the target site as defined by GENCODE March 2017 v26lift37. NTCs were tested against all genes within 1 Mb of any target site.

Tests with two sources of potential false positives were excluded: In the pilot experiment, we identified inflation of NTCs when testing them against genes highly impacted by perturbing-gRNA in our library (for example, NTCs associated with targets of our TSS and globin LCR controls). This was due to subtle yet detectable nonrandom associations of gRNA-groups with other gRNA-

groups across cells, potentially due to slight bottlenecking at the transduction level (400,000 cells transduced for 1,119 pilot library vs 2.5 million transduced for 5,779 scaled library). To exclude this source of inflation in the pilot dataset, we used Fisher's exact test to identify when an NTC was nonrandomly assorted with a perturbing-gRNA (adjusted P-value < 0.01 & odds ratio > 1). Then, any test of an NTC against a gene within 1 Mb of that gRNA's gRNA-group was excluded from further analytical steps.

We noted Monocle was susceptible to inflating P-values when a gene was highly expressed but only in few cells. Three of our 381 TSS controls fell into this category. To avoid this problem, we excluded outlier genes that were expressed in < 20,000 cells in either the high-MOI 47,650-cell dataset and/or the scaled 207,324-cell dataset, and with log10(total UMIs / cells with a UMI) >0.2 greater than predicted by a spline fit generated via smooth.spline() with spar=0.85 to limit overfitting (35 genes total). Remaining tests were filtered to those that decreased expression of the target gene.

Then, an empirical P-value was defined for each gene-gRNA-group pair test as: [(the number of NTCs with a smaller P-value than that test's raw P-value) + 1] divided by [the total number of NTCs tests + 1].

These empirical P-values were Benjamini-Hochberg corrected, and those < 0.1 were kept for 10% empirical FDR sets.

Use of 3.5% empirical FDR to initially select enhancer-gene pairs from the pilot study: We originally used an alternative method to call the original 145 enhancer-gene pairs from the original pilot dataset (a universal cutoff of the P-value at which the proportion of passing NTC-tests/total NTC-tests was 10% of the proportion of passing candidate enhancer tests/total candidate enhancer tests). However, upon further discussion and review of the eQTL literature, we revised our method to the one defined above. This original threshold corresponded to a 3.5% empirical FDR rate, as defined above.

Inclusive vs. high confidence enhancer-gene pairs: The only requirement of enhancer-gene pairs in the inclusive set was that they passed a 10% empirical FDR in the scaled experiment. To be included in the high confidence set, enhancer-gene pairs either had to be replicated at a 10% empirical FDR in the pilot dataset, or (if a candidate enhancer was unique to the scaled experiment) both gRNAs had to be individually associated with >10% repression of the gene.

Analyses to evaluate reproducibility between gRNA: To evaluate reproducibility between gRNAs, we subset the 377 pairs (two sets of gRNA pairs targeting the same candidate enhancers in the scaled experiment) to pairs where both pairs negatively repressed at least one target gene (no significance requirement). 20 of the 377 did not meet this criteria. Then, we ranked all tested genes by average repression between the two gRNA pairs, and kept the top ranked gene for each pair. The repression levels of each type of gRNA pair on this top-ranked gene are plotted in Fig. 3.5B, regardless of significance.

Intracellular abundance of gRNA and dCas9-KRAB transcript does not correlate with effect size: As both the dCas9-BFP-KRAB and the sgOPTI-CROP-seq construct transcripts are poly-A tagged, we are able to test if there is an association between the CRISPR components' UMI counts and transcript abundance of a targeted gene. For the 441 candidate enhancers in a high confidence pair, we subsetted to the cells that held a guide targeting each enhancer. Within this set of cells, we tested for a significant association between the expression of the target gene and the UMI count of the dCas9-BFP-KRAB or the guides (adjusting for total cell UMI count). Of the 470 enhancer-gene pairs, only 2 and 10 had any significant (adjusted P-value < 0.01; 7 and 27 for adjusted P-value < 0.05) association with dCas9-BFP-KRAB count or guide count respectively (0.4% and 2% or 1.5% and 5.7% of tests for each adjusted P-value threshold respectively). Based on this, we conclude there is not evidence for a substantial effect of dCas9-BFP-KRAB or guide counts on the observed effect size for a given enhancer-gene pair.

Quantifying gRNA abundance: In the process of assigning gRNAs to cells, we had already quantified the number of reads and UMIs associated with gRNA-cell pairs. These counts were used as is for the above analysis.

Quantifying dCas9-BFP-KRAB in cells: We constructed a bowtie2 (Langmead and Salzberg 2012) index for a reference including both the PuroR transcript from the sgOPTI-CROP-seq vector (extending from PuroR to the 3' LTR encoding the guide sequence as N's) and dCas9-BFP-KRAB (including the 3' LTR). Note that both gRNA and dCas9 transcripts were included in this analysis because several regions are identical within the 3' UTR of the transcripts encoded by these two constructs. We then took all the unmapped reads from the unbiased (cell) libraries and converted

them back into fastq format adding the final cell ID and UMI from cellranger into the read name for use downstream. We mapped these reads to the reference above using bowtie2 using the command "bowtie2 -p 8 --n-ceil 20 --np 0 -x <reference> -U <fastq input> -S <bam output>". We then took only reads that map uniquely to the dCas9 contig with mapq of 30 or greater and enumerated the number of UMIs and total reads seen for each cell / barcode pair dCas9.

In each case, we tested for associations between the gRNA/dCas9 counts and the abundance of each high confidence hit in our screen, only within cells that had a guide to the target. This was done using our modified version of differentialGeneTest as described above. Note that in this case we observed that size factors typically used to account for variation in total UMI counts across cells did not appear to sufficiently correct for the strong correlation between the counts of two transcripts (the gRNA transcript / dCas9 and the target) that results from variation in total UMI counts across cells. This initially resulted in residual associations that indicated increased gRNA transcript / Cas9 resulted in higher target expression. To account for this, we added an additional term to both the full and reduced model "~log10(total_umis)" and set all size factors to 1. This is the model from which we report the above results.

Individual replication by CRISPRi singletons: To replicate a enhancer-gene pair's phenotype outside of the pooled mapping format, we prepared small pools of gRNAs re-targeting 15 high-confidence candidate enhancers or the TSSs of their respective paired-target genes. These enhancer-gene pairs were chosen from the following requirements: candidate enhancer tested in both the pilot and at-scale study (replicated between both); target gene in upper 50% of expression of all paired genes; target gene had no strong cancer associations or growth phenotypes.

Additionally, we chose 6 candidate enhancers that were not paired with any target gene using the following requirements: tested in both the pilot and at-scale screen; empirical P-values for any cis gene > 0.5 in both experiments; overlapping H3K27Ac ChIP-seq peak is in the top half of all the peaks that overlap the entire at-scale library (thus to be comparable with our paired enhancers); and within 1 Mb of a K562 expressed gene.

The two original gRNAs and two new gRNAs (making up the top 4 ranked on-target activity per candidate enhancer, filtering out those with high off-target scores using Flashfry (McKenna and Shendure 2018); exception is candidate enhancer chr11.4680 where only 3 gRNAs passed these quality filters) were used for each respective pool, for a total of 4 gRNAs in the pool. The two original gRNAs were used for the TSS controls (plus two more alternative TSS gRNAs in the cases of NMU, GYPC, PTGER3, and PRKCB). These small gRNA pools were cloned into the CRISPRi-optimized-CROP-seq vector (as described above, except in the case of e-NMU targeting pool, which was cloned by ordering two reverse complement single stranded oligos and annealing them together into px459 (CRISPR-Reagent-Description_Rev20140509.pdf, (Cong et al. 2013)). House-made lentiviral preps from these gRNA pools were transduced at low MOI into the K562-dCas9-BFP-KRAB line, and cultured for 10 days under puromycin selection before two technical replicates of total RNA were collected from each sample (RNeasy Mini Kit, Qiagen).

Bulk RNA-seq libraries were prepared from each replicate via a TruSeq mRNA kit (400 ng input, Illumina, TruSeq RNA Sample Prep Kit v2 RS-122-2002 or TruSeq Stranded mRNA Library Prep 20020595), and sequenced on a NextSeq 500 (total two 150-cycle kits cycling 80/80/6 in mid output mode for e-NMU, e-PRKCB, e-GYPC, e-PTGER3; total two 75-cycle kits cycling 40/40/8

in high output for all others; aiming for 10-20 million reads/sample). Gene-level quantifications and differential expression tests were performed via kallisto (Bray et al. 2016) and sleuth (Pimentel et al. 2017). Repression percentages were calculated from the kallisto transcript per million output table (normalized by size factors): (mean between the two replicates / mean between all-non targeting samples). To note, targeting the TSS of CITED2 did not seem to successfully repress CITED2's expression, though this is potentially due to inaccuracy of 1 of 2 technical replicates for this sample. The 3 that matched direction and magnitude of effect but were not significant in a test of differential expression potentially were not detectable due to lack of power, as we sequenced only two RNA replicates per sample.

To note: we additionally generated singleton datasets for chr6:34191315-34191338 (paired with HMGA1 in the pilot screen), but did not include this in analysis as it did not reproduce between the pilot and at-scale screen, and thus was not part of our high confidence enhancer-gene pair set.

Validation by sequence deletions: To generate monoclonal sequence lines of three candidate enhancers, we designed protospacers to flank the DHSs targeted in e-NMU, e-GLUL, and e-CITED2. Spacers were order as single stranded oligos and then amplified (KHF, 5-GTGGAAAGGACGAAACACCg-3, 5-gctaTTTCtagctctaaaac-3, 55°C tm, 15 second extension; followed by clean-up via Zymo Research DNA Clean & Concentrator) to be made double stranded for Gibson Assembly cloning (50 ng prepared vector : 0.66 ng prepared insert) into the Cas9- and gRNA- expression vector px459 (Ran et al. 2013), expressing both the gRNA and a cassette of Cas9-2A-puromycin resistance; NEBuilder® HiFi DNA Assembly Cloning Kit). Some e-NMU

targeting oligos were cloned by annealing two complementary oligos together followed by ligation into px459, in the method of CRISPR-Reagent-Description_Rev20140509 (Cong et al. 2013).

We transiently transfected the small px459 pools into the K562+dCas9-KRAB cell line using the Neon nucleofection system (500,000 cells per library, 10 uL tips, 500 ng of plasmid, pulse voltage 1450–pulse width 10–pulse number 3; ThermoFisher). Beginning 24 hours after transfection, cells were selected with 1 ug/mL puromycin for 48-72 hours, then single-cell sorted into 96 well plates using a FACSAria II (Becton Dickinson). To finally achieve clones that harbored fully homozygous deletions of e-NMU, this process was repeated on an initial set of heterozygous clones using a second round of flanking gRNAs.

After 3-4 weeks of growth, gDNA was extracted by concentrating cells into 20 uL of media, and adding 40 uL of house-made Quick Extract buffer (EB + 4 mg/mL proK + 0.45% Tween20), followed by 65°C for 6 minutes and 98°C for 2 minutes. 1 uL of this gDNA extract was used for genotyping PCRs (Kapa2G Robust PCR kit, 35 cycles 60°C-HS-3 minute extensions).

Two rounds of genotyping PCRs were performed. First, clones were screened with primers flanking the deletion to identify clones that harbored a deletion on at least one allele. Second, to confirm homozygosity, primers internal to the deleted region were used to identify candidates that still harbored wildtype alleles (Fig. 3.9A). Clones that harbored full deletions with no remaining wildtype alleles were submitted to bulkRNA-sequencing (Fig. 3.6E-G). Two technical replicates of RNA were extracted from each monoclonal line (RNeasy Mini Kit, Qiagen), bulkRNA-seq libraries prepared via a TruSeq mRNA kit (400 ng input, Illumina, TruSeq Stranded mRNA

Library Prep 20020595), and sequenced on a NextSeq 500 (one 75-cycle kits cycling 40/40/8 in high output for monoclonal samples; aiming for 10-20 million reads/sample). Gene-level quantifications were performed as for the CRISPRi singletons, and reduction percentages calculated from kallisto transcript per million output table (normalized by size factors): (mean of all replicates per candidate enhancer) / (mean between all-non targeting samples).

Phenotyping e-NMU perturbations by flowFISH: Cells harboring e-NMU CRISPRi perturbations were generated as in Individual replication by CRISPRi singletons. A heterogeneous population of cells harboring full e-NMU deletions was generated as in Validation by sequence deletions: (though without single-cell clone sorting). A heterogeneous population of cells harboring scanning deletions across e-NMU was generated by cloning and transfecting 19 gRNAs targeted every ~100 bp across the e-NMU locus as described above in Validation: sequence deletions"

Fluorophore labeled complementary probes to NMU transcript were designed on and ordered from <https://www.molecularinstruments.com/>. The ‘non-targeting’ probes were scrambled versions of the original NMU-targeting probes (to preserve sequence features such as GC content). RNA flowFISH was performed according to Molecular Instruments' in situ HCR v3.0 protocol (Choi et al. 2018), which we have described again here: Cells were by resuspending in 4% formaldehyde to reach 10^6 cells/mL, and fixing for 1 hour. Formaldehyde was then removed, cells were washed four times in PBST (1x PBS + 0.1% Tween 20), and then resuspended in 70% ethanol. For labeling, cells were first washed twice with PBST, and then pre-hybridized by incubating at 37C for 30 minutes in 30% probe hybridization buffer (30% formamide, 4x sodium chloride sodium citrate (SSC), 9 mM citric acid, 0.1% Tween 20, 50 ug/mL heparin, 1x Denhardt’s solution, and

10% low MW dextran sulfate). Cells were then incubated overnight at 37℃ in a final 4 nM probe solution (prepared by adding 2 pmol each probe (a mix of 1 uL of 2 uM stock per each probe) + 100 uL of 30% probe hybridization buffer). Cells were then repeatedly resuspended in 30% probe wash buffer and incubated for 10 minutes at 37℃, for a total of four washes. Cells were then resuspended in 5x SSCT (5x SSC + 0.1% Tween 20) and incubated at room temperature for 5 minutes before amplification.

For amplification, cells are resuspended in amplification buffer (5x SSC + 0.1% Tween 20 + 10% low MW dextran sulfate) and pre-amplified by incubating for 30 minutes at room temperature. 15 pmol of each fluorescently labeled hairpin was snap-cooled by heating 5 uL of 3 uM stock in hairpin storage buffer (Molecular Instruments) to 95℃ and then cooling for 30 minutes to room temperature in a dark drawer. Snap-cooled hairpins were then mixed with amplification buffer, added to the sample for a final concentration of 60 nM, and then incubated overnight (>12 hours) at room temperature in the dark. Cells were washed six times by resuspension in 5x SSCT, resuspended in 500 ul 2x SSC, and incubated for 30 minutes at room temperature with 0.5 uL Vybrant Dye Cycle Orange (DNA stain).

For sorting, the cells are first gated based on size and granularity using forward versus side scatter to discriminate between debris and cells. Cells in G0/G1 stage are then selected using DNA dye (Vybrant Dye Cycle Orange).Cells are then sorted into low, medium, or high bins of NMU expression using AF647 (Becton Dickinson; ~500,000 cells for the full deletion low NMU bin, ~1,000,000 cells for all other bins).

To reverse cross-link the sorted samples, cell pellets were resuspended in 500 ul of elution buffer (4 mL H2O + 500 ul 10% SDS + 500 ul NaHCO3 -1M) + 30 ul of NaCl (5M) and incubated overnight at 65°C. 8 ul of RNase (10 mg/mL) was added to each sample, mixed by inversion, and incubated at 37°C for 2 hours. 4 ul of Proteinase K (20 mg/mL) was added, mixed by inversion, and incubated for 2 hours at 55°C. gDNA was extracted by phenol chloroform, ethanol precipitated, and resuspended in Qiagen elution buffer.

PCR to identify e-NMU genotype enrichments in each of the NMU expression bins (Fig. 3.9B-C) was performed using Kapa2G Robust (e-NMU outer PCR: F primer 5'TCCAACCCCTCAACTTGTT3' Reverse primer 5'TGCCTTCTCTGCCTTTCATT3'; anneal 60°C, extension time 1:50) on 10 ng of gDNA. PCRs were spiked with SybrGreen, and monitored on a qPCR to allow removal before overamplification to prevent excessive PCR biases. 1 uL of each PCR reaction was run on a 6% TBE polyacrylamide gel (Invitrogen) for 35 minutes at 180 V and stained with Sybr Gold for visualization. Replicate PCRs are represented by different lanes in Fig. 3.9B-C.

Aggregate analysis of enhancer-gene pairs: The high confidence enhancer-gene pairs were used for these analyses unless otherwise noted. Details of empirical FDR and the significance thresholds used to call enhancer-gene pairs can be found above in Calling hits from differential expression test results. Singleton re-testing and validations of enhancer-gene pairs used to functionally test if the data met the assumptions of these statistical methods can be found above in Replication of enhancer-gene pairs as singletons.

Distance between perturbation and target gene: Distance was calculated between the GENCODE March 2017 v26lift37 annotated TSS of the perturbed gene and the middle of the originally targeted open chromatin region (if targeting a candidate enhancer, ENCFF001UWQ) or the GENCODE-annotated TSS of the originally targeted transcript (if targeting a TSS). To note, in Fig. 3.10A and to calculate the median distance, we have only used enhancers that are upstream of the target gene, as the length of the gene body would confound distance-to-TSS measurements for downstream enhancer-gene pairs.

Expression distributions: Average expression of each transcript was defined as mean UMI counts per cell in the 47,650 or 207,324 cell scaled dataset. K562 expressed genes were defined as at least one read in 0.525% of cells in the same dataset.

ChIP-seq strength quintile analysis and logistic regression classifier: All candidate enhancers targeted in each library were bedtools-intersected with 170 ChIP-seq of histone-associated marks (ENCODE Project Consortium 2012)), broken into quintiles of the 7th “signalValue” column (peak strength, usually representing overall average enrichment in the region), and the rates of enhancer-gene pairs identified in each quintile were used. In addition to average phyloP conservation score per candidate enhancer, these were used to fit both independent and multivariate logistic regression classifiers using the glm() function with binomial family. We calculated fold changes for how likely a candidate enhancers was paired by: 1 + (((odds ratio - 1) * highest quintile ChIP-seq value) - ((odds ratio - 1) * lowest quintile ChIP-seq value)).

Motif enrichment in enhancers and promoters: Using the AME tool (Analysis of Motif Enrichment) from the MEME suite (McLeay and Bailey 2010), enhancer analysis: we compared motifs enriched in the 600 candidate enhancers in the inclusive set of 664 pairs as compared to all 5,779 in the at-scale library; promoter analysis: compared motifs enriched the 1 Kb upstream of the TSS (~promoter) of the 479 genes in the inclusive 664 pairs as compared to the 1 Kb of promoters of all K562 expressed genes within 1 Mb of a tested candidate enhancer. Parameters were set to default, and Hocomoco Human v11 (core) (Kulakovskiy et al. 2013) was used as the motif library.

Motifs of TF couples across paired promoters and enhancer: To test if pairs of transcription factor (TF) motifs were enriched for co-presence across paired promoters and enhancers, we first identified 179 TFs that were expressed in K562s and had high quality motifs in Hocomoco. Using the FIMO tool (Find Individual Motif Occurrences) from the MEME suite, we annotated all 600 candidate enhancers and the promoters of all 479 genes (1 Kb upstream of the TSS) in the inclusive set of 664 pairs. Motifs in the bottom quartile of how often seen in a promoter were excluded for lack of power. Then, we looped through all possible pairs of 179 TFs in the enhancer (TFe) x 179 TFs in the promoter (TFp), and for each TFe x TFp pair, performed a Fisher's Exact test on contingency tables designed as follows:

For the promoters of 479 paired genes: TFp in promoter or TFp not in promoter vs. Promoter paired with an enhancer that contains TFe or Promoter not paired with an enhancer that contains TFe

For the 600 paired enhancers: TFe in enhancer or TFe not in enhancer vs. Enhancer paired with an enhancer that contains TFp or Enhancer not paired with an enhancer that contains TFp

The six TFe x TFp co-enriched couples that had a Benjamini Hochberg corrected P-value < 0.1 for both the 479 paired promoter analysis and the 600 paired enhancers analysis were described in the main text.

ChIP-seq of TF couples across paired promoters and enhancer: Bedtools was used to mark when a paired enhancer or promoter in the 664 inclusive dataset overlapped a ChIP-seq peak from ENCODE generated K562 datasets were used. ChIP-seq datasets that that were in the bottom quartile of how-often-overlapping with a paired enhancer or promoter were excluded for power (leaving 168 TFe and 166 TFp). Analysis was then performed the same as in the TFe x TFp motif analysis (Fisher's Exact Test, adjusted P-value < 0.1, pair required to be enriched when looping through both enhancers and then through promoters, TFe and TFp required to be different).

Functional annotation enrichment: We used the Piano package (Väremo, Nielsen, and Nookaew 2013) to perform functional annotation enrichment from the 'all pathways' Gene Ontology (http://download.baderlab.org/EM_Genesets/June_20_2014/Human/June_20_2014_versions.txt). The 10,560 K562-expressed genes within 1 Mb of a perturbing-gRNA were used as our background dataset, and randomly sampled from genes with expression greater than one standard deviation below the mean of our 353 targeted genes was used as the comparison set of "expression matched controls" (Fig. 3.10C).

Hi-C analysis: We used the in situ Hi-C dataset for K562 cells from Rao et al, Cell 2014, using the MAPQ 0 threshold and KR normalization, at 5 Kb resolution. We first created shuffled control loci pairs by starting with the set of enhancer-gene TSS pairs, and randomly shuffling the oriented distances between enhancer-TSS pairs, keeping either the enhancers or the TSSs intact. The rare cases where shuffling resulted in an invalid chromosomal coordinate were excluded. For each set of loci pairs, we identified the TADs (as defined in Rao et al, Cell 2014 using Arrowhead) encompassing each loci pair. For overlapping domains, we used the farthest domain boundary on each side of the loci pair. We omitted loci pairs that were not encompassed by any TADs from further analysis. We then extracted the normalized Hi-C counts for each loci pair, along with those for all other bins representing interactions at the same genomic distance within the same TAD, and calculated its fractional rank (scaled from 0 to 1, with 1 representing the highest interaction frequency). Finally, the distributions of fractional ranks were plotted and compared. In addition to comparing interactions within TADs, we also compared loci pairs to other bins within 200 Kb or 1 Mb of each loci pair.

Analyses for multiplexability of CRISPRi within cells - low versus high MOI comparisons: In order to confirm the efficacy of repression in our high MOI experiments (pilot library MOI = ~15 and at-scale library MOI = ~28), we sought to compare the degree of repression observed in each of these experiments to that observed in our low MOI control experiment (pilot library MOI = ~1). We took all gene-target site differential expression tests passing a 10% empirical FDR in any one of the three experiments (as evaluated independently in each screen). We used this set rather than our final hit list to ensure that we were not biasing our comparison by excluding tests that would

be independently called by any one screen but not the others, although we note that the results of the same set of analyses using our final set of hits are very similar.

For each of these tests, we calculated the observed fold changes of repression (where 1 is no change and 0 is complete loss of expression) for each screen and then calculated the following ratios: (pilot high MOI fold change) / (pilot low MOI fold change) and (at-scale fold change) / (pilot low MOI fold change), using a pseudocount of 0.01. As we found it potentially confusing that a higher value of these ratios represents worse efficacy of repression in the high MOI experiments, we considered making these ratios of percent repression (1 - fold change). However, as this value could be negative in some cases (where the fold change was greater than one in one of the screens), this was not compatible with display on a log scale. Therefore, in all plots showing such ratios, we are actually showing the inverse of the fold change ratios described above, which should approximately represent the ratios of percent repression without producing any negative values. Thus, in our plots and reported summary statistics, values less than one represent cases where more repression was observed in the low MOI control.

Despite the distributions of the ratios described above being centered at one, which indicates largely equivalent repression in high and low MOI experiments, there was a left tail, representing a smaller number of tests with reduced estimated efficacy in the high MOI experiments. We reasoned that this could be an artifact of these genes being more lowly expressed and/or being represented by fewer cells given the sparse sampling of the pilot library in the low MOI experiment. In either case we might tend to underestimate the amount of transcript remaining after repression or at the very least the estimates would be substantially noisier, resulting in an artifactual

tail. To confirm the lower expression levels genes in the observed tail, we took all tests falling in the first quartile of each distribution and compared the expression of these genes (average expression for the pilot low MOI experiments in the group of cells without the relevant gRNA; calculated by exponentiating the intercept from the differential expression test, which in the pilot high differential test is the estimated expression in UMI counts for the group of cells without the relevant gRNA). We further scaled these values by the total number of cells observed for each gRNA group in the pilot low MOI experiment to examine the combined effect of representation and expression level, which both contribute to what we expect is simply less robust estimation of fold change. We note that this scaling does not appreciably impact the overall distributions in this case.

Power Simulations: In order to predict the impact of multiplexing on the power of enhancer-gene pair screens, we developed a simulation framework. First, using single-cell RNA-seq data collected from the pilot 47,650 K562 cells, we estimated a dispersion function that relates the mean expression of a gene to its dispersion estimate (one of the two parameters required for the negative binomial distribution) calling the Monocle2 functions estimateSizeFactors and estimateDispersions. This function is typically used in differential expression testing to shrink dispersion estimates, but here we use it to estimate dispersion values for simulated transcripts. This dispersion function is then extracted from the CellDataset object output by Monocle2 and used as input to our simulations.

Next, we chose relevant ranges for each of the parameters varied in our simulation: the MOI, total cell count, effect size (fraction repressed by CRISPRi), and mean expression level of the gene

being tested. By examining the range of expression values observed in our data, we chose to simulate expression data for genes having mean expression values (size parameter of the negative binomial distribution) of 0.01, 0.1, 0.32, 1.0, 3.16, and 10.0 UMIs (0.10, 0.32 and 1.00 used respectively as low, medium, and high in Fig. 3.1B) to provide a range of representative values.

We simulated MOIs at several values from 0.3 to 50, a range which includes the MOIs estimated from our own enhancer-gene pair screens. For each MOI, we calculate the expected number of cells containing a given guide by assuming a Poisson distribution of lentiviral delivery, zero-truncating the distribution to account for drug selection for cells that contain a guide transcript, and rescaling the probability distribution of guide counts accordingly. Perfect library uniformity was assumed to obtain the expected number of cells containing a given guide and the number of cells that do not contain that guide. Effect sizes of CRISPRi repression were chosen using estimates from the literature and were simulated at several values between 10% to 90% percent repression of the average expression level of the target transcript (size parameter input to the negative binomial distribution).

Finally, we simulated several values of total cells included in the experiment ranging from 35,000 to 300,000 cells (45,000 cells shown in Fig. 3.1B). Expression data from transcripts corresponding to 100 samplings per set of parameters were generated for the populations of cells containing the gRNA and not containing the gRNA respectively. Our expression data simulation assumed a negative binomial distribution with the appropriate size parameter for the cells with and without the gRNA, and a dispersion value estimated using the dispersion function described above given the starting mean expression level being simulated. For each set of parameters, the simulated

transcripts were subjected to a differential expression test performed between cells with and without the gRNA assigned using our modified version of the Monocle2 function differentialGeneTest as described above (see Differential Expression Tests). P-values were obtained and corrected assuming an average number of 20 tests per group in the library to approximate the number of genes contained within 1 Mb on either side of each gRNA-group and the impact of multiple testing. The rate of tests falling below a adjusted P-value of 0.05 were tabulated at each set of parameters to make power curves.

Quantify errors in gRNA backbone as described in Method Details: "gRNA-library cloning - special note", Related to Fig. 3.14: To quantify the rate of mismatches and indel lengths in the gRNA backbones for each library, we extracted the backbone portion of the gRNA transcript for each read in our gRNA transcript enrichment libraries and aligned it to the expected reference, (gtttAagagctaTGCTGGAAACAGCAtagcaagttTaaat), using semi-global version of the Needleman-Wunsch algorithm implemented by RecNW (Yahi et al. 2018). Mismatch and indel counts were made within the hairpin portion of the backbone (we initially screened backbone bases 8 to 31 downstream of the spacer), to restrict to bases that would be the most likely to have some if any functional impact. However, it should be noted that the overwhelming majority of all indels were small deletions observed in bases 8 to 14 or so; thus, rates provided in Fig. 3.14A are limited to these 7 bp. For the pilot-gRNA libraries, where we had a shorter cDNA read length that does not cover the entire hairpin, so we simply quantified mismatches and indels in the 8 to 14 bp window (which again contained the overwhelming majority of all indels in our at-scale gRNA library). For each target-UMI pair in each cell, we averaged the observed mismatch and indel counts/lengths to get a consensus over all reads with a given UMI. We then averaged the statistics

derived from UMIs for each target-cell assignment to get a final set of statistics for each. Each average was rounded to the nearest integer for plotting. This allowed us to quantify rates across screens and also examine how any changes in effect sizes correlated with effect sizes.

tSNE clustering of each dataset to check for biological distortions: We tested for enrichment of gRNAs in specific tSNE-based clusters of the at-scale single cell transcriptome dataset, to identify any perturbed targets that resulted in stronger changes to global expression, presumably mediated through trans effects of the target gene. For the at-scale dataset, we subsetted to genes that were expressed in at least 0.5% of cells and 50,000 cells were randomly sampled. We then processed the dataset using Seurat (Butler et al. 2018). We removed cells with greater than 10% mitochondrial transcripts, ran NormalizeData, and found the top 5,000 variable genes using FindVariableGenes. Using these top 5,000 variable genes as input we then ran ScaleData, regressing out the percent of each cell's transcriptome accounted for by mitochondrial genes. We then computed 100 PCs using RunPCA (weighting PCs by variance explained), which were used as input to both the FI-tSNE method using RunTSNE and Louvain clustering at a resolution of 0.5 using FindClusters. Fisher's Exact tests were performed to test for a perturbed target's enrichment in each cluster. 8 TSS controls and 6 candidate enhancers were enriched within specific clusters (odds ratio > 5, adjusted P-value < 0.01). However, even in these cases, only 10% of cells in which the target is perturbed actually fall into the cluster in which they are found to be enriched. Thus, this is not expected to compromise the screen, as in order to be a chronic source of false positives, the gRNAs targeting these global-change genes would have to be non-randomly associated with other gRNAs in the library.

Data and software availability: The accession number for the sequencing data (single cell RNA-seq and bulkRNA-seq) and processed data files is GEO: GSE120861 (metadata file), and GSM3417251 to GSM3417303 (actual datasets).

Chapter 4. THE PATH TO A COMPREHENSIVE, USEFUL CATALOG OF FUNCTIONALLY CHARACTERIZED HUMAN ENHANCERS

4.1 Towards A Comprehensive Catalog of Human Enhancers

The human gene catalog is essentially complete, but we lack an equivalently vetted inventory of bona fide human enhancers. Over one million candidate enhancers have been nominated by biochemical annotations; however, only a handful of these are validated and paired to their target gene(s). How might we generate a comprehensive catalog of human enhancers-gene pairs?

Each of the CRISPR and non-CRISPR technologies available today will be critical to this goal, but each requires further improvement. A shift towards single-cell biochemical assays will allow comprehensive profiling of cell-type specific enhancers, and development of sc-ChIP or or sc-HiChIP could greatly advance high resolution enhancer annotation. MPRAs require laborious delivery to diverse cell lines to test enhancers across representative human cell types, which is potentially addressable by *in vivo* single-cell MPRAs within a model organism. CRISPR screens show unprecedented promise in high-throughput endogenous enhancer-gene pair testing, but like MPRAs, they require laborious application to many human cell lines or tissues within a whole organism. Screens also require further testing across perturbation modes to characterize their sensitivity and specificity. In total, the agreement rates across biochemical annotation, MPRA activity, and CRISPR screens is still unknown but remains a critical comparison to track.

I argue an enhancer's minimal validation must entail deletion in the cell line or tissue of relevance, with corresponding observation of allele-specific reduction in a target gene's transcript. Yet since even this standard is low throughput, a solution is to present tiers of enhancer classification based on different levels of supporting data. Multiple sets of classification requirements will encompass enhancers that act outside of a "one size fits all" definition.

We propose criteria for cataloguing candidate enhancers as weakly, moderately, or strongly supported for enhancer activity (Fig. 4.1). If not validated *in vivo* or by clean deletion, "strongly supported" enhancers will be supported by agreement of all available data in a specific cell type: the expected set of biochemical annotations, an MPRA based demonstration of activity, and a genomic perturbation resulting in change in expression of target gene. Moderately supported enhancers will have subsets of this evidence, and weakly supported will have single pieces of evidence. Overall, in order for such a catalog to be useful, the majority of candidates must not only be classified as enhancers, but also confidently linked to 1+ target gene(s).

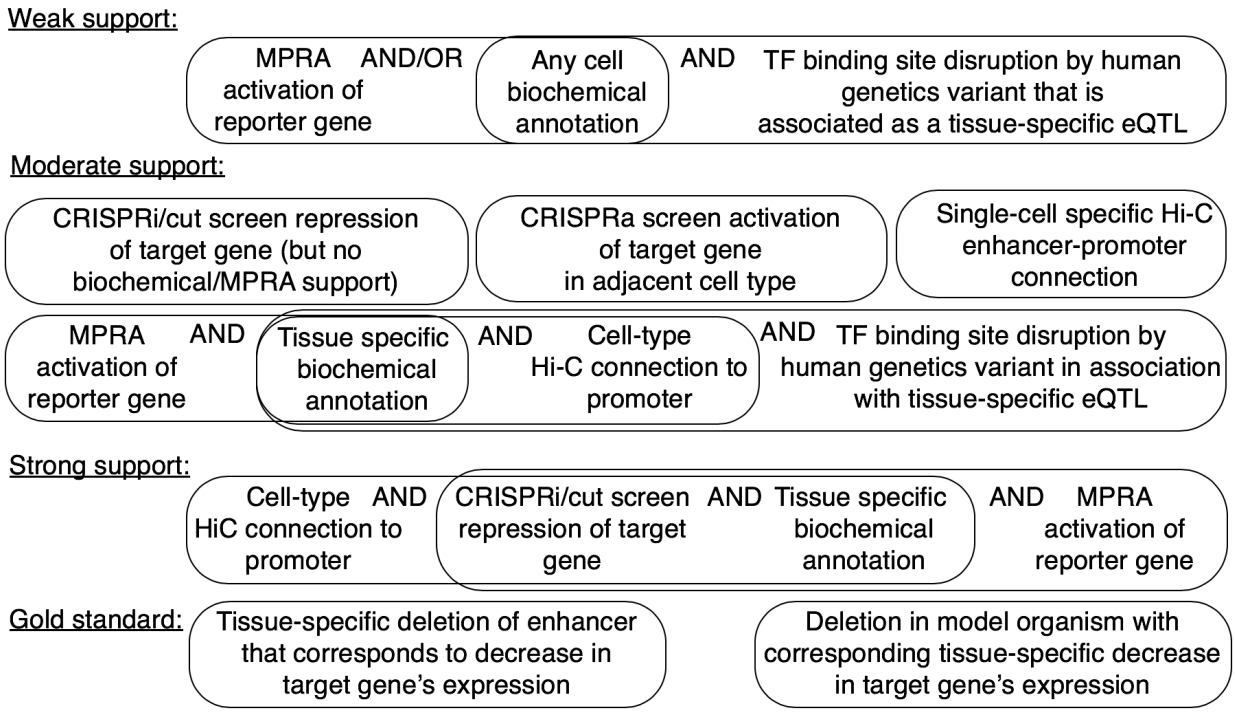


Figure 4.1. Proposed classification criteria for enhancer activity.

Long-term, it may not be realistic to test every candidate enhancer by each functional approach in every cell type of interest. Current approaches to *predict* enhancer-gene pairs vary from simple assignment to the nearest expressed gene to models informed by chromatin state and/or CRISPR screens (Pliner et al. 2018; Zeng, Wu, and Jiang 2018; C. P. Fulco et al. 2019). A perhaps optimistic hope is that from further functional cataloguing we learn the complicated rules for which enhancers are real and which genes they target (e.g. identifying sets of 1D/3D biochemical marks and associated strengths that support a specific degree of confidence that a particular element would validate if tested, and moreover what its target(s) are). As previously discussed in this manuscript, gene regulatory circuits likely act not by a single rulebook but by a diverse set of highly specific, complicated, and overlapping guidelines. This next generation of functional assays has the potential to explore this diversity and inform us of these rules.

4.2 From Coarse to Fine-grained Understanding of the Effects of Human Noncoding Variants

A catalogue of all human enhancers and their target gene(s) is still an insufficient goal for the next generation of functional genomics. How do we facilitate the discovery and characterization of disease-relevant noncoding variants that fall within these enhancers? Though evidence points to an important role in pathogenesis (Maurano et al. 2012; Zhang et al. 2018), the extent to which noncoding variants contribute to disease is still unknown. Understanding this is a major priority of modern human genomics and gene regulation.

The path towards functionally characterizing noncoding variants will differ in rare versus common disease. Rare disease studies can operate under a relatively simple model that disruption of a *cis* enhancer causes misregulation of a single Mendelian disease gene. This presents a clearer path to characterizing variants (e.g. disease gene is often known; mutations are rare or *de novo*). Contrastingly, common disease studies have a murkier task. Multiple hypotheses could explain how noncoding variants (each with small effect sizes or interactive effect sizes, acting alone or in combination) contribute to overall common disease predisposition. Disease genes are often unknown. Linkage disequilibrium limits the resolution with which large human population association studies can map common variants.

4.2.1 *Characterizing Noncoding Variants in Common Disease*

Though the molecular mechanism for pathogenesis is relatively simple, the list of previously validated distal noncoding rare disease mutations is relatively short. One strategy to identify

more variants would be to first perform a CRISPR monogenic enhancer screen for all rare disease genes (Gasperini et al. 2017) followed by saturation mutagenesis across all gene-paired enhancers in all disease-relevant cell types. MPRAs, saturation genome editing (Findlay et al. 2014) and base editing (Hess et al. 2017) are approaching the throughput required. Yet, even if Mendelian disease genes are prioritized by potential for enhancer-control based on the number of cases unsolved by clinical exome or genic sequencing (Chong et al. 2015), the scale needed to provide functional data for every potential clinically identified noncoding variant is still likely prohibitive. Many computational models exist to predict a noncoding variant's pathogenicity (Rojano et al. 2018; Kircher and Shendure 2015) but altogether have inconsistent performance. Further mutagenesis and enhancer-gene pair datasets could improve their accuracy.

4.2.2 *Characterizing Noncoding Variants in Rare Disease*

Disentangling the role of variants in common disease is a complicated challenge. Thousands of noncoding variants have been identified in haplotypes associated with a common disease phenotype (Maurano et al. 2012; Gusev et al. 2014). Our understanding of these haplotypes is minimal. How do we increase the number of true noncoding variants that have been validated as drivers of common disease? Are these common noncoding variants contributing to disease by simple *cis* enhancer disruption or a different mechanism?

Functional rather than statistical studies may reveal the answer. Gallagher and Chen-Plotkin report that as of 2016 there had been 84 functional studies to functionally dissect GWAS hits, usually by reporter assays or genome/epigenome editing (Gallagher and Chen-Plotkin 2018). However, these studies usually operate under the simple model as if one disease-associated haplotype contains one

causal mutation. A next step is to use these studies to consider multiple contributing variants either from within a haplotype (e.g. “multiple enhancer variant” hypothesis [Corradin et al. 2014]) or across the genome (e.g. “omnigenic” model [E. A. Boyle, Li, and Pritchard 2017]). Possible future experiments could include: further layering of MPRA phenotyped variants (Tewhey et al. 2018; Ulirsch et al. 2016) upon CRISPR screened enhancer-gene pairs (one-haplotype, one causal mutation model); use genome engineering to generate all possible combinations of the variants in a disease-associated haplotype, and phenotype the synthetic alleles upon an associated target gene (multiple enhancer variant hypothesis [Corradin et al. 2014]); or synthetically inducing small effect disease-associated perturbations at genome-wide loci to identify a core set of affected genes (omnigenic model [E. A. Boyle, Li, and Pritchard 2017]).

It is likely that all of these hypotheses underlie the role of common noncoding variants in pathogenesis. As the field’s current level of understanding is still nascent, any reasonably successful level of scalable empirical measurement will be helpful in advancing efforts to map causal variants underlying common diseases and link them to the gene(s) through which they mediate their effects.

4.3 Outstanding Questions in Enhancer Biology

Efforts to catalog enhancer-gene pairs and the impact of noncoding variants are critical for furthering our understanding of human gene regulation. Such datasets could address many of the outstanding questions of enhancer biology. I detail four prominent mysteries in the following text.

Foremost, how does an enhancer pick its target gene? In the enhancer-gene pair screen in chapter 3, we observed that enhancers target their closest expressed gene only two thirds of the time. What

is the mechanism of choice in the remaining third of enhancer-gene pairs? What biochemistry and molecular biology determines this specificity? Though many hypothesized models exist, no causal mechanisms have emerged as genome-wide consensus. This concept remains mysterious enough to the point that it is reasonable to ask if promoters "pick" their enhancer(s), instead of the other way around. The sparsity of available enhancer-gene pairs has prevented the differentiation between any described mechanism as a genome-wide phenomenon or a circuit unique to an individual locus. A large-scale enhancer-gene pair catalog would further our understanding of the ubiquity or diversity of the rules of enhancer-gene choice.

The role of 3D chromatin structure in gene regulation is another highly contested mystery. The recent advances of chromatin conformation mapping have pointed to a key role for chromatin structure. However, though ablation of key chromatin structural proteins (CTCF, cohesin) disrupts looping, it induces effects of variable strength upon the transcriptome (Rao et al. 2017, Nora et al. 2017, Kubo et al. 2017). Though proximity correlates with interacting enhancer-promoter pairs, is this a causal mechanism of enhancer activity or a residual feature? Does this vary by cell type? And how do enhancers influence 3D chromatin structure and vice versa?

Third, how do individual enhancers coordinate together within a gene regulation circuit? Early studies in developing mouse and drosophila point at redundancy as a widespread phenomenon (Cannavò et al. 2016, Osterwalder et al. 2018). How often do enhancers act redundantly around the genome? Does the prevalence of this differ between developing or terminally differentiated cells? In such cases, what is the compensatory mechanism for redundant enhancers when one

enhancer is ablated? What differs between enhancers that act redundantly or independently? The increased throughput of perturbational screens promises to address this.

Last, the biochemistry of an enhancer's activity upon the promoter still remains opaque. How exactly does an enhancer act upon a promoter to enhance gene expression? Enhancers are widely thought to simply increase the concentration of transcriptional co-factors at the promoter, but what are the exact biochemical rules? A litany of hypotheses are available (reviewed in Furlong and Levine, 2018) and much extensive biochemistry, microscopy, and molecular biology will be required to sift through them all.

The paths to solving these outstanding mysteries are overlapping, and insight into one will reveal mechanisms of another. An enhancer-gene pair catalog will aid achievement of these goals, as further perturbational data and functionally supported enhancer-gene pairs allows for exploration of each hypothesis at a genome-wide scale.

4.4 Closing Remarks

These current advancements in enhancer annotation and screening have already produced a new generation of enhancer datasets. Yet, these nascent results currently disagree. Why have so many sites been biochemically annotated as candidate enhancers (~80% of the genome [ENCODE Project Consortium 2012]), but functional screens (both MPRA and CRISPR based) are returning such low hit rates (~10% of tested loci)? Is this the methods or is this the reality? Further studies - especially in the fledgling single-cell and CRISPR screen fields - are required to identify consensus or sources of error across methods.

Additionally, why are there still so few validated noncoding variants for common disease? Have we simply not put in enough effort in characterizing these variants? Is there an underlying mechanism of pathogenesis that the field has not yet discovered? Future work is required to attempt functional variant validation at scale and dissect the many potential underlying mechanisms.

However, based on the emerging technologies reviewed here, we present two major goals as still worthwhile and possibly within reach: First, a functionally supported comprehensive map of human enhancers, with knowledge of each element's target gene(s) and ideally cell type specificity. Second, a prediction of the cell type-specific consequences of an arbitrary SNV in any of these elements on *cis* gene expression. Progress towards these goals throughout the next era of human genomics will significantly advance our understanding of the human genome and its role in disease.

The human genome functions uniquely across the thousands of cell types it generates. The specific activity profiles of our 21,000 genes and millions of candidate enhancers creates a combinatorial functional space almost as vast as our galaxy. In this dissertation, I have presented my humble effort towards understanding a tiny drop of this miraculous molecular universe.

BIBLIOGRAPHY

Adamson, Britt, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, et al. 2016. "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response." Cell 167 (7): 1867–82.e21.

Akhtar, Waseem, Johann de Jong, Alexey V. Pindyurin, Ludo Pagie, Wouter Meuleman, Jeroen de Ridder, Anton Berns, Lodewyk F. A. Wessels, Maarten van Lohuizen, and Bas van Steensel. 2013. "Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel." Cell 154 (4): 914–27.

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, et al. 2014. "An Atlas of Active Enhancers across Human Cell Types and Tissues." Nature 507 (7493): 455–61.

Aparicio-Prat, Estel, Carme Arnan, Ilaria Sala, Núria Bosch, Roderic Guigó, and Rory Johnson. 2015. "DECKO: Single-Oligo, Dual-CRISPR Deletion of Genomic Elements Including Long Non-Coding RNAs." BMC Genomics 16 (October): 846.

Axel, R., H. Cedar, and G. Felsenfeld. 1973. "Synthesis of Globin Ribonucleic Acid from Duck-Reticulocyte Chromatin in Vitro." Proceedings of the National Academy of Sciences of the United States of America 70 (7): 2029–32.

Banerji, J., L. Olson, and W. Schaffner. 1983. "A Lymphocyte-Specific Cellular Enhancer Is Located Downstream of the Joining Region in Immunoglobulin Heavy Chain Genes." Cell 33 (3): 729–40.

Banerji, J., S. Rusconi, and W. Schaffner. 1981. "Expression of a Beta-Globin Gene Is Enhanced by Remote SV40 DNA Sequences." Cell 27 (2 Pt 1): 299–308.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. 2007. "High-Resolution Profiling of Histone Methylations in the Human Genome." Cell 129 (4): 823–37.

Blow, Matthew J., David J. McCulley, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2010. "ChIP-Seq Identification of Weakly Conserved Heart Enhancers." Nature Genetics 42 (9): 806–10.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." Bioinformatics 30 (15): 2114–20.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. 2008. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome." Cell 132 (2): 311–22.

Boyle, Alan P., Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, et al. 2012. "Annotation of Functional Variation in Personal Genomes Using RegulomeDB." Genome Research 22 (9): 1790–97.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. "An Expanded View of Complex Traits: From Polygenic to Omnigenic." Cell 169 (7): 1177–86.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Erratum: Near-Optimal Probabilistic RNA-Seq Quantification." Nature Biotechnology 34 (8): 888.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” Nature Methods 10 (12): 1213–18.

Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015. “Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation.” Nature 523 (7561): 486–90.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. “Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species.” Nature Biotechnology 36 (5): 411–20.

Byrne, Susan M., Luis Ortiz, Prashant Mali, John Aach, and George M. Church. 2015. “Multi-Kilobase Homozygous Targeted Gene Replacement in Human Induced Pluripotent Stem Cells.” Nucleic Acids Research 43 (3): e21.

Cannavò, Enrico, Pierre Khoueiry, David A. Garfield, Paul Geeleher, Thomas Zichner, E. Hilary Gustafson, Lucia Ciglar, Jan O. Korbel, and Eileen E.M. Furlong. 2016. “Shadow Enhancers are Pervasive Features of Developmental Regulatory Networks.” Curr Biol 26(1): 38-51.

Canver, Matthew C., Daniel E. Bauer, Abhishek Dass, Yvette Y. Yien, Jacky Chung, Takeshi Masuda, Takahiro Maeda, Barry H. Paw, and Stuart H. Orkin. 2014. “Characterization of Genomic Deletion Efficiency Mediated by Clustered Regularly Interspaced Palindromic Repeats (CRISPR)/Cas9 Nuclease System in Mammalian Cells.” Journal of Biological Chemistry. https://doi.org/10.1074/jbc.m114.564625.

Canver, Matthew C., Elenoe C. Smith, Falak Sher, Luca Pinello, Neville E. Sanjana, Ophir Shalem, Diane D. Chen, et al. 2015. “BCL11A Enhancer Dissection by Cas9-Mediated in Situ Saturating Mutagenesis.” Nature 527 (7577): 192–97.

Chen, Baohui, Luke A. Gilbert, Beth A. Cimini, Joerg Schnitzbauer, Wei Zhang, Gene-Wei Li, Jason Park, et al. 2013. “Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System.” Cell 155 (7): 1479–91.

Chen, Sidi, Neville E. Sanjana, Kaijie Zheng, Ophir Shalem, Kyungheon Lee, Xi Shi, David A. Scott, et al. 2015. “Genome-Wide CRISPR Screen in a Mouse Model of Tumor Growth and Metastasis.” Cell. https://doi.org/10.1016/j.cell.2015.02.038.

Chen, W., A. McKenna, J. Schreiber, Y. Yin, and V. Agarwal. 2018. “Massively Parallel Profiling and Predictive Modeling of the Outcomes of CRISPR/Cas9-Mediated Double-Strand Break Repair.” bioRxiv. https://www.biorxiv.org/content/10.1101/481069v1.abstract.

Choi, Harry M. T., Maayan Schwarzkopf, Mark E. Fornace, Aneesh Acharya, Georgios Artavanis, Johannes Stegmaier, Alexandre Cunha, and Niles A. Pierce. 2018. “Third-Generation in Situ Hybridization Chain Reaction: Multiplexed, Quantitative, Sensitive, Versatile, Robust.” Development 145 (12). https://doi.org/10.1242/dev.165753.

Chong, Jessica X., Kati J. Buckingham, Shalini N. Jhangiani, Corinne Boehm, Nara Sobreira, Joshua D. Smith, Tanya M. Harrell, et al. 2015. “The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities.” American Journal of Human Genetics 97 (2): 199–215.

Claussnitzer, Melina, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, et al. 2015. “FTO Obesity Variant Circuitry and

Adipocyte Browning in Humans." New England Journal of Medicine. https://doi.org/10.1056/nejmoa1502214.

Coetzee, Simon G., Suhn K. Rhie, Benjamin P. Berman, Gerhard A. Coetzee, and Houtan Noushmehr. 2012. "FunciSNP: An R/bioconductor Tool Integrating Functional Non-Coding Data Sets with Genetic Association Studies to Identify Candidate Regulatory SNPs." Nucleic Acids Research 40 (18): e139.

Cong, Le, F. Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D. Hsu, et al. 2013. "Multiplex Genome Engineering Using CRISPR/Cas Systems." Science 339 (6121): 819–23.

Consortium, Gtex, and GTEx Consortium. 2017. "Genetic Effects on Gene Expression across Human Tissues." Nature. https://doi.org/10.1038/nature24277.

Coppola, Candice J., Ryne C Ramaker, and Eric M. Mendenhall. 2016. "Identification and Function of Enhancers in the Human Genome." Human Molecular Genetics 25 (R2): R190–97.

Corradin, Olivia, Alina Saiakhova, Batool Akhtar-Zaidi, Lois Myeroff, Joseph Willis, Richard Cowper-Sal lari, Mathieu Lupien, Sanford Markowitz, and Peter C. Scacheri. 2014. "Combinatorial Effects of Multiple Enhancer Variants in Linkage Disequilibrium Dictate Levels of Gene Expression to Confer Susceptibility to Common Traits." Genome Research 24 (1): 1–13.

Cusanovich, Darren A., Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. 2015. "Multiplex Single Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing." Science 348 (6237): 910–14.

Cusanovich, Darren A., Andrew J. Hill, Delasa Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, et al. 2018. "A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility." Cell 174 (5): 1309–24.e18.

Cusanovich, Darren A., James P. Reddington, David A. Garfield, Riza M. Daza, Delasa Aghamirzaie, Raquel Marco-Ferreres, Hannah A. Pliner, et al. 2018. "The Cis-Regulatory Dynamics of Embryonic Development at Single-Cell Resolution." Nature 555 (7697): 538–42.

Datlinger, Paul, André F. Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C. Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. 2017. "Pooled CRISPR Screening with Single-Cell Transcriptome Readout." Nature Methods 14 (3): 297–301.

De Santa, Francesca, Iros Barozzi, Flore Mietton, Serena Ghisletti, Sara Polletti, Betsabeh Khoramian Tusi, Heiko Muller, Jiannis Ragoussis, Chia-Lin Wei, and Gioacchino Natoli. 2010. "A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers." PLoS Biology 8 (5): e1000384.

Diao, Yarui, Rongxin Fang, Bin Li, Zhipeng Meng, Juntao Yu, Yunjiang Qiu, Kimberly C. Lin, et al. 2017. "A Tiling-Deletion-Based Genetic Screen for Cis-Regulatory Element Identification in Mammalian Cells." Nature Methods 14 (6): 629–35.

Diao, Yarui, Bin Li, Zhipeng Meng, Inkyung Jung, Ah Young Lee, Jesse Dixon, Lenka Maliskova, Kun-Liang Guan, Yin Shen, and Bing Ren. 2016. "A New Class of Temporarily Phenotypic Enhancers Identified by CRISPR/Cas9-Mediated Genetic Screening." Genome Research 26 (3): 397–405.

Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, et al. 2016. “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.” Cell 167 (7): 1853–66.e17.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” Nature. https://doi.org/10.1038/nature11082.

Doench, John G., Ella Hartenian, Daniel B. Graham, Zuzana Tothova, Mudra Hegde, Ian Smith, Meagan Sullender, Benjamin L. Ebert, Ramnik J. Xavier, and David E. Root. 2014. “Rational Design of Highly Active sgRNAs for CRISPR-Cas9-Mediated Gene Inactivation.” Nature Biotechnology 32 (12): 1262–67.

Döring, Andreas, David Weese, Tobias Rausch, and Knut Reinert. 2008. “SeqAn an Efficient, Generic C++ Library for Sequence Analysis.” BMC Bioinformatics 9 (January): 11.

Eagen, Kyle P. 2018. “Principles of Chromosome Architecture Revealed by Hi-C.” Trends in Biochemical Sciences 43 (6): 469–78.

Eisenberg, Eli, and Erez Y. Levanon. 2013. “Human Housekeeping Genes, Revisited.” Trends in Genetics: TIG 29 (10): 569–74.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” Nature 489 (7414): 57–74.

Ernst, Jason, and Manolis Kellis. 2012. “ChromHMM: Automating Chromatin-State Discovery and Characterization.” Nature Methods 9 (3): 215–16.

Essletzbichler, Patrick, Tomasz Konopka, Federica Santoro, Doris Chen, Bianca V. Gapp, Robert Kralovics, Thijn R. Brummelkamp, Sebastian M. B. Nijman, and Tilmann Bürckstümmer. 2014. “Megabase-Scale Deletion Using CRISPR/Cas9 to Generate a Fully Haploid Human Cell Line.” Genome Research 24 (12): 2059–65.

Fang, Rongxin, Miao Yu, Guoqiang Li, Sora Chee, Tristin Liu, Anthony D. Schmitt, and Bing Ren. 2016. “Mapping of Long-Range Chromatin Interactions by Proximity Ligation-Assisted ChIP-Seq.” Cell Research 26 (12): 1345–48.

Findlay, Gregory M., Evan A. Boyle, Ronald J. Hause, Jason C. Klein, and Jay Shendure. 2014. “Saturation Editing of Genomic Regions by Multiplex Homology-Directed Repair.” Nature 513 (7516): 120–23.

Finucane, Hilary K., Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, et al. 2015. “Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics.” Nature Genetics 47 (11): 1228–35.

Flyamer, Ilya M., Johanna Gassler, Maxim Imakaev, Hugo B. Brandão, Sergey V. Ulianov, Nezar Abdennur, Sergey V. Razin, Leonid A. Mirny, and Kikuë Tachibana-Konwalski. 2017. “Single-Nucleus Hi-C Reveals Unique Chromatin Reorganization at Oocyte-to-Zygote Transition.” Nature 544 (7648): 110–14.

Forsberg, M., and G. Westin. 1991. “Enhancer Activation by a Single Type of Transcription Factor Shows Cell Type Dependence.” The EMBO Journal 10 (9): 2543–51.

Fulco, Charles P., Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R. Grossman, Elizabeth M. Perez, Michael Kane, Brian Cleary, Eric S. Lander, and Jesse M. Engreitz. 2016. “Systematic Mapping of Functional Enhancer-Promoter Connections with CRISPR Interference.” Science 354 (6313): 769–73.

Fulco, C. P., J. Nasser, T. R. Jones, G. Munson, and D. T. Bergman. 2019. “Activity-by-Contact Model of Enhancer Specificity from Thousands of CRISPR Perturbations.” bioRxiv. https://www.biorxiv.org/content/10.1101/529990v1.abstract.

Fullwood, Melissa J., Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L. Orlov, et al. 2009. “An Oestrogen-Receptor-Alpha-Bound Human Chromatin Interactome.” Nature 462 (7269): 58–64.

Fullwood, M. J., and Y. Ruan. 2009. “ChIP-based Methods for the Identification of Long-range Chromatin Interactions.” Journal of Cellular Biochemistry. https://onlinelibrary.wiley.com/doi/abs/10.1002/jcb.22116.

Furlong, Eileen E. M., and Michael Levine. 2018. “Developmental Enhancers and Chromosome Topology.” Science 361 (6409): 1341–45.

Fu, Rong, Irene Ceballos-Picot, Rosa J. Torres, Laura E. Larovere, Yasukazu Yamada, Khue V. Nguyen, Madhuri Hegde, et al. 2014. “Genotype–phenotype Correlations in Neurogenetics: Lesch-Nyhan Disease as a Model Disorder.” Brain: A Journal of Neurology 137 (5): 1282–1303.

Gallagher, Michael D., and Alice S. Chen-Plotkin. 2018. “The Post-GWAS Era: From Association to Function.” American Journal of Human Genetics 102 (5): 717–30.

Gambone, Julia E., Stephanie S. Dusaban, Roxana Loperena, Yuji Nakata, and Susan E. Shetzline. 2011. “The c-Myb Target Gene Neuromedin U Functions as a Novel Cofactor during the Early Stages of Erythropoiesis.” Blood 117 (21): 5733–43.

Ganapathi, Mythily, Pragya Srivastava, Sushanta Kumar Das Sutar, Kaushal Kumar, Dipayan Dasgupta, Gajinder Pal Singh, Vani Brahmachari, and Samir K. Brahmachari. 2005. “Comparative Analysis of Chromatin Landscape in Regulatory Regions of Human Housekeeping and Tissue Specific Genes.” BMC Bioinformatics 6 (May): 126.

Gasiunas, Giedrius, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. 2012. “Cas9-crRNA Ribonucleoprotein Complex Mediates Specific DNA Cleavage for Adaptive Immunity in Bacteria.” Proceedings of the National Academy of Sciences of the United States of America 109 (39): E2579–86.

Gasperini, Molly, Gregory M. Findlay, Aaron McKenna, Jennifer H. Milbank, Choli Lee, Melissa D. Zhang, Darren A. Cusanovich, and Jay Shendure. 2017. “CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions.” American Journal of Human Genetics 101 (2): 192–205.

Gasperini, Molly, Andrew J. Hill, José L. McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D. Zhang, Dana Jackson, et al. 2019. “A Genome-Wide Framework for Mapping Gene Regulation via Cellular Genetic Screens.” Cell, January. https://doi.org/10.1016/j.cell.2018.11.029.

Gasperini, Molly, Lea Starita, and Jay Shendure. 2016. “The Power of Multiplexed Functional Analysis of Genetic Variants.” Nature Protocols 11 (10): 1782–87.

Ghavi-Helm, Yad, Felix A. Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen E. M. Furlong. 2014. “Enhancer Loops Appear Stable during Development and Are Associated with Paused Polymerase.” Nature 512 (7512): 96–100.

Gillies, S. D., S. L. Morrison, V. T. Oi, and S. Tonegawa. 1983. “A Tissue-Specific Transcription Enhancer Element Is Located in the Major Intron of a Rearranged Immunoglobulin Heavy Chain Gene.” Cell 33 (3): 717–28.

Gross, D. S., and W. T. Garrard. 1988. “Nuclease Hypersensitive Sites in Chromatin.” Annual Review of Biochemistry 57: 159–97.

Grossman, Sharon R., Xiaolan Zhang, Li Wang, Jesse Engreitz, Alexandre Melnikov, Peter Rogov, Ryan Tewhey, et al. 2017. “Systematic Dissection of Genomic Features

Determining Transcription Factor Binding and Enhancer Function." Proceedings of the National Academy of Sciences of the United States of America 114 (7): E1291–1300.

Gu, Bo, Tomek Swigut, Andrew Spencley, Matthew R. Bauer, Mingyu Chung, Tobias Meyer, and Joanna Wysocka. 2018. "Transcription-Coupled Changes in Nuclear Mobility of Mammalian Cis-Regulatory Elements." Science 359 (6379): 1050–55.

Gusev, Alexander, S. Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J. Vilhjálmsson, Han Xu, Chongzhi Zang, et al. 2014. "Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases." American Journal of Human Genetics 95 (5): 535–52.

Hainer, Sarah J., Ana Boskovic, Oliver J. Rando, and Thomas G. Fazzio. n.d. "Profiling of Pluripotency Factors in Individual Stem Cells and Early Embryos." https://doi.org/10.1101/286351.

Hanahan, D. 1985. "Heritable Formation of Pancreatic Beta-Cell Tumours in Transgenic Mice Expressing Recombinant Insulin/simian Virus 40 Oncogenes." Nature 315 (6015): 115–22.

Hebbes, T. R., A. L. Clayton, A. W. Thorne, and C. Crane-Robinson. 1994. "Core Histone Hyperacetylation Co-Maps with Generalized DNase I Sensitivity in the Chicken Beta-Globin Chromosomal Domain." The EMBO Journal. https://doi.org/10.1002/j.1460-2075.1994.tb06451.x.

Hebbes, T. R., A. W. Thorne, and C. Crane-Robinson. 1988. "A Direct Link between Core Histone Acetylation and Transcriptionally Active Chromatin." The EMBO Journal 7 (5): 1395–1402.

Heintzman, Nathaniel D., Rhona K. Stuart, Gary Hon, Yutao Fu, Christina W. Ching, R. David Hawkins, Leah O. Barrera, et al. 2007. "Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome." Nature Genetics 39 (3): 311–18.

Hesselberth, Jay R., Xiaoyu Chen, Zhihong Zhang, Peter J. Sabo, Richard Sandstrom, Alex P. Reynolds, Robert E. Thurman, et al. 2009. "Global Mapping of Protein-DNA Interactions in Vivo by Digital Genomic Footprinting." Nature Methods 6 (4): 283–89.

Hess, Gaelen T., Josh Tycko, David Yao, and Michael C. Bassik. 2017. "Methods and Applications of CRISPR-Mediated Base Editing in Eukaryotic Genomes." Molecular Cell 68 (1): 26–43.

Hill, Andrew J., José L. McFaline-Figueroa, Lea M. Starita, Molly J. Gasperini, Kenneth A. Matreyek, Jonathan Packer, Dana Jackson, Jay Shendure, and Cole Trapnell. 2018. "On the Design of CRISPR-Based Single-Cell Molecular Screens." Nature Methods 15 (4): 271–74.

Hnisz, Denes, Brian J. Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A. Sigova, Heather A. Hoke, and Richard A. Young. 2013. "Super-Enhancers in the Control of Cell Identity and Disease." Cell 155 (4): 934–47.

Hoffman, Michael M., Orion J. Buske, Jie Wang, Zhiping Weng, Jeff A. Bilmes, and William Stafford Noble. 2012. "Unsupervised Pattern Discovery in Human Chromatin Structure through Genomic Segmentation." Nature Methods 9 (5): 473–76.

Hong, Joung-Woo, David A. Hendrix, and Michael S. Levine. 2008. "Shadow Enhancers as a Source of Evolutionary Novelty." Science 321 (5894): 1314.

Horlbeck, Max A., Luke A. Gilbert, Jacqueline E. Villalta, Britt Adamson, Ryan A. Pak, Yuwen Chen, Alexander P. Fields, et al. 2016. "Compact and Highly Active next-Generation

Libraries for CRISPR-Mediated Gene Repression and Activation." eLife 5 (September). https://doi.org/10.7554/eLife.19760.

Hsu, Patrick D., David A. Scott, Joshua A. Weinstein, F. Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, et al. 2013. "DNA Targeting Specificity of RNA-Guided Cas9 Nucleases." Nature Biotechnology 31 (9): 827–32.

Huang, Yung-Hsin, Jianzhong Su, Yong Lei, Lorenzo Brunetti, Michael C. Gundry, Xiaotian Zhang, Mira Jeong, Wei Li, and Margaret A. Goodell. 2017. "DNA Epigenome Editing Using CRISPR-Cas SunTag-Directed DNMT3A." Genome Biology. https://doi.org/10.1186/s13059-017-1306-z.

Ikuta, T., and Y. W. Kan. 1991. "In Vivo Protein-DNA Interactions at the Beta-Globin Gene Locus." Proceedings of the National Academy of Sciences of the United States of America 88 (22): 10188–92.

Inoue, Fumitaka, and Nadav Ahituv. 2015. "Decoding Enhancers Using Massively Parallel Reporter Assays." Genomics 106 (3): 159–64.

Inoue, Fumitaka, Martin Kircher, Beth Martin, Gregory M. Cooper, Daniela M. Witten, Michael T. McManus, Nadav Ahituv, and Jay Shendure. 2017. "A Systematic Comparison Reveals Substantial Differences in Chromosomal versus Episomal Encoding of Enhancer Activity." Genome Research 27 (1): 38–52.

Jacob, F., and J. Monod. 1961. "Genetic Regulatory Mechanisms in the Synthesis of Proteins." Journal of Molecular Biology 3 (June): 318–56.

Jacobs, Lois, and Robert DeMars. 1984. "Chemical Mutagenesis with Diploid Human Fibroblasts." In Handbook of Mutagenicity Test Procedures, 321–56. Elsevier.

Jain, Dhawal, Sandro Baldi, Angelika Zabel, Tobias Straub, and Peter B. Becker. 2015. "Active Promoters Give Rise to False Positive 'Phantom Peaks' in ChIP-Seq Experiments." Nucleic Acids Research. https://doi.org/10.1093/nar/gkv637.

Jaitin, Diego Adhemar, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van Oudenaarden, and Ido Amit. 2016. "Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq." Cell 167 (7): 1883–96.e15.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity." Science 337 (6096): 816–21.

John, Sam, Peter J. Sabo, Robert E. Thurman, Myong-Hee Sung, Simon C. Biddie, Thomas A. Johnson, Gordon L. Hager, and John A. Stamatoyannopoulos. 2011. "Chromatin Accessibility Pre-Determines Glucocorticoid Receptor Binding Patterns." Nature Genetics 43 (3): 264–68.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." Science 316 (5830): 1497–1502.

Kearns, Nicola A., Hannah Pham, Barbara Tabak, Ryan M. Genga, Noah J. Silverstein, Manuel Garber, and René Maehr. 2015. "Functional Annotation of Native Enhancers with a Cas9-Histone Demethylase Fusion." Nature Methods 12 (5): 401–3.

Kim, Tae-Kyung, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, et al. 2010. "Widespread Transcription at Neuronal Activity-Regulated Enhancers." Nature 465 (7295): 182–87.

Kircher, Martin, and Jay Shendure. 2015. "Running Spell-Check to Identify Regulatory Variants." Nature Genetics 47 (8): 853–55.

Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." Nature Genetics 46 (3): 310–15.

Klann, Tyler S., Joshua B. Black, Malathi Chellappan, Alexias Safi, Lingyun Song, Isaac B. Hilton, Gregory E. Crawford, Timothy E. Reddy, and Charles A. Gersbach. 2017. "CRISPR-Cas9 Epigenome Editing Enables High-Throughput Screening for Functional Regulatory Elements in the Human Genome." Nature Biotechnology 35 (6): 561–68.

Klein, Jason C., Wei Chen, Molly Gasperini, and Jay Shendure. 2018. "Identifying Novel Enhancer Elements with CRISPR-Based Screens." ACS Chemical Biology 13 (2): 326–32.

Klein, Jason C., Aidan Keith, Vikram Agarwal, Timothy Durham, and Jay Shendure. 2018. "Functional Characterization of Enhancer Evolution in the Primate Lineage." Genome Biology 19 (1): 99.

Korkmaz, Gozde, Rui Lopes, Alejandro P. Ugalde, Ekaterina Nevedomskaya, Ruiqi Han, Ksenia Myacheva, Wilbert Zwart, Ran Elkon, and Reuven Agami. 2016. "Functional Genetic Screens for Enhancer Elements in the Human Genome Using CRISPR-Cas9." Nature Biotechnology 34 (2): 192–98.

Kosicki, Michael, Kärt Tomberg, and Allan Bradley. 2018. "Repair of Double-Strand Breaks Induced by CRISPR-Cas9 Leads to Large Deletions and Complex Rearrangements." Nature Biotechnology, July. https://doi.org/10.1038/nbt.4192.

Kubo, Naoki, Haruhiko Ishii, David Gorkin, Franz Meitinger, Xiong Xiong, Rongxin Fang, Tristin Liu, et al. 2017. "Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells." bioRxiv. https://doi.org/10.1101/118737.

Kulakovskiy, Ivan V., Yulia A. Medvedeva, Ulf Schaefer, Artem S. Kasianov, Ilya E. Vorontsov, Vladimir B. Bajic, and Vsevolod J. Makeev. 2013. "HOCOMOCO: A Comprehensive Collection of Human Transcription Factor Binding Sites Models." Nucleic Acids Research 41 (Database issue): D195–202.

Kumasaka, Natsuhiko, Andrew J. Knights, and Daniel J. Gaffney. 2016. "Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq." Nature Genetics 48 (2): 206–13.

Kwon, Deborah Y., Ying-Tao Zhao, Janine M. Lamonica, and Zhaolan Zhou. 2017. "Locus-Specific Histone Deacetylation Using a Synthetic CRISPR-Cas9-Based HDAC." Nature Communications 8 (May): 15315.

Laat, Wouter de, and Denis Duboule. 2013. "Topology of Mammalian Developmental Enhancers and Their Regulatory Landscapes." Nature 502 (7472): 499–506.

Lai, Binbin, Weiwu Gao, Kairong Cui, Wanli Xie, Qingsong Tang, Wenfei Jin, Gangqing Hu, Bing Ni, and Keji Zhao. 2018. "Publisher Correction: Principles of Nucleosome Organization Revealed by Single-Cell Micrococcal Nuclease Sequencing." Nature 564 (7735): E17.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. "The Human Transcription Factors." Cell. https://doi.org/10.1016/j.cell.2018.01.029.

Lancaster, Madeline A., and Juergen A. Knoblich. 2014. "Organogenesis in a Dish: Modeling Development and Disease Using Organoid Technologies." Science 345 (6194): 1247125.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." Nature Methods 9 (4): 357–59.

Lee, D. S., C. Luo, J. Zhou, S. Chandran, and A. Rivkin. 2018. “Single-Cell Multi-Omic Profiling of Chromatin Conformation and DNA Methylome.” bioRxiv. https://www.biorxiv.org/content/10.1101/503235v1.abstract.

Lee, D. Y., J. J. Hayes, D. Pruss, and A. P. Wolffe. 1993. “A Positive Role for Histone Acetylation in Transcription Factor Access to Nucleosomal DNA.” Cell 72 (1): 73–84.

Lei, Yong, Xiaotian Zhang, Jianzhong Su, Mira Jeong, Michael C. Gundry, Yung-Hsin Huang, Yubin Zhou, Wei Li, and Margaret A. Goodell. 2017. “Targeted DNA Methylation in Vivo Using an Engineered dCas9-MQ1 Fusion Protein.” Nature Communications. https://doi.org/10.1038/ncomms16026.

Lesch, Michael, and William L. Nyhan. 1964. “A Familial Disorder of Uric Acid Metabolism and Central Nervous System Function.” The American Journal of Medicine. https://doi.org/10.1016/0002-9343(64)90104-4.

Levings, Padraic P., and Jörg Bungert. 2002. “The Human Beta-Globin Locus Control Region.” European Journal of Biochemistry / FEBS 269 (6): 1589–99.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” Science 326 (5950): 289–93.

Li, Mulin Jun, Lily Yan Wang, Zhengyuan Xia, Pak Chung Sham, and Junwen Wang. 2013. “GWAS3D: Detecting Human Regulatory Variants by Integrative Analysis of Genome-Wide Associations, Chromosome Interactions and Histone Modifications.” Nucleic Acids Research 41 (Web Server issue): W150–58.

Lindblad-Toh, Kerstin, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, et al. 2011. “A High-Resolution Map of Human Evolutionary Constraint Using 29 Mammals.” Nature 478 (7370): 476–82.

Lin, Yingxin, Shila Ghazanfar, Dario Strbenac, Andy Wang, Ellis Patrick, Terence Speed, Jean Yang, and Pengyi Yang. 2017. “Housekeeping Genes, Revisited at the Single-Cell Level.” BioRxiv, December, 229815.

Liu, X. Shawn, Hao Wu, Xiong Ji, Yonatan Stelzer, Xuebing Wu, Szymon Czauderna, Jian Shu, Daniel Dadon, Richard A. Young, and Rudolf Jaenisch. 2016. “Editing DNA Methylation in the Mammalian Genome.” Cell 167 (1): 233–47.e17.

Liu, Yuwen, Shan Yu, Vineet K. Dhiman, Tonya Brunetti, Heather Eckart, and Kevin P. White. 2017. “Functional Assessment of Human Enhancer Activities Using Whole-Genome STARR-Sequencing.” Genome Biology 18 (1): 219.

Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, et al. 2015. “Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions.” Cell 161 (5): 1012–25.

Lupo, Angelo, Elena Cesaro, Giorgia Montano, Diana Zurlo, Paola Izzo, and Paola Costanzo. 2013. “KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions.” Curr Genomics 14(4):268-278.

MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, et al. 2017. “The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog).” Nucleic Acids Research 45 (D1): D896–901.

Mali, Prashant, Luhan Yang, Kevin M. Esvelt, John Aach, Marc Guell, James E. DiCarlo, Julie E. Norville, and George M. Church. 2013. “RNA-Guided Human Genome Engineering via Cas9.” Science 339 (6121): 823–26.

Maricque, Brett B., Hemangi G. Chaudhari, and Barak A. Cohen. 2018. “A Massively Parallel Reporter Assay Dissects the Influence of Chromatin Structure on Cis-Regulatory Activity.” Nature Biotechnology, November. https://doi.org/10.1038/nbt.4285.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, et al. 2012. “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.” Science. https://doi.org/10.1126/science.1222794.

Ma, Wenxiu, Ferhat Ay, Choli Lee, Gunhan Gulsoy, Xinxian Deng, Savannah Cook, Jennifer Hesson, et al. 2015. “Fine-Scale Chromatin Interaction Maps Reveal the Cis-Regulatory Landscape of Human lincRNA Genes.” Nature Methods 12 (1): 71–78.

McKenna, Aaron, and Jay Shendure. 2018. “FlashFry: A Fast and Flexible Tool for Large-Scale CRISPR Target Design.” BMC Biology 16 (1): 74.

McLeay, Robert C., and Timothy L. Bailey. 2010. “Motif Enrichment Analysis: A Unified Framework and an Evaluation on ChIP Data.” BMC Bioinformatics 11 (April): 165.

Melnikov, Alexandre, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, et al. 2012. “Systematic Dissection and Optimization of Inducible Enhancers in Human Cells Using a Massively Parallel Reporter Assay.” Nature Biotechnology 30 (February): 271.

Mercola, M., X. Wang, J. Olsen, and K. Calame. 1983. “Transcriptional Enhancer Elements in the Mouse Immunoglobulin Heavy Chain Locus.” Science. https://doi.org/10.1126/science.6306772.

Mikkelsen, Tarjei S., Manching Ku, David B. Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, et al. 2007. “Genome-Wide Maps of Chromatin State in Pluripotent and Lineage-Committed Cells.” Nature 448 (7153): 553–60.

Mirchandani, Pitu B., and Richard L. Francis. 1990. Discrete Location Theory. Wiley Series in Discrete Mathematics and Optimization. Wiley.

Monnat, R. J. 2009. “Protocol for HPRT Mutagenesis Analyses.”

Moreau, P., R. Hen, B. Wasylyk, R. Everett, M. P. Gaub, and P. Chambon. 1981. “The SV40 72 Base Repair Repeat Has a Striking Effect on Gene Expression Both in SV40 and Other Chimeric Recombinants.” Nucleic Acids Research 9 (22): 6047–68.

Morley, Michael, Cliona M. Molony, Teresa M. Weber, James L. Devlin, Kathryn G. Ewens, Richard S. Spielman, and Vivian G. Cheung. 2004. “Genetic Analysis of Genome-Wide Variation in Human Gene Expression.” Nature 430 (7001): 743–47.

Mumbach, Maxwell R., Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. 2016. “HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture.” Nature Methods 13 (11): 919–22.

Nagano, Takashi, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. 2013. “Single-Cell Hi-C Reveals Cell-to-Cell Variability in Chromosome Structure.” Nature 502 (7469): 59–64.

Nagano, Takashi, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. 2017. “Cell-Cycle Dynamics of Chromosomal Organization at Single-Cell Resolution.” Nature 547 (7661): 61–67.

Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. “Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre.” Nature 485 (7398): 381–85.

Nora, Elphège P., Anton Goloborodko, Anne-Laure Valton, Johan H. Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A. Mirny, and Benoit G. Bruneau. 2017. "Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization." Cell 169 (5): 930–44.e22.

Osterwalder, Marco, Iros Barozzi, Virginie Tissières, Yoko Fukuda-Yuzawa, Brandon J. Mannion, Sarah Y. Afzal, Elizabeth A. Lee, et al. 2018. "Enhancer Redundancy Provides Phenotypic Robustness in Mammalian Development." Nature 554 (7691): 239–43.

Overbeek, Megan van, Daniel Capurso, Matthew M. Carter, Matthew S. Thompson, Elizabeth Frias, Carsten Russ, John S. Reece-Hoyes, et al. 2016. "DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks." Molecular Cell 63 (4): 633–46.

Patwardhan, Rupali P., Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, et al. 2012. "Massively Parallel Functional Dissection of Mammalian Enhancers in Vivo." Nature Biotechnology 30 (3): 265–70.

Patwardhan, Rupali P., Choli Lee, Oren Litvin, David L. Young, Dana Pe'er, and Jay Shendure. 2009. "High-Resolution Analysis of DNA Regulatory Elements by Synthetic Saturation Mutagenesis." Nature Biotechnology 27 (12): 1173–75.

Pennacchio, Len A., Nadav Ahituv, Alan M. Moses, Shyam Prabhakar, Marcelo A. Nobrega, Malak Shoukry, Simon Minovitsky, et al. 2006. "In Vivo Enhancer Analysis of Human Conserved Non-Coding Sequences." Nature 444 (7118): 499–502.

Pimentel, Harold, Nicolas L. Bray, Suzette Puente, Páll Melsted, and Lior Pachter. 2017. "Differential Analysis of RNA-Seq Incorporating Quantification Uncertainty." Nature Methods 14 (7): 687–90.

Pliner, Hannah A., Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan, et al. 2018. "Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data." Molecular Cell 71 (5): 858–71.e8.

Pollard, Katherine S., Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. 2010. "Detection of Nonneutral Substitution Rates on Mammalian Phylogenies." Genome Research 20 (1): 110–21.

Ptashne, M. 1988. "How Eukaryotic Transcriptional Activators Work." Nature 335 (6192): 683–89.

Ptashne, Mark. 1967. "Specific Binding of the λ Phage Repressor to λ DNA." Nature 214 (5085): 232–34.

Qiu, Xiaojie, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. 2017. "Single-Cell mRNA Quantification and Differential Analysis with Census." Nature Methods 14 (3): 309–15.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." Bioinformatics 26 (6): 841–42.

Rajagopal, Nisha, Sharanya Srinivasan, Kameron Kooshesh, Yuchun Guo, Matthew D. Edwards, Budhaditya Banerjee, Tahin Syed, Bart J. M. Emons, David K. Gifford, and Richard I. Sherwood. 2016. "High-Throughput Mapping of Regulatory DNA." Nature Biotechnology 34 (2): 167–74.

Ramani, Vijay, Xinxian Deng, Ruolan Qiu, Kevin L. Gunderson, Frank J. Steemers, Christine M. Disteche, William S. Noble, Zhijun Duan, and Jay Shendure. 2017. "Massively Multiplex Single-Cell Hi-C." Nature Methods 14 (3): 263–66.

Ran, F. Ann, Patrick D. Hsu, Jason Wright, Vineeta Agarwala, David A. Scott, and Feng Zhang. 2013. “Genome Engineering Using the CRISPR-Cas9 System.” Nature Protocols 8 (11): 2281–2308.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” Cell 159 (7): 1665–80.

Rao, Suhas S. P., Su-Chen Huang, Brian Glenn St Hilaire, Jesse M. Engreitz, Elizabeth M. Perez, Kyong-Rim Kieffer-Kwon, Adrian L. Sanborn, et al. 2017. “Cohesin Loss Eliminates All Loop Domains.” Cell 171 (2): 305–20.e24.

Reid, L. H., R. G. Gregg, O. Smithies, and B. H. Koller. 1990. “Regulatory Elements in the Introns of the Human HPRT Gene Are Necessary for Its Expression in Embryonic Stem Cells.” Proceedings of the National Academy of Sciences of the United States of America 87 (11): 4299–4303.

Rice, P., I. Longden, and A. Bleasby. 2000. “EMBOSS: The European Molecular Biology Open Software Suite.” Trends in Genetics: TIG 16 (6): 276–77.

Rincón-Limas, D. E., D. A. Krueger, and P. I. Patel. 1991. “Functional Characterization of the Human Hypoxanthine Phosphoribosyltransferase Gene Promoter: Evidence for a Negative Regulatory Element.” Molecular and Cellular Biology 11 (8): 4157–64.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” Nature 518 (7539): 317–30.

Robertson, Gordon, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, et al. 2007. “Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing.” Nature Methods 4 (8): 651–57.

Rojano, Elena, Pedro Seoane, Juan A. G. Ranea, and James R. Perkins. 2018. “Regulatory Variants: From Detection to Predicting Impact.” Briefings in Bioinformatics, June. https://doi.org/10.1093/bib/bby039.

Rotem, Assaf, Oren Ram, Noam Shoresh, Ralph A. Sperling, Alon Goren, David A. Weitz, and Bradley E. Bernstein. 2015. “Single-Cell ChIP-Seq Reveals Cell Subpopulations Defined by Chromatin State.” Nature Biotechnology 33 (11): 1165–72.

Sabo, Peter J., Michael S. Kuehn, Robert Thurman, Brett E. Johnson, Ericka M. Johnson, Hua Cao, Man Yu, et al. 2006. “Genome-Scale Mapping of DNase I Sensitivity in Vivo Using Tiling DNA Microarrays.” Nature Methods 3 (7): 511–18.

Sanborn, Adrian L., Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, et al. 2015. “Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes.” Proceedings of the National Academy of Sciences of the United States of America 112 (47): E6456–65.

Sanjana, Neville E., Ophir Shalem, and Feng Zhang. 2014. “Improved Vectors and Genome-Wide Libraries for CRISPR Screening.” Nature Methods 11 (8): 783–84.

Sanjana, Neville E., Jason Wright, Kaijie Zheng, Ophir Shalem, Pierre Fontanillas, Julia Joung, Christine Cheng, Aviv Regev, and Feng Zhang. 2016. “High-Resolution Interrogation of Functional Elements in the Noncoding Genome.” Science 353 (6307): 1545–49.

Schleif, R. 1992. “DNA Looping.” Annual Review of Biochemistry 61: 199–223.

Schmidt, D., M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown, A. Marshall, C. Kutter, et al. 2010. “Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding.” Science. https://doi.org/10.1126/science.1186176.

Schmitt, Anthony D., Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L. Tan, Yun Li, et al. 2016. “A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome.” Cell Reports 17 (8): 2042–59.

Serfling, Edgar, Maria Jasin, and Walter Schaffner. 1985. “Enhancers and Eukaryotic Gene Transcription.” Trends in Genetics: TIG 1 (January): 224–30.

Sexton, Tom, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. 2012. “Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome.” Cell 148 (3): 458–72.

Shalem, Ophir, Neville E. Sanjana, Ella Hartenian, Xi Shi, David A. Scott, Tarjei Mikkelson, Dirk Heckl, et al. 2014. “Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells.” Science 343 (6166): 84–87.

Shlyueva, Daria, Gerald Stampfel, and Alexander Stark. 2014. “Transcriptional Enhancers: From Properties to Genome-Wide Predictions.” Nature Reviews. Genetics 15 (4): 272–86.

Simeonov, Dimitre R., Benjamin G. Gowen, Mandy Boontanrart, Theodore L. Roth, John D. Gagnon, Maxwell R. Mumbach, Ansuman T. Satpathy, et al. 2017. “Discovery of Stimulation-Responsive Immune Enhancers with CRISPR Activation.” Nature 549 (7670): 111–15.

Smith, Cory J., Oscar Castanon, Khaled Said, Verena Volf, Parastoo Khoshakhlagh, Amanda Hornick, Raphael Ferreira, et al. 2019. “Enabling Large-Scale Genome Editing by Reducing DNA Nicking.” bioRxiv. https://doi.org/10.1101/574020.

Spitz, François, and Eileen E. M. Furlong. 2012. “Transcription Factors: From Enhancer Binding to Developmental Control.” Nature Reviews. Genetics 13 (9): 613–26.

Stalder, J., A. Larsen, J. D. Engel, M. Dolan, M. Groudine, and H. Weintraub. 1980. “Tissue-Specific DNA Cleavages in the Globin Chromatin Domain Introduced by DNAase I.” Cell 20 (2): 451–60.

Stranger, Barbara E., Stephen B. Montgomery, Antigone S. Dimas, Leopold Parts, Oliver Stegle, Catherine E. Ingle, Magda Sekowska, et al. 2012. “Patterns of Cis Regulatory Variation in Diverse Human Populations.” PLoS Genetics 8 (4): e1002639.

Stunnenberg, Hendrik G., International Human Epigenome Consortium, and Martin Hirst. 2016. “The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery.” Cell 167 (7): 1897.

Sullivan, W. T. 1993. “The Salvation of Doug.” Generations . http://hocking.biology.ualberta.ca/locke.hp/dougandbill.htm.

Tanabe, Osamu, David McPhee, Shoko Kobayashi, Yannan Shen, William Brandt, Xia Jiang, Andrew D. Campbell, et al. 2007. “Embryonic and Fetal Beta-Globin Gene Repression by the Orphan Nuclear Receptors, TR2 and TR4.” The EMBO Journal 26 (9): 2295–2306.

Tan, Longzhi, Dong Xing, Chi-Han Chang, Heng Li, and X. Sunney Xie. 2018. “Three-Dimensional Genome Structures of Single Diploid Human Cells.” Science 361 (6405): 924–28.

Tewhey, Ryan, Dylan Kotliar, Daniel S. Park, Brandon Liu, Sarah Winnicki, Steven K. Reilly, Kristian G. Andersen, et al. 2018. “Direct Identification of Hundreds of Expression-Modulating Variants Using a Multiplexed Reporter Assay.” Cell 172 (5): 1132–34.

Teytelman, Leonid, Deborah M. Thurtle, Jasper Rine, and Alexander van Oudenaarden. 2013. "Highly Expressed Loci Are Vulnerable to Misleading ChIP Localization of Multiple Unrelated Proteins." Proceedings of the National Academy of Sciences of the United States of America 110 (46): 18602–7.

Thakore, Pratiksha I., Anthony M. D'Ippolito, Lingyun Song, Alexias Safi, Nishkala K. Shivakumar, Ami M. Kabadi, Timothy E. Reddy, Gregory E. Crawford, and Charles A. Gersbach. 2015. "Highly Specific Epigenome Editing by CRISPR-Cas9 Repressors for Silencing of Distal Regulatory Elements." Nature Methods 12 (12): 1143–49.

Thanos, Dimitris, and Tom Maniatis. 1995. "Virus Induction of Human IFNβ Gene Expression Requires the Assembly of an Enhanceosome." Cell 83 (7): 1091–1100.

The GTEx Consortium. 2015. "The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." Science 348 (6235): 648–60.

Thurman, Robert E., Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, et al. 2012. "The Accessible Chromatin Landscape of the Human Genome." Nature 489 (7414): 75–82.

Trapnell, Cole. 2015. "Defining Cell Types and States with Single-Cell Genomics." Genome Research 25 (10): 1491–98.

Tsai, Shengdar Q., Zongli Zheng, Nhu T. Nguyen, Matthew Liebers, Ved V. Topkar, Vishal Thapar, Nicolas Wyvekens, et al. 2015. "GUIDE-Seq Enables Genome-Wide Profiling of off-Target Cleavage by CRISPR-Cas Nucleases." Nature Biotechnology 33 (2): 187–97.

Tuan, Dorothy, William Solomon, Qiliang Li, and Irving M. London. 1985. "The' Beta-like-Globin' Gene Domain in Human Erythroid Cells." Proceedings of the National Academy of Sciences 82 (19): 6384–88.

Ulirsch, Jacob C., Satish K. Nandakumar, Li Wang, Felix C. Giani, Xiaolan Zhang, Peter Rogov, Alexandre Melnikov, et al. 2016. "Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits." Cell 165 (6): 1530–45.

Vanhille, Laurent, Aurélien Griffon, Muhammad Ahmad Maqbool, Joaquin Zacarias-Cabeza, Lan T. M. Dao, Nicolas Fernandez, Benoit Ballester, Jean Christophe Andrau, and Salvatore Spicuglia. 2015. "High-Throughput and Quantitative Assessment of Enhancer Activity in Mammals by CapStarr-Seq." Nature Communications. https://doi.org/10.1038/ncomms7905.

Van Loo, Peter, and Peter Marynen. 2009. "Computational Methods for the Detection of Cis-Regulatory Modules." Briefings in Bioinformatics 10 (5): 509–24.

Väremo, Leif, Jens Nielsen, and Intawat Nookaew. 2013. "Enriching the Gene Set Analysis of Genome-Wide Data by Incorporating Directionality of Gene Expression and Combining Statistical Hypotheses and Methods." Nucleic Acids Research 41 (8): 4378–91.

Visel, Axel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2009. "ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers." Nature 457 (7231): 854–58.

Visel, Axel, Simon Minovitsky, Inna Dubchak, and Len A. Pennacchio. 2007. "VISTA Enhancer Browser—a Database of Tissue-Specific Human Enhancers." Nucleic Acids Research 35 (suppl_1): D88–92.

Vockley, Christopher M., Anthony M. D'Ippolito, Ian C. McDowell, William H. Majoros, Alexias Safi, Lingyun Song, Gregory E. Crawford, and Timothy E. Reddy. 2016. "Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome." Cell 166 (5): 1269–81.e19.

Vockley, Christopher M., Cong Guo, William H. Majoros, Michael Nodzenski, Denise M. Scholtens, M. Geoffrey Hayes, William L. Lowe Jr, and Timothy E. Reddy. 2015. "Massively Parallel Quantification of the Regulatory Effects of Noncoding Genetic Variation in a Human Cohort." Genome Research 25 (8): 1206–14.

Vojta, Aleksandar, Paula Dobrinić, Vanja Tadić, Luka Bočkor, Petra Korać, Boris Julg, Marija Klasić, and Vlatka Zoldoš. 2016. "Repurposing the CRISPR-Cas9 System for Targeted DNA Methylation." Nucleic Acids Research 44 (12): 5615–28.

Wakabayashi, Aoi, Jacob C. Ulirsch, Leif S. Ludwig, Claudia Fiorini, Makiko Yasuda, Avik Choudhuri, Patrick McDonel, Leonard I. Zon, and Vijay G. Sankaran. 2016. "Insight into GATA1 Transcriptional Activity through Interrogation of Cis Elements Disrupted in Human Erythroid Disorders." Proceedings of the National Academy of Sciences of the United States of America 113 (16): 4434–39.

Wang, Tim, Jenny J. Wei, David M. Sabatini, and Eric S. Lander. 2014. "Genetic Screens in Human Cells Using the CRISPR-Cas9 System." Science 343 (6166): 80–84.

Wang, Xinchen, Liang He, Sarah M. Goggin, Alham Saadat, Li Wang, Nasa Sinnott-Armstrong, Melina Claussnitzer, and Manolis Kellis. 2018. "High-Resolution Genome-Wide Functional Dissection of Transcriptional Regulatory Regions and Nucleotides in Human." Nature Communications. https://doi.org/10.1038/s41467-018-07746-1.

Ward, Lucas D., and Manolis Kellis. 2012. "HaploReg: A Resource for Exploring Chromatin States, Conservation, and Regulatory Motif Alterations within Sets of Genetically Linked Variants." Nucleic Acids Research 40 (Database issue): D930–34.

Weedon, Michael N., Ines Cebola, Ann-Marie Patch, Sarah E. Flanagan, Elisa De Franco, Richard Caswell, Santiago A. Rodríguez-Seguí, et al. 2014. "Recessive Mutations in a Distal PTF1A Enhancer Cause Isolated Pancreatic Agenesis." Nature Genetics 46 (1): 61–64.

Wei, Kai, Rishabh Iyer, and Jeff Bilmes. 2015. "Submodularity in Data Subset Selection and Active Learning." In International Conference on Machine Learning, 1954–63.

Weintraub, H., and M. Groudine. 1976. "Chromosomal Subunits in Active Genes Have an Altered Conformation." Science 193 (4256): 848–56.

Whyte, Warren A., David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. 2013. "Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes." Cell 153 (2): 307–19.

Williamson, Iain, Ragnhild Eskeland, Laura A. Lettice, Alison E. Hill, Shelagh Boyle, Graeme R. Grimes, Robert E. Hill, and Wendy A. Bickmore. 2012. "Anterior-Posterior Differences in HoxD Chromatin Topology in Limb Development." Development 139 (17): 3157–67.

Williamson, Iain, Laura A. Lettice, Robert E. Hill, and Wendy A. Bickmore. 2016. "Shh and ZRS Enhancer Colocalisation Is Specific to the Zone of Polarising Activity." Development 143 (16): 2994–3001.

Won, Hyejung, Luis de la Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, Jerry Huang, Carli K. Opland, Michael J. Gandal, et al. 2016. "Chromosome Conformation Elucidates Regulatory Relationships in Developing Human Brain." Nature 538 (7626): 523–27.

Worsley Hunt, Rebecca, and Wyeth W. Wasserman. 2014. "Non-Targeted Transcription Factors Motifs Are a Systemic Component of ChIP-Seq Datasets." Genome Biology 15 (7): 412.

Wright, Jason B., and Neville E. Sanjana. 2016. "CRISPR Screens to Discover Functional Noncoding Elements." Trends in Genetics: TIG 32 (9): 526–29.

Xie, Shiqi, Anne Cooley, Daniel Armendariz, Pei Zhou, and Gary C. Hon. 2018. “Frequent sgRNA-Barcode Recombination in Single-Cell Perturbation Assays.” PloS One 13 (6): e0198635.

Xie, Shiqi, Jialei Duan, Boxun Li, Pei Zhou, and Gary C. Hon. 2017. “Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells.” Molecular Cell 66 (2): 285–99.e5.

Yahi, Alexandre, Tuuli Lappalainen, Pejman Mohammadi, and Nicholas Tatonetti. 2018. “RecNW: A Fast Pairwise Aligner for Targeted Sequencing.” bioRxiv, July, 371989.

Yang, Yaping, Donna M. Muzny, Jeffrey G. Reid, Matthew N. Bainbridge, Alecia Willis, Patricia A. Ward, Alicia Braxton, et al. 2013. “Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders.” The New England Journal of Medicine 369 (16): 1502–11.

Zabidi, Muhammad A., Cosmas D. Arnold, Katharina Schernhuber, Michaela Pagani, Martina Rath, Olga Frank, and Alexander Stark. 2014. “Enhancer–core-Promoter Specificity Separates Developmental and Housekeeping Gene Regulation.” Nature 518 (December): 556.

Zeng, Wanwen, Mengmeng Wu, and Rui Jiang. 2018. “Prediction of Enhancer-Promoter Interactions via Natural Language Processing.” BMC Genomics 19 (Suppl 2): 84.

Zhang, Guanxiong, Jian Shi, Shiwei Zhu, Yujia Lan, Liwen Xu, Huating Yuan, Gaoming Liao, et al. 2018. “DiseaseEnhancer: A Resource of Human Disease-Associated Enhancer Catalog.” Nucleic Acids Research 46 (D1): D78–84.

Zhou, B., S. S. Ho, S. U. Greer, X. Zhu, J. M. Bell, and J. G. Arthur. 2018. “Comprehensive, Integrated, and Phased Whole-Genome Analysis of the Primary ENCODE Cell Line K562.” BioRxiv. https://www.biorxiv.org/content/early/2018/06/17/192344.abstract.

Zhou, Yuexin, Shiyou Zhu, Changzu Cai, Pengfei Yuan, Chunmei Li, Yanyi Huang, and Wensheng Wei. 2014. “High-Throughput Screening of a CRISPR/Cas9 Library for Functional Genomics in Human Cells.” Nature 509 (7501): 487–91.

Zhu, Shiyou, Wei Li, Jingze Liu, Chen-Hao Chen, Qi Liao, Ping Xu, Han Xu, et al. 2016. “Genome-Scale Deletion Screening of Human Long Non-Coding RNAs Using a Paired-Guide RNA CRISPR-Cas9 Library.” Nature Biotechnology 34 (12): 1279–86.

VITA

Molly Gasperini grew up in Shoreline, Washington in the shadow of the Aurora Village Costco. She graduated from Shorewood High School and went on to earn an honors degree in molecular biology at the University of Washington. As an undergraduate, she was a student researcher in Mary-Claire King's human genetics lab. Following graduation, she moved to Boston to try Dunkin Donuts for the first time and work as a research technician in Louis Kunkel's muscular dystrophy lab. In 2014, she returned home to join the UW Genome Sciences graduate program and enjoy craft cappuccinos once more. She is a proud daughter, sister, niece, aunt, life partner, and cat mom.