Title
Genome-scale discovery of protein-coding and lncRNA gene function with CRISPRi and CRISPRa

Author
Horlbeck, Max A.

Publication Date
2017

Peer reviewed|Thesis/dissertation

Genome-scale discovery of protein-coding and lncRNA gene function with CRISPRi and CRISPRa

by

Max A. Horlbeck

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

To my parents, Susan and Gernot Horlbeck, and my partner, Darienne Myers, for their constant support, patience, and love

Acknowledgements

It has been my great fortune to have conducted my thesis work in the lab of Jonathan Weissman at UCSF. Seven years ago, the work of the Weissman lab caught my attention as an undergraduate and led me to ask for a summer research job. To my great surprise and pleasure, Jonathan accepted my request. That this summer experience alone convinced me to move across the country and commit to an eight-year program is a testament to the genius, creativity, excitement, and collegiality with which Jonathan and his lab operates. As a mentor, Jonathan is wise and empathic, and it is only through his vision, intellectual curiosity, and ability to bring together highly effective and complementary teams that this thesis work came to fruition. I will always keep in mind the example he sets as I continue my career.

The ultimate pleasure of graduate school has been the ability to work with the extremely talented and dynamic members of the Weissman lab. In particular, Luke Gilbert has been a second mentor, an equal partner-in-crime, and a good friend throughout the entirety of grad school. It is rare that two people are able to work together for four years with a constant, open exchange of ideas and without conflict, and I truly believe that together we have worked at a far higher level than the sum of our parts. I also owe a debt of gratitude to Martin Kampmann and Michael Bassik, who first took me on as a summer student and taught me everything I needed to know to undertake this thesis work, and have remained collaborators and friends over the years. While the full list of the Weissman lab colleagues and friends I would like to thank could take several more pages, I particularly want to thank Britt Adamson, Jacqueline Villalta, Joshua Dunn, Michelle Chan, Marcos Hein und Jost, Alex Fields, Tom Norman, Matt Larson, Calvin Jan, Owen Chen, Dan Santos, Liz Costa, and Min Cho. I would also like to thank Christopher Reiger, Manny De Vara, and Joan Kanter for their hard work and support.

My thesis work was also shaped by the valuable input of the members my thesis and qualifying exam committees: Daniel Lim, Barbara Panning, Wendell Lim, and Nadav Ahituv. Their insight and expertise were critical in shaping a major part of my thesis work – so much so that the picture of the board I took after the challenging and intellectually stimulating qualifying exam served as a blueprint for my studies on lncRNAs over the following years.

The collaborative and open environment at UCSF and around the San Francisco Bay Area has also been an indispensable asset in the work described here. I would like to thank Stanely Qi, Marvin Tanenbaum, and Ron Vale for their contributions to the initial development of CRISPRi/a screening. I greatly appreciate Lea Witkowsky and Robert Tjian, first for their openness in presenting their exciting, unpublished data, and then for jumping enthusiastically into a collaboration to reconcile and publish our complementary data sets. In developing the lncRNA screening project, John Liu was an unstoppable force and the ideal collaborator from the very first day we discussed the project in the kitchen we shared as housemates, and I eagerly anticipate working together with him long into the future. And outside his role as John's advisor and close collaborator, Daniel Lim has been a personal role model as a scientist, physician, and good person. I would also like to thank Howard Chang and Seung Woo Cho for their early enthusiasm and continued contributions to the lncRNA screening efforts. Finally, Mo Mandegar and Michael Olvera in Bruce Conklin's lab have been generous partners in two-way exchanges of ideas, reagents, and data, and Bruce himself has been a fantastic supporter and mentor.

UCSF has been a fantastic environment in which to work and learn thanks to a large number of people. From the MSTP, I thank Jana Toutolmin, Catherine Norton, and Geri Ehli for their administrative support and welcoming attitude, and Mark Anderson and especially Kevin Shannon for their leadership and guidance. I thank the Biophysics program, particularly Tanja

Kortemme, Rebecca Brown, and Nicole Flowers, for their assistance and for the many excellent retreats. Most importantly, I thank my MSTP classmates, who truly are a second family and have served as inspirations and partners in both science and life. I look forward to staying close with this group as we advance through our careers together.

I am deeply grateful for the love and support of Darienne Myers, who challenges and shapes my scientific ideas, provides encouragement when problems arose, reminds me to put down work every once in a while, and takes me on adventures around the world.

Finally, I would like to thank Gernot and Susan Horlbeck for their life long support and encouragement, stoking my passion for life sciences throughout with biographies of Charles Darwin, scientific summer camps, and even allowing me to join the local rescue squad despite it waking us all up with midnight pages. I would never be writing this without you.

Chapter 2 is reprinted largely as it appears in

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, Qi LS, Kampmann M, Weissman JS. Cell. 2014 Oct 23;159(3):647-61. doi: 10.1016/j.cell.2014.09.029.

LAG and MAH contributed equally to this work. BA, LAG, MAH, MK, JSW were primarily responsible for the conception, design, and interpretation of the experiments and wrote the manuscript. BA, YC, LAG, MAH, MK, JEV, and EHW conducted experiments. LAG cloned dCas9 chimeras and sgRNA expression constructs, constructed cell lines, carried out tiling screens, and conducted validation experiments. MAH designed libraries, carried out tiling and genome-scale screens, and analyzed screen data. BA constructed the inducible cell line and

conducted all inducible experiments. CG and HLP contributed to CTx-DTA studies. BP contributed to XIST studies. MCB and LSQ contributed to the conception and interpretation of the experiments.

Chapter 3 is reprinted largely as it appears in

Horlbeck MA, Witkowsky LB, Guglielmi B, Replogle JM, Gilbert LA, Villalta JE, Torigoe SE, Tjian R, Weissman JS. Elife. 2016 Mar 17;5. pii: e12677. doi: 10.7554/eLife.12677.

MAH and LBW contributed equally to this work. MAH: Conceived of and conducted the in vivo data analysis and experiments, Helped write this report. LBW: Conceived of and conducted the in vitro experiments and analysis, Helped write this report. BG: Contributed to conception and interpretation of the in vitro experiments, Edited this report. JMR: Contributed analysis of in vivo tiling screen data. LAG: Conducted tiling screens. JEV: Conducted tiling screens. SET: Contributed to chromatin remodeling experiments. RT: Supervised the in vitro experiments carried out by LBW, Helped write this report. JSW: Supervised the in vivo experiments carried out by MAH, Helped write this report.

Chapter 4 is reprinted largely as it appears in

Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M, Weissman JS. Elife. 2016 Sep 23;5. pii: e19760. doi: 10.7554/eLife.19760.

MAH: Conceived of and conducted algortihm development and data analysis, contributed to genome-scale screening experiments, and wrote this report. LAG: Contributed to sgRNA library development and genome-scale screening experiments, and helped write this report. JEV: Contributed to sgRNA library development, generated sgRNA libraries, and contributed

technical assistance. BA: Conducted genome-scale screening experiments. RAP: Generated sgRNA libraries. YC: Contributed technical assistance. APF: Contributed to algorithm development. CYP: Supervised sgRNA library generation and conducted genome-scale screening experiments. JEC: Contributed to algorithm development and supervised sgRNA library generation. MK: Contributed to and supervised algorithm development. JSW: Conceived of and supervised the study, and helped write this report.

Chapter 5 is reprinted largely as it appears in

Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y, Mandegar MA, Olvera MP, Gilbert LA, Conklin BR, Chang HY, Weissman JS, Lim DA. Science. 2017 Jan 6;355(6320). pii: aah7111. doi: 10.1126/science.aah7111.

S.J.L and M.A.H. contributed equally to this work. M.A.H., S.J.L., J.S.W., D.A.L., and H.Y.C. conceived the project, interpreted the data, and wrote the manuscript. M.A.H. and S.J.L designed sgRNA library, performed RNA-seq, qPCR, growth assays, and analyzed data. S.J.L. performed screens in U87 and HeLa cells, 4C-seq, and machine learning. M.A.H. performed screens in K562, HEK293T, and iPS cells. S.W.C. performed screens and validation experiments in MCF7 and MDA-MB-231 cells. H.S.B., M.M., and F.J.A. performed qPCR, growth assays, and ASO experiments. D.H. performed ChIP-seq. J.E.V. cloned sgRNA libraries. M.Y.C. and Y.C. performed validation experiments. M.A.M., M.P.O., and B.R.C. contributed to iPSC experiments and performed iPSC protein-coding screens. L.A.G. contributed to data interpretation and project conception.

Additional acknowledgements are included in each chapter. Supplementary tables listed throughout may be accessed through the corresponding chapter citations listed above.

Genome-scale discovery of protein-coding and lncRNA gene function with CRISPRi and CRISPRa

Max A. Horlbeck

Abstract

The genome sequencing efforts of the past decade have cataloged the universe of protein-coding and non-coding transcripts produced by the human genome. A central challenge now is to understand how these many thousands of genes act to mediate the vast array of cellular processes involved in either normal or disease states. To accomplish this, we require specific and scaleable tools to manipulate the expression of individual genes and measure their functional contribution in a given cellular context. In this thesis, I describe the initial development and subsequent refinement of genome-scale genetic screening technologies based on CRISPR-mediated interference and activation. I present the results of several screens for protein-coding genes that modulate cell proliferation and toxin susceptibility. Finally, I describe the application of CRISPR interference in the systematic repression of long non-coding RNA genes, a large class of genes of which very few genes have any known function. The screens described here identify nearly 500 lncRNA genes that modify cell growth in at least one of seven diverse cell lines, and furthermore highlight the exquisite cell type-specificity of lncRNA function.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Each of the trillions of cells in the human body performs a set of specific functions and undergoes a distinct lifecycle of division, differentiation, and death – all while operating from a copy of the genome identical in sequence to every other cell. Following from the publication of the sequence of the human genome, numerous large-scale sequencing projects (*1*-*5*) were launched to understand and catalog how this phenotypic complexity arises from a shared sequence. Through these sequencing projects we now have a comprehensive annotation of the specific sets of genes in the genome transcribed in a given cell type, as well as the regulatory mechanisms that enforce this transcriptional program(*6*), but our understanding of the functional role of most genes in the genome lags far behind. This gap in understanding is due in part to the abundance and relative ease of technologies available for "observational genomics" as opposed to "functional genomics." To remedy this, we require scalable and precise tools for systematically manipulating the expression of individual genes in order to characterize their function, and must apply these tools broadly across the spectrum of cell types found in the human body.

In this thesis, I describe the development, refinement, and application of two such genome-scale screening tools based on CRISPR-mediated interference and activation (CRISPRi and CRISPRa, respectively). The CRISPRi and CRISPRa screening platforms enable systematic and specific repression or overexpression of targeted genes in a by building upon two foundational technologies: the CRISPR gene editing system and pooled genetic screening.

CRISPR (Clustered Regularly Interspaced Palindromic Repeats) systems are a diverse group of adaptive genome defense systems employed by bacteria to recognize and excise viral genomes, and their discovery and recent application to the editing of other genomes has revolutionized biology (reviewed in (*7*)). While preceding methods (e.g. zinc finger nucleases

and transcription activator-like effector nucleases) also enabled site-specific gene editing functionality, CRISPR is unique and transformative because it can be targeted to the genome simply by providing a single guide RNA (sgRNA) that contains a ~20 base pair sequence complementary to the target site. This feature not only makes design of specific CRISPR reagents trivial for the average experimentalist (to first approximation) but also dovetails well with our current capabilities in highly multiplexed DNA synthesis and sequencing (*8*) that are critical in the pooled screening methods described below.

Shortly after the initial publications describing the application of the *S. pyogenes* CRISPR effector protein Cas9 to gene editing in human cells (*9*-*11*), Qi and colleagues described a nuclease-dead variant (dCas9) that was capable of stably binding to DNA at the target site – and sterically interfering with bacterial and eukaryotic transcription when targeted to a gene body – but could not cut DNA (*12*). dCas9 has proven to be a versatile DNA-binding platform for a host of applications, including live cell DNA imaging (*13*), single base editing (*14*), and epigenome editing (*15*, *16*), the basis for the work presented here. In particular, Gilbert and colleagues developed versions of dCas9 fused to broadly active transcriptional repressors or activators which were highly active in mammalian cells with minimal off-target effects (*17*, *18*). These initial results showed that CRISPRi was generally more effective than RNA interference (RNAi) with little to no off-target effects typically seen with RNAi, while CRISPRa allowed for robust and programmable overexpression not previously achievable. Coupled with properly targeted sgRNAs, we expected these CRISPRi and CRISPRa to be powerful and complementary tools for systematically manipulating gene expression.

In order to apply these reagents systematically, we turned to the pooled screening approach originally developed for RNAi (*19*-*22*). In brief, pooled screening utilizes highly

multiplexed DNA synthesis to generate pre-defined complex libraries of ~10^3-10^6 unique sequences (e.g. sgRNAs for CRISPR). This library is then introduced into a population of cells using lentivirus at low multiplicity of infection, thus stably integrating one unique sequence into the genome of the host cell. This sequence then acts both as a targeting molecule (again, an sgRNA targeting (d)Cas9 to the target gene) and as a molecular barcode that can be read out by deep sequencing of the library. Therefore, by simply taking samples of the cell population over the course of days of continuous cell culture, in untreated and drug/toxin-treated cell populations, or at different reporter gene expression levels, one can infer the phenotype caused by a particular sgRNA by measuring the enrichment or disenrichment of that sgRNA sequence in the selected populations. The approach is advantageous compared to systems where each perturbation is separately tested in an array because it requires only the ability to perform cell culture and molecular biology at medium scale, without the need for advanced robotics, and each perturbation is performed in the same pool, resulting in an internally-controlled experiment free from plate or batch effects (*23*, *24*). The main limitations of the pooled screening method are the requirements for large numbers of cells to maintain statistical robustness (*22*) and the relatively low complexity readouts of the screen (e.g. pro-/anti-growth, toxin sensitivity/resistance) (*25*). Certain approaches are currently being developed to expand the richness of screen readouts, including coupling CRISPR-based perturbations to single-cell sequencing (*26*) and generating scalable genetic interaction maps for mammalian cells (*21*, *27*).

With CRISPRi and CRISPRa as versatile tools for manipulating gene expression, and the established RNAi pooled screening methods as a template for systematically and scalably applying these tools, I sought to create screening platforms to enable targeted repression and overexpression screening and thus define the function of any gene in the human genome. In

Chapter 2, I describe the development of a complex library designed to establish sgRNA design rules for CRISPRi and CRISPRa, the use of the resulting design principles to create genome-scale platforms for CRISPRi/a screening, and analysis and validation of initial CRISPRi/a screens for cell growth and toxin susceptibility.

Analysis of the results from the screens described in Chapter 2, along with additional screen data from collaborative projects, suggested that my initial sgRNA design rules did not fully capture the parameters governing sgRNA efficacy and that improved predictions for highly active sgRNAs could be used to generate substantially improved libraries. In Chapter 3, I describe the discovery that nucleosome occupancy at the sgRNA target site is strongly anticorrelated with sgRNA activity for CRISPRi and for CRISPR nuclease. In conjunction with *in vitro* results from the Tjian lab, I show that this effect is due to nucleosomes creating direct physical impediment to Cas9 access to DNA. In the work presented in Chapter 4, I leverage the nucleosome occupancy findings and other observations to generate a machine learning model for predicting sgRNA efficacy for CRISPRi and CRISPRa with high accuracy. I validate this by creating next-generation genome-scale libraries that exhibit greatly improved discrimination of positive control sgRNAs as well as a substantially greater fraction of highly active sgRNAs, resulting in state-of-the-art tools for controlling gene expression in both mouse and human cells.

Finally, I sought to use these technologies to identify novel functions for genes that produce long non-coding RNAs (lncRNAs), transcripts longer than 200bp that are processed much like messenger RNAs but do not appear to code for proteins. While tens of thousands of lncRNA genes have been discovered through observational genomics (*28*), only a small fraction have been shown to mediate important functions (*29*). In Chapter 5, I describe the design of a genome-scale CRISPRi library targeting lncRNA genes and the results from screens for genes

that modulate cell proliferation across seven diverse human cell lines. These screens identify 499 genes that modify robust cell growth and demonstrate the exquisite cell type-specificity of lncRNA function. These results underscore both the functional relevance of certain lncRNA genes as well as the importance of applying large-scale, unbiased, and systematic approaches in discovering gene functions.

References

1. S. Djebali *et al.*, Landscape of transcription in human cells. **489**, 101–108 (2012).

2. J. Harrow *et al.*, GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*. **22**, 1760–1774 (2012).

3. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. **489**, 57–74 (2012).

4. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.*, A promoter-level mammalian expression atlas. **507**, 462–470 (2014).

5. P. Carninci *et al.*, The transcriptional landscape of the mammalian genome. *Science*. **309**, 1559–1563 (2005).

6. V. Amin *et al.*, Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. *Nat Commun*. **6**, 6370–6370 (2015).

7. J. A. Doudna, E. Charpentier, Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*. **346**, 1258096–1258096 (2014).

8. M. C. Bassik *et al.*, Rapid creation and quantitative monitoring of high coverage shRNA libraries. *Nature Methods*. **6**, 443–445 (2009).

9. M. Jinek *et al.*, RNA-programmed genome editing in human cells. *eLife*. **2**, e00471 (2013).

10. L. Cong *et al.*, Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* (2013), doi:10.1126/science.1231143.

11. P. Mali *et al.*, RNA-Guided Human Genome Engineering via Cas9. *Science* (2013), doi:10.1126/science.1232033.

12. L. S. Qi *et al.*, Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. **152**, 1173–1183 (2013).

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

14. A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris, D. R. Liu, Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. (2016), doi:10.1038/nature17946.

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

16. P. Perez-Pinera *et al.*, RNA-guided gene activation by CRISPR-Cas9-based transcription

factors. *Nature Methods*. **10**, 973–976 (2013).

17. L. A. Gilbert *et al.*, CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. **154**, 442–451 (2013).

18. M. E. Tanenbaum, L. A. Gilbert, L. S. Qi, J. S. Weissman, R. D. Vale, A protein-tagging system for signal amplification in gene expression and fluorescence imaging. **159**, 635–646 (2014).

19. P. J. Paddison *et al.*, A resource for large-scale RNA-interference-based screens in mammals. **428**, 427–431 (2004).

20. F. Stegmeier, G. Hu, R. J. Rickles, G. J. Hannon, S. J. Elledge, A lentiviral microRNA-based system for single-copy polymerase II-regulated RNA interference in mammalian cells. *Proc Natl Acad Sci U S A*. **102**, 13212–13217 (2005).

21. M. C. Bassik *et al.*, A Systematic Mammalian Genetic Interaction Map Reveals Pathways Underlying Ricin Susceptibility. **152**, 909–922 (2013).

22. M. Kampmann, M. C. Bassik, J. S. Weissman, Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. *Proc Natl Acad Sci U S A*. **110**, E2317–E2326 (2013).

23. M. Schuldiner, S. Collins, J. Weissman, N. Krogan, Quantitative genetic analysis in Saccharomyces cerevisiae using epistatic …. *Methods* (2006).

24. M. Schuldiner, N. Krogan, J. Weissman, A strategy for extracting and analyzing large-scale quantitative epistatic …. *Genome biology* (2006).

25. P. Liberali, B. Snijder, L. Pelkmans, Single-cell and multivariate approaches in genetic perturbation screens. *Nat. Rev. Genet.* **16**, 18–32 (2015).

26. B. Adamson *et al.*, A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*. **167**, 1867–1882.e21 (2016).

27. A. Roguev *et al.*, Quantitative genetic-interaction mapping in mammalian cells. *Nature Methods*. **10**, 432–437 (2013).

28. I. Ulitsky, D. P. Bartel, lincRNAs: genomics, evolution, and mechanisms. *Cell*. **154**, 26–46 (2013).

29. M. Sauvageau *et al.*, Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife*. **2**, e01749 (2013).

Chapter 2

Genome-scale CRISPR-mediated control of gene repression and activation

Summary

While the catalog of mammalian transcripts and their expression levels in different cell types and disease states is rapidly expanding, our understanding of transcript function lags behind. We present a robust technology enabling systematic investigation of the cellular consequences of repressing or inducing individual transcripts. We identify rules for specific targeting of transcriptional repressors (CRISPRi), typically achieving 90-99% knockdown with minimal off-target effects, and activators (CRISPRa) to endogenous genes via endonuclease-deficient Cas9. Together they enable modulation of gene expression over a ~1000-fold range. Using these rules, we construct genome-scale CRISPRi and CRISPRa libraries, each of which we validate with two pooled screens. Growth-based screens identify essential genes, tumor suppressors and regulators of differentiation. Screens for sensitivity to a cholera-diphtheria toxin provide broad insights into the mechanisms of pathogen entry, retro-translocation and toxicity. Our results establish CRISPRi and CRISPRa as powerful tools that provide rich and complementary information for mapping complex pathways.

Introduction

Dramatic advances in sequencing technology have catalogued a universe of transcribed loci—greatly exceeding the number of canonical protein-coding open reading frames (ORFs)—which collectively are responsible for carrying out the instructions encoded by the genome (Djebali et al., 2012). A central challenge now is to understand the biological role of these transcripts and how quantitative differences in their expression define cellular states in normal development and in disease. Despite intense efforts, the function of many protein-coding genes remains poorly defined. Even less is known about the biological roles of most non-canonical transcripts such as enhancer RNAs, upstream antisense RNAs, lncRNAs, or other intergenic RNAs (Cech and Steitz, 2014). Efforts to address this deficiency in our knowledge would be greatly aided by techniques that are capable of dynamically and precisely controlling the expression of individual transcripts.

One way to explore the function of genes is to disrupt their expression through repression. The dominant tool for programmed knockdown of mRNAs is RNA interference (RNAi) (Chang et al., 2006). However, RNAi has pervasive problems with off-target effects, which can be especially confounding in the context of large-scale screens (Adamson et al., 2012; Jackson et al., 2003; Sigoillot et al., 2012). Additionally, because RNAi is mediated by cytoplasmic argounaute proteins, gene silencing through this approach is best suited to depletion of cytosolic mRNA targets.

An alternative emerging strategy is the use of programmable genome editing methods that permanently delete or modify DNA using designable, sequence-specific endonucleases such as zinc finger, transcription activator-like effector (TALE) nucleases, or CRISPR (clustered regularly interspaced short palindromic repeats)/Cas9 (CRISPR-associated protein 9) proteins

(Gaj et al., 2013; Sander and Joung, 2014). A series of elegant studies recently exploited the readily programmable nature of Cas9, in which the specificity is determined by a short guide (sg)RNA, to enable genome-scale loss-of-function screens (Koike-Yusa et al., 2014; Shalem et al., 2014; Wang et al., 2014). These studies established CRISPR-mediated cutting as a powerful screening technology complementary to RNAi and haploid mutagenesis screens (Carette et al., 2009). Nonetheless, screening approaches based on genome editing are currently focused on loss-of-function studies involving irreversible frameshift disruptions, limiting their utility for the study of essential genes and long noncoding RNAs. Additionally, double-stranded DNA breaks can be cytotoxic (Huang et al., 1996; Jackson, 2002). Finally, indels formed from error-prone DNA repair are often short and in-frame, which could limit the ability to disable all of the alleles of a gene.

A programmable DNA binding protein that can recruit an effector domain to turn transcription on and off in a dynamic and quantitative manner offers, in principle, a more flexible tool for interrogating the many transcripts in complex genomes. Pioneering experiments with designed chimeric zinc finger and TALE proteins fused to transcription effector domains demonstrate that such an approach can modulate transcription of endogenous genes (Beerli et al., 1998, 2000; Zhang et al., 2011). However, as each transcript target requires a unique fusion protein, expanding these methods to genome-scale is arduous.

Recently, we and others have used catalytically inactive Cas9 (dCas9) fusion proteins guided by gene-specific sgRNAs to localize effector domains to specific DNA sequences to either repress (CRISPRi) or activate (CRISPRa) transcription of target genes (Gilbert et al., 2013; Sander and Joung, 2014). To date, a small number of sgRNAs have been tested, leaving unanswered whether CRISPRi/a is a feasible strategy for globally interrogating gene function

and, if so, how best to target a gene to activate or repress transcription while minimizing off-target effects.

Here, we describe the development and application of a method for high-specificity, genome-scale modulation of transcription of endogenous genes in human cells using CRISPRi/a. To accomplish this, we first performed a saturating screen in which we tested the activity of every unique sgRNA broadly tiling around the transcription start sites (TSSs) of 49 genes known to modulate cellular susceptibility to ricin (Bassik et al., 2013). From this, we extracted distinct rules for regions where either CRISPRi or CRISPRa maximally changes the expression of endogenous genes, as well as rules for predicting off-target effects, providing an algorithm to design two genome-scale libraries targeting each gene with 10 sgRNAs. We validated these libraries by screening for genes that control cell growth and response to a chimeric cholera/diphtheria fusion toxin (CTx-DTA) (Guimaraes et al., 2011). These experiments demonstrate that our CRISPRi/a screening platform is robust, showing high reproducibility and activity with undetectable intrinsic toxicity. More generally, we establish that transcriptional repression is inducible, reversible, and can target essential genes. We demonstrate that we can use CRISPRi and CRISPRa to control transcript levels for endogenous genes across a wide dynamic range. We also provide extensive evidence that properly designed CRISPRi reagents are highly specific. As such, these methods represent transformative tools for defining transcript function across the breadth of transcripts encoded by the human genome.

Results

A High-Throughput Tiling Screen Defines Rules for CRISPRi Activity at Endogenous Genes

CRISPRi can repress transcription by directly blocking RNA polymerase activity (dCas9) or through effector domain-mediated transcriptional silencing (dCas9-KRAB) (Gilbert et al., 2013; Qi et al., 2013). In order to better understand and optimize CRISPRi activity, we used a pooled high-throughput screen to define rules that determine CRISPRi repression of endogenous genes. We targeted 49 genes that we had previously shown to modulate cellular susceptibility to the AB toxin ricin (Bassik et al., 2013). The extent of gene repression for these genes typically has a monotonic relationship with the ricin resistance phenotype, allowing us to use a ricin resistance score calculated by monitoring sgRNA frequencies in a pooled screen to indirectly measure transcriptional repression.

Using massively parallel oligonucleotide synthesis, we generated a library of sgRNAs that tile the DNA in a 10-kilobase window around the TSS of these 49 genes (54,810 total sgRNAs) (Bassik et al., 2009) (Figure 1A). We also included 1,000 negative control sgRNAs derived from scrambled sequences corresponding to the same windows.

We packaged this tiling library of sgRNAs into lentiviral particles and transduced K562 human myeloid leukemia cells stably expressing dCas9 or a dCas9-KRAB fusion protein, which we have previously described (Gilbert et al., 2013). We harvested populations of cells expressing the library either at the outset of the experiment, after growth under standard conditions, or following ricin treatment. We then counted the frequency of each sgRNA in the library in each sample using deep sequencing to determine how each sgRNA in the library modulates cell growth and cellular susceptibility to ricin phenotypes. We defined these phenotypes

quantitatively as gamma (γ) and rho (ρ), respectively (See Figure S1A and (Kampmann et al., 2013)).

Many sgRNAs potently repress gene expression, as evidenced by their impact on ricin sensitivity (Figure 1B and Figure S2A). Plotting this data for all 49 genes showed that active sgRNAs cluster around or just downstream from the TSS of each gene for dCas9-KRAB and dCas9, respectively (Figure 1C). We saw that strong CRISPRi activity is obtained by targeting dCas9-KRAB to a window of DNA from -50 to +300 bp relative to the TSS of a gene, with a maximum in the ~50-100 bp region just downstream of the TSS (Figure 1C-D). This suggested that optimal activity leverages the combined activity of dCas9 interference along with repression from the KRAB domain. We also observed that sgRNAs with protospacer lengths of 18-21 base pairs were significantly more active than sgRNAs containing longer protospacers (Figure S2B). Nucleotide homopolymers had a strongly negative effect on sgRNA activity (Figure S2D). However, neither the DNA strand that was targeted nor the sgRNA GC content across a broad range strongly correlated with sgRNA activity (Figure S2C and S2E).

To evaluate the feasibility of genome-scale genetic screens based on CRISPRi, we compared the strength of phenotypes obtained with CRISPRi to our previously published shRNA data. We applied the rules described above to data from our sgRNA tiling library to select all sgRNAs predicted to be highly active and then randomly subsampled sets of 10 or 24 sgRNAs. We calculated a normalized phenotype z-score by dividing mean phenotypes for each gene by the standard deviation of sgRNA phenotypes from the non-targeting control set (Figure S1B). We see significant ricin phenotypes for each of the 49 genes. Moreover, in virtually every case the normalized ricin phenotype z-score or p-value is stronger (in many cases far stronger) than

seen with a comparably-sized shRNA library (generated by sub-sampling our published data) (Figure 1E and Figure S2F).

CRISPRi Transcriptional Silencing is Highly Sensitive to Mismatches between the Target DNA Site and the sgRNA

To assess CRISPRi off-target activity at endogenous genes, we selected a set of 30 sgRNAs from our tiling library (6 sgRNAs/gene targeting 5 genes). For each of these sgRNAs, we tested the activity of a series of derivative sgRNAs with a variable number and position of mismatches (Figure 2). This experiment allowed us to measure the relative amount of gene repression for sgRNAs with or without mismatch base pairing targeting the same DNA locus. We found that even a single mismatch at the 3' end of the protospacer decreased CRISPRi activity on average, while combinations of mismatches that pass our off-target filter abolished activity (Figure 2, Figure S3, and Extended Experimental Procedures). From this analysis, we concluded that properly designed CRISPRi sgRNAs have minimal off-target transcriptional repression activity.

A High-Throughput Tiling Screen Defines Rules for CRISPRa Activity at Endogenous Genes

We recently developed an improved CRISPRa method, termed sunCas9, in which expression of a single sgRNA with one binding site is sufficient to robustly activate transcription (Tanenbaum et al.) In the sunCas9 system, a single dCas9 fusion protein bound to DNA recruits multiple copies of the activating effector domain, thus amplifying our ability to induce transcription (Figure 3A).

To define rules for optimal CRISPRa sgRNA design, we used our tiling library, which targets genes capable of modulating cellular sensitivity to ricin. We previously showed for several of the genes in this tiling library that knockdown and plasmid overexpression resulted in opposite ricin phenotypes (Bassik et al., 2013). For example, knockdown of *SEC23B* sensitized cells to ricin, whereas *SEC23B* overexpression desensitized cells to ricin. These observations suggested that we should be able to observe reversed phenotypes in this tiling screen arising from CRISPRa activity.

We transduced K562 cells stably expressing the sunCas9 system (Figure 3A) with the sgRNA tiling library and screened for ricin phenotypes as described for CRISPRi above. Analysis of data for individual genes or averaged data for all 49 genes demonstrated that many sgRNAs for each gene affected ricin resistance (Figure 3B and Figure S4A-B). Our negative control sgRNAs showed very little activity and were not correlated between biological replicate screens, suggesting that CRISPRa activity is specific. We observed a peak of active sgRNAs for CRISPRa at -400 to -50 bp upstream from the TSS (Figure 3B). This activity pattern fits with a model in which each VP16 domain can bind the mediator complex and recruit basal transcription machinery, activating transcription when spaced appropriately from a TSS (Mittler et al., 2003). With this system, we have shown we can turn on genes that are poorly expressed and increase the expression of well-expressed genes (Figure 3E). Overall, our CRISPRi/a tiling screens provide rules for how CRISPRi/a controls expression of endogenous genes.

An Allelic CRISPRi/a Series of Transcript Activation and Repression Shows that Protein Abundance Dynamically Modulates the Cellular Response to Ricin

For many genes, we do not know how the relative abundance of the encoded protein relates to its function. We observed a marked anti-correlation in our ricin screens between CRISPRa phenotypes and CRISPRi phenotypes for individual genes (Figure 3C). As the genes targeted by our tiling library were selected based on a knockdown phenotype, all genes showed phenotypes in the CRISPRi screen, but only a subset showed phenotypes in the CRISPRa screen.

To validate results from both the CRISPRi and CRISPRa tiling screens, we selected an allelic series of sgRNAs by phenotype from the screen and re-tested each sgRNA individually (38 sgRNAs targeting 4 genes). Our results show that our CRISPRi/a screens produced reliable phenotype scores, robustly reproduced upon re-testing, and that CRISPRi/a can activate and repress the transcription of endogenous genes over a wide dynamic range (up to ~1000-fold) (Figure 3D-E), enabling systematic interrogation of how gene dosage controls cellular functions of interest.

A Robust and Highly Specific Genome-Scale CRISPRi Screening Platform

The results of our tiling CRISPRi screen established our ability to pick active sgRNAs with low off-target activity and provided a set of rules enabling us to design a robust genome-scale sgRNA library. We chose a library size of 10 sgRNAs/gene for the following reasons. Over half of the sgRNAs conforming to these rules gave clear ricin phenotypes. For a library with 10 sgRNAs/gene, 94% of the genes would thus have 2 or more highly active sgRNAs. Finally, computational sub-sampling of the phenotypic data from our tiling library data to 10 sgRNAs/gene and calculation of z-scores for hit genes indicated that a library with 10 sgRNAs/gene would reliably detect hit genes (Figure 1E).

We synthesized and cloned a genome-scale CRISPRi sgRNA library targeting 15,977 human protein-coding genes (10 sgRNAs/TSS, targeting 20,898 TSS) with 11,219 non-targeting control sgRNAs for a total of 206,421 sgRNAs (Table S2). To evaluate this library, we first screened for genes essential for cell growth in K562 cells. Briefly, K562 cells stably expressing dCas9-KRAB were transduced in replicate with the entire genome-scale library, and each replicate was grown for 10 days at a minimum library coverage of 3,750 cells/sgRNA in a single spinner flask.

To characterize our screening methodology and library design, we examined the correlation between screen replicates. Individual sgRNAs reproducibly showed dramatic depletion (up to 256-fold) over a 10-day screen, demonstrating that individual sgRNAs can have profound effects on cell growth (Table S2) (Figure 4A). The distribution of our negative-control sgRNAs was very narrow with little correlation between replicates (Spearman R = 0.036), suggesting that the off-target activity of these controls is very low (Figure 4A). Indeed, 99.7% of our negative controls had no detectable activity. The observed specificity is consistent with our previously published RNA-seq data (Gilbert et al., 2013).

To further explore the prevalence of off-target effects, we examined two classes of genes not expected to show any on-target activity in our screen: olfactory receptors and genes on the Y chromosome. The sgRNAs that target these genes were designed and picked in the same manner as the rest of library; however, olfactory receptors should not be expressed in this cell type and, as K562 cells are derived from a female donor, sgRNAs that target genes on the Y chromosome lack a DNA target. As with the negative controls, these genes show no phenotype on average and exhibit very little correlation between replicates (Spearman R = 0.057 for olfactory genes and -0.052 for Y-targeting) (Figure 4A). We also observed no evidence of non-specific toxicity due to

expression of dCas9-KRAB and our sgRNA library in K562 cells, suggesting that dCas9 bound to the genome is not toxic under these conditions (Figure 4B). Thus, CRISPRi is highly specific and non-toxic.

To identify hit genes in this screen, we used a metric of average growth phenotype (γ) for the top three sgRNAs for each gene (see Experimental Procedures and Table S3). Among the top hits were genes involved in essential cellular functions, including translation, transcription and DNA replication (Figure 4C and Figure S5A) (Huang et al., 2009a, 2009b), thus validating our approach as a screening platform.

A Genome-Scale CRISPRa Screening Platform

The results of our CRISPRa tiling screen established our ability to confidently measure gene phenotypes resulting from inducing expression with single sgRNAs. As with the CRISPRi tiling screens, our data enabled the development of a set of rules that allowed construction of a genome-scale CRISPRa library. Many of these rules overlapped with those of CRISPRi (e.g., sgRNA length and sequence preferences). A key difference is that the optimal window for targeting sgRNAs for CRISPRa lies upstream of the TSS (Figure 3B). We therefore constructed an independent CRISPRa library, designing 10 sgRNAs between 400 to 50 base pairs upstream of each TSS for 15,977 human genes, along with 5,968 non-targeting control sgRNAs, for a total of 198,810 sgRNAs.

We evaluated our CRISPRa platform in a screen for genes that affect cell growth when induced in K562 cells constitutively expressing the sunCas9 system. Replicate screens were conducted as described above. The magnitude of growth defects seen in our CRISPRa screen was comparable to that of the above CRISPRi screen, although fewer sgRNAs caused a growth

phenotype (Figure S5B and Table S2). We analyzed control sgRNAs with no genomic target or Y chromosome targets and found minimal phenotypes, which lacked substantial correlation between experimental replicates (Spearman R = 0.155 and R = 0.010, respectively), indicating that the phenotype distribution observed in non-targeting controls was primarily a result of stochastic noise rather than off-target effects. Furthermore, the fraction of cells expressing sgRNAs and the sunCas9 system was stable over the course of the experiment, indicating that there was no general toxicity associated with the CRISPRa platform (Figure 4B). These data suggest that, like CRISPRi, CRISPRa is specific and non-toxic.

Defining Regulators of Survival and Differentiation in Human Cells by CRISPRa

We then investigated the genes whose induction caused cells to deplete over the course of our CRISPRa screen. We scored genes by the average γ of the three most active sgRNAs as above, and compared these phenotypes to those observed in the CRISPRi screen (Figure 4D and Table S3). The results from the two screens had little overlap, suggesting that few genes are both essential and toxic upon induction, and that wild-type expression levels of genes are generally optimal for K562 growth. Whereas CRISPRi hits are naturally limited to expressed genes, CRISPRa hits included genes across a broad range of endogenous expression levels (Figure S5D). We observed that the majority of genes that inhibited growth in the CRISPRa screen fell into three overlapping classes.

The first class was tumor suppressor genes: 18 of the top 50 genes, including six of the top seven, are known to have potent tumor suppressor activity (Vogelstein et al., 2013; Zhao et al., 2013). These genes include p53-related protein *TP73*, cell cycle inhibitors *CDKN1C* (p57) and *CDKN1A* (p21), apoptotic factors *BAK1* and *BCL2L11* (BIM), and chromatin remodeling

factor *ARID1A* (Figure 4D and Table S4). Gene set enrichment analysis (GSEA) confirmed this observation, highlighting several genes important in the intrinsic pathway of apoptosis or in chronic myeloid leukemia (CML) homeostasis consistent with the origin of K562 cells as a clonal isolate from a CML blast crisis (ATCC) (Figure 4E). Similarly, top gene ontology annotations included "positive regulation of apoptosis" and "regulation of cell cycle" (Figure S5C). While tumor suppressors are classically considered to be mutated early in cancer progression (Vogelstein et al., 2013), these results demonstrate than many potential tumor suppressor genes remain functional but down-regulated, and suggest that CRISPRa can be used to pinpoint intact pathways and vulnerabilities in tumor cells.

Transcription factor families with well-established roles in tissue development and differentiation represent another class of growth hits, accounting for 16 of the top 50 genes (K562 cells have known potential to undergo erythroid differentiation). These genes include CCAAT/Enhancer-binding proteins (CEBP), Homeobox genes, Forkhead box genes, Ikaros family zinc finger proteins, and hematopoietic differentiation factor *SPI1* (PU.1) (Figure 4D and Table S4) (Spitz and Furlong, 2012). This observation is reflected in enriched annotations relating to multicellularity, cell differentiation, and development (Figure S5C).

The complementary nature of the CRISPRi and CRISPRa screens is nicely illustrated by results from two gene pairs (*SPI1*/*GATA1* and *CEBPA*/*CEBPG*) in which one member of each pair inhibits the function of the other. *GATA1* and *CEBPG* were strong hits in the CRISPRi screen, consistent with their roles as inhibitors of myeloid differentiation. By contrast, both *SPI1* and *CEBPA* were robust hits in our CRISPRa activation screen. These observations are consistent with the inhibitory functions of *SPI1* and *CEBPA*: silencing of *CEBPA* leads to de-

repression of *CEBPG* (Alberich-Jordà et al., 2012) and the protein encoded by *SPI1* (PU.1) is a direct binding partner of GATA-1 and inhibits its transcriptional activity (Zhang et al., 2000).

Finally, several hit genes have key roles in mitosis. *PLK4* controls centrosome duplication, and overexpression of the gene in U2OS cells leads to increased centriole number (Habedanck et al., 2005). The proteins encoded by *KIF18B* and *KIF2C* form a complex that destabilizes microtubules during mitosis (Tanenbaum et al., 2011).

Overall, the results from our paired CRISPRi/a growth screens demonstrate that complementary information can be obtained by loss- and gain-of-function genetic screens, and highlight the utility of the platform for future studies into tumor biology and cell differentiation.

Dynamically Controlling Gene Expression with CRISPRi

The ability to reversibly tune the expression of select transcripts would be a powerful tool for evaluating transcript function. To evaluate the applicability of CRISPRi to this purpose, we cloned a lentiviral expression construct that places an optimized KRAB-dCas9 fusion protein under the control of a doxycycline-inducible promoter (Figure 5A-B). Induced expression of KRAB-dCas9 robustly depletes transcript levels from sgRNA-targeted genes (Figure 5C and S5E). To further assess dynamic control of CRISPRi, we inducibly repressed several genes identified in our genome-scale CRISPRi growth screen (Figure S5G). Cells that express sgRNAs targeting these essential genes showed almost no growth phenotype in the absence of doxycycline, but rapidly and robustly disappeared from the population upon addition of doxycycline (Figure 5D). Additionally, gene repression and resulting phenotypes were reversible (Figure 5C and S5E-F), indicating that KRAB-dCas9 does not create a permanently repressive chromatin state at targeted promoters.

To test our ability to dynamically control expression of essential genes on a larger scale, we cloned a sublibrary targeting 426 manually curated genes (10 sgRNAs / TSS or 4,923 targeting sgRNAs plus 750 non-targeting controls). This library was transduced into K562 cells stably expressing our inducible KRAB-dCas9 fusion protein, and cell growth effects were then evaluated in the presence and absence of doxycycline. Only 4 sgRNAs were depleted strongly in the absence of doxycycline; however, with induction of KRAB-dCas9, many sgRNAs were strongly depleted (Figure 5E). Negative control sgRNAs again produced a narrow distribution of phenotypes with little correlation between biological replicates with or without doxycycline. Additionally, we found no evidence that targeted KRAB-dCas9 generally impedes cell growth (Figure 5F). Taken together, these results demonstrate CRISPRi is non-toxic, inducible and reversible.

A Genome-Scale CRISPRi Screen Reveals Pathways and Complexes that Govern Response to Cholera and Diphtheria Toxin

To test the performance of our CRISPRi approach for detecting genes controlling a more complex cellular phenotype, we performed a genome-scale CRISPRi screen for genes that modulate sensitivity to a chimeric toxin composed of the diphtheria toxin catalytic A subunit covalently linked to cholera toxin (CTx-DTA, Figure 6A). This chimera had been previously developed to provide a growth readout for cholera intoxication (Guimaraes et al., 2011). Some aspects of both cholera and diphtheria toxin entry and toxicity are well characterized, but open questions remain. The cell surface receptor for cholera toxin is the GM1a ganglioside (Van Ness et al., 1980). After endocytosis, the toxin traffics via the Golgi to the endoplasmic reticulum (ER), from which it retro-translocates into the cytosol, possibly through the ER-associated

degradation (ERAD) machinery. Once in the cytosol, the DTA moiety ADP-ribosylates the diphthamide residue in Elongation Factor 2, halting translation and killing the cell (Figure 6A).

K562 cells expressing the CRISPRi sgRNA library and dCas9-KRAB were either grown under standard conditions or treated with several pulses of CTx-DTA over the course of 10 days. We observed highly correlated enrichment and depletion of many sgRNAs between replicates, indicating that CRISPRi can identify genes that modulate both resistance and sensitivity to a selective pressure (Table S2).

We ranked genes by the average phenotype of their three strongest sgRNAs (Table S3, Figure 6B, and Figure S6). GSEA revealed that KEGG pathways enriched for top protective hit genes were "Infection with *Vibrio cholerae*" and "Glycosphingolipid biosynthesis, ganglio-series" (Figure S7B), while gene sets for top sensitizing genes included "ribosome" and "proteasome" (Figure S7B). Since the diphtheria toxin catalytic subunit inhibits translation, depletion of the ribosome can be expected to sensitize cells to the toxin. Disruption of the proteasome also sensitizes cells to CTx-DTA, suggesting that the cytosolic toxin is a substrate for proteasomal degradation. Taken together, the unbiased GSEA analysis provides support for the high specificity in hit gene identification by our CRISPRi approach.

We further defined the 50 hits with the strongest protective effect and the 50 hits with the strongest sensitizing effect as "top hits" (all of these are far outside of the range seen with otherwise matched negative control sgRNAs). We characterized these genes by assigning them to cellular pathways and protein complexes according to their previously characterized roles (Figure 6B and Figure S6). Our CRISPRi screen identified a protective effect of knockdown for all top hits recovered in the previously published haploid mutagenesis screen (white stars in Figure 6B). The two top pathways identified by haploid mutagenesis as modulating cellular

sensitivity to CTx-DTA are the diphthamide biosynthetic pathway (required to generate eEF-2-diphthamide, the target of diphtheria toxin) and the ganglioside biosynthetic pathway (required to produce GM1a, the cell-surface receptor for cholera toxin). Our screen also identified many additional core components of each pathway. While knockdown of all hits in the diphthamide biosynthesis pathway had a protective effect, the results for ganglioside biosynthesis genes showed a more complex pattern: knockdown of enzymes involved in the production of GM1a were protective, whereas knockdown of enzymes that catalyze the production of other gangliosides (including GM1b) was sensitizing. These results provide genetic confirmation that GM1a is the relevant cell-surface receptor for CTx-DTA and more broadly illustrate the value of being able to reliably detect both sensitizing and protective genes to dissect biological pathways.

Many of the top hits are components of cellular pathways and protein complexes previously identified in experiments to be important for retrograde trafficking and retro-translocation of other toxins such as ricin and Shiga toxin (Bassik et al., 2013; Smith et al., 2009). Retro-translocation of the catalytic chain of CTx has been proposed to be mediated by the ER-associated degradation (ERAD) pathway, although this pathway was not identified in previous genetic screens. Consistent with this proposed role for the ERAD machinery, knockdown of members of the ERAD E3 ubiquitin ligase complex, *SYVN1* (encoding Hrd1) and *SEL1L* (the mammalian homolog of yeast Hrd3) rendered cells resistant to CTx-DTA. Factors that mediate cytosolic degradation of ERAD substrates (in particular *UBXN4*, also known as UBXD2 or erasin, and the proteasome) appeared as sensitizing hits, suggesting that they may reduce cytosolic levels of the toxin's catalytic subunit in WT cells.

To validate the suggested role of the identified ERAD factors in toxin retro-translocation from the ER to the cytosol, we quantified the amount of CTx chains in the cytosol and membrane

fractions. As expected, *SEL1L* knockdown resulted in a dramatic reduction of cytosolic CTx-A1, whereas levels in the membrane fraction were much less affected (Figure 7A-C). By contrast, knockdown of *B4GALNT1*, an enzyme required for the synthesis of the CTx receptor GM1a, resulted in a nearly complete absence of CTx chains from both the cytosolic and the membrane fraction (Figure 7A-C).

An open question in CTx biology is how the toxin traverses the Golgi network (Wernick et al., 2010). Our screen revealed that COG and GARP complexes, which tether late endosomes to the trans-Golgi network or modulate intra-Golgi retrograde transport (Bonifacino and Rojas, 2006) are critical host factors for CTx-DTA. These and other complexes and pathways we identify here (Figure 6B), including several involved in RNA processing, had not previously been linked to cholera toxin biology, highlighting the potential of CRISPRi as a discovery platform. Importantly, many top hits—even those not previously implicated in cholera or diphtheria pathogenesis—were tightly clustered in well-defined protein complexes and pathways. For several of these, the vast majority of components were hits, suggesting that CRISPRi screens can approach saturation.

Potent Phenotypes and Knockdown Levels Achieved by the Genome-Scale CRISPRi Library

To validate the results from this screen, we re-tested sgRNAs that putatively modulate cellular response to CTx-DTA in mechanistically diverse ways. For each sgRNA, we quantified both the ricin phenotypes as well as the change in abundance of the targeted transcript by qPCR. Our re-test experiments were highly correlated with data from the primary screen (Figure 7D). In our validation experiments for the tiling ricin screen and the genome-scale CTx-DTA screen, the

activities of 71 out of 72 sgRNAs were robustly confirmed and were highly correlated (R^2=0.879) with the results obtained in the primary screen. Finally, analysis of mRNA levels by qPCR data showed robust repression, with ~80-99% knockdown for each sgRNA and at least 90% for every gene (Figure 7E).

A Genome-Scale CRISPRa Screen of Cholera-Diphtheria Toxin Complements and Extends CRISPRi Results

To further explore the biological insights gained from CRISPRa screening, we performed a genome scale CRISPRa screen for genes that modulate sensitivity to CTx-DTA (Table S2-3). As with the CRISPRi screen, GSEA revealed the specificity of the detected hits (Figure S7B). For some of the top hits, CRISPR-mediated transcriptional activation and repression caused opposite phenotypes (e.g. enzymes in ganglioside biosynthesis, Figure 6C), similar to what we observed for genes controlling ricin sensitivity (Figure 3C).

CRISPRa also revealed additional and highly complementary information as illustrated by analysis of glycosphingolipid biosynthesis pathways. Induction of enzymes in the neolacto branch of sphingolipid biosynthesis protected cells from CTx-DTA (Figure 6C and Figure S7A,C). This pathway is a parallel branch to the ganglioside branch, which produces the CTx-DTA receptor GM1a. Our findings suggest that upregulation of the neolacto branch diverts the common precursor lactosylceramide away from the ganglioside branch. Similarly, upregulation of the sulfatide-generating enzyme *GAL3ST1* has a protective effect, presumably by diverting ceramide from the sphingolipid to the cerebroside pathway (Figure 6C). These results highlight the capacity of CRISPRa to complement CRISPRi by querying the consequences of upregulating pathways that may otherwise be inactive.

Effective Knockdown of non-coding RNAs

Finally, we investigated whether CRISPRi was able to repress the transcription of long non-coding RNAs (lncRNAs), a class of transcripts that have been difficult to systematically perturb by other methods (Bassett et al., 2014). Using our CRISPRi library design algorithm, we selected and cloned up to three sgRNAs each targeting six characterized lncRNAs (*GAS5*, *H19*, *MALAT1*, *NEAT1*, *TERC*, *XIST*) (Geisler and Coller, 2013) with good evidence of expression in K562 cells. We transduced the sgRNAs into cells expressing dCas9-KRAB and quantified the amount of transcript knockdown by qPCR. We achieved >80% knockdown for all but one of the lncRNA genes tested (Figure 7F). Overall, more than 50% of the sgRNAs yielded >85% knockdown. We confirmed the strong repression of *XIST* by RNA fluorescence *in situ* hybridization (FISH) and observed no residual expression along the X chromosome (Figure 7G). These results demonstrate that CRISPRi can effectively repress lncRNA expression, enabling future systematic studies of non-coding gene function.

Discussion

Here, we establish CRISPRi and CRISPRa as robust tools for systematically manipulating transcription of endogenous genes in human cells. We demonstrate that CRISPRi/a can be used to rapidly screen for both loss-of-function and gain-of-function phenotypes in a pooled format. We identify both known and unexpected genes that control growth of K562 cells or that modulate sensitivity to a toxin (CTx-DTA). We also show that we can use CRISPRi/a to create allelic series of gene expression, spanning a broad range from ~100-fold repression to ~10-fold induction, allowing us to define how the abundance of a protein or transcript relates to its function.

A key feature of CRISPRi is the low incidence of off-target effects, as evidenced by the near-absence of activity for three large and distinct classes of negative control sgRNAs in our genome-scale CRISPRi library. This feature simplifies validation and interpretation of screening results. The observed specificity likely stems from two distinct properties of our system. First, CRISPRi/a complexes bound outside a narrow window around the TSS largely fail to modulate transcription; this dramatically shrinks the sequence space across the genome where off-target binding could produce significant off-target activity. Second, CRISPRi activity is highly sensitive to mismatches between the sgRNA and target DNA, suggesting that off-target binding of dCas9 observed in ChIP-seq experiments is too transient to impact transcription (Duan et al., 2014; Kuscu et al., 2014; Wu et al., 2014).

CRISPRa screening provides a new approach for exploring the diversity of transcripts across complex genomes. Gene activation has been used to dissect the limiting component of a biochemical process, identify the molecular target of a drug, or activate key rate-limiting steps in a pathway (Davis et al., 1987; Rine et al., 1983). Recently, a combinatorial cDNA

overexpression screen identified genes that, when co-expressed, reprogram fibroblasts into pluripotent stem cells (Takahashi and Yamanaka, 2006). CRISPRa should greatly accelerate similar searches for combinations of factors with emergent properties. In addition, CRISPRa will likely provide insight into cellular pathways where redundancy hampers loss-of-function genetic approaches. CRISPRa will also enable the exploration of cellular states in which otherwise inactive pathways are induced, and thereby reveal functional coupling within complex cellular networks and suggest potential therapeutic strategies.

Our ability to control transcription with high specificity simplifies the analysis and validation of high-throughput screening data. The genome-scale CRISPRi and CRISPRa libraries described here contain 10 sgRNAs per TSS. The resulting library size allows each to be screened in a population of 200 million cells, which can be easily grown in a single spinner flask. Furthermore, the observed high specificity and an improved understanding of rules governing sgRNA activity should enable us to create even more compact sgRNA libraries. Additionally, an sgRNA library designed to activate or repress a broader range of transcripts in the human genome could reveal the function of many non-canonical RNAs encoded in the human genome. As most non-coding transcripts are nuclear and lack an open reading frame, methods that directly modulate transcription are optimally suited for interrogating the function of these RNAs (Derrien et al., 2012).

Systematic genetic interaction (GI) maps are powerful tools for revealing gene functions within pathways or complexes (Bassik et al., 2013; Boone et al., 2007; Costanzo et al., 2010). A CRISPRa GI map or a combined CRISPRi/a GI map could yield rich novel biology and help elucidate how networks of proteins dictate cellular function. More generally, quantitative methods of turning on and off one or multiple transcripts represent a critical tool for

understanding how expression of the genes encoded in our genomes controls cell function and fate.

Experimental Procedures

CRISPRi/a Libraries

Tiling libraries sgRNAs were designed targeting 49 genes (see Figure 1E) previously identified in shRNA screens as having ricin resistance phenotype. All possible sgRNAs within a 10kb window around the gene TSS and meeting certain criteria were included (see Extended Experimental Procedures). Negative controls were designed based on scrambled sequences from these 10kb windows and filtered by the same criteria as targeting sgRNAs.

Genome-scale CRISPRi/a libraries Genes were selected from the entire set of protein coding genes, although a subset of genes with a RPKM of 0 in a K562 cell RNA-seq expression data set were excluded. sgRNAs conforming to rules including low predicted off-targets and minimal length (see Figure S2 and Extended Experimental Procedures) were selected from a window of -50 to +300bp (CRISPRi) or -400 to -50bp (CRISPRa) with respect to the TSS. Negative controls we designed in the same way based on scrambled sequence derived from the same window of several hundred genes.

Library Cloning Oligonucleotides encoding sgRNAs designed as described above were synthesized as pooled libraries. These were then cloned into lentiviral vectors for expression from a U6 promoter (see Extended Experimental Procedures).

Cell Line Construction

For constitutive and inducible CRISPRi screens, polyclonal cells expressing dCas9/KRAB fusion proteins driven from an SFFV or TRE3G promoter, respectively, were generated by viral

transduction. For CRISPRa screens, a clonal cell line expressing dCas9-SunTag and a scFV-sfGFP-VP64 fusion was generated (See Extended Experimental Procedures).

Growth and Toxin Screens

Cells were grown at minimum library coverage of 1,000 for tiling screens and 3,750 for genome-scale screens. For growth screens cells were grown in spinner flasks and harvested at 0 and 10 days after puromycin selection. For toxin screens, cells were treated with pulses of ricin or CTx-DTA (Bassik et al., 2013; Guimaraes et al., 2011) and harvested when sufficient selective pressure relative to untreated cells had been applied. Briefly, DNA was isolated, the cassette encoding the sgRNA was amplified by PCR, and relative sgRNA abundance was determined by next generation sequencing as previously described (Bassik et al., 2013; Kampmann et al., 2014).

Acknowledgments

The authors thank M. Tanenbaum, J. Tsai, K. Kostova, J. Zalatan, W. Lim, S. Weissman, J. Doudna, and R. Vale for unpublished reagents, technical advice and helpful discussion. L.S.Q. acknowledges support from the UCSF Center for Systems and Synthetic Biology. This work was supported by NIH P50 GM102706 (J.S.W.), NIH P50 GM081879 (L.S.Q., E.H.W.), NIH U01 CA168370 and NIH R01 DA036858 (J.S.W.), as well as the Howard Hughes Medical Institute (J.E.V., Y.C., M.K., M.A.H., J.S.W.). B.A. is an HHMI Fellow of the Damon Runyon Cancer Research Foundation (DRG-[2182-14]). L.A.G. is a Fellow of the Leukemia and Lymphoma Society. M.A.H. is supported by the UCSF Medical Scientist Training Program. M.K. is supported by NCI/NIH Pathway to Independence Award K99CA181494.

Figures

Figure 1

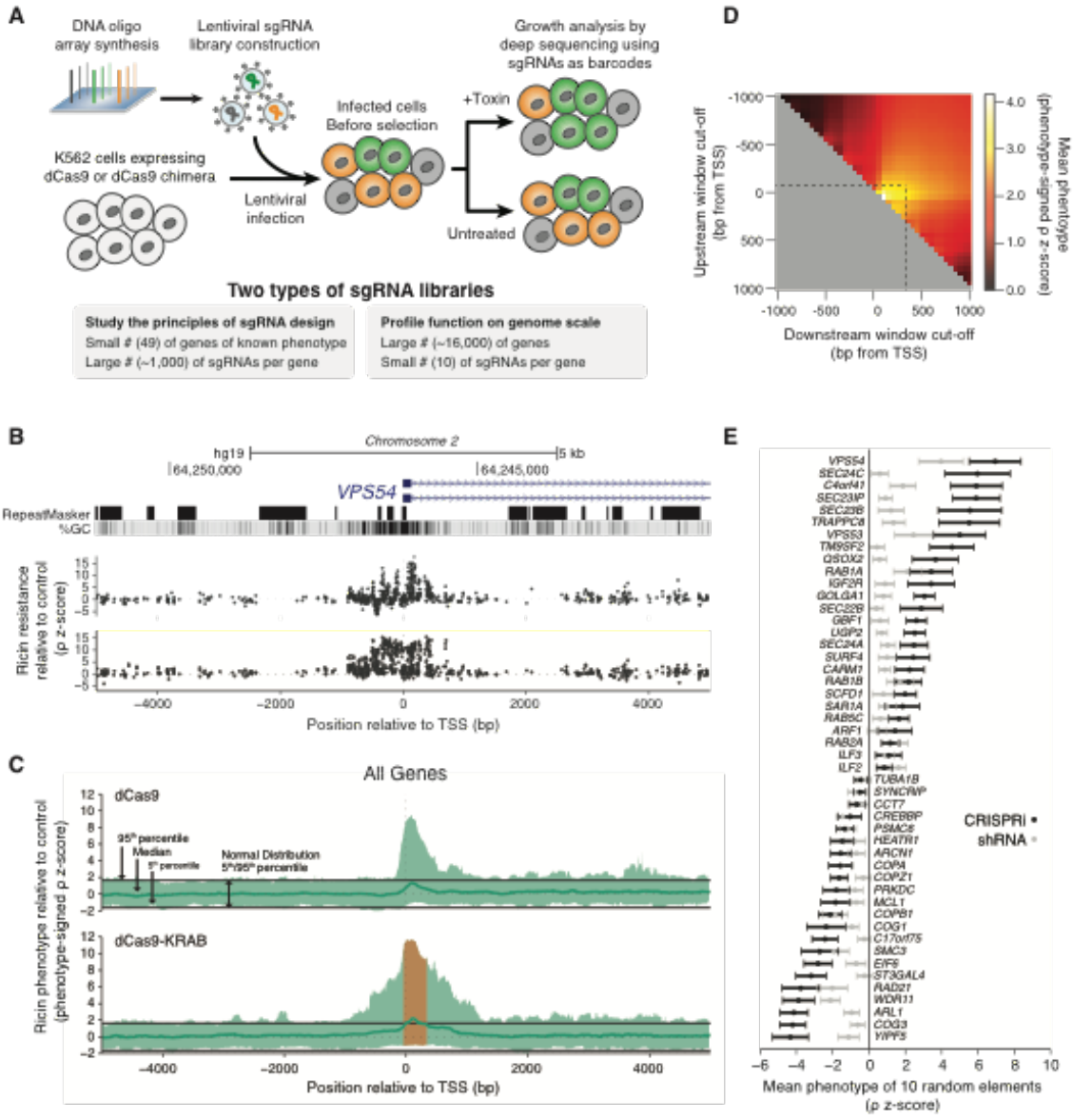


Figure 1. A Tiling sgRNA Screen Defines Rules for CRISPRi Activity at Endogenous Genes in Human Cells

(A) Massively parallel determination of growth or toxin resistance phenotypes caused by sgRNAs in mammalian cells expressing dCas9 or dCas9 fusion constructs. **(B)** UCSC genome browser tracks showing the genomic organization, GC content, and repetitive elements around the TSS of a representative gene, *VPS54*, across a 10kb window targeted by the tiling sgRNA library. sgRNA ricin resistance phenotypes (as z-scores, see Figure S1 and Experimental Procedures) in dCas9 and dCas9-KRAB expressing K562 cells are depicted in black on the top and bottom, respectively. See also Figure S2A for more examples. **(C)** Sliding-window analysis of all 49 genes targeted in a tiling sgRNA library. Green line: median sgRNA activity in a defined window for all genes. Orange region: observed average window of maximum CRISPRi activity. Data displayed as a phenotype signed z-score, excluding all guides longer than 22bp. **(D)** CRISPRi activity for all 49 genes in defined windows relative to the TSS of each gene. **(E)** Ricin resistance phenotypes, comparing CRISPRi sgRNAs selected by our rules to RNAi, for genes previously established to cause ricin resistance phenotypes when knocked down by RNAi. Mean ± SD phenotype-signed z-score of 100 sets of 10 randomly subsampled sgRNAs or shRNAs. See also Figure S2F.

Figure 2

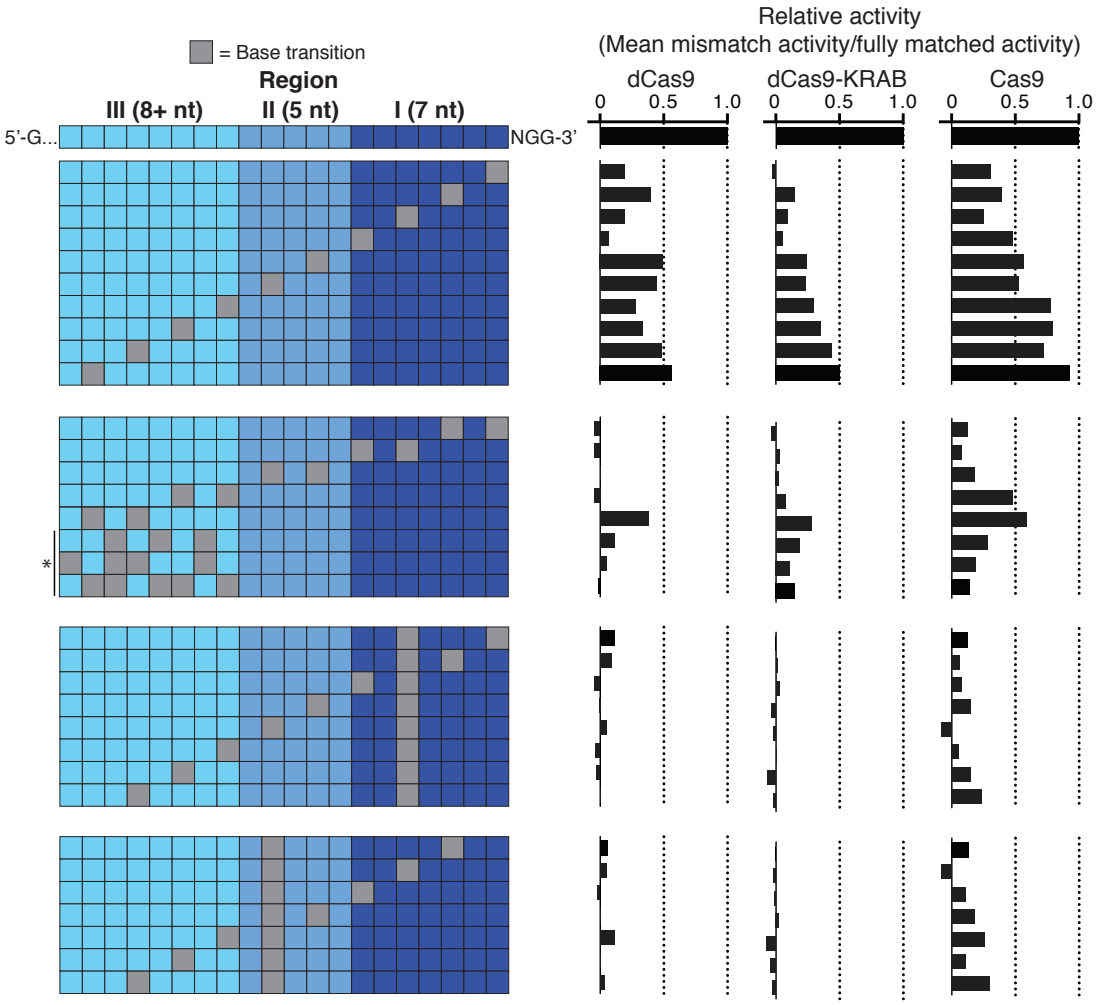

= Base transition
Region
III (8+ nt)
II (5 nt)
I (7 nt)
5'-G...
NGG-3'
*
Relative activity
(Mean mismatch activity/fully matched activity)
dCas9
dCas9-KRAB
Cas9
0
0.5
1.0

Figure 2. CRISPRi Activity is Highly Sensitive to Mismatches Between the sgRNA and DNA sequence

On- and off-target activity of dCas9, dCas9-KRAB and Cas9 for sgRNAs with a varying number and position of mismatches. Off-target activity of sgRNAs with mismatches is displayed as percent of the on-target activity for the corresponding sgRNA without mismatches. Asterisk indicates sgRNAs with 3, 4, or 5 mismatches randomly distributed across region 3 of the sgRNA sequence. Data is displayed for each mismatch position as the mean of all sgRNAs with that mismatch; see Figure S3 for individual sgRNA activities. sgRNAs were included in the analysis only if the fully matched guide was highly active (phenotype-signed z-score ≥ 4); N=5 for dCas9, N=11 for dCas9-KRAB, and N=10 for Cas9.

Figure 3

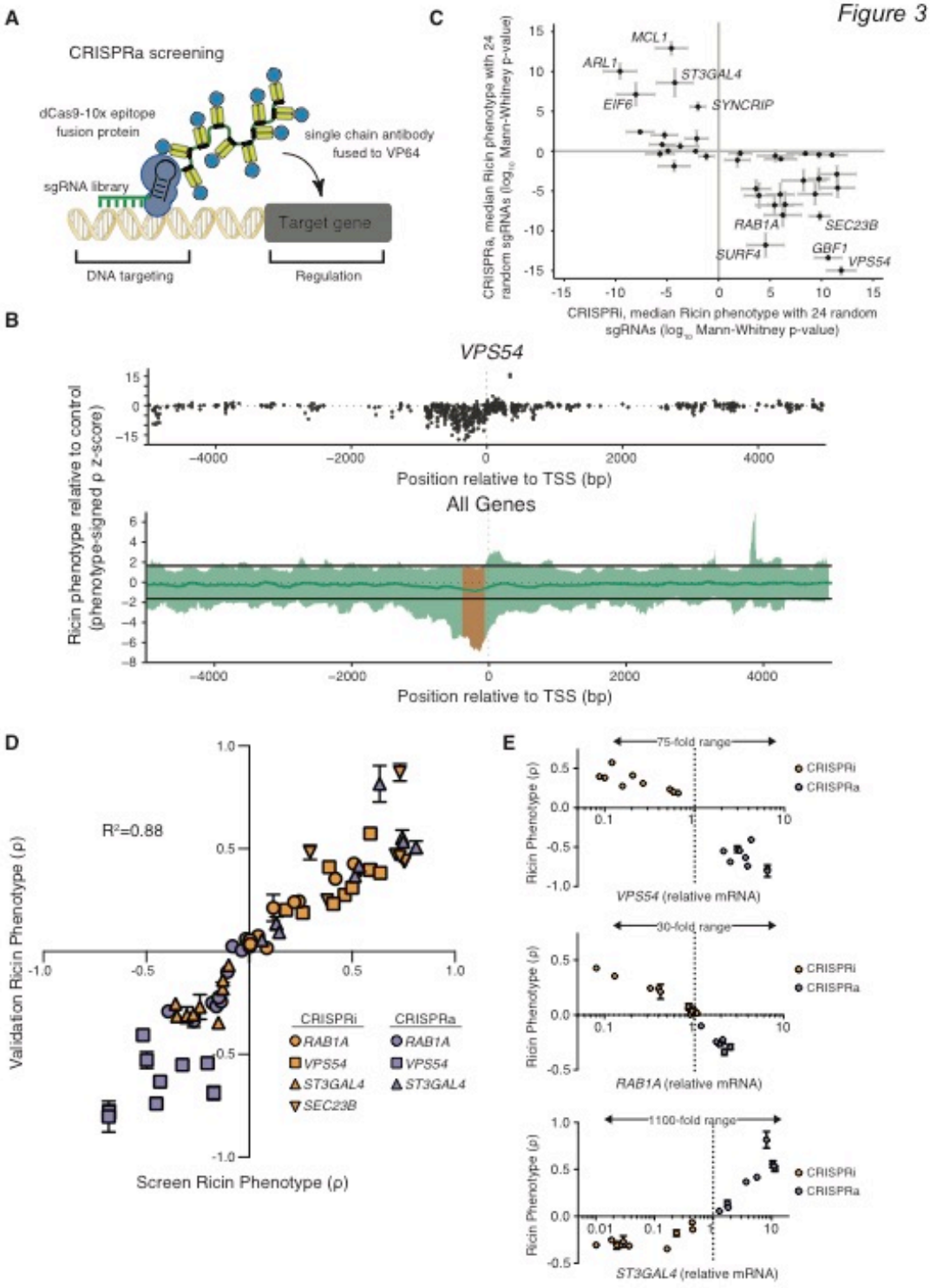

A
CRISPRa screening
dCas9-10x epitope fusion protein
single chain antibody fused to VP64
sgRNA library
Target gene
DNA targeting
Regulation
C
MCL1
ARL1
ST3GAL4
EIF6
SYNCRIP
RAB1A
SEC23B
SURF4
GBF1
VPS54
CRISPRa, median Ricin phenotype with 24 random sgRNAs (log10 Mann-Whitney p-value)
CRISPRi, median Ricin phenotype with 24 random sgRNAs (log10 Mann-Whitney p-value)
B
VPS54
Ricin phenotype relative to control (phenotype-signed ρ z-score)
Position relative to TSS (bp)
All Genes
Position relative to TSS (bp)
D
R²=0.88
Validation Ricin Phenotype (ρ)
Screen Ricin Phenotype (ρ)
CRISPRi
RAB1A
VPS54
ST3GAL4
SEC23B
CRISPRa
RAB1A
VPS54
ST3GAL4
E
75-fold range
Ricin Phenotype (ρ)
VPS54 (relative mRNA)
30-fold range
RAB1A (relative mRNA)
1100-fold range
ST3GAL4 (relative mRNA)
CRISPRi
CRISPRa

Figure 3. A Tiling sgRNA Screen Defines Rules for CRISPRa Activity at Endogenous Genes in Human Cells

(A) A schematic of the dCas9-SunTag + scFV-VP64 + sgRNA system for CRISPRa. **(B)** Activity of sgRNAs in K562 cells stably expressing each component of CRISPRa, as a function of the distance of the sgRNA site to the TSS of the targeted gene (Phenotype-signed z-scores; therefore, negative values represent opposite results than from knockdown). Top, sgRNAs targeting *VPS54*; Bottom, sliding-window analysis of all 49 genes targeted by our tiling library in green. Green line, median activity; orange, window of maximal activity. Guides longer than 22bp were excluded. See also Figure S4. **(C)** CRISPRa phenotypes and CRISPRi (dCas9-KRAB) phenotypes are anti-correlated for select genes. For each gene, a Mann-Whitney p-value is calculated using CRISPRi/a sgRNA activity relative to a negative control distribution for 24 sub-sampled sgRNAs. Mean ± SD p-value of 100 randomly sub-sampled sets is displayed. **(D)** CRISPRi knockdown and CRISPRa activation of the same gene can have opposing effects on ricin resistance in both primary screens and single sgRNA validation experiments (mean ± standard deviation of 3 replicates). **(E)** Modulation of expression levels for 3 genes by CRISPRi and CRISPRa as quantified by qPCR plotted against the ricin resistance phenotype (mean ± standard deviation of 3 replicates) measured for each sgRNA.

Figure 4

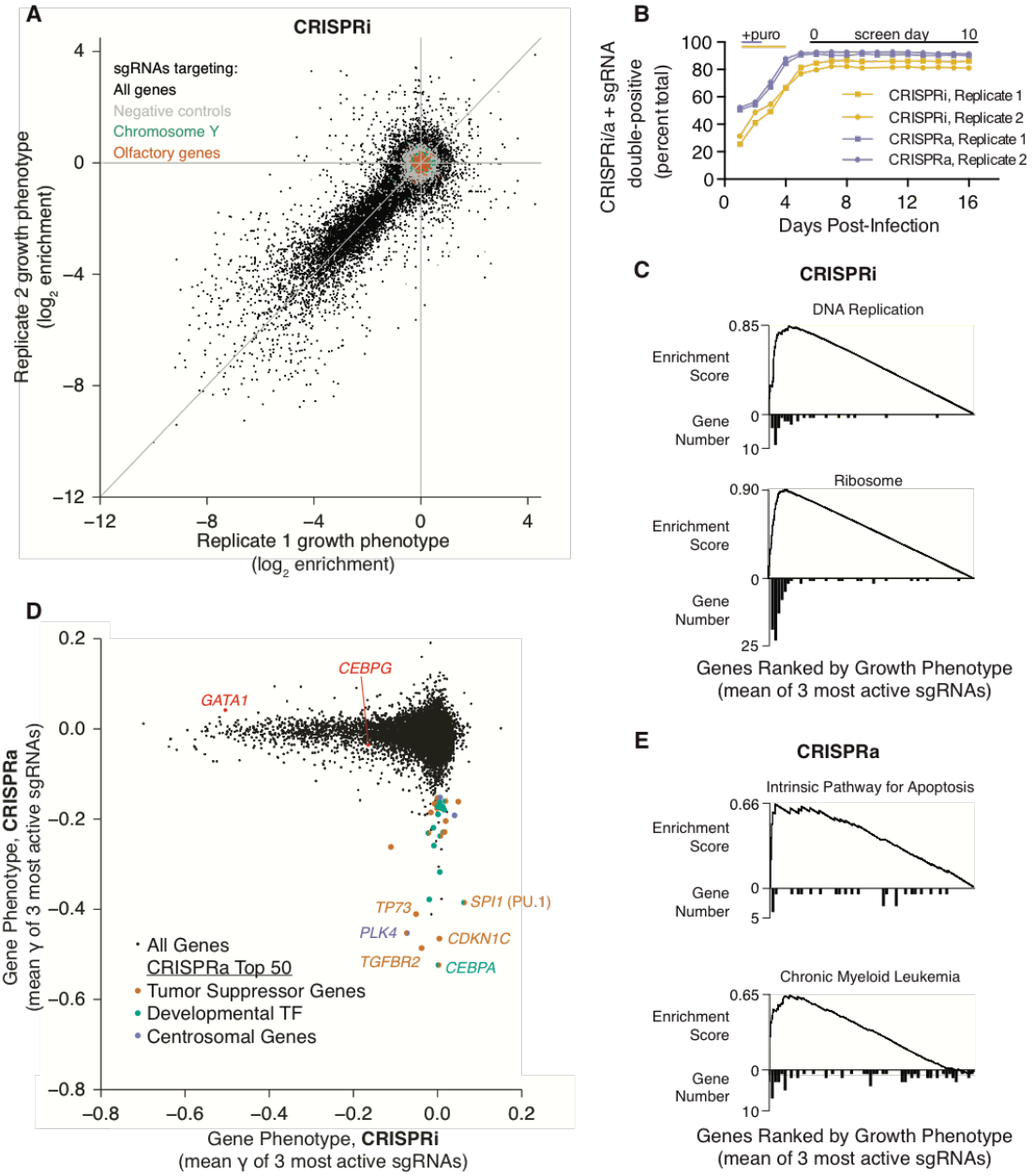

A
CRISPRi
sgRNAs targeting:
All genes
Negative controls
Chromosome Y
Olfactory genes
Replicate 2 growth phenotype (log2 enrichment)
Replicate 1 growth phenotype (log2 enrichment)
B
+puro
0
screen day
10
CRISPRi/a + sgRNA double-positive (percent total)
CRISPRi, Replicate 1
CRISPRi, Replicate 2
CRISPRa, Replicate 1
CRISPRa, Replicate 2
Days Post-Infection
C
CRISPRi
DNA Replication
Ribosome
Enrichment Score
Gene Number
Genes Ranked by Growth Phenotype (mean of 3 most active sgRNAs)
D
GATA1
CEBPG
TP73
SPI1 (PU.1)
PLK4
CDKN1C
TGFBR2
CEBPA
All Genes
CRISPRa Top 50
Tumor Suppressor Genes
Developmental TF
Centrosomal Genes
Gene Phenotype, CRISPRa (mean γ of 3 most active sgRNAs)
Gene Phenotype, CRISPRi (mean γ of 3 most active sgRNAs)
E
CRISPRa
Intrinsic Pathway for Apoptosis
Chronic Myeloid Leukemia
Enrichment Score
Gene Number
Genes Ranked by Growth Phenotype (mean of 3 most active sgRNAs)

Figure 4. Genome-Scale CRISPRi and CRISPRa Screens Reveal Genes Controlling Cell Growth

(A) sgRNA phenotypes from a genome-scale CRISPRi screen for growth in human K562 cells (black). Three classes of negative control sgRNAs are color-coded: non-targeting sgRNAs (grey), sgRNAs targeting Y-chromosomal genes (green) and sgRNAs targeting olfactory genes (orange). **(B)** Co-expression of sgRNAs and dCas9-KRAB or dCas9-SunTag + scFV-VP64 is not toxic in K562 cell lines over 16 days. **(C)** Gene set enrichment analysis (GSEA) for hits from the CRISPRi screen. A histogram of gene distribution is shown under the GSEA curve. **(D)** CRISPRi versus CRISPRa gene phenotypes for genome-scale growth screens (black). For the 50 genes in the CRISPRa screen with the most negative growth phenotype, each gene was annotated and labeled based on evidence of activity as a tumor suppressor (orange), developmental transcription factor (green), or in regulation of the centrosome (purple). Two additional CRISPRi hit genes that are discussed in the text are labeled in red. See Table S4 for annotations and references. **(E)** GSEA for hits from the CRISPRa growth screen. A histogram of gene distribution is shown under the GSEA curve.

Figure 5

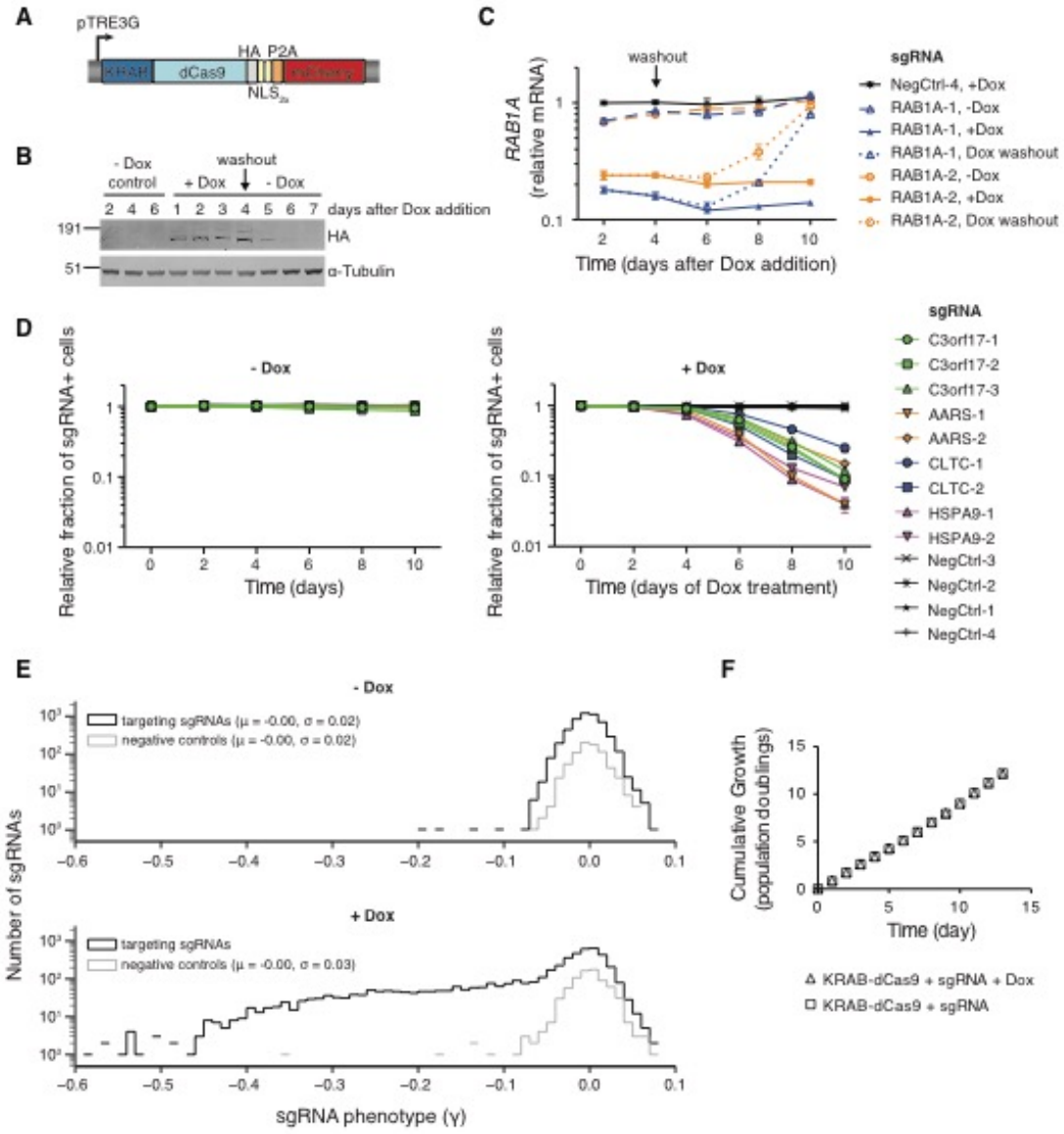

A
pTRE3G
HA P2A
KRAB
dCas9
NLS
B
- Dox control
washout
+ Dox
- Dox
2 4 6 1 2 3 4 5 6 7 days after Dox addition
191
HA
51
α-Tubulin
C
washout
RAB1A (relative mRNA)
Time (days after Dox addition)
sgRNA
NegCtrl-4, +Dox
RAB1A-1, -Dox
RAB1A-1, +Dox
RAB1A-1, Dox washout
RAB1A-2, -Dox
RAB1A-2, +Dox
RAB1A-2, Dox washout
D
- Dox
+ Dox
Relative fraction of sgRNA+ cells
Time (days)
Time (days of Dox treatment)
sgRNA
C3orf17-1
C3orf17-2
C3orf17-3
AARS-1
AARS-2
CLTC-1
CLTC-2
HSPA9-1
HSPA9-2
NegCtrl-3
NegCtrl-2
NegCtrl-1
NegCtrl-4
E
- Dox
targeting sgRNAs (μ = -0.00, σ = 0.02)
negative controls (μ = -0.00, σ = 0.02)
+ Dox
targeting sgRNAs
negative controls (μ = -0.00, σ = 0.03)
Number of sgRNAs
sgRNA phenotype (γ)
F
Cumulative Growth (population doublings)
Time (day)
KRAB-dCas9 + sgRNA + Dox
KRAB-dCas9 + sgRNA

Figure 5. CRISPRi Gene Silencing is Inducible, Reversible, and Non-Toxic

(A) Expression construct encoding an inducible KRAB-dCas9 fusion protein. **(B)** Western blot analysis of inducible KRAB-dCas9 in the absence, presence, and after washout of doxycycline. **(C)** Relative *RAB1A* expression levels (as quantified by qPCR) in inducible CRISPRi K562 cells transduced with *RAB1A*-targeting sgRNAs in the absence, presence, and after washout of doxycycline. Mean ± standard error of technical replicates (N=2) normalized to control cells (assayed in the presence of doxycycline) from the day 2 time point. **(D)** Competitive growth assays performed with inducible CRISPRi K562 cells transduced with the indicated sgRNAs in the presence and absence of doxycycline. Data is represented as the mean ± standard deviation of replicates (N=3). See also Figure S5G. **(E)** A CRISPRi sublibrary screen for effects on cell growth was performed with inducible CRISPRi K562 cells in the presence and absence of doxycycline. **(F)** Cumulative growth curves from the sublibrary screen represented in (E) show no bulk changes to growth caused by induction of KRAB-dCas9. Mean ± standard deviation of replicate infections each screened in duplicate.

Figure 6

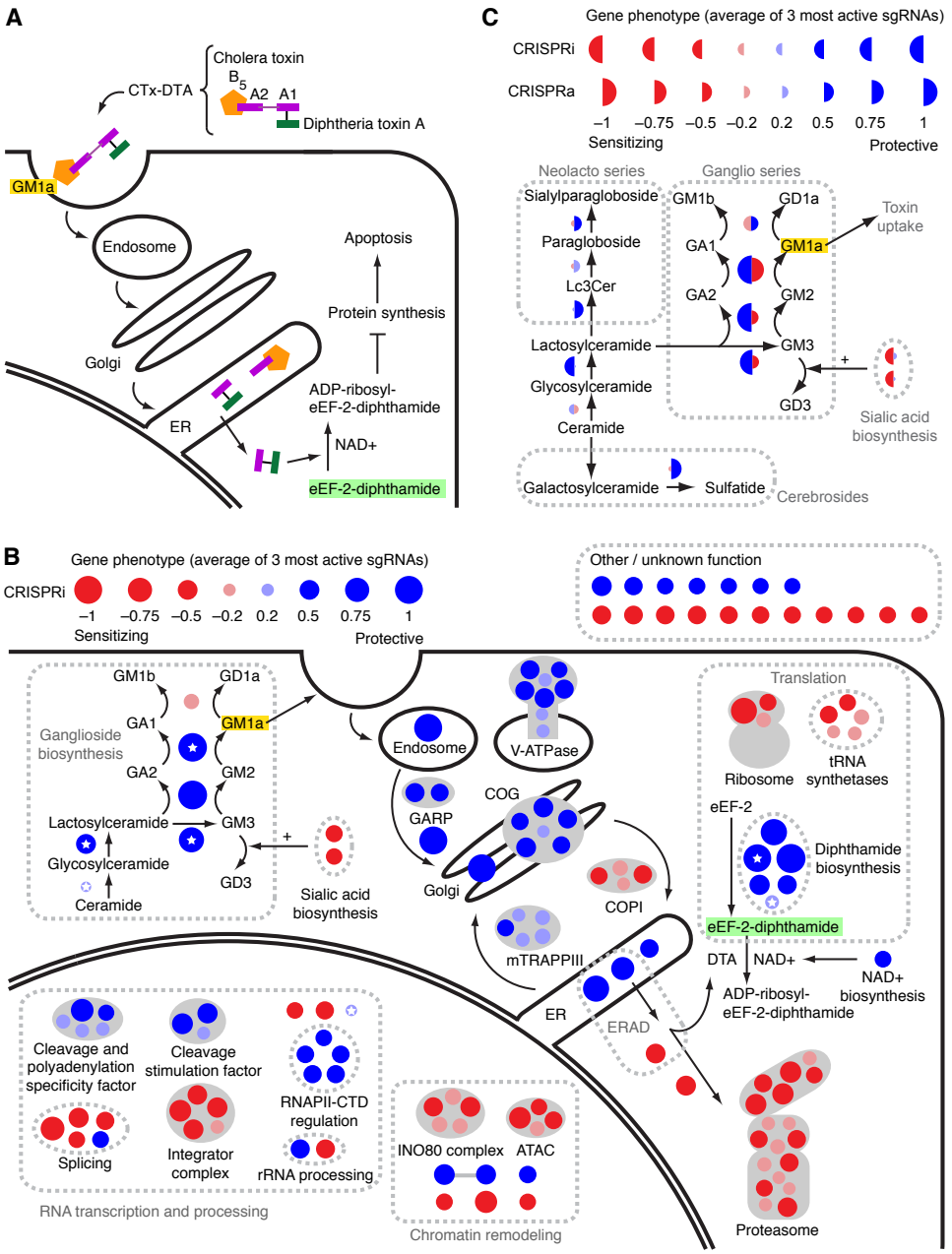

A
Cholera toxin
B5
CTx-DTA
A2
A1
Diphtheria toxin A
GM1a
Endosome
Apoptosis
Protein synthesis
Golgi
ADP-ribosyl-
eEF-2-diphthamide
ER
NAD+
eEF-2-diphthamide
C
Gene phenotype (average of 3 most active sgRNAs)
CRISPRi
CRISPRa
–1
–0.75
–0.5
–0.2
0.2
0.5
0.75
1
Sensitizing
Protective
Neolacto series
Ganglio series
Sialylparagloboside
Paragloboside
Lc3Cer
GM1b
GD1a
GA1
GM1a
GA2
GM2
GM3
GD3
Toxin
uptake
Lactosylceramide
Glycosylceramide
Ceramide
Sialic acid
biosynthesis
Galactosylceramide
Sulfatide
Cerebrosides
B
Gene phenotype (average of 3 most active sgRNAs)
CRISPRi
–1
–0.75
–0.5
–0.2
0.2
0.5
0.75
1
Sensitizing
Protective
Other / unknown function
GM1b
GD1a
GA1
GM1a
Ganglioside
biosynthesis
GA2
GM2
Lactosylceramide
GM3
Glycosylceramide
GD3
Ceramide
Sialic acid
biosynthesis
Endosome
V-ATPase
GARP
COG
Golgi
COPI
mTRAPPIII
ER
ERAD
Translation
Ribosome
tRNA
synthetases
eEF-2
Diphthamide
biosynthesis
eEF-2-diphthamide
DTA
NAD+
NAD+
biosynthesis
ADP-ribosyl-
eEF-2-diphthamide
Cleavage and
polyadenylation
specificity factor
Cleavage
stimulation factor
RNAPII-CTD
regulation
Splicing
Integrator
complex
rRNA processing
RNA transcription and processing
INO80 complex
ATAC
Chromatin remodeling
Proteasome

Figure 6. Genome-Scale CRISPRi and CRISPRa Screens Reveal Known and New Pathways and Complexes Governing the Response to a Cholera-Diphtheria Fusion Toxin (CTx-DTA)

(A) Model for CTx-DTA binding, retrograde trafficking, retro-translocation and cellular toxicity. **(B)** Overview of top hit genes detected by the CTx-DTA screen. Dark red and blue circles: Top 50 sensitizing and protective hits, respectively. Light red and blue circles: further hits that fall into the same protein complexes or pathways as top 50 hits. Circle area is proportional to phenotype strength. White stars denote genes identified in a previous haploid mutagenesis screen (Guimaraes et al., 2011). See also Figure S6 for hit gene names. (**C**) CRISPRi and CRISPRa hits in sphingolipid metabolism. Display as in (B), except that the left and right sides of each circle represent the phenotypes in the CRISPRi and CRISPRa screens, respectively.

Figure 7

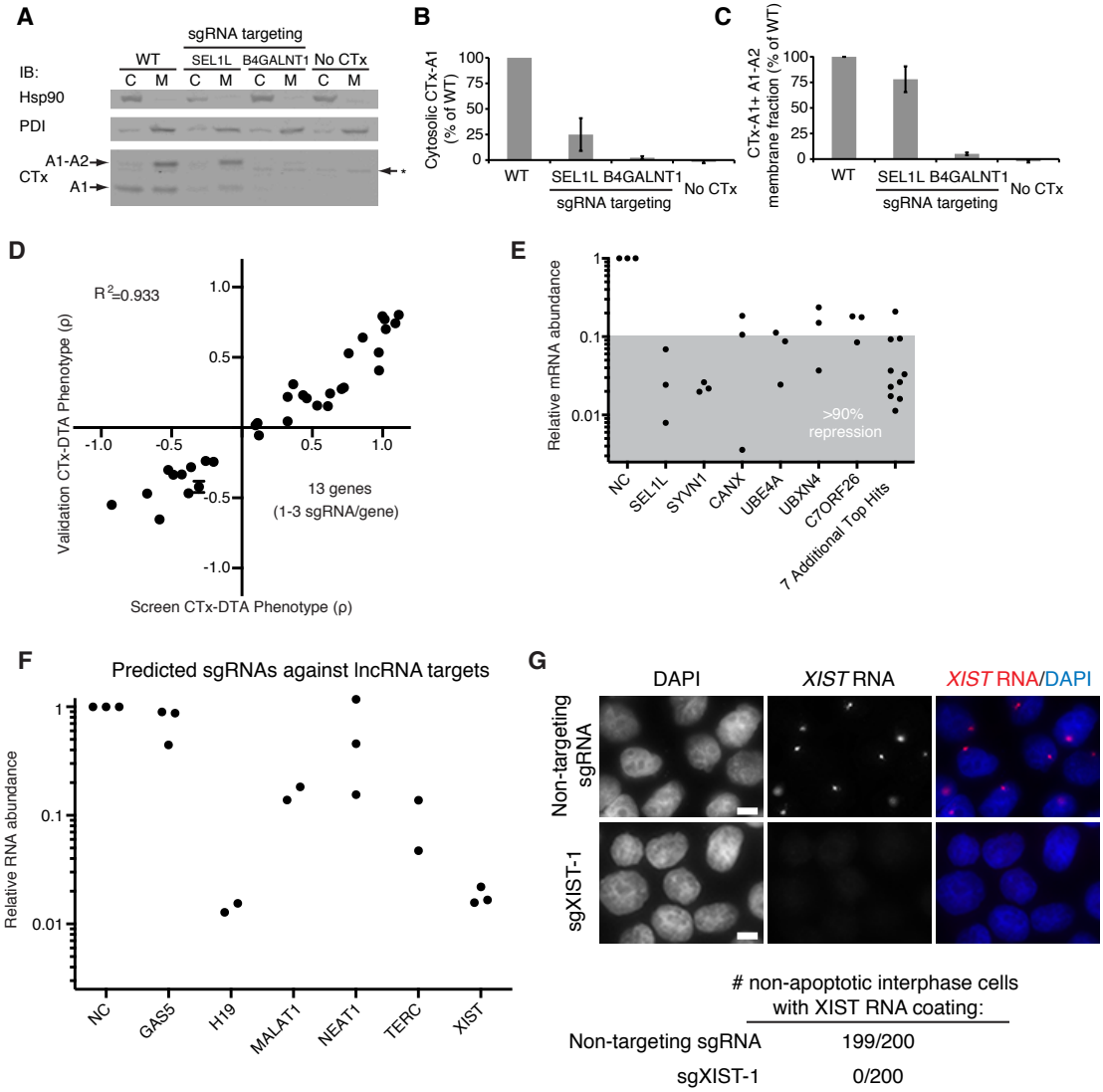

A
sgRNA targeting
WT
SEL1L
B4GALNT1
No CTx
IB:
C M C M C M C M
Hsp90
PDI
A1-A2
CTx
A1
B
Cytosolic CTx-A1 (% of WT)
100
75
50
25
0
WT
SEL1L B4GALNT1
sgRNA targeting
No CTx
C
CTx-A1+ A1-A2 membrane fraction (% of WT)
D
R2=0.933
Validation CTx-DTA Phenotype (ρ)
Screen CTx-DTA Phenotype (ρ)
13 genes
(1-3 sgRNA/gene)
E
Relative mRNA abundance
>90% repression
NC
SEL1L
SYVN1
CANX
UBE4A
UBXN4
C7ORF26
7 Additional Top Hits
F
Predicted sgRNAs against lncRNA targets
Relative RNA abundance
NC
GAS5
H19
MALAT1
NEAT1
TERC
XIST
G
DAPI
XIST RNA
XIST RNA/DAPI
Non-targeting sgRNA
sgXIST-1
non-apoptotic interphase cells with XIST RNA coating:
Non-targeting sgRNA 199/200
sgXIST-1 0/200

Figure 7. CRISPRi Strongly Represses Gene Expression of Both Protein-Coding and Non-Coding Genes, Resulting in Reproducible Phenotypes

(A-C) Cells expressing a negative control sgRNA or an sgRNA targeting *SEL1L* or *B4GALNT1* were incubated with cholera toxin and fractionated to quantify cholera toxin present in the cytosolic and membrane fractions by Western blot. *B4GALNT1* repression blocks toxin uptake whereas *SEL1L* repression prevents toxin retro-translocation from the membrane fraction to the cytosol. **(D)** Validation of CTx-DTA screen phenotypes with single sgRNA re-test experiments. Data is represented as the mean ± standard deviation of replicates (N=3). **(E)** CRISPRi knockdown of 13 hit genes (28 sgRNAs; sgRNAs correspond to 7D) identified in the CTx-DTA screen was quantified by qPCR. The gray shaded region denotes sgRNAs showing at least 90% knockdown for each gene. **(F)** CRISPRi knockdown of 6 lncRNA genes was quantified by qPCR. 2-3 sgRNAs computationally predicted to target each gene were cloned and transduced into K562 cells expressing dCas9-KRAB. (**G**) K562 cells expressing dCas9-KRAB were transduced with either a non-targeting sgRNA or an sgRNA targeting the *XIST* locus (sgXIST-1). The cells were then stained with DAPI and an RNA FISH probe for the *XIST* transcript. 200 non-apoptotic interphase cells in each condition were scored for *XIST* RNA coating. *XIST* is undetectable in cells transduced with sgXIST-1. Scale bar represents 5μm.

Supplementary Figures

Figure S1

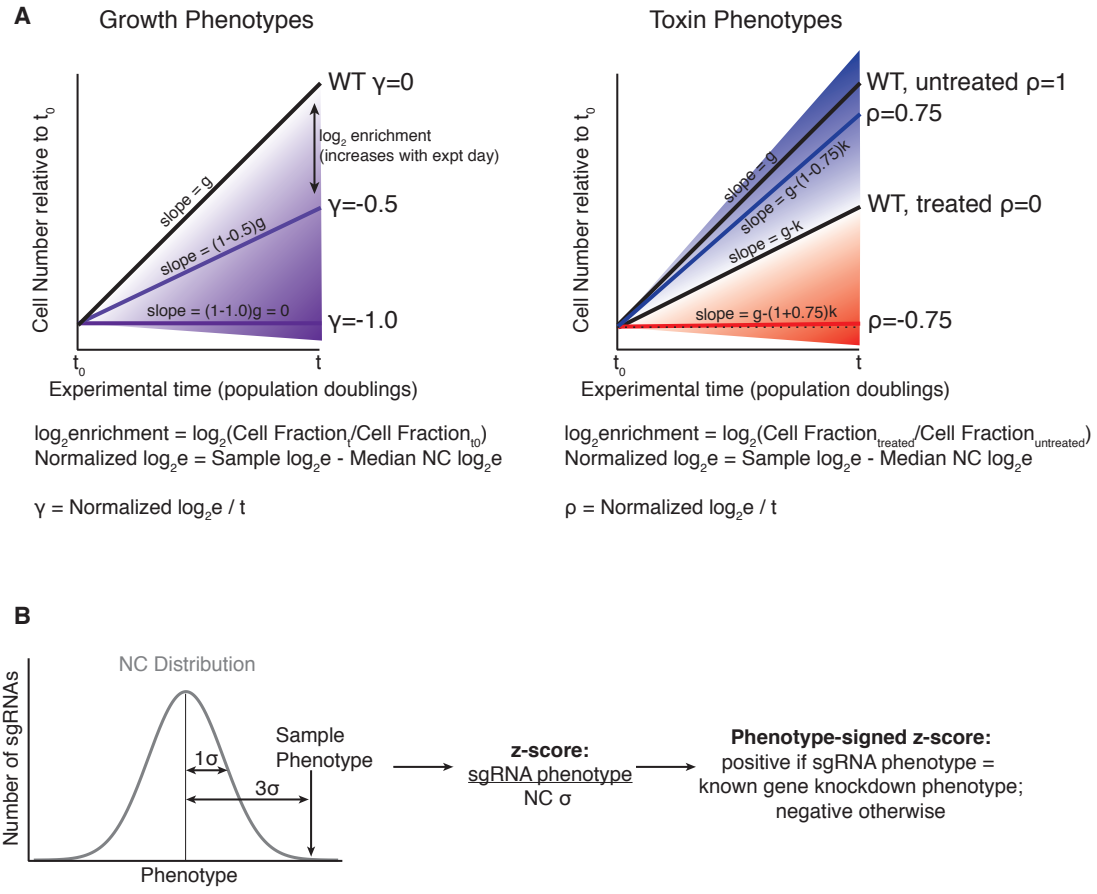


Figure S1. Mathematical Framework for Quantifying sgRNA Phenotype and Activity.

(A) The formulas for deriving cellular phenotype from measurements of cell fraction in a population at discrete timepoints. Cells with a given genotype have an intrinsic growth rate g, often expressed as cell doublings per day. The $\log_2$ enrichment ($\log_2 e$) of cells with a specific genotype in a population can be calculated from the fraction of cells in the population at the endpoint t versus that fraction at t0. In order to express this as the growth rate relative to wild-type (γ), $\log_2 e$ is normalized to the median $\log_2 e$ of the negative control set and then divided by t. Similarly, the phenotype of cells exposed to a selective pressure (e.g. toxin treatment) can calculated from the $\log_2 e$ of treated and untreated populations to obtain ρ, which is +1 for completely resistant cells and -1 for cells with 2-fold sensitivity to the pressure relative to wild-type. **(B)** In order to quantify the strength of individual sgRNA activity relative to noise in the experiment, sgRNA phenotype was divided by the standard deviation of negative control phenotypes to yield the z-score. For analyses of sgRNA strength in tiling screens, in which genes had known knockdown phenotypes for either sensitivity or resistance, z-scores were re-signed to give positive values where the phenotypes agreed with the expected phenotype and negative values otherwise.

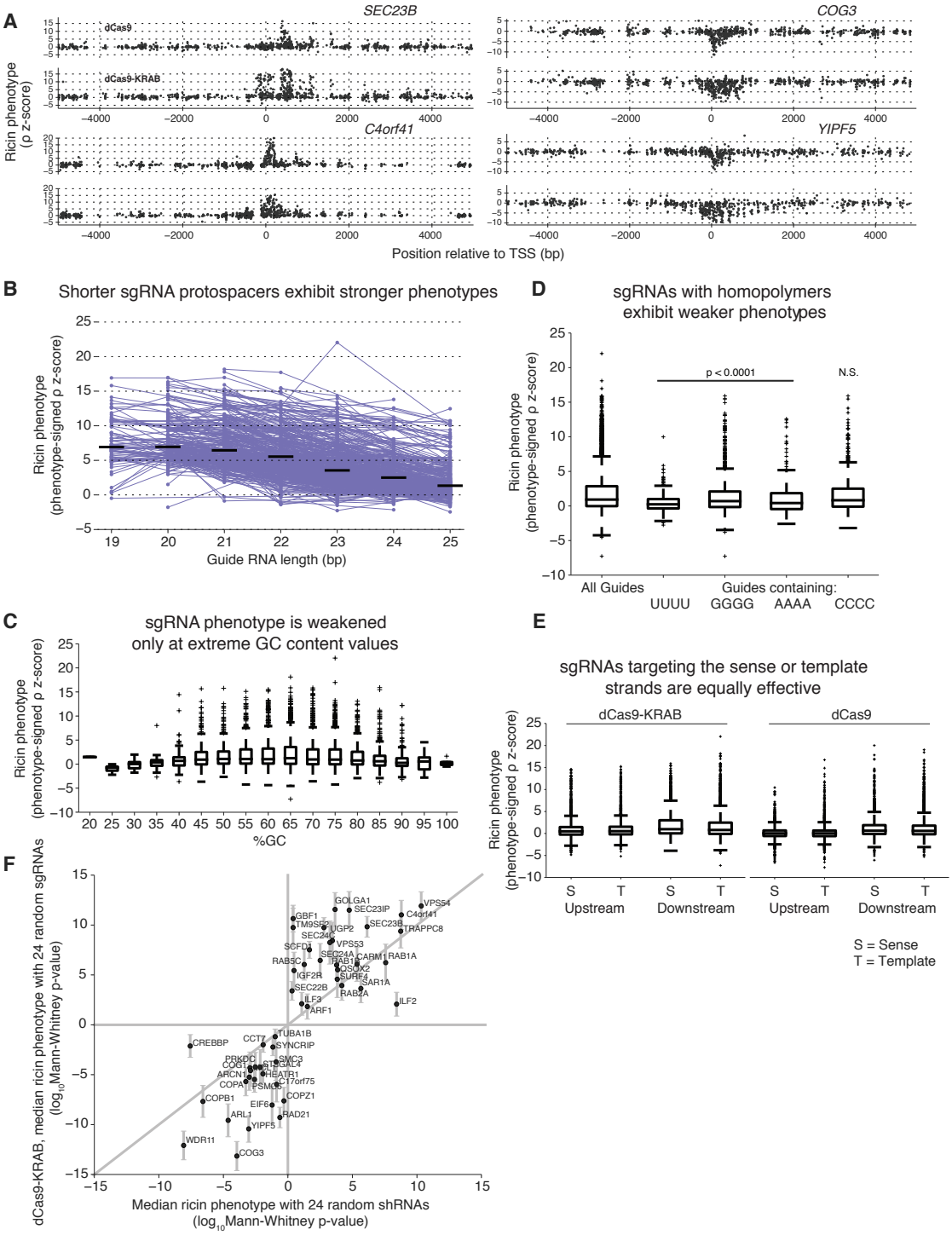

A
SEC23B
COG3
C4orf41
YIPF5
dCas9
dCas9-KRAB
Ricin phenotype
(ρ z-score)
Position relative to TSS (bp)
B
Shorter sgRNA protospacers exhibit stronger phenotypes
Ricin phenotype
(phenotype-signed ρ z-score)
Guide RNA length (bp)
C
sgRNA phenotype is weakened
only at extreme GC content values
%GC
D
sgRNAs with homopolymers
exhibit weaker phenotypes
p < 0.0001
N.S.
All Guides
Guides containing:
UUUU
GGGG
AAAA
CCCC
E
sgRNAs targeting the sense or template
strands are equally effective
dCas9-KRAB
dCas9
S
T
Upstream
Downstream
S = Sense
T = Template
F
dCas9-KRAB, median ricin phenotype with 24 random sgRNAs
(log10Mann-Whitney p-value)
Median ricin phenotype with 24 random shRNAs
(log10Mann-Whitney p-value)
GOLGA1
SEC23IP
VPS54
C4orf41
GBF1
TM9SF2
UGP2
SEC23B
TRAPPC8
SEC24A
VPS53
SCFD1
SEC24A
CARM1
RAB1A
RAB1A
RAB5C
GOSR2
IGF2R
SURF4
GAR1A
SEC22B
RAB2A
ILF3
ILF2
ARF1
CREBBP
CCT7
TUBA1B
SYNCRIP
PRKDC
SMC3
COPB1
STRAP
ARCN1
HEATR1
COPA
C17orf75
PSMC6
COPZ1
COPB1
EIF6
ARL1
RAD21
YIPF5
WDR11
COG3

Figure S2. Highly Active CRISPRi sgRNAs Are Close to the TSS, Short, and Do Not Contain Nucleotide Homopolymers.

(A) For several example genes, the phenotypes observed for sgRNAs expressed in dCas9 or dCas9-KRAB cells as a function of their position with respect to the TSS are depicted. Each point is an sgRNA. **(B)** Shorter sgRNAs have, on average, significantly higher activity. Each point represents an sgRNA, with lines connecting related sgRNAs that target the same PAM site but have increasing protospacer base pair length. Black lines represent the median activity for sgRNAs of specific lengths. sgRNAs are depicted if there are multiple sgRNAs targeting the same site and at least one sgRNA at that site is highly active (phenotype-signed z-score ≥ 5). **(C)** The presence of homopolymers (AAAA, GGGG, UUUU) within an sgRNA reduces activity on average. **(D)** sgRNA sequences with very high or very low GC content are less active. **(E)** The DNA strand targeted by an sgRNA has no effect on activity. **(F)** A comparison of a subset of sgRNAs selected based on CRISPRi activity rules versus our previously published shRNA library. For each gene, the Mann-Whitney p-value was calculated using sgRNA or shRNA activity relative to a negative control distribution using 24 shRNAs or 24 sgRNAs. The 24 shRNAs and sgRNAs were randomly selected 100 times from the shRNA library and sgRNAs meeting position and length rules, respectively, and the median and SD $\log_{10}$ p-value are displayed.

Figure S3

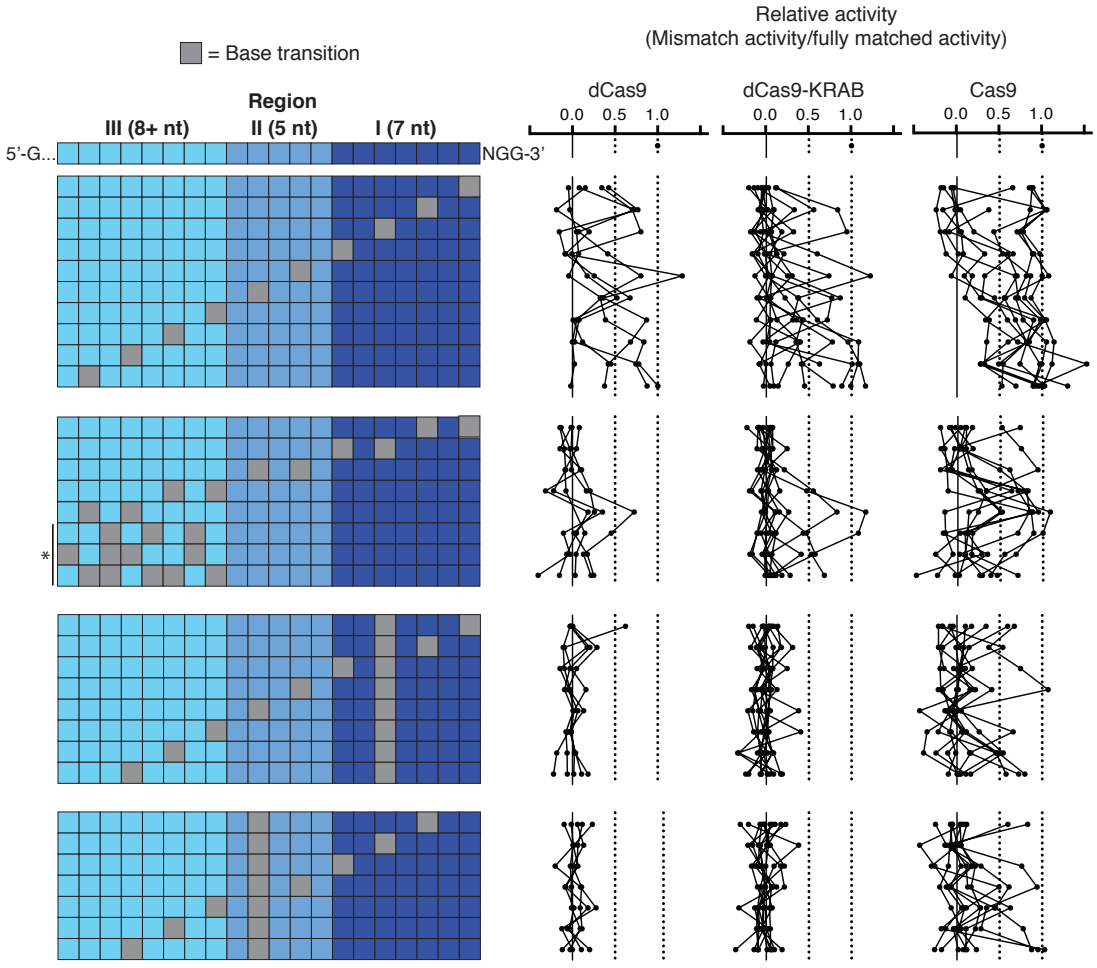

= Base transition
Region
III (8+ nt)
II (5 nt)
I (7 nt)
5'-G...
NGG-3'
*
Relative activity
(Mismatch activity/fully matched activity)
dCas9
dCas9-KRAB
Cas9
0.0 0.5 1.0

Figure S3. CRISPRi Activity is Highly Sensitive to Mismatches Between the sgRNA and DNA sequence

The on- and off- target activity of dCas9, dCas9-KRAB and Cas9 was measured for a series of sgRNAs with a varying number and position of mismatches. Each sgRNA is a point with the related mismatch series connected by lines. The measured off-target activity of each sgRNA with one or more mismatch is displayed as percent of the on-target activity for the corresponding sgRNA with 0 mismatches. The sgRNA series denoted with a star represents sgRNAs with 3, 4, or 5 mismatch base pairs randomly distributed across region 3 of the sgRNA sequence. sgRNAs were included in the analysis only if the fully matched guide was highly active (phenotype-signed z-score >= 4); N=5 for dCas9, 11 for dCas9-KRAB, and 10 for Cas9.

Figure S4

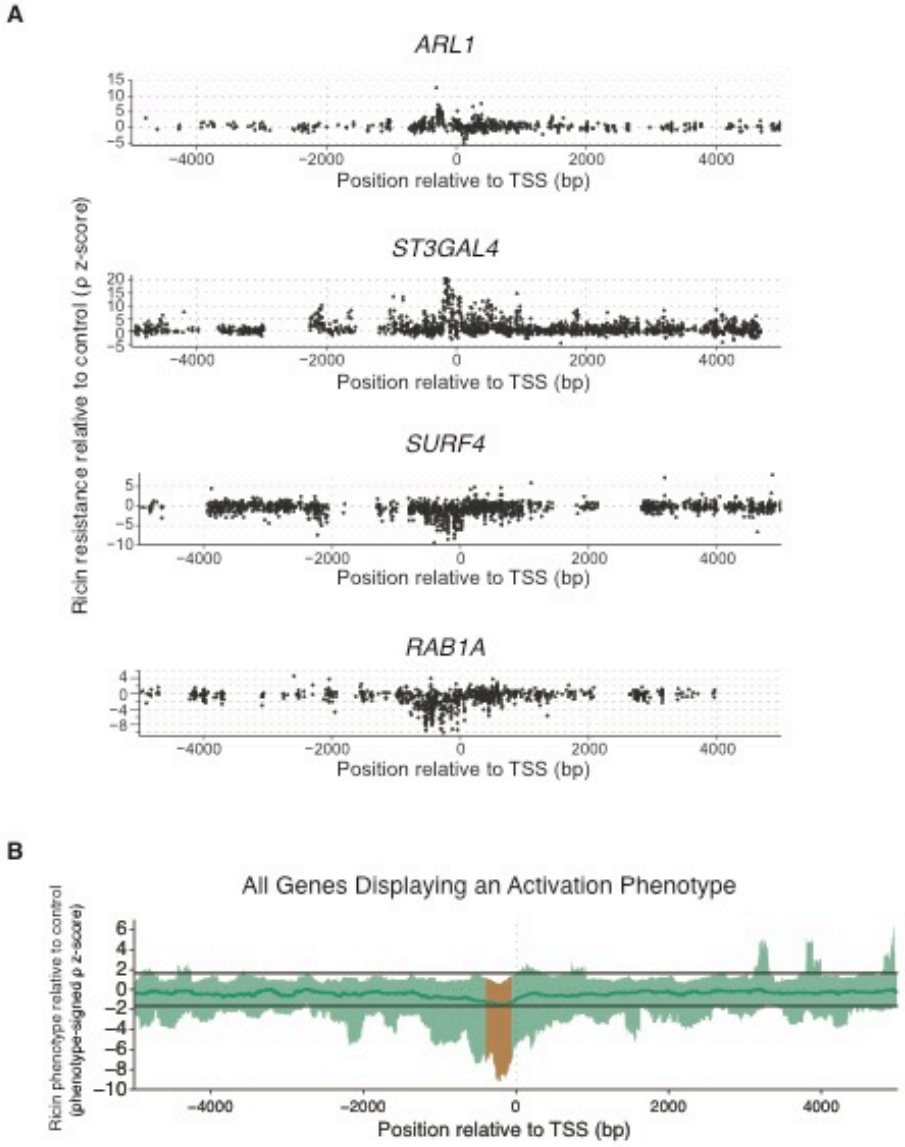

A
ARL1
ST3GAL4
SURF4
RAB1A
Ricin resistance relative to control (ρ z-score)
Position relative to TSS (bp)
B
All Genes Displaying an Activation Phenotype
Ricin phenotype relative to control (phenotype-signed ρ z-score)
Position relative to TSS (bp)

Figure S4. A Tiling sgRNA Screen Defines Rules for CRISPRa Activity at Endogenous Genes in Human Cells

(A) The activity of sgRNAs in a CRISPRa cell line as a function of the distance of the sgRNA site to the TSS of the targeted gene for four example genes. Top, ARL1 and ST3GAL4 activation results in ricin resistance; bottom, SURF4 and RAB1A activation results in ricin sensitivity. **(B)** A sliding-window average only for genes with a significant CRISPRa ricin resistance phenotype targeted by our test library is shown in blue. The median activity is shown with a blue line while the window of maximal activity is shown in red. The data is displayed as a phenotype-signed z-score.

Figure S5

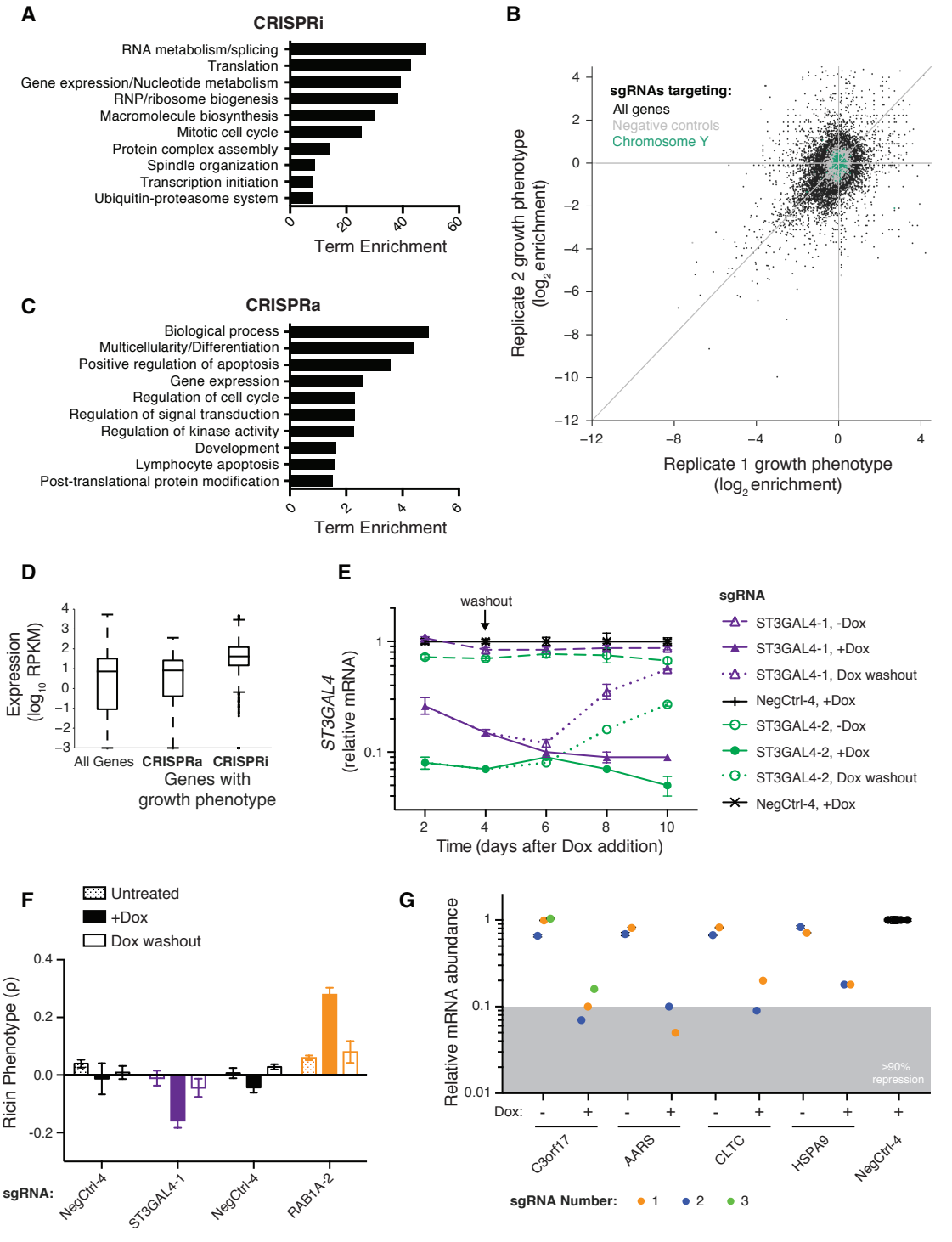

A
CRISPRi
RNA metabolism/splicing
Translation
Gene expression/Nucleotide metabolism
RNP/ribosome biogenesis
Macromolecule biosynthesis
Mitotic cell cycle
Protein complex assembly
Spindle organization
Transcription initiation
Ubiquitin-proteasome system
0
20
40
60
Term Enrichment
B
sgRNAs targeting:
All genes
Negative controls
Chromosome Y
Replicate 2 growth phenotype (log2 enrichment)
Replicate 1 growth phenotype (log2 enrichment)
C
CRISPRa
Biological process
Multicellularity/Differentiation
Positive regulation of apoptosis
Gene expression
Regulation of cell cycle
Regulation of signal transduction
Regulation of kinase activity
Development
Lymphocyte apoptosis
Post-translational protein modification
Term Enrichment
D
Expression (log10 RPKM)
All Genes
CRISPRa
CRISPRi
Genes with growth phenotype
E
washout
ST3GAL4 (relative mRNA)
Time (days after Dox addition)
sgRNA
ST3GAL4-1, -Dox
ST3GAL4-1, +Dox
ST3GAL4-1, Dox washout
NegCtrl-4, +Dox
ST3GAL4-2, -Dox
ST3GAL4-2, +Dox
ST3GAL4-2, Dox washout
NegCtrl-4, +Dox
F
Untreated
+Dox
Dox washout
Ricin Phenotype (ρ)
sgRNA:
NegCtrl-4
ST3GAL4-1
NegCtrl-4
RAB1A-2
G
Relative mRNA abundance
≥90% repression
Dox:
C3orf17
AARS
CLTC
HSPA9
NegCtrl-4
sgRNA Number: 1 2 3

Figure S5. CRISPRi/a Screens Reveal Genes Controlling Cell Growth And CRISPRI Can Inducibly and Rapidly Repress Transcription

(A) The top 10 DAVID annotation clusters identified in our CRISPRi growth screen are strongly enriched for known essential cellular processes. **(B)** Growth phenotypes from a genome-scale CRISPRa screen carried out in duplicate in K562 cells. Two classes of negative control sgRNAs are shown: non-targeting sgRNAs (grey) and sgRNAs targeting Y-chromosomal genes (green). **(C)** The top 10 DAVID annotation clusters identified in our CRISPRa growth screen. **(D)** RNA-sequencing RPKM (reads per kilo base per million) from K562 cells plotted for three gene sets: all genes and those genes identified to have activation or repression growth phenotypes by CRIPSRa or CRISPRi, respectively. **(E)** *ST3GAL4* expression levels (relative to *ACTB,* as quantified by qPCR) in inducible CRISPRi K562 cells transduced with *ST3GAL4*-targeting sgRNAs in the absence, presence, and after washout of doxycycline. Data is normalized to control cells (assayed in the presence of doxycycline) at each time point and represented as the mean of technical replicates ± standard error. **(F)** Reversibility of ricin sensitivity (ST3GAL4-1) and resistance (RAB1A-2) phenotypes induced by CRISPRi. Inducible CRISRPi cells transduced with the indicated sgRNAs were challenged with 0.4 ng / mL ricin toxin before and after 2 days of doxycycline treatment, as well as after removal of doxycycline from tissue culture media and following an 8 day recovery period. Rho values were calculated as described. Data represent the mean of ricin challenged replicates (N=3) ± standard deviation. **(G)** CRISPRi knockdown (mRNA levels relative to *ACTB*) of hit genes identified in the genome-scale growth screen (9 sgRNAs) was quantified by qPCR after two days of doxycycline treatment. The gray shaded region denotes sgRNAs showing at least 90% knockdown for each gene. See also Figure 5D.

Figure S6

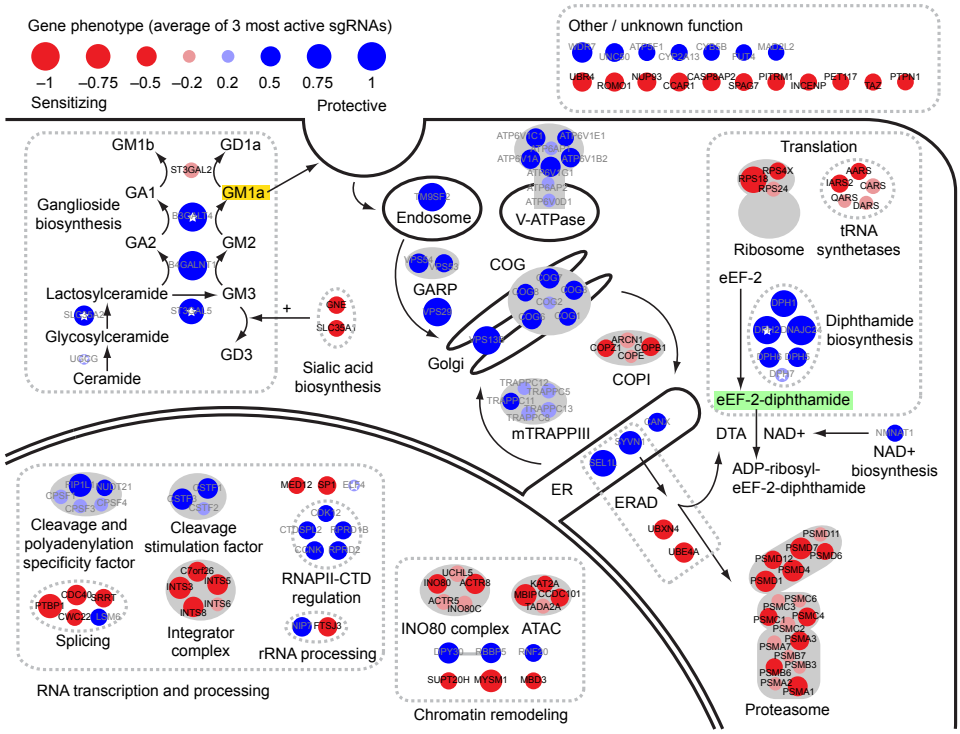

Gene phenotype (average of 3 most active sgRNAs)
−1
−0.75
−0.5
−0.2
0.2
0.5
0.75
1
Sensitizing
Protective
Other / unknown function
Ganglioside biosynthesis
GM1b
GD1a
GA1
GM1a
GA2
GM2
Lactosylceramide
GM3
Glycosylceramide
GD3
Ceramide
Sialic acid biosynthesis
Endosome
V-ATPase
GARP
COG
Golgi
COPI
mTRAPPIII
ER
ERAD
Translation
Ribosome
tRNA synthetases
eEF-2
Diphthamide biosynthesis
eEF-2-diphthamide
DTA
NAD+
NAD+ biosynthesis
ADP-ribosyl-eEF-2-diphthamide
Cleavage and polyadenylation specificity factor
Cleavage stimulation factor
RNAPII-CTD regulation
Splicing
Integrator complex
rRNA processing
RNA transcription and processing
INO80 complex
ATAC
Chromatin remodeling
Proteasome

Figure S6. A Genome-Scale CRISPRi Screen Reveals Known and New Pathways and Complexes Governing the Sensitivity to a Cholera-Diphtheria Fusion Toxin.

Same representation as Figure 6B, with added gene names.

Figure S7

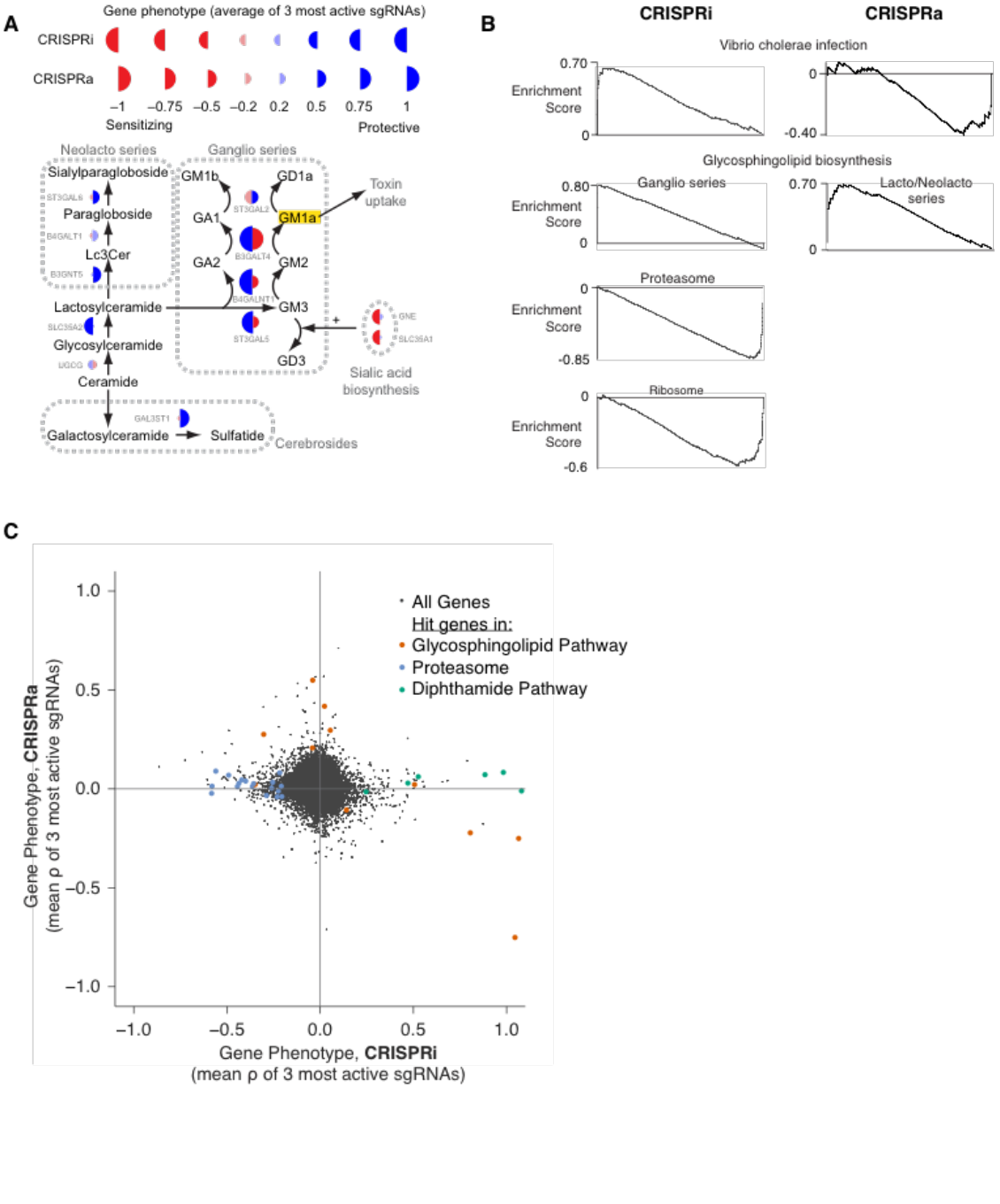

A
Gene phenotype (average of 3 most active sgRNAs)
CRISPRi
CRISPRa
−1 −0.75 −0.5 −0.2 0.2 0.5 0.75 1
Sensitizing
Protective
Neolacto series
Ganglio series
Sialylparagloboside
Paragloboside
Lc3Cer
Lactosylceramide
Glycosylceramide
Ceramide
GM1b
GD1a
GA1
GM1a
GA2
GM2
GM3
GD3
Toxin uptake
Sialic acid biosynthesis
Galactosylceramide
Sulfatide
Cerebrosides
B
CRISPRi
CRISPRa
Vibrio cholerae infection
Enrichment Score
0.70
0
-0.40
Glycosphingolipid biosynthesis
Ganglio series
Lacto/Neolacto series
0.80
0.70
Proteasome
-0.85
Ribosome
-0.6
C
All Genes
Hit genes in:
Glycosphingolipid Pathway
Proteasome
Diphthamide Pathway
Gene Phenotype, CRISPRa
(mean ρ of 3 most active sgRNAs)
Gene Phenotype, CRISPRi
(mean ρ of 3 most active sgRNAs)
1.0
0.5
0.0
−0.5
−1.0

Figure S7. Complementary Insights from CRISPRi and CRISPRa Screens for Sensitivity to a Cholera-Diphtheria Fusion Toxin

(A) CRISPRi and CRISPRa hits in sphingolipid metabolism. Same representation as Figure 6C, with added gene names **(B)** Gene set enrichment analysis of top hits from genome-scale CRISPRi and CRISPRa screens for sensitivity to CTx-DTA. (**C**) CRISPRi versus CRISPRa gene CTx-DTA phenotypes. Phenotypes for genome-scale CTx-DTA CRISPRi and CRISPRa screens (black) were calculated based on the average resistance phenotype (ρ) of the three most active sgRNAs. Hit genes from either screen that were featured in Figure 6 as part of the glycosphingolipid biosynthesis pathway, proteasome, or diphthamide biosynthesis pathway are labele as indicated.

Supplementary Tables

Table S1. sgRNA Sequences and qPCR Primer Pairs Used in Validation Experiments, Related to Figures 3, 5, and 7.

Sequences of the sgRNA protospacers (Tab 1) and qPCR primers (Tab 2) used for individual validation experiments in this study.

Table S2. Genome-Scale Library sgRNA Sequences and Phenotypes, Related to Figures 4 and 6.

sgRNA IDs, transcripts targeted, protospacer sequences, and growth (γ) and CTx-DTA sensitivity (ρ) phenotypes for the genome-scale CRISPRi (Tab 1) and CRISPRa libraries (Tab 2). Phenotypes are average of two biological replicates.

Table S3. Gene Phenotypes from Genome-Scale Screens, Related to Figures 4 and 6.

Gene phenotypes for growth (γ) and CTx-DTA (ρ) phenotypes calculated from the 3 most active sgRNAs as described. Phenotypes are average of two biological replicates.

Table S4. Annotation of the Top 50 CRISPRa Growth Hits, Related to Figure 4.

Categorization of the 50 genes with the most negative growth phenotype (γ). Genes were annotated as tumor suppressor genes, developmental transcription factors, and/or involved in centrosomal processes, as determined by published surveys of frequently mutated genes in cancer, individual reports, and gene family membership.

References

Adamson, B., Smogorzewska, A., Sigoillot, F.D., King, R.W., and Elledge, S.J. (2012). A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. Nat. Cell Biol. *14*, 318–328.

Alberich-Jordà, M., Wouters, B., Balastik, M., Shapiro-Koss, C., Zhang, H., Di Ruscio, A., DiRuscio, A., Radomska, H.S., Ebralidze, A.K., Amabile, G., et al. (2012). C/EBPγ deregulation results in differentiation arrest in acute myeloid leukemia. J. Clin. Invest. *122*, 4490–4504.

Bassett, A.R., Akhtar, A., Barlow, D.P., Bird, A.P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A.C., Gingeras, T.R., Haerty, W., et al. (2014). Considerations when investigating lncRNA function in vivo. eLife *3*, e03058.

Bassik, M.C., Lebbink, R.J., Churchman, L.S., Ingolia, N.T., Patena, W., LeProust, E.M., Schuldiner, M., Weissman, J.S., and McManus, M.T. (2009). Rapid creation and quantitative monitoring of high coverage shRNA libraries. Nat. Methods *6*, 443–445.

Bassik, M.C., Kampmann, M., Lebbink, R.J., Wang, S., Hein, M.Y., Poser, I., Weibezahn, J., Horlbeck, M.A., Chen, S., Mann, M., et al. (2013). A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. Cell *152*, 909–922.

Beerli, R.R., Segal, D.J., Dreier, B., and Barbas, C.F. (1998). Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. Proc. Natl. Acad. Sci. U. S. A. *95*, 14628–14633.

Beerli, R.R., Dreier, B., and Barbas, C.F. (2000). Positive and negative regulation of endogenous genes by designed transcription factors. Proc. Natl. Acad. Sci. U. S. A. *97*, 1495–1500.

Bonifacino, J.S., and Rojas, R. (2006). Retrograde transport from endosomes to the trans-Golgi network. Nat. Rev. Mol. Cell Biol. *7*, 568–579.

Boone, C., Bussey, H., and Andrews, B.J. (2007). Exploring genetic interactions and networks with yeast. Nat. Rev. Genet. *8*, 437–449.

Carette, J.E., Guimaraes, C.P., Varadarajan, M., Park, A.S., Wuethrich, I., Godarova, A., Kotecki, M., Cochran, B.H., Spooner, E., Ploegh, H.L., et al. (2009). Haploid genetic screens in human cells identify host factors used by pathogens. Science *326*, 1231–1235.

Cech, T.R., and Steitz, J.A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. Cell *157*, 77–94.

Chang, K., Elledge, S.J., and Hannon, G.J. (2006). Lessons from Nature: microRNA-based shRNA libraries. Nat. Methods *3*, 707–714.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science *327*, 425–431.

Davis, R.L., Weintraub, H., and Lassar, A.B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell *51*, 987–1000.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long

noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Duan, J., Lu, G., Xie, Z., Lou, M., Luo, J., Guo, L., and Zhang, Y. (2014). Genome-wide identification of CRISPR/Cas9 off-targets in human genome. Cell Res.

Gaj, T., Gersbach, C.A., and Barbas, C.F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. *31*, 397–405.

Geisler, S., and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat. Rev. Mol. Cell Biol. *14*, 699–712.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell *154*, 442–451.

Guimaraes, C.P., Carette, J.E., Varadarajan, M., Antos, J., Popp, M.W., Spooner, E., Brummelkamp, T.R., and Ploegh, H.L. (2011). Identification of host cell factors required for intoxication through use of modified cholera toxin. J. Cell Biol. *195*, 751–764.

Habedanck, R., Stierhof, Y.-D., Wilkinson, C.J., and Nigg, E.A. (2005). The Polo kinase Plk4 functions in centriole duplication. Nat. Cell Biol. *7*, 1140–1146.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Huang, L.C., Clarkin, K.C., and Wahl, G.M. (1996). Sensitivity and selectivity of the DNA damage sensor responsible for activating p53-dependent G1 arrest. Proc. Natl. Acad. Sci. U. S. A. *93*, 4827–4832.

Jackson, S.P. (2002). Sensing and repairing DNA double-strand breaks. Carcinogenesis *23*, 687–696.

Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P.S. (2003). Expression profiling reveals off-target gene regulation by RNAi. Nat. Biotechnol. *21*, 635–637.

Kampmann, M., Bassik, M.C., and Weissman, J.S. (2013). Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *110*, E2317–E2326.

Kampmann, M., Bassik, M.C., and Weissman, J.S. (2014). Functional genomics platform for pooled screening and generation of mammalian genetic interaction maps. Nat. Protoc. *9*, 1825–1847.

Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M.D.C., and Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat. Biotechnol. *32*, 267–273.

Kuscu, C., Arslan, S., Singh, R., Thorpe, J., and Adli, M. (2014). Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nat. Biotechnol. *32*, 677–683.

Mittler, G., Stühler, T., Santolin, L., Uhlmann, T., Kremmer, E., Lottspeich, F., Berti, L., and Meisterernst, M. (2003). A novel docking site on Mediator is critical for activation by VP16 in mammalian cells. EMBO J. *22*, 6494–6504.

Van Ness, B.G., Howard, J.B., and Bodley, J.W. (1980). ADP-ribosylation of elongation factor 2 by diphtheria toxin. Isolation and properties of the novel ribosyl-amino acid and its hydrolysis products. J. Biol. Chem. *255*, 10717–10720.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183.

Rine, J., Hansen, W., Hardeman, E., and Davis, R.W. (1983). Targeted selection of recombinant clones through gene dosage effects. Proc. Natl. Acad. Sci. U. S. A. *80*, 6750–6754.

Sander, J.D., and Joung, J.K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. Nat. Biotechnol. *32*, 347–355.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

Sigoillot, F.D., Lyman, S., Huckins, J.F., Adamson, B., Chung, E., Quattrochi, B., and King, R.W. (2012). A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens. Nat. Methods *9*, 363–366.

Smith, R.D., Willett, R., Kudlyk, T., Pokrovskaya, I., Paton, A.W., Paton, J.C., and Lupashin, V.V. (2009). The COG complex, Rab6 and COPI define a novel Golgi retrograde trafficking pathway that is exploited by SubAB toxin. Traffic Cph. Den. *10*, 1502–1517.

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

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Tanenbaum, M.E., Macurek, L., van der Vaart, B., Galli, M., Akhmanova, A., and Medema, R.H. (2011). A complex of Kif18b and MCAK promotes microtubule depolymerization and is negatively regulated by Aurora kinases. Curr. Biol. CB *21*, 1356–1365.

Tanenbaum, M.E., Gilbert, L.A., Qi, L.S., Weissman, J.S., and Vale, R.D. A versatile protein tagging system for signal amplification in single molecule imaging and gene regulation. Revis.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

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

Wernick, N.L.B., Chinnapen, D.J.-F., Cho, J.A., and Lencer, W.I. (2010). Cholera toxin: an intracellular journey into the cytosol by way of the endoplasmic reticulum. Toxins *2*, 310–325.

Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E., Konermann, S., Chen, S., et al. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat. Biotechnol. *32*, 670–676.

Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G.M., and Arlotta, P. (2011). Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. Nat. Biotechnol. *29*, 149–153.

Zhang, P., Zhang, X., Iwama, A., Yu, C., Smith, K.A., Mueller, B.U., Narravula, S., Torbett, B.E., Orkin, S.H., and Tenen, D.G. (2000). PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. Blood *96*, 2641–2648.

Zhao, M., Sun, J., and Zhao, Z. (2013). TSGene: a web resource for tumor suppressor genes. Nucleic Acids Res. *41*, D970–D976.

Chapter 3

Nucleosomes Impede Cas9 Access to DNA *in vivo* and *in vitro*

Abstract

The prokaryotic CRISPR (Clustered Regularly Interspaced Palindromic Repeats)-associated protein, Cas9, has been widely adopted as a tool for editing, imaging, and regulating eukaryotic genomes. However, our understanding of how to select single-guide RNAs (sgRNAs) that mediate efficient Cas9 activity is incomplete, as we lack insight into how chromatin impacts Cas9 targeting. To address this gap, we analyzed large-scale genetic screens performed in human cell lines using either nuclease-active or nuclease-dead Cas9 (dCas9). We observed that highly active sgRNAs for Cas9 and dCas9 were found almost exclusively in regions of low nucleosome occupancy. *In vitro* experiments demonstrated that nucleosomes in fact directly impede Cas9 binding and cleavage, while chromatin remodeling can restore Cas9 access. Our results reveal a critical role of eukaryotic chromatin in dictating the targeting specificity of this transplanted bacterial enzyme, and provide rules for selecting Cas9 target sites distinct from and complementary to those based on sequence properties.

Introduction

CRISPR (Clustered Regularly Interspaced Palindromic Repeats) prokaryotic adaptive immune systems have yielded transformative tools for manipulating eukaryotic genomes. Most notably, the CRISPR associated Cas9 protein from *Streptococcus pyogenes*, together with a single chimeric guide RNA (sgRNA), provides a programmable endonuclease that has revolutionized our ability to edit genomes [1]. Cas9 has been further modified to a nuclease-dead form (dCas9) to provide a programmable DNA binding protein that can be fused to effector domains, making it possible to turn on or off targeted genes, mark specific genomic loci with fluorescent proteins, or alter epigenetic marks [1-8]. A central challenge in implementing these tools is identifying effective and specific sgRNAs. While much of the effort to define relevant rules has focused on the sequence of the target site and sgRNA [9-12], these only partially predict Cas9 activity and suggest that additional determinants likely exist.

Chromatin structure may represent a key parameter governing Cas9 efficacy in eukaryotic cells. CRISPR evolved in archea and bacteria [13] and is likely not optimized to explore and modify large, chromatin-bound eukaryotic genomes, a hypothesis supported by several studies that pointed to a correlation between rates of DNA binding and cleavage in regions of open chromatin as measured by DNase I hypersensitivity [9, 14, 15]. Additionally, recent single-molecule imaging studies have shown that dCas9 explores euchromatin more frequently than it does heterochromatin [16].

We hypothesized that nucleosomes, the basic unit of chromatin structure, are an important impediment to Cas9 recognition. Here, we addressed this hypothesis *in vivo* by leveraging large datasets collected from over 30 sgRNA tiling and genome-scale genetic screens [17]. We found that regions of high nucleosome occupancy *in vivo*, as determined by MNase-seq

(micrococcal nuclease sequencing) [18, 19], were strongly depleted of highly active sgRNAs for CRISPR interference (CRISPRi) [3, 17] and nuclease-active Cas9. We complemented these results with *in vitro* experiments demonstrating that formation of a nucleosome provides a direct and profound block to dCas9 binding and Cas9 cleavage. Despite this strong barrier to Cas9 activity, we found that addition of a chromatin remodeling enzyme to chromatinized DNA *in vitro* can restore Cas9 access to nucleosomal DNA, highlighting one route by which CRISPR may still be able to modify chromatin *in vivo*. Our results reveal a fundamental aspect of the mechanism by which this transplanted bacterial enzyme interacts with eukaryotic chromatin, and provide a new dimension for selecting highly active sgRNAs.

Results

CRISPRi activity is periodic and out-of-phase with nucleosome positioning

In order to study features of chromatin that affect CRISPRi activity, we integrated data from whole-genome screens testing for a wide range of phenotypes performed with a previously described CRISPRi library targeting each gene with ~10 sgRNAs [17]. We selected 30 screens performed in the cell line K562 expressing dCas9-KRAB (M.L., B.A., B.B., C.Y.P., M.K., Y.C., J.F., and J.N., personal communication) and set a threshold for high-confidence hit genes, which allowed us to assess the relative strength of phenotypes for the sgRNAs targeting these genes. Specifically, we analyzed 18,380 sgRNAs targeting 1,539 genes, and generated "activity scores" by normalizing sgRNA phenotypes to the average of the 3 sgRNAs with the strongest phenotypes for each gene (Figure 1A and Table S1). To assess how CRISPRi activity varies with respect to the transcription start site (TSS), we plotted the average sgRNA activity score as a function of distance to the FANTOM-annotated TSS [20] (Figure 1B). This analysis revealed a robust, periodic pattern of activity, with peaks at ~190bp intervals relative to the TSS. This periodicity was highly reminiscent of patterns previously described for nucleosomes [21]. Indeed, analysis of K562 MNase-seq data from the ENCODE consortium [19, 22] revealed that the average nucleosome signal was strongly anti-correlated with CRISPRi activity (Figure 1B), suggesting that high nucleosome occupancy leads to low CRIPSRi activity.

To further explore this inverse relationship between CRISPRi activity and nucleosome organization, we exploited the previous observation that nucleosome phasing is more pronounced in well-expressed genes [19]. We grouped genes by their expression in K562 [22, 23] and analyzed CRISPRi activity and nucleosome occupancy within each group. In support of

a connection, we found that peak to trough amplitudes of both features were larger for highly expressed genes (Figures 1C,D).

We also analyzed the MNase signal at each sgRNA target site to determine whether nucleosome occupancy could directly explain variation in CRISPRi activity between individual sgRNAs (Figure 1E). Consistent with the hypothesis that nucleosomes exclude dCas9, almost all of the highly active sgRNAs targeted sites corresponded with low MNase-seq signal. However, not all sgRNAs targeting sites with low nucleosome occupancy were highly active, matching previous findings that sgRNA and target DNA sequence features also influence efficiency of the CRISPR system [9-12]. To exclude the possibility that differences in sequence features alone between nucleosome-bound and -free regions could explain the periodicity of CRISPRi activity (rather than the presence of nucleosomes per se), we performed linear regression using MNase signal, sgRNA length [12, 17], and a validated sgRNA sequence scoring algorithm [10]. We found that each parameter individually correlated with sgRNA activity ($p < 10^{-58}$ for each; Figure 1F). Importantly, correction for sequence and length features had minimal impact on the ability of the MNase signal to predict CRISPRi activity ($p < 10^{-85}$ after correction). Indeed, a linear fit of all three features provided a still stronger correlation ($p < 10^{-221}$; Figure 1F, far right column), suggesting that incorporating nucleosome occupancy in future sgRNA design algorithms could significantly improve predictive value. We have recently developed a comprehensive algorithm for predicting highly active CRISPRi sgRNAs that, by accounting for nucleosome positioning, higher order sequence features, and non-linear relationships in these parameters, shows even greater correlation with this dataset that was already enriched for active sgRNAs using our original CRISPRi library design principles [17] (cross-validation R^2=0.32; Horlbeck et al., manuscript in preparation [24]).

Nuclease-active Cas9 activity anti-correlates with nucleosome occupancy

In order to generate a dataset for evaluating the effect of nucleosome positioning on nuclease-active Cas9 in K562 cells, we took advantage of our previously described library densely tiling sgRNAs in 10kb windows around the TSS of 49 genes known to modulate susceptibility to the toxin ricin [17, 25] and tested Cas9-expressing cells for resistance or sensitivity to ricin (Table S2). We observed phenotypes consistent with the expected knockdown phenotype primarily in coding sequences (CDS) but also in some promoter regions, consistent with recent results showing that modifications introduced by Cas9 can disrupt *cis*-regulatory regions [26]. Analysis of CDS-targeting sgRNAs revealed that strong phenotypes were found predominantly in regions of low MNase-seq signal (Figure 2A), although this relationship was less pronounced than for CRISPRi. This may be due to decreased phasing of nucleosomes within the gene body [21]. When we analyzed all sgRNAs in the library, thus incorporating information from *cis*-regulatory regions where nucleosomes are well phased, we found the effect of nucleosome position to be even stronger (Figure 2-figure supplement 1). Although nucleosome occupancy was predictive of CRISPR activity independent of sgRNA sequence features, a much stronger correlation was obtained when all features were considered (Figure 2B). Therefore, nucleosome organization likely represents an important feature for CRISPR sgRNA design and should be considered a key contributing factor in interpreting future tiling mutagenesis experiments of coding and non-coding regions.

Nucleosomes are sufficient to fully block cleavage by Cas9 *in vitro*

Our *in vivo* experiments reveal a strong anti-correlation between Cas9 mediated

downstream phenotypic outputs and nucleosome positioning, but do not directly report on the ability of Cas9 to access nucleosomal DNA. Factors other than Cas9 access could contribute to the observed correlation in our data. For example, (d)Cas9 may bind or cut equally well in nucleosome-bound and un-bound regions, but may exert the observed modulation of gene expression through interference with transcriptional pausing, splicing, regulatory looping, or binding of important regulatory factors, processes which also correlate with nucleosome organization [27-33].

To determine whether (d)Cas9 activity is indeed directly affected by the presence of nucleosomes, we turned to a purified *in vitro* reconstituted system. Mouse histones were recombinantly expressed, purified, and assembled into a nucleosome using 147bp of the Widom 601 positioning sequence (Figures 3A,B) [34]. Conveniently, the 601 sequence contains numerous NGG protospacer adjacent motifs (PAMs) required for Cas9 recognition, spanning the full length of the DNA at different helical positions, allowing us to test the effects of position and solvent accesibility within the nucleosome (Figure 3C). We first tested the ability of Cas9 to cleave nucleosomal DNA. We pre-loaded purified Cas9-HaloTag with *in vitro* transcribed sgRNA, then introduced either naked 601 DNA or that same DNA assembled into a nucleosome (Figures 3A,D and Figure 3-figure supplement 1A). Fluorescent labeling of the DNA allowed us to visualize the cleavage products on a denaturing Urea-PAGE gel. In agreement with our *in vivo* data, the nucleosome protected its DNA from cleavage by Cas9, and complete protection from cleavage was observed to be independent of target position within the nucleosome (Figures 3E,F and Figure 3-figure supplement 1B,C). While our manuscript was under review, a study by Hinz and colleagues reported the similar *in vitro* finding that Cas9 nuclease activity is inhibited within the nucleosome but not at adjacent linker sequences[35]. Previous single-molecule and

biochemical studies have established that nucleosomal DNA undergoes transient unwrapping or breathing at the entry and exit sites, creating a gradient of accessibility along the nucleosome [36-41]. This property is often credited for the observed position dependent binding patterns of many transcription factors and DNA binding proteins. Interestingly, our data suggests that the cleavage activity of Cas9 *in vitro* is not detectably influenced by this effect, although less stable nucleosomes than the Widom 601 nucleosome or target sites closer to the nucleosome edge may exhibit more breathing and thus more accessibility than those tested here.

Cas9 is unable to bind nucleosomal DNA *in vitro*

Cleavage of DNA by Cas9 has been described as a stepwise process [15, 16, 42] in which Cas9 must first scan for PAMs, unzip the DNA duplex, and fully pair the guide RNA and target DNA prior to cleavage. While our *in vitro* data show that full pairing and cleavage is prevented by the presence of a nucleosome, we wondered if binding without cleavage, especially at the more accessible ends of the nucleosome, might still occur. Additionally, the first step in binding, PAM recognition, may be governed by helical location within the nucleosome as well as its proximity to the more dynamic ends. Histones make contacts with the DNA at every helical turn, thus a PAM may fall on the outside of the nucleosome, exposing it to solvent, or on the inside at the DNA-histone interface. To test the influence of target location within the nucleosome on Cas9 binding, we used an electrophoretic mobility shift assay to monitor binding of dCas9 to Cy3 end-labeled 601 DNA (Figures 4A,B), either free or assembled into a nucleosome containing Alexa Fluor 647-labeled histone H2B (Figure 4-figure supplement 1A). Consistent with our cleavage results, binding by dCas9 was abolished by the presence of the nucleosome, regardless of the targeted dCas9 binding site (Figure 4C-E and Figure 4-figure supplement

2A,B).

To better understand how a nucleosome might impede Cas9 binding, we aligned the available crystal structures of DNA-bound Cas9 and the structure of the 601 nucleosome ([43], PDB ID 3LZ0; [44], PDB ID 4UN3). We superimposed the target DNA in the Cas9 crystal structure with the DNA in the nucleosome structure at a site where the two DNA paths gave the best fit. The resulting combined structure reveals that the Cas9 protein poses significant steric clashes with the histones (Figure 4F). Given the extent of overlapping densities in the two structures, it seems unlikely that the histones and Cas9 could co-occupy the same piece of DNA. Additionally, it may be important to note that unlike other DNA binders such as transcription factors, binding by Cas9 constitutes melting of target DNA, which may pose an additional barrier to binding on a nucleosome. This hypothesis leaves two possible outcomes of targeting Cas9 to nucleosomal DNA: either Cas9 is capable of displacing histones in order to engage nucleosomal DNA, or it is excluded altogether. Our data support the latter conclusion.

The chromatin remodeling enzyme yChd1 can restore access to nucleosomal DNA *in vitro*

The nucleosome landscape of eukaryotic chromatin is dictated by both intrinsic DNA sequence preferences as well as extrinsic factors such as chromatin remodeling enzymes [21, 45, 46]. In order to model how these dynamics affect Cas9 access to nucleosomal DNA *in vitro*, we turned to a chromatinized plasmid system. We dialyzed plasmid DNA containing a single 601 nucleosome positioning sequence with a sub-saturating quantity of purified histone octamers, and confirmed the quality of the resulting chromatin assemblies by MNase digestion (Figures 5A,B). We tested whether a nucleosome was positioned at the 601 sequence using restriction enzyme accessibility mapping, and found that sites within the 601 sequence were well protected

from digestion, suggesting high nucleosome occupancy, while sites immediately adjacent were not, suggesting precise positioning (Figure 5C). To recapitulate the effects of chromatin remodeling *in vitro*, we used a purified, truncated form of the Snf2-like chromatin remodeling enzyme Chd1 from *Saccharomyces cerevisiae* (yChd1), which had previously been shown to mediate nucleosome sliding in an ATP-dependent manner without additional co-factors [47, 48]. To confirm yChd1 activity on our chromatinized plasmid, we used the frequent cutter, HaeIII, to digest the chromatin in the presence or absence of the remodeler. Upon addition of yChd1, we observed a shift toward lower molecular weight bands, indicative of dimished protection at HaeIII sites while still maintaining a chromatinized state (Figure 5D).

We next sought to test whether yChd1 could affect Cas9's ability to access nucleosomal DNA. Before addition of the remodeler to our chromatinized plasmid, we found that sites within the 601 nucleosome were strongly protected from cleavage by Cas9, consistent with our mononucleosome results (Figure 5E-G, PAM sites without remodeler). However, when the 601 nucleosome was remodeled by yChd1, as indicated by a loss of protection from restriction enzyme cleavage (approaching protection levels similar to those in the linker region), Cas9 cleavage efficiency was restored to around 80% of the corresponding naked plasmid control (Figures 5E-G). Notably, the percent protection at the EcoRI site adjacent to the positioned nucleosome did not decrease upon addition of yChd1, demonstrating that the decrease in protection along the 601 sequence was mediated by the nucleosome displacement activity of yChd1 rather than by a non-specific effect on cleavage efficiency. While our data with the chromatinized plasmid system confirm our findings that a well-positioned nucleosome provides a profound block to Cas9 cleavage, our further finding that chromatin remodeling restores access

to nucleosomal DNA provides one potential mechanism by which Cas9 may efficiently modify broad portions of eukaryotic genomes. This plasmid model could be further exploited to assay the activity of Cas9 at nucleosome-free and boundary sites, and thus derive biophysical parameters governing Cas9-chromatin interactions.

Discussion

Despite its swift success as a repurposed tool for gene editing, imaging, and transcription modulation, the ability of this prokaryotic CRISPR/Cas9 system to effectively navigate eukaryotic chromatin has remained poorly understood. Here, we show that the nucleosome, the basic unit of chromatin, poses a strong barrier to Cas9, both *in vitro* and *in vivo*. By masking ~147bp of DNA, the nucleosome effectively reduces the size of the eukaryotic genome available to Cas9. Previous studies using ChIP-seq to assay Cas9 binding have shown that off-target binding at PAM plus seed sequences more frequently occurs in regions of open chromatin [15, 49]. Our data expand upon these findings to show that the discrete pattern of nucleosome organization is able to modulate the efficiency of Cas9 binding and cleavage at on-target sites. The practical implications of these observations are underscored by our finding that accounting for nucleosome occupancy offers a significant improvement in predictive power for sgRNA design.

While our data show that nucleosomes strongly protect their DNA from Cas9 binding and cleavage *in vitro*, their organization in cells is not static. Transient displacement of nucleosomes occurs during replication, remodeling, and transcription. By adding the chromatin remodeling enzyme yChd1 to nucleosomes *in vitro*, we demonstrate that this displacement can in fact restore Cas9 access to DNA. However, despite brief exposure of nucleosomal DNA during remodeling and various other cellular processes, we still observe a clear anti-correlation between Cas9 activity and nucleosome occupancy *in vivo*, suggesting that the barrier to Cas9 target recognition exists even in a cellular environment. Indeed, the balance between nucleosome disruption, turnover, and repositioning in the cell leads to the average level of occupancy and positioning at each site observed by MNase-seq [19, 21]. Thus, our data suggest that it is likely the overall

effect of this average nucleosome positioning that leads to the observed anti-correlation with Cas9/sgRNA activity. It is important to note, however, that there is likely a fundamental difference between applications that use dCas9 versus nuclease-active Cas9. Knock-down of transcription by CRISPRi likely requires persistent binding by dCas9 in order to continually block transcription, and would be largely ineffective during S-phase when transcription is globally shut down. In contrast, to make a genomic edit, Cas9 must succeed in cleaving DNA only once, and could potentially take advantage of nucleosome turnover during replication. The role these differences play in the dependence of Cas9 on nucleosome position is still not clear. We expect, however, that nucleosome position and occupancy will be of particular concern to applications that use the nuclease-dead Cas9 and require sustained binding. Future investigations into the role of cell cycle and nucleosome disruption may provide an additional piece to our understanding of the mechanism of Cas9 in eukaryotic cells. Furthermore, nucleosome organization represents only one aspect of eukaryotic chromatin, and thus, our results contribute a first step in understanding and exploiting how chromatin affects Cas9 activity in order to enable more sophisticated and precise rules for targeting Cas9.

Methods

Analysis of CRISPRi sgRNA activity

The K562 dCas9-KRAB-BFP cell line was obtained from [17] and had been constructed from K562 cells obtained from ATCC. The resulting cell line tested negative for mycoplasma (MycoAlert Kit, Lonza, Basel, Switzerland) in regular screenings, and cytogenetic profiling by array comparative genomic hybridization (not shown) closely matched previous characterizations of the K562 cell line [50]. Data from 30 published [17] and unpublished screens (M.L., B.A., B.B., C.Y.P., M.K., Y.C., J.F., and J.N., personal communication), conducted using the CRISPRi sgRNA library described in Gilbert 2014 [17] in K562 cells constitutively expressing dCas9-KRAB-BFP, were processed through a standardized pipeline adapted from Bassik et al. 2013 [25], and Kampmann et al. 2013 [51]. Briefly, sgRNA phenotypes were calculated as the $\log_2$ enrichment of sequencing read counts between two conditions (e.g. initial and final timepoints for growth screens, untreated and treated for drug/toxin screens) and normalized to cell doubling differences where appropriate. Most screens were conducted in duplicate, and sgRNA phenotypes from the duplicates were averaged. To determine hit genes, each gene was given an effect size (average of strongest 3 sgRNA phenotypes by absolute value) and a confidence value (Mann-Whitney p-value of all ~10 sgRNAs compared to negative controls), and hits were selected using a score that integrates effect size and statistical confidence (| effect Z score * $\log_{10}$ p-value | $\geqq$20 in any screen). For genes with multiple TSS, each TSS was analyzed separately and the gene was assigned the highest score. Finally, sgRNA phenotypes were extracted for hit genes from the screen in which the gene scored as a hit and normalized to the average of the strongest 3 phenotypes to generate the "sgRNA activity score."

sgRNA positions were defined as the genomic coordinate of the 3' G of the NGG PAM (all genomic coordinates referenced in this text are from hg19). TSS positions were determined from the FANTOM5 project annotation (Riken) (http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/TSS_human.bed.gz; accessed March 2, 2015), using the downstream genomic coordinate of the corresponding "p1@gene" BED file entry. All local averages were calculated in 50bp windows centered around the indicated point. As sgRNA lengths including the PAM were ~24bp and position was calculated relative to PAM, a window size of 50bp captures all sgRNAs that directly neighbor the center point at either the 3' or 5' end. In order to quantify amplitude, the local averages were first smoothed using a low-pass Butterworth filter (N = 4, Wn = 0.03; SciPy signal processing module) and then peaks and troughs were calculated by determining the local maxima and minima, respectively. As described for Figure 1D, Peak 0 was defined as the local maximum closest to the TSS, peak 1 was defined as the next maximum downstream of peak 0, and troughs were defined as the minima immediately downstream of the respective peaks.

MNase-seq data analysis

K562 MNase-seq data was obtained from the ENCODE consortium as processed continuous signal data (BigWig file format; accession number ENCFF000VNN, Michael Snyder lab, Stanford University). sgRNA target site signal was calculated as the average signal at all positions between the 5' end of the sgRNA and the 3' end of the PAM.

RNA-seq data analysis

K562 RNA-seq data was obtained from the ENCODE consortium as transcript

quantifications (accession number ENCFF485YKK, Thomas Gingeras lab, Cold Spring Harbor Laboratories). Genes were assigned expression levels in units of FPKM according to their highest-expressed transcript.

Linear regression

Sequence score was calculated by passing the specified 30bp target site for each sgRNA to the on_target_score_calculator.py script provided by Doench et al., 2015 (accessed October 9, 2015) [10]. This sequence score, MNase signal as calculated above, and sgRNA length (base pairs in protospacer and PAM) were each compared to sgRNA activity scores by Pearson correlation. Linear fits of the specified parameters were computed using multidimensional linear regression (Sci-kit learn linear_model package), and correction of the activity scores for sequence and length features was performed by subtracting the predicted scores based on the combined fit.

Ricin tiling screen

Screens for ricin susceptibility were performed essentially as previously described [17]. Briefly, K562 cells constitutively expressing Cas9-BFP from an SFFV (spleen focus-forming virus) promoter were transduced with our previously described pooled sgRNA tiling library packaged into lentivirus for a multiplicity of infection below 1. Duplicate screens were infected and subsequently treated independently. Infected cells were allowed to recover for 2 days, then selected with 0.75μg/mL puromycin (Tocris) for 2 days, and finally allowed to recover from puromycin treatment for 2 days. Cells were then cultured for 19 days and were either treated with three pulses of 0.5ng/mL ricin administered for 24 hours and followed by re-suspension in

fresh media, or passaged untreated. Genomic DNA was harvested from the endpoint untreated and treated samples and processed for high-throughput Illumina sequencing as previously described [17]. Screens were conducted at a minimum library coverage of 1,000 cells per sgRNA, and sequenced to a median depth of ~500 reads per sgRNA. Phenotypes were calculated as $\log_2$ enrichments of read counts between untreated and treated conditions, normalized to cell doubling differences, and averaged between duplicates. Phenotype-signed Z scores were calculated by dividing all scores by the standard deviation of negative control phenotypes and then multiplying phenotypes by -1 for sgRNAs targeting genes shown to produce sensitizing phenotypes upon knockdown [25] such that positive values represent "expected" phenotypes.

Protein Purification

Mouse histones H2B(T115C), and H4 were recombinantly expressed in BL21(DE3)pLysS cells from expression plasmids gifted by Dr. Karolin Luger. Expression and purification of mH4 was conducted as previously described by the Luger lab [52, 53], while mH2B(T115C) was expressed and purified as described by the Cairns lab [54] with the following exception: after purification, histones were dialyzed against multiple changes of double distilled water and 1mM β-mercaptoethanol (BME) before lyophilizing for storage. Purified recombinant mouse (*Mus Musculus*) histones H2A and H3 were gifts from Dr. Karolin Luger.

The labeling mutant, mH2B(T115C) was fluorescently labeled with ~5-fold molar excess Alexa Fluor 647 C_2 maleimide dye (Thermo Fisher Scientific, Waltham, MA, USA) as follows. 2 mg of lyophilized mH2B(T115C) was dissolved at 2 mg/mL in labeling buffer (7 M GuHCl, 50 mM Tris-HCl pH 7.5, 1 mM TCEP) and nutated at room temperature for 30 minutes. 1 mg of the dye was then dissolved in 50 µl of anhydrous DMF under Argon gas, and approximately half of

the dye solution was slowly mixed with the dissolved mH2B(T115C) in the dark at room temperature to begin the labeling reaction. After nutating the reaction for 1 hour, the rest of the dissolved dye was slowly added and the reaction was moved to 4ºC overnight in the dark. In the morning, the reaction was quenched by adding over 100-fold molar excess of BME.

Histone octamers were refolded and purified as previously described [52]. Specifically, 110 nmoles each mH3 and mH4 were refolded with 130 nmoles each mH2A and mH2B(T115C). The resulting octamer was concentrated using a 10,000 MWCO Spin-X UF concentrator (Corning, Tewksbury, MA, USA), then purified on a Superdex 200 HR (10/30) column (GE Life Sciences, Pittsburgh, PA, USA) using an Akta Explorer FPLC (GE Life Sciences, Pittsburgh, PA, USA) at 0.2 mL/min. Selected fractions were concentrated, flash frozen, and stored at -80ºC until use.

A new purification scheme was conceived to achieve exceptionally high purity (d)Cas9 (Figure 4 – figure supplement 3). Nuclease active *S. pyogenes* Cas9-HaloTag was recombinantly expressed and purified from BL21(DE3)pLysS-Rosetta cells (Novagen/EMD Millipore, Darmstadt, Germany) using the expression plasmid pET302-6His-wtCas9-Halo-NLS, while the nuclease dead *S. pyogenes* dCas9(D10A,H840A)-HaloTag was recombinantly expressed and purified from BL21(DE3)pLysS cells using the expression plasmid pET302-6His-dCas9-Halo [16]. Bacterial cultures were grown in Terrific Broth II (MP Biomedicals, Santa Ana, CA, USA) at 37ºC until an OD_{600} reached 0.4. Cultures were then transferred to an ice bath for ~15 minutes until an OD_{600} reached 0.5, at which point expression was induced with 0.2mM IPTG, and the cultures were moved to an 18ºC shaker for 16 hours. Cells were harvested at 3000xg for 20 minutes, then resuspended in lysis buffer (500 mM NaCl, 50 mM Hepes pH 7.6, 5% glycerol, 1% Triton X-100, 10 mM imidazole, 1 mM benzamadine, 2.3 µg/mL aprotinin, 0.5 mM PMSF,

and 1 tablet per 50 mL of Protease Inhibitor Cocktail (Roche, Basel, Switzerland). Cells were lysed using sonication on ice at 50% duty cycle, power 8, 30 seconds on, 1 minute off (Misonix/ Qsonica, LLC, Newtown, CT, USA). The lysed cells were then ultracentrifuged at 4º, 40K rpms, for 40 minutes to remove cell debris. The supernatant was then allowed to bind to Ni-NTA agarose resin (Qiagen, Hilden, Germany) by nutating at 4ºC for 30 minutes. The resin containing bound (d)Cas9 was poured into a mini column (Bio-Rad, Hercules, CA, U.S.A.) and washed with 10 column volumes (CV) of lysis buffer, and 5 CVs of 20 mM Imidazole buffer (same as lysis buffer but with 20 mM imidazole). Elution of the (d)Cas9 was achieved using 250 mM Imidazole buffer (same as lysis buffer but with 250 mM imidazole), and fractions were checked on an SDS-PAGE gel. Chosen fractions were pooled and diluted to a starting NaCl concentration of 200 mM using Buffer A (0 M NaCl, 50 mM Hepes pH 7.6, 5% glycerol, 1 mM DTT, 0.5 mM PMSF). A 5 mL HiTrap Q-HP (GE Life Sciences, Pittsburgh, PA, USA) and a 5 mL HiTrap SP-HP (GE Life Sciences, Pittsburgh, PA, USA) column were attached in tandem (Q first in line) and equilibrated on an Akta FPLC at 10% Buffer B (same as Buffer A except with 2 M NaCl). The pooled (d)Cas9 was filtered, then loaded onto the tandem columns at 2 mL/min. The columns were washed with 10% Buffer B until A_{280} and A_{260} returned to baseline, at which point the Q column was removed and (d)Cas9 was eluted from the SP column using a gradient from 10% to 50% Buffer B over 10 CVs. Fractions were chosen using SDS-PAGE, pooled, and dialyzed into storage buffer (200 mM NaCl, 50 mM Hepes pH 7.6, 5% glycerol, 1 mM DTT) (Figure 4 – figure supplement 3). Aliquots were flash frozen in liquid nitrogen and stored at -80ºC.

Target DNA purification

To make the fluorescent DNA used in this study, the 601 DNA sequence was amplified by PCR from plasmid pBSIISK+601 (parent 601 sequence plasmid gifted by K. Luger) with the primers listed below. Two different PCR products were produced; one labeling the Watson strand of the 601 sequence with a 5' Cy3 dye (IDT, Coralville, IA, USA), and the other labeling the Crick strand. Large scale (~2 mL) PCR reactions using in-house produced Pfu DNA polymerase were ethanol precipitated before loading onto a 20cm x 20cm 12% native TBE-PAGE gel for DNA purification. The PCR product was cut out of the gel, and the gel slice was crushed and soaked in 0.3 M Sodium Acetate pH 5.2 with multiple buffer changes. The pooled extract was ethanol precipitated, resuspended in 10 mM Tris-HCl pH 7.5, 10 mM NaCl, and stored at -20º until use.

Watson Primer pair:

LW021: 5'- /5Cy3/ATCGGATGTATATATCTGACACGTGC -3'

LW022: 5'- ATCTGAGAATCCGGTGCCG -3'

Crick Primer pair:

LW025: 5'- /5Cy3/ATCTGAGAATCCGGTGCCGAG -3'

LW030: 5'- ATCGGATGTATATATCTGACACGTGC -3'

sgRNA Production for *in vitro* experiments

sgRNA was produced by T7 transcription of a short DNA oligo template. To create this template, two oligos, one containing the T7 promoter and DNA target sequence, and the other encoding the invariable scaffolding of the sgRNA, were annealed and filled in using a single

PCR cycle. The DNA template was ethanol precipitated, then transcribed using the T7 Quick High Yield RNA Synthesis Kit (New England Biolabs, Ipswich, MA, USA) according to vendor instructions. The resulting sgRNAs were purified via 6% Urea-PAGE gel. The correct band was cut out and crushed and soaked in 0.3M NaOAc. After ethanol precipitation, and resuspension in double distilled water, aliquots were stored at -80ºC until use. SgRNAs targeting the 601 sequence included 20bp of complimentarity 5' to the targeted PAM, with the exception of PAM 4, which has room for only 19bp of complimentarity. SgRNAs were used against PAMs 1, 3-15, and 17-20. PAM 2 did not have a long enough target sequence available within the 601. The non-sense guides used in this study contained the target sequence 5'-ACATGTTGATTTCCTGAAA-3' or 5'-GATTTCACCTCTCAGCGCAT-3'.

Template oligonucleotides:

5'-
TTAATACGACTCACTATAGNNNNNNNNNNNNNNNNNNNNGTTTTAGAGCTAGAAAT
AGC -3'
5'-
AAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTA
ACTTGCTATTTCTAGCTCTAAAAC -3'

Nucleosome assembly

Nucleosomes were assembled by salt gradient dialysis as previously described [52] using 50 µL home made dialysis buttons and ~1 µM DNA. The Watson and the Crick labeled 601

DNA were assembled into separate nucleosomes. The best DNA:octamer molar ratio for optimal nucleosome assembly was selected by titrating octamer. The most homogenous assembly as judged by 6% Native TBE-PAGE gel was chosen for future assays.

Cleavage and binding assays

(d)Cas9 cleavage and binding assays were both conducted in the same manner, exchanging dCas9 for wtCas9 during binding reactions. First, the complete ribonucleoprotein complex was formed by incubating 5-fold molar excess of the chosen sgRNA with (d)Cas9 at 37ºC for 10 minutes. Next, the DNA substrate (either naked DNA or nucleosome) was added to 40 nM (5-fold less than Cas9), and the reactions were returned to 37º for an hour. Reaction buffer contained 20 mM Hepes pH 7.5, 100 mM NaCl, 5 mM $MgCl_2$, 1 mM EDTA, and importantly, 2.5 mg/mL insulin (Roche, Basel, Switzerland). Importantly, we found that including a nonspecific protein such as insulin prevents the nucleosomes from falling apart and getting lost to surfaces during binding and cleavage assays (Figure 4 – figure supplement 1B). Additionally, inclusion of insulin reduces nonspecific protein-protein interactions and aggregation. Binding reactions also contained 14.6% sucrose to allow direct loading onto a native gel without a loading buffer. At 1 hour, binding reactions were loaded onto a pre-run 6% native TBE-PAGE gel at 4ºC, while cleavage reactions were stopped using a 5X stop buffer (250 mM EDTA, 2% SDS), and prepared to load onto a 10% Urea-PAGE gel by adding 2X loading buffer (95% formamide, 20% DMSO, 5 mM EDTA, 0.025% Orange G) and heating to 95ºC for 5 minutes before snap cooling on ice. Binding and cleavage reactions were repeated at least twice each for most of the PAMs targeted. Gels were imaged using a PharosFX Plus (Bio-Rad, Hercules, CA, U.S.A.) and quantified using Image Lab (Bio-Rad, Hercules, CA, U.S.A.). For

binding and cleavage reactions, gels were scanned in the Cy3 fluorescence channel, while Alexa Fluor 647 fluorescence was also imaged for binding reactions. For cleavage reactions, labeling of either the Watson or Crick strand was chosen such that the fluorophore was always attached to the strand complimentary to the sgRNA used.

Structure alignment and solvent accessibility

Molecular graphics and analyses were produced using the UCSF Chimera package (supported by NIGMS P41-GM103311) [55]. The crystal structures for the 601 nucleosome [43] (PDBID 3LZ0) and the Cas9-sgRNA-DNA ternary complex [44] (PDBID 4UN3) were superimposed by aligning the DNA path in both structures using MatchMaker and manual manipulation. The solvent accessible surface area of the DNA in the 601 nucleosome structure was computed using residue areaSAS.

Reconstitution of nucleosomes on a plasmid

Supercoiled plasmid was produced using a Qiagen Maxi prep kit (Hilden, Germany). Histone octamers were prepared from purified histones as described above. DNA and histone octamer were mixed at a weight ratio of 1:0.8 (DNA to octamer) at a concentration of 20µg of DNA in 1X TE with 2M NaCl. 50µL homemade dialysis buttons were used to dialyse the solution by a step-wise gradient in 500mL at room temperature. The gradient was as follows; 1.5 hr in 1M NaCl in TE, 2 hr in 0.8M NaCl in TE, 1.5 hr in 0.6M NaCl in TE, and 2 hr in 0.05M NaCl in TE**.** The resulting chromatin was stored at 4ºC until use.

Micrococcal nuclease digestion assays of chromatin assembly reactions

MNase (Sigma-Aldrich, St. Louis, USA) was resuspended at 200 units/mL in water and then disluted at 1:100, 1:500, or 1:1000 in a solution of 1X MNase reaction buffer (50mM Tris-HCl pH7.9 and 5mM $CaCl_2$ dihydrate), 0.1 mg/mL insulin, and 10% glycerol. Chromatinized plasmid assemblies in 1X MNase buffer were added to each MNase dilution. As a control, an equal amount of DNA as supercoiled plasmid was added to the lowest MNase dilution. Each reaction was incubated for 11 minutes at room temperature, then stopped in a solution with final concentrations of 20mM EGTA, 200mM NaCl, 1% SDS, 20 mg/mL GlycoBlue (Thermo Fisher Scientific, Waltham, MA, USA), and 13.4 mg/mL Proteinase K (Roche, Basel, Switzerland) and incubated at 37ºC for 30 minutes. Reactions were Phenol:Chloroform:Isoamyl Alcohol extracted, ethanol precipitated, and run on a 1.3% agarose gel in 0.5X TBE. 3µg 123bp DNA ladder (Thermo Fisher Scientific, Waltham, MA, USA) was used as a standard.

Restriction enzyme accessibility assays

Chromatin assembly or supercoiled plasmid with equivalent amounts of DNA were added to restriction enzyme master mixes for final concentrations of 1X CutSmart buffer (New England Biolabs, Ipswich, MA, USA), 6 ng/µL DNA, and 0.4 U/µL of the indicated restriction enzyme (New England Biolabs). Reactions were incubated for 1 hour at NEB recommended temperatures. Reactions were stopped in final concentrations of 15mM EDTA, 0.75% SDS, 150mM NaCl, and 15mg/mL Proteinase K and incubated at 37ºC for 30 minutes. DNA was extracted with Phenol:Chloroform:Isoamyl Alcohol, and ethanol precipitated, then resuspended in 1X CutSmart buffer and linearized by digesting with 1 U/µL DraIII-HF (New England Biolabs) at 37ºC for 1 hour. The full reactions were run on an agarose gel and quantified as above.

Chromatin remodeling assays

Reactions were set up with 40mM NaCl, 0.1 mg/mL BSA, 25mM Tris-Acetate pH 7.5, 10mM Mg-Acetate, 1mM DTT, 1.2mM rATP, and 6 ng/ μL chromatinized or supercoiled DNA. For restriction enzyme accessibility, the indicated restriction enzyme was added at 0.5 units/mL. For Cas9 accessiblity, Cas9/sgRNA ribonucleoproteins assembled as described above were added at 15.62nM. Chromatin remodeling enzyme yChd1$_{\Delta NC}$ (gift of Dr. Ashok Patel and Dr. Gregory Bowman) was added to the indicated reactions at 0.2 μM final concentration. Reactions were incubated at 27ºC for 1 hour, then processed as with the restriction enzyme accessibility assay above.

Acknowledgments

We would like to thank Dr. Karolin Luger, Pamela Dyer, and Dr. Uma Muthurajan for reagents and training in preparing and working with nucleosomes, and Dr. Ashok Patel and Dr. Gregory Bowman for the kind gift of yChd1. We would also like to thank Manuel Leonetti, Dr. Britt Adamson, Ben Barsi-Rhyne, Dr. Chong Y. Park, Dr. Martin Kampmann, Yuwen Chen, Dr. Jonathan Friedman, and Dr. Jodi Nunnari for generously sharing unpublished screening data for determination of sgRNA activity. Additionally, we would like to thank members of the Tjian and Weissman labs, in particular, Dr. Elisa Zhang, Dr. Liangqi Xie, and Chiahao Tsui for their help with *in vitro* reagents, and Dr. Alex Fields and Joshua Dunn, for helpful discussions and assistance.

Figures

Figure 1

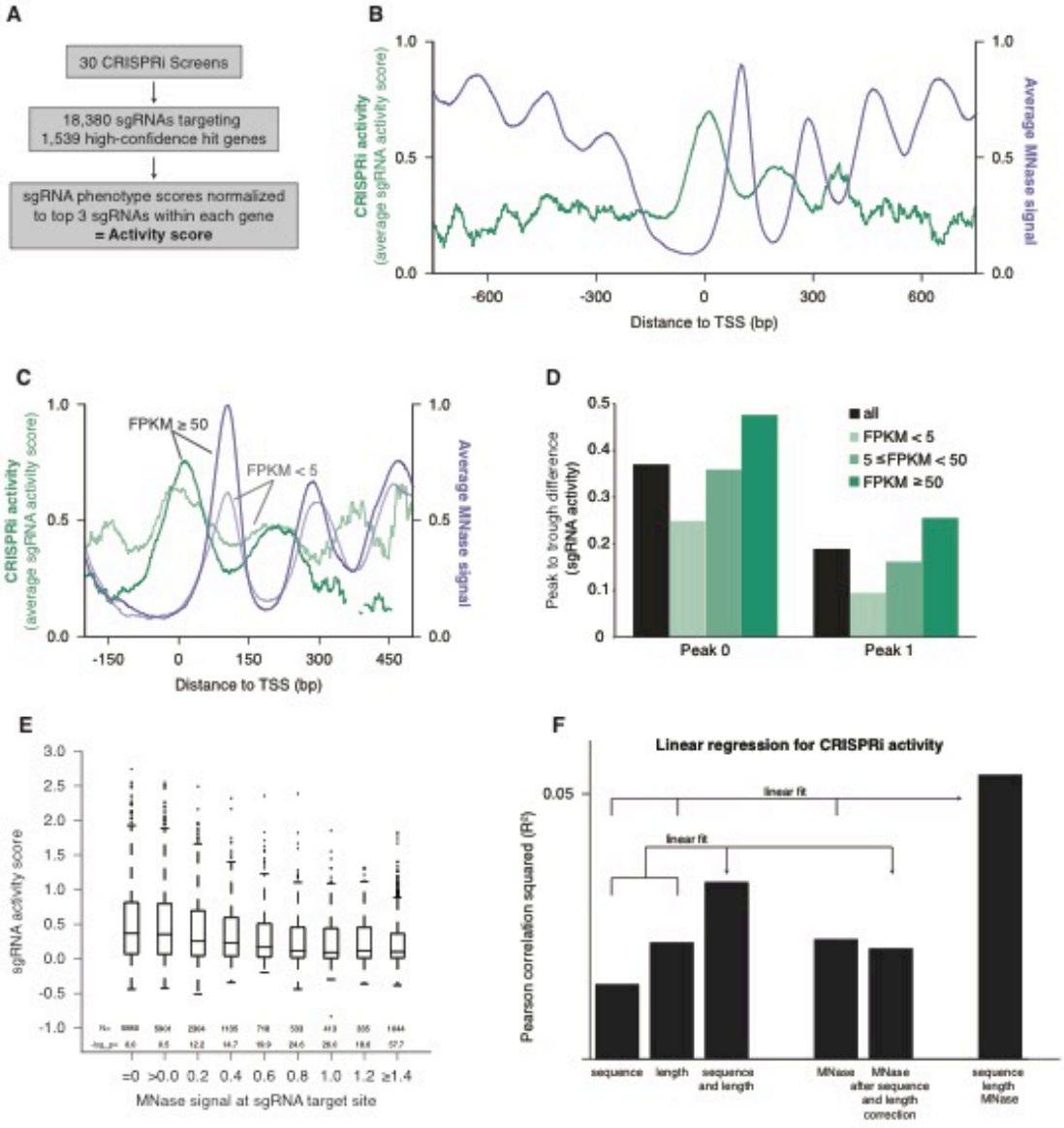


Figure 1. CRISPRi activity anti-correlates with nucleosome occupancy

(A) Workflow for generating CRISPRi activity scores from pooled genetic screens. The resulting values are distributed around 0 for inactive sgRNAs and around 1 for highly active sgRNAs. (B) Average CRISPRi activity and MNase-seq signal relative to the TSS. Green line represents average CRISPRi activity score of all sgRNAs within a 50bp window around each position. Purple line represents the K562 MNase-seq signal at each position averaged across all genes analyzed. (C) Average CRISPRi activity and MNase-seq signal for genes grouped by expression value. Genes were grouped into lower expression (light lines; N=240), higher expression (heavy lines; N=368), and medium expression (omitted for clarity; N=930), and analyzed as in (B). Expression values were obtained as fragments per kilobase million (FPKM) from ENCODE K562 RNA-seq data. Average activity at positions with fewer than 10 sgRNAs within the 50bp window was not calculated. (D) Quantification of the amplitude of periodic CRISPRi activity. Peak and trough coordinates were obtained by calculating the local maxima and minima of the activity traces from analyses in (B) and (C). Peak 0 was defined as the local maximum closest to the TSS, peak 1 was defined as the next maximum downstream of peak 0, and troughs were defined as the minima immediately downstream of the respective peaks. (E) CRISPRi activity and target site nucleosome occupancy for individual sgRNAs. Target site nucleosome occupancy was calculated from the average MNase-seq signal at all genomic coordinates across the length of the sgRNA protospacer and the protospacer adjacent motif (PAM). sgRNAs were then binned by the target site nucleosome occupancy, displayed as box-and-whisker plots, and labeled according to the minimum value within the bin except where indicated. P-values were calculated by a two-tailed Mann-Whitney test comparing each bin to the =0.0 bin. (F) Linear regression for CRISPRi activity. The squared Pearson correlation was calculated for the sgRNA activity scores

compared to the indicated individual parameters (bars 1, 2, and 4) or linear fits of multiple parameters (bars 3 and 6). sgRNA activity scores were corrected for sequence and length features (bar 5) by subtracting the linear fit of those two features.

Figure 2

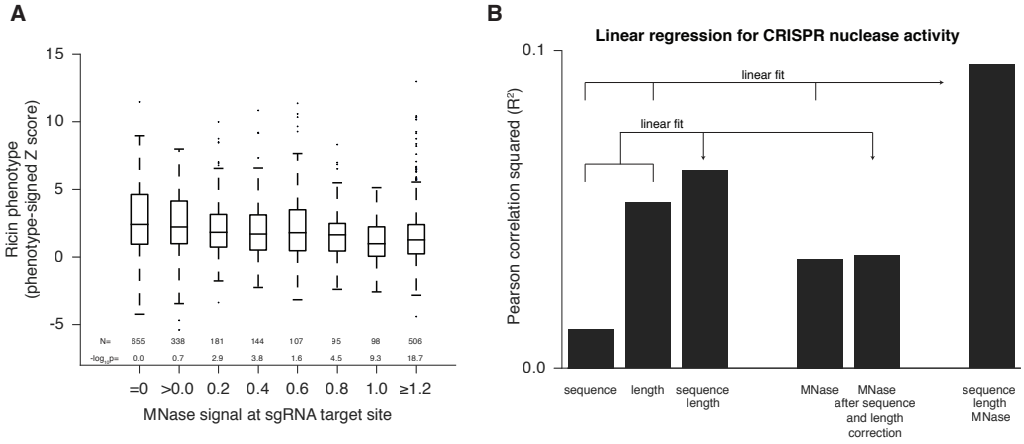

A
Ricin phenotype
(phenotype-signed Z score)
15
10
5
0
-5
N=
555
338
181
144
107
95
98
506
-log10p=
0.0
0.7
2.9
3.8
1.6
4.5
9.3
18.7
=0
>0.0
0.2
0.4
0.6
0.8
1.0
≥1.2
MNase signal at sgRNA target site
B
Linear regression for CRISPR nuclease activity
Pearson correlation squared (R²)
0.1
0.0
linear fit
linear fit
sequence
length
sequence length
MNase
MNase after sequence and length correction
sequence length MNase

Figure 2. Cas9 nuclease activity anti-correlates with nucleosome occupancy

(A) Cas9 nuclease phenotypes and target site nucleosome occupancy for individual sgRNAs targeting CDS regions. Ricin susceptibility phenotypes for each sgRNA are expressed as a Z score and are positive if the phenotype matches the expected knockdown phenotype. Target site nucleosome occupancy was calculated as in Figure 1E. sgRNAs were then binned by the target site nucleosome occupancy, displayed as box-and-whisker plots, and labeled according to the minimum value within the bin except where indicated. P-values were calculated by a two-tailed Mann-Whitney test comparing each bin to the =0.0 bin. (B) Linear regression for Cas9 nuclease phenotypes. The squared Pearson correlation was calculated for the sgRNA activity scores compared to the indicated individual parameters (bars 1, 2, and 4) or linear fits of multiple parameters (bars 3 and 6). sgRNA activity scores were corrected for sequence and length features (bar 5) by subtracting the linear fit of those two features.

Figure 2 – figure supplement 1

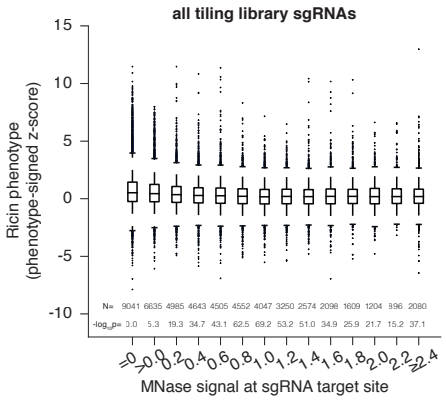


Figure 2—figure supplement 1. Cas9 nuclease activity anti-correlates with nucleosome occupancy at all target sites

Cas9 nuclease phenotypes and target site nucleosome occupancy for individual sgRNAs. As in Figure 2A, but incorporating all CDS- and non-CDS-targeting sgRNAs in the ricin-susceptibility gene tiling library.

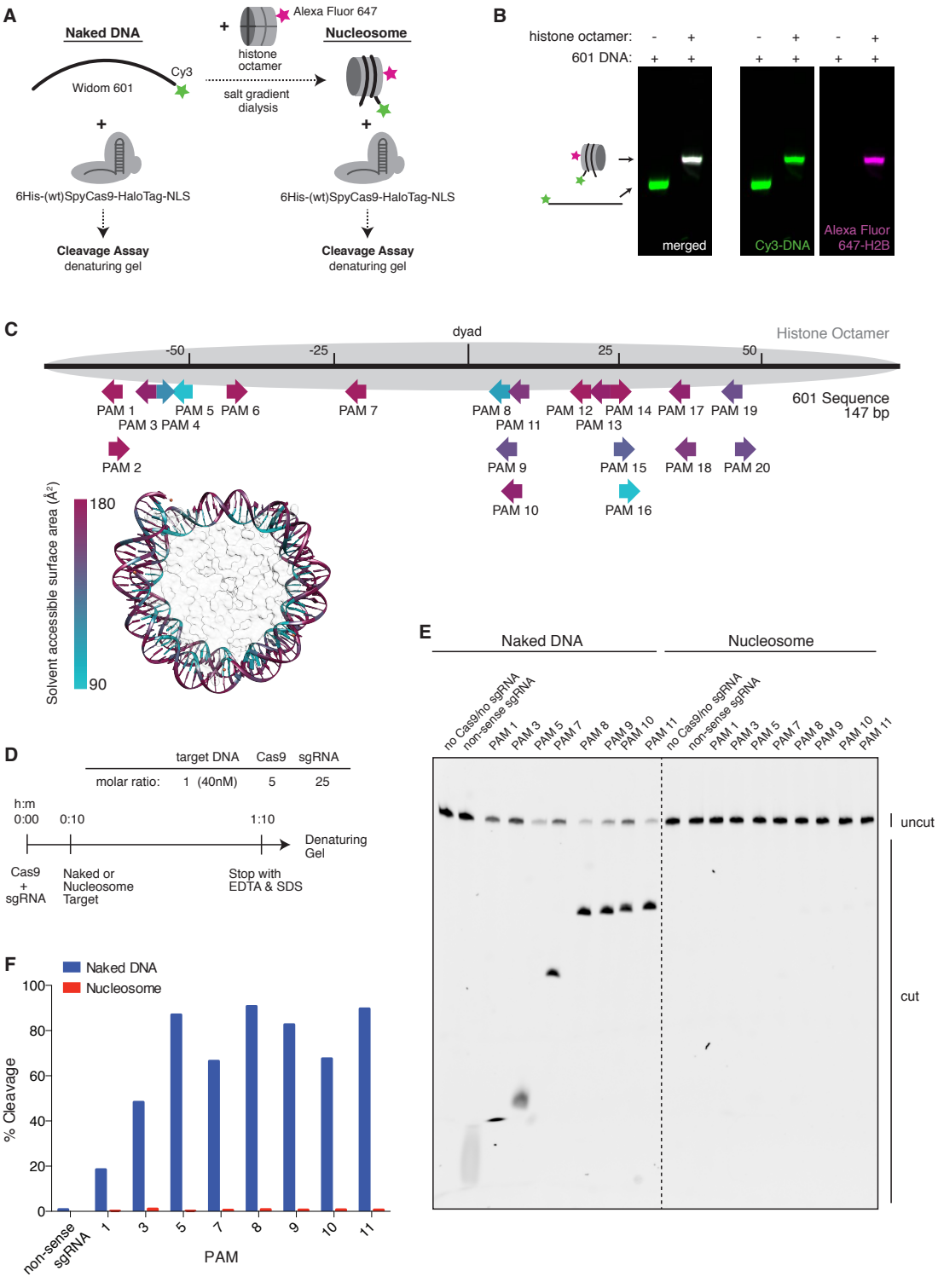

A
Naked DNA
Widom 601
Cy3
Alexa Fluor 647
histone octamer
salt gradient dialysis
Nucleosome
6His-(wt)SpyCas9-HaloTag-NLS
Cleavage Assay
denaturing gel
B
histone octamer:
601 DNA:
merged
Cy3-DNA
Alexa Fluor 647-H2B
C
dyad
Histone Octamer
-50
-25
25
50
601 Sequence 147 bp
PAM 1
PAM 3
PAM 4
PAM 5
PAM 6
PAM 7
PAM 8
PAM 11
PAM 12
PAM 13
PAM 14
PAM 17
PAM 19
PAM 2
PAM 9
PAM 10
PAM 15
PAM 16
PAM 18
PAM 20
Solvent accessible surface area (Å²)
180
90
D
target DNA
Cas9
sgRNA
molar ratio:
1 (40nM)
5
25
h:m
0:00
0:10
1:10
Denaturing Gel
Cas9 + sgRNA
Naked or Nucleosome Target
Stop with EDTA & SDS
E
Naked DNA
Nucleosome
no Cas9/no sgRNA
non-sense sgRNA
PAM 1
PAM 3
PAM 5
PAM 7
PAM 8
PAM 9
PAM 10
PAM 11
uncut
cut
F
Naked DNA
Nucleosome
% Cleavage
100
80
60
40
20
0
non-sense sgRNA
1
3
5
7
8
9
10
11
PAM

Figure 3. Cas9 nuclease activity is blocked by the presence of a nucleosome *in vitro*

(A) Schematic of the experimental setup for *in vitro* cleavage assays. Mononucleosomes were assembled by salt gradient dialysis of purified mouse histone octamer with the minimal nucleosome positioning sequence, Widom 601 (147bp). Prior to assembly, DNA was 5'-end-labeled with Cy3 and histone H2B was fluorescently labeled using an introduced cysteine (T115C) coupled to Alexa Fluor 647. Purified Halo-tagged Cas9 with *in vitro* transcribed sgRNAs were added to naked DNA or assembled nucleosomes, and DNA cleavage products were visualized using a denaturing gel imaged for Cy3-DNA fluorescence. See also Figure 3–figure supplement 1A. (B) Confirmation that fully occupied, well-positioned nucleosomes were assembled. After assembly using salt gradient dialysis, the produced nucleosomes were visualized using a native PAGE gel imaged in the Cy3 and Alexa Fluor 647 channels. Full incorporation of the free DNA into a nucleosome occupying a single position on the DNA is indicated by the presence of a single shifted band containing all of the detectable Cy3-DNA and Alexa Fluor 647-H2B signal. (C) Available PAMs and solvent accessibility of the 601 nucleosome positioning sequence. (Above) A schematic of the 601 sequence. The location of the histone octamer when assembled into a nucleosome is indicated by the gray oval. The location of PAMs within the double-stranded sequence are indicated with arrows spanning the 3 bp of the PAM, pointing in the 5' to 3' orientation of the NGG motif. The arrows are colored according to solvent accessibility at the center of the PAM as calculated from the crystal structure of the 601 nucleosome (PDBID 3LZ0). (Below) Crystal structure of the 601 nucleosome. For clarity, the surface area of the histones in the crystal structure has been made transparent. The DNA in the crystal structure is colored according to solvent accessibility using the same color scale as the PAMs above. Residues colored teal are less accessible, while residues colored fuchsia are more

accessible by solvent. (D) Experimental conditions and timeline for cleavage assays. (E) Denaturing PAGE gel showing results of a cleavage assay targeting the indicated PAMs. Cleavage reactions containing naked DNA were loaded on the left half of the gel, while reactions containing nucleosomes were loaded on the right. The DNA was imaged via a Cy3 fluorophore attached to the 5' end of the sgRNA-complimentary strand. A negative control was conducted with an sgRNA that had no sequence complementarity to the 601 sequence used (non-sense guide). See also Figure 3–figure supplement 1B,C for additional controls. (F) Quantification of the gel in (D). For each lane, percent cleavage was determined by calculating the percent of the total band signal corresponding to cleaved DNA.

Figure 3 – figure supplement 1

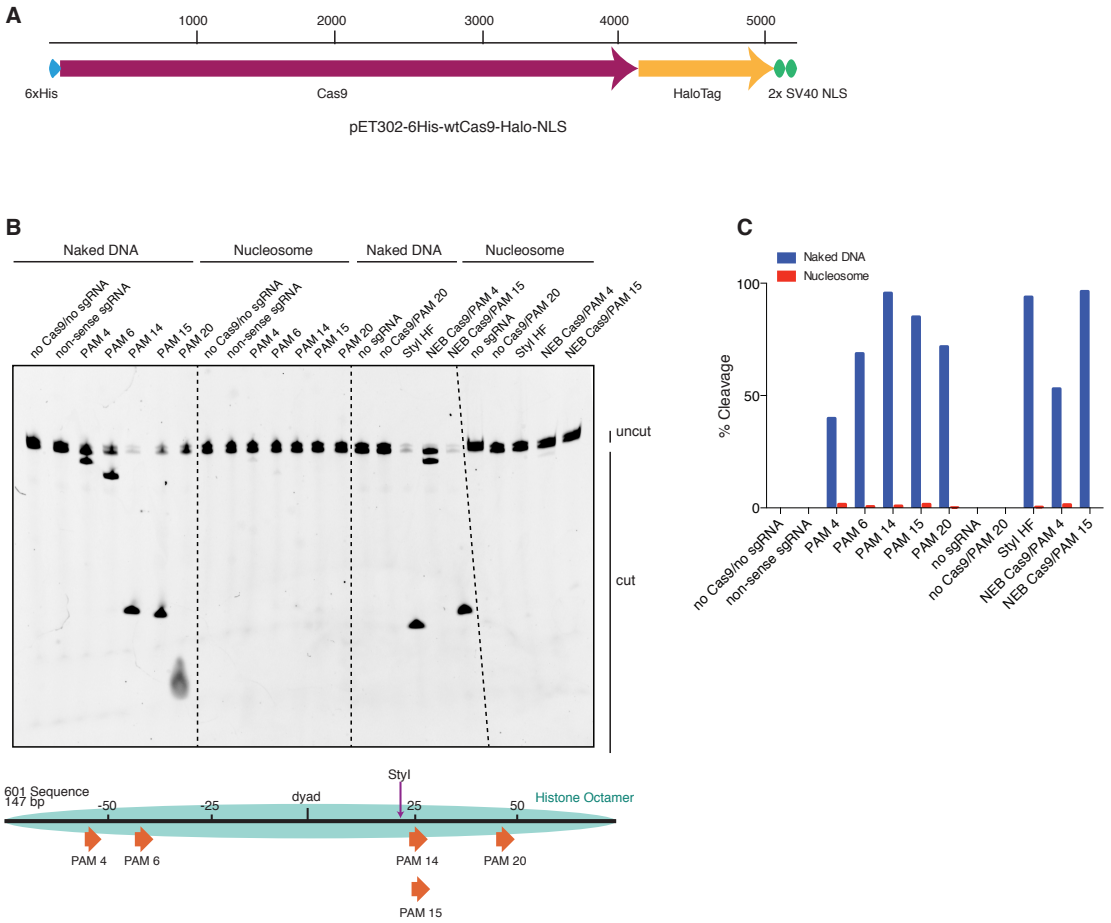


Figure 3—figure supplement 1. HaloTagged Cas9 activity is indistinguishable from untagged Cas9

(A) Diagram of the CDS region in the Cas9 expression plasmid used in this study. (B) Cleavage assay comparing the HaloTagged Cas9 construct used in this study with an untagged Cas9 commercially purchased from New England Biolabs (NEB). Both forms of Cas9 were incubated with either naked DNA or the same DNA assembled into a nucleosome (see Figure 3A,C). A positive control used the restriction enzyme, StyI-HF (from NEB), to target a sequence at a location within the DNA known to be fully protected upon assembly into a nucleosome. Unless explicitly labaled as NEB, all constructs of Cas9 used were HaloTagged.

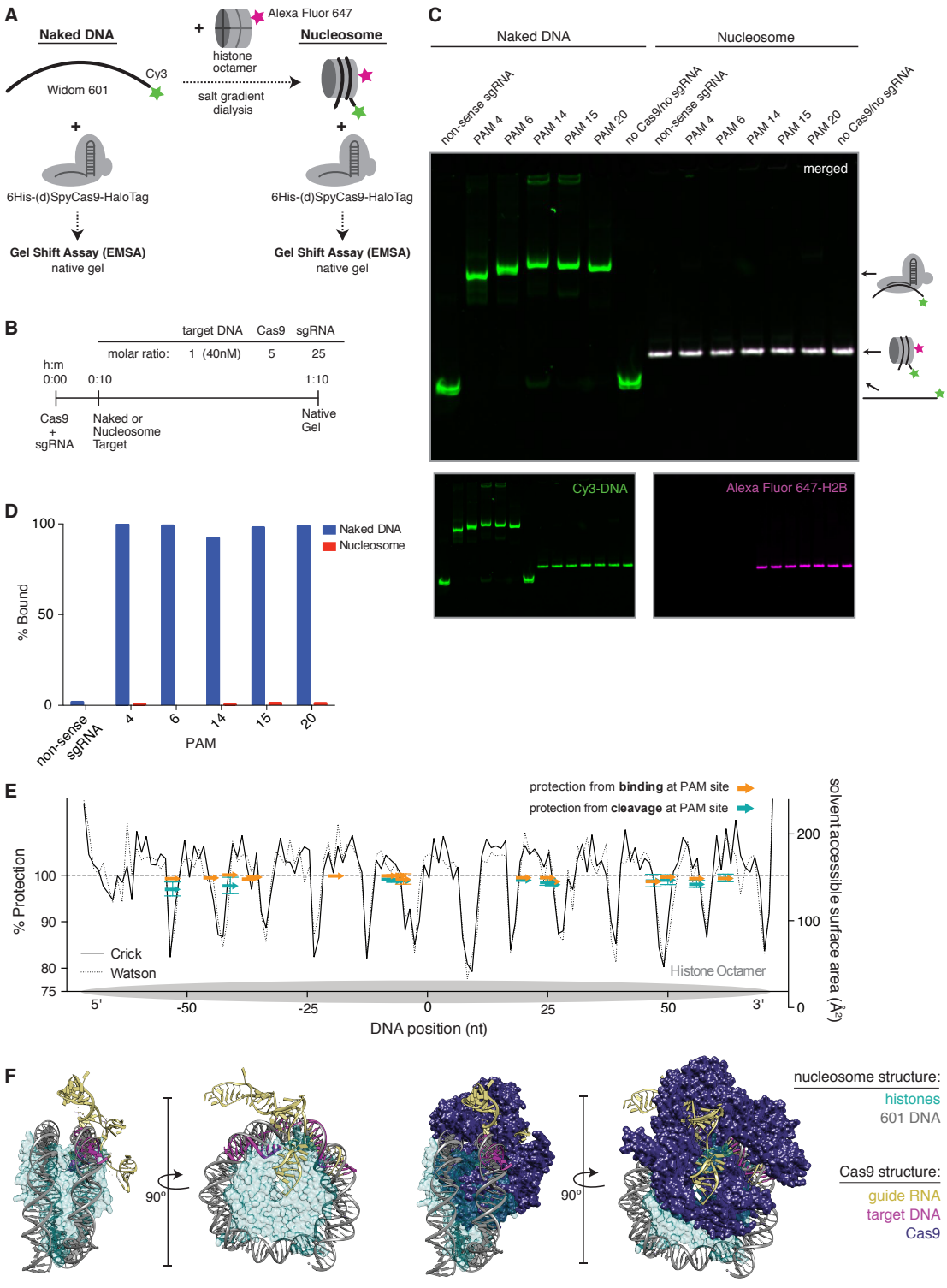

A
Naked DNA
Widom 601
Cy3
6His-(d)SpyCas9-HaloTag
Gel Shift Assay (EMSA)
native gel
Alexa Fluor 647
histone octamer
salt gradient dialysis
Nucleosome
6His-(d)SpyCas9-HaloTag
Gel Shift Assay (EMSA)
native gel
B
target DNA
Cas9
sgRNA
molar ratio:
1 (40nM)
5
25
h:m
0:00
0:10
1:10
Cas9 + sgRNA
Naked or Nucleosome Target
Native Gel
C
Naked DNA
Nucleosome
non-sense sgRNA
PAM 4
PAM 6
PAM 14
PAM 15
PAM 20
no Cas9/no sgRNA
merged
Cy3-DNA
Alexa Fluor 647-H2B
D
% Bound
Naked DNA
Nucleosome
non-sense sgRNA
PAM
E
protection from binding at PAM site
protection from cleavage at PAM site
% Protection
solvent accessible surface area (Å²)
Crick
Watson
Histone Octamer
DNA position (nt)
F
90°
nucleosome structure:
histones
601 DNA
Cas9 structure:
guide RNA
target DNA
Cas9

Figure 4. dCas9 is unable to bind nucleosomal DNA *in vitro*

(A) Schematic of the experimental setup for *in vitro* binding assays. Either naked DNA or assembled nucleosomes were incubated with catalytically dead Cas9 (d-SpyCas9-Halo), and binding was assessed by an electrophoretic mobility shift assay (EMSA). (B) Experimental conditions and timeline for binding assays. (C) A native PAGE gel showing the results of an EMSA in which dCas9 was targeted to the indicated PAMs on either naked or nucleosomal DNA. Gels were scanned for fluorescence from Cy3 on the DNA (green) and Alexa Fluor 647 on histone H2B (magenta). The two color channels were merged to identify the location of intact nucleosomes (white). See also Figure 4–figure supplement 1 and 3 for reagent preparation and experimental conditions, and Figure 4–figure supplement 2 for comparison with wtCas9 binding. (D) Quantification of the gel in (C). Percent bound was determined by calculating the percent of the total band signal in each lane corresponding to Cas9-bound target as determined by a shift in mobility within the gel. (E) Summary of binding and cleavage results for each PAM tested. The ability of Cas9 to bind or cleave nucleosomal DNA at a targeted PAM is displayed as percent protection by the nucleosome, and was calculated by taking the ratio of binding or cleavage on nucleosomal DNA to that on naked DNA. While only the largest error bars are visible, replicates were performed for 15 of the 30 data points and are displayed with error bars showing standard deviation from the mean. The DNA positions plotted correspond to the three nucleotides of the targeted PAM. In order to compare percent protection from binding and cleavage with solvent accessibility, the PAMs are overlaid with a plot of the solvent accessible surface area for each strand (Watson or Crick) of DNA in the 601 nucleosome structure. The percent protection at each PAM, as well as the solvent accessibility were plotted so that the 5’ end of each DNA strand begins at the left of the graph, where position 0 indicates the dyad. (F) Structural

assessment of the ability of Cas9 to bind nucleosomal DNA. Superposition of the Cas9-guideRNA-DNA crystal structure [41] onto the 601 nucleosome crystal structure [40] was achieved by alignment of the DNA path in both structures. (Left) To better view the alignment of the Cas9 target DNA with the nucleosomal DNA, the Cas9 protein density has been removed. (Right) After alignment of the DNA, inclusion of the Cas9 protein density reveals extensive steric clashes with the histones. Histone surface area was made partially transparent to better reveal the overlapping densities with Cas9.

Figure 4 – figure supplement 1

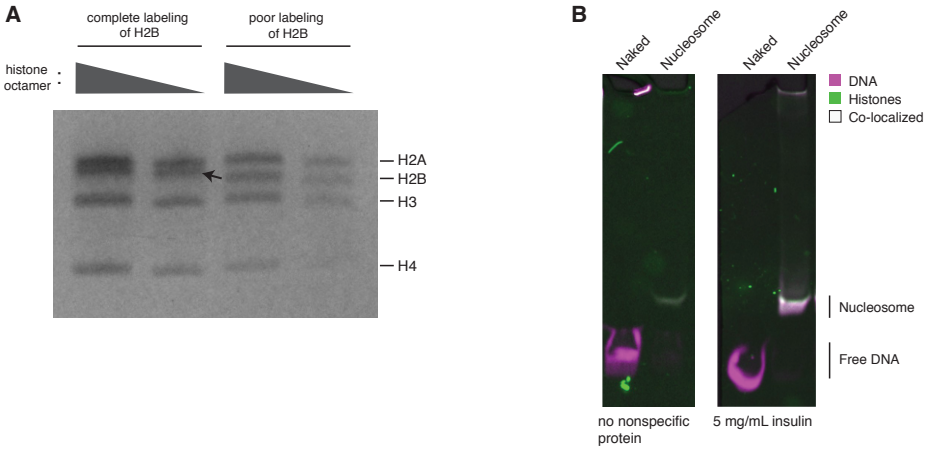


Figure 4—figure supplement 1. Quality control

(A) H2B labeling is near complete. An SDS-PAGE gel after refolding and purifying the histone octamer containing fully labeled versus poorly labeled H2B. The arrow indicates a mobility shift of H2B corresponding to a fully labeled band. Each histone is present at equimolar ratios as indicated by PageBlue protein stain. (B) Nucleosome stability in Cas9 binding and cleavage assays is ensured by including a nonspecific protein in solution. On the left, a native PAGE gel showing naked DNA and nucleosomes under Cas9 binding and cleavage reaction conditions without a nonspecific protein in solution. On the right, the same conditions plus inclusion of a nonspecific competitor protein, insulin.

Figure 4 – figure supplement 2

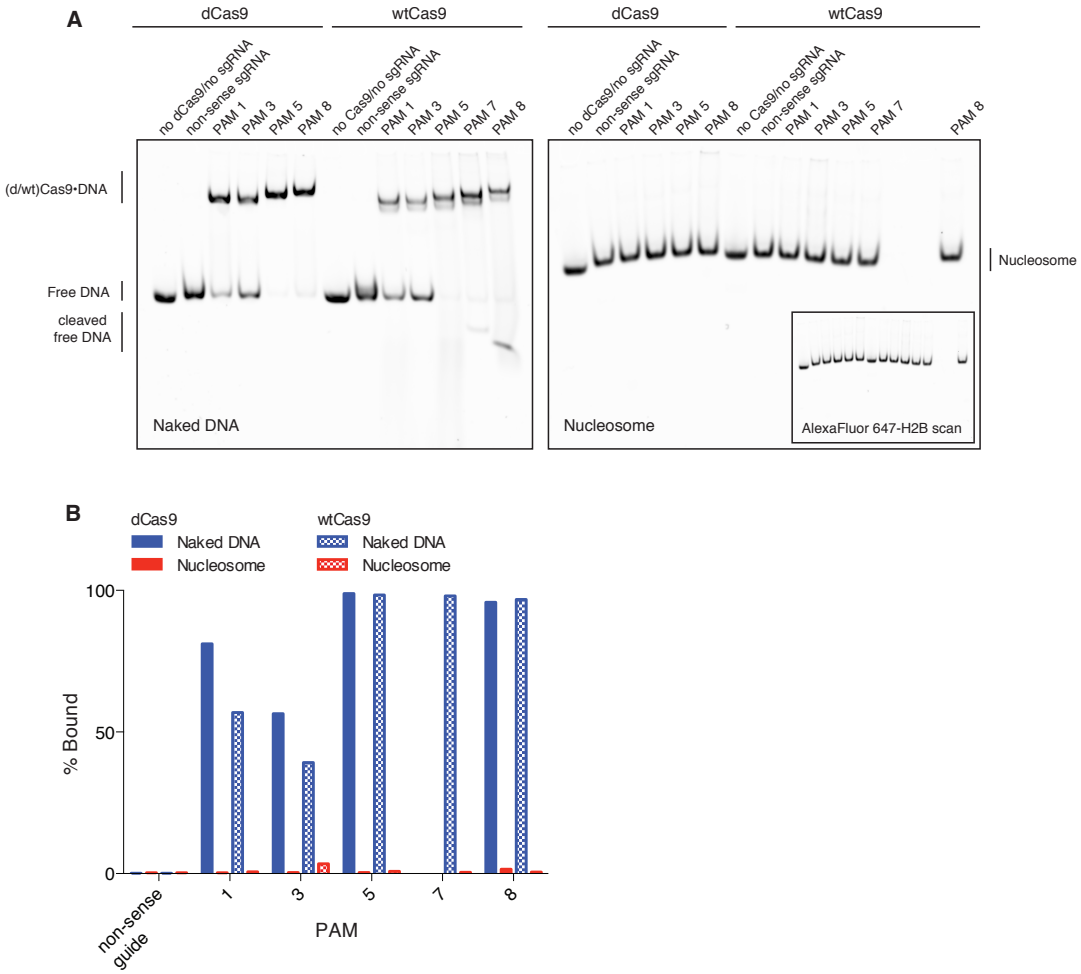


Figure 4— figure supplement 2. DNA binding by dCas9 is also representative of wtCas9 binding

(A) Two native PAGE gels showing the results of an EMSA binding assay comparing dCas9 and wtCas9. Binding to naked DNA is shown in the gel on the left, while binding to nucleosomes is shown in the gel on the right. Both gels were imaged in the Cy3-DNA channel, while the gel on the right was also imaged in the Alexa Fluor 647 – H2B channel (inset image). Binding conditions for both dCas9 and wtCas9 were identical, and were as described in the methods. (B) Quantification of the Native PAGE gel in (A).

Figure 4 – figure supplement 3

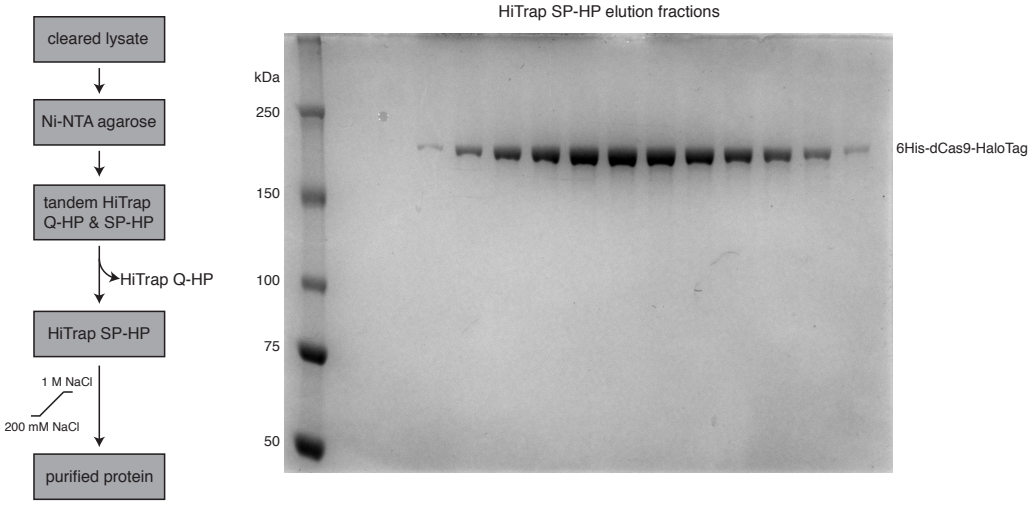


Figure 4— figure supplement 3. (d)Cas9 purification strategy

(d)Cas9 purification scheme (left). SDS-PAGE gel stained with PageBlue (Life Technologies, Carlsbad, CA) showing the elution fractions from the HiTrap SP-HP column during purification of 6His-dCas9-HaloTag (right).

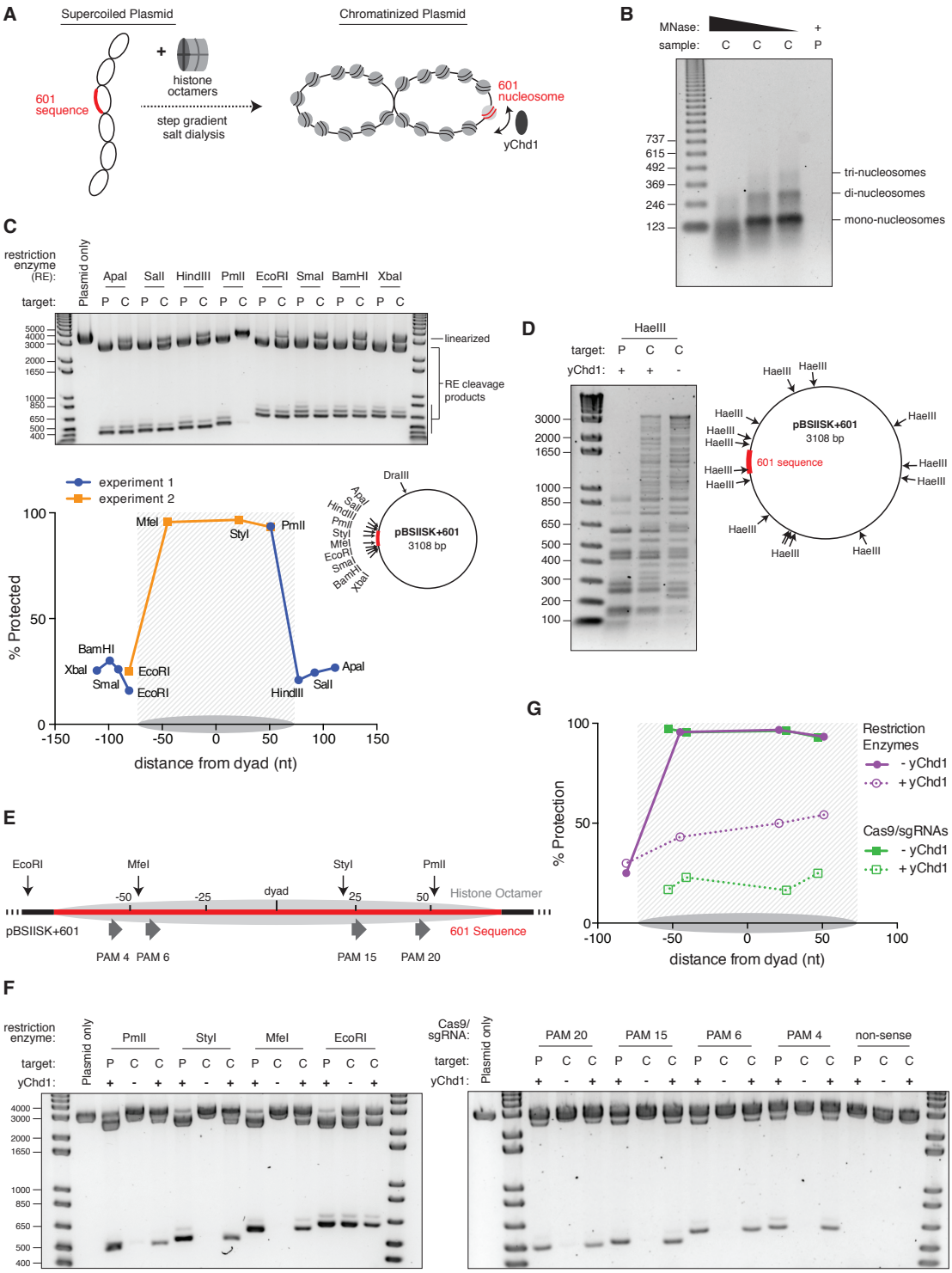

A
Supercoiled Plasmid
Chromatinized Plasmid
601 sequence
histone octamers
step gradient salt dialysis
601 nucleosome
yChd1
B
MNase:
sample: C C C P
737
615
492
369
246
123
tri-nucleosomes
di-nucleosomes
mono-nucleosomes
C
restriction enzyme (RE):
target:
Plasmid only
ApaI SalI HindIII PmlI EcoRI SmaI BamHI XbaI
P C P C P C P C P C P C P C P C
5000
4000
3000
2000
1650
1000
850
650
500
400
linearized
RE cleavage products
experiment 1
experiment 2
% Protected
100
50
0
-150 -100 -50 0 50 100 150
distance from dyad (nt)
MfeI
StyI
PmlI
BamHI
XbaI
SmaI
EcoRI
EcoRI
HindIII
SalI
ApaI
DraIII
pBSIISK+601
3108 bp
D
HaeIII
target: P C C
yChd1: + + -
3000
2000
1650
1000
850
650
500
400
300
200
100
pBSIISK+601
3108 bp
601 sequence
HaeIII
E
EcoRI
MfeI
StyI
PmlI
-50
-25
dyad
25
50
Histone Octamer
pBSIISK+601
601 Sequence
PAM 4
PAM 6
PAM 15
PAM 20
G
% Protection
100
50
0
-100 -50 0 50 100
distance from dyad (nt)
Restriction Enzymes
- yChd1
+ yChd1
Cas9/sgRNAs
- yChd1
+ yChd1
F
restriction enzyme:
Plasmid only
PmlI StyI MfeI EcoRI
target:
yChd1:
P C C P C C P C C P C C
+ - + + - + + - + + - +
4000
3000
2000
1650
1000
850
650
500
400
Cas9/ sgRNA:
Plasmid only
PAM 20 PAM 15 PAM 6 PAM 4 non-sense
target:
yChd1:
P C C P C C P C C P C C P C C
+ - + + - + + - + + - + + - +

Figure 5. Nucleosomes within chromatinized DNA can block cleavage by Cas9, but a chromatin remodeling factor can restore Cas9 access.

(A) Schematic of the experimental setup. Supercoiled plasmid containing the 601 sequence inserted into a pBlueScript II SK (+) backbone (pBSIISK+601) was chromatinized by step gradient salt dialysis in the presence of histone octamer. Purified yeast Chd1 (yChd1) remodeling factor was used to test the effect of ATP-dependent remodeling factors on Cas9 access to nucleosomal DNA. (B) Quality assessment of the chromatinized plasmid used in this study. Titrated amounts of Micrococcal Nuclease (MNase) were incubated with the chromatinized plasmid and the resulting pattern of protection by assembled nucleosomes was visualized on a 1.3% agarose gel post-stained with ethidium bromide (EtBr). As a control, the supercoiled plasmid was also incubated with the lowest concentration of MNase. (C) A restriction enzyme accessibility assay (REAA) was used to assess the occupancy and position of the nucleosome assembled at the 601 sequence within the chromatinized plasmid. A panel of unique restriction enzyme sites spanning the 601 sequence were incubated with either the supercoiled plasmid, or the chromatinized plasmid. Cleavage was stopped, and protein was removed by incubation with proteinase K followed by Phenol:Chloroform:Isoamyl Alcohol extraction and ethanol precipitation. (Top) The resulting DNA was then linearized using DraIII, and the level of cleavage by the restriction enzyme panel was visualized on a 1% agarose gel post-stained with EtBr. The label “P” represents supercoiled plasmid, while “C” represents chromatinized plasmid. (Bottom right) The location of the restriction sites used are indicated on a diagram of the plasmid. (Bottom left) After quantification of the gel, the percent protection from cleavage experienced in the chromatinized plasmid was plotted versus the location of the cleavage sites on the top strand of the 601 sequence. Experiment 1 refers to the REAA experiment shown in the

gel above, while experiment 2 refers to the REAA experiment without remodeler shown in Figure 5F. The grey shading indicates the borders of the 601 sequence, and the grey oval represents the corresponding nucleosome. (D) REAA experiment using the frequent cutter, HaeIII, to assess the remodeling activity around the chromatinized plasmid by the purified yChd1 chromatin remodeler. The resulting banding patterns were visualized on a 1.5% agarose gel post-stained with EtBr. Low molecular weight fragments indicate a high degree of HaeIII accessibility, while higher weight bands indicate protection from digestion. (E) Diagram showing the location of the restriction enzyme cleavage sites and the PAMs targeted by Cas9/sgRNA in the experiment shown in F and G. (F) An accessibility assay was performed essentially as in C using either restriction enzymes or Cas9/sgRNAs in the presence or absence of the remodeler yChd1. The level of cleavage by the restriction enzyme panel (left) or Cas9/sgRNAs (right) was visualized on a 1.3% agarose gel post-stained with EtBr. A negative control was conducted with an sgRNA that had no sequence complementarity to the plasmid used (non-sense guide). The concentration of yChd1 used was the same as in panel D. (G) Quantification of the gels shown in F. Percent protection from cleavage of the chromatinized plasmid in the presence or absence of the chromatin remodeler was calculated relative to the percent cleavage in the corresponding supercoiled plasmid control, and plotted at the location of the restriction enzyme cleavage sites or the center of the PAMs with respect to the 601 dyad.

Supplementary Tables

Table S1. CRISPRi sgRNA annotations, activity scores, and target site MNase signal, related to Figure 1

Table S2. Ricin tiling library sgRNA annotations, phenotype scores, and target site MNase signal, related to Figure 2

References

1. Doudna JA, Charpentier E (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science 346:1258096–1258096. doi: 10.1126/science.1258096
2. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA (2013) Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell 152:1173–1183. doi: 10.1016/j.cell.2013.02.022
3. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, Lim WA, Weissman JS, Qi LS (2013) CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell 154:442–451. doi: 10.1016/j.cell.2013.06.044
4. Maeder ML, Linder SJ, Cascio VM, Fu Y, Ho QH, Joung JK (2013) CRISPR RNA-guided activation of endogenous human genes. Nature Methods 10:977–979. doi: 10.1038/nmeth.2598
5. Chen X-W, Wang H, Bajaj K, Zhang P, Meng Z-X, Ma D, Bai Y, Liu H-H, Adams E, Baines A, Yu G, Sartor MA, Bin Zhang, Yi Z, Lin J, Young SG, Schekman R, Ginsburg D, Hobbs H (2013) SEC24A deficiency lowers plasma cholesterol through reduced PCSK9 secretion. eLife 2:e00444. doi: 10.7554/eLife.00444
6. Ma H, Naseri A, Reyes-Gutierrez P, Wolfe SA, Zhang S, Pederson T (2015) Multicolor CRISPR labeling of chromosomal loci in human cells. Proc Natl Acad Sci U S A 112:3002–3007. doi: 10.1073/pnas.1420024112
7. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, Gersbach CA (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes

from promoters and enhancers. Nature biotechnology 33:510–517. doi: 10.1038/nbt.3199

8. Kearns NA, Pham H, Tabak B, Genga RM, Silverstein NJ, Garber M, Maehr R (2015) Functional annotation of native enhancers with a Cas9-histone demethylase fusion. Nature Methods 12:401–403. doi: 10.1038/nmeth.3325

9. Chari R, Mali P, Moosburner M, Church GM (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nature Methods 12:823–826. doi: 10.1038/nmeth.3473

10. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nature biotechnology 32:1262–U130. doi: 10.1038/nbt.3026

11. Wang T, Wei JJ, Sabatini DM, Lander ES (2014) Genetic screens in human cells using the CRISPR-Cas9 system. Science 343:80–84. doi: 10.1126/science.1246981

12. Xu H, Xiao T, Chen C-H, Li W, Meyer CA, Wu Q, Wu D, Cong L, Zhang F, Liu JS, Brown M, Liu XS (2015) Sequence determinants of improved CRISPR sgRNA design. Genome Research 25:1147–1157. doi: 10.1101/gr.191452.115

13. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJM, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV (2015) An updated evolutionary classification of CRISPR-Cas systems. Nature Reviews Microbiology. doi: 10.1038/nrmicro3569

14. Singh R, Kuscu C, Quinlan A, Qi Y, Adli M (2015) Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. Nucleic Acids Research. doi:

10.1093/nar/gkv575

15. Wu X, Scott DA, Kriz AJ, Chiu AC, Hsu PD, Dadon DB, Cheng AW, Trevino AE, Konermann S, Chen S, Jaenisch R, Zhang F, Sharp PA (2014) Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nature biotechnology 32:670–676. doi: 10.1038/nbt.2889

16. Knight SC, Xie L, Deng W, Guglielmi B, Witkowsky LB, Bosanac L, Zhang ET, Beheiry El M, Masson J-B, Dahan M, Liu Z, Doudna JA, Tjian R (2015) Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science 350:823–826. doi: 10.1126/science.aac6572

17. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, Qi LS, Kampmann M, Weissman JS (2014) Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell 159:647–661. doi: 10.1016/j.cell.2014.09.029

18. Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ (2006) Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. Genome Research 16:1505–1516. doi: 10.1101/gr.5560806

19. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A (2011) Determinants of nucleosome organization in primary human cells. Nature 474:516–520. doi: 10.1038/nature10002

20. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Haberle V, Lassman T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple

CA, Ishizu Y, Young RS, Francescatto M, Alam I, Albanese D, Altschuler GM, Arakawa T, Archer JAC, Arner P, Babina M, Rennie S, Balwierz PJ, Beckhouse AG, Pradhan-Bhatt S, Blake JA, Blumenthal A, Bodega B, Bonetti A, Briggs J, Brombacher F, Burroughs AM, Califano A, Cannistraci CV, Carbajo D, Chen Y, Chierici M, Ciani Y, Clevers HC, Dalla E, Davis CA, Detmar M, Diehl AD, Dohi T, Drabløs F, Edge ASB, Edinger M, Ekwall K, Endoh M, Enomoto H, Fagiolini M, Fairbairn L, Fang H, Farach-Carson MC, Faulkner GJ, Favorov AV, Fisher ME, Frith MC, Fujita R, Fukuda S, Furlanello C, Furino M, Furusawa J-I, Geijtenbeek TB, Gibson AP, Gingeras T, Goldowitz D, Gough J, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M, Hara M, Harbers M, Harshbarger J, Hasegawa A, Hasegawa Y, Hashimoto T, Herlyn M, Hitchens KJ, Ho Sui SJ, Hofmann OM, Hoof I, Hori F, Huminiecki L, Iida K, Ikawa T, Jankovic BR, Jia H, Joshi A, Jurman G, Kaczkowski B, Kai C, Kaida K, Kaiho A, Kajiyama K, Kanamori-Katayama M, Kasianov AS, Kasukawa T, Katayama S, Kato S, Kawaguchi S, Kawamoto H, Kawamura YI, Kawashima T, Kempfle JS, Kenna TJ, Kere J, Khachigian LM, Kitamura T, Klinken SP, Knox AJ, Kojima M, Kojima S, Kondo N, Koseki H, Koyasu S, Krampitz S, Kubosaki A, Kwon AT, Laros JFJ, Lee W, Lennartsson A, Li K, Lilje B, Lipovich L, Mackay-Sim A, Manabe R-I, Mar JC, Marchand B, Mathelier A, Mejhert N, Meynert A, Mizuno Y, de Lima Morais DA, Morikawa H, Morimoto M, Moro K, Motakis E, Motohashi H, Mummery CL, Murata M, Nagao-Sato S, Nakachi Y, Nakahara F, Nakamura T, Nakamura Y, Nakazato K, van Nimwegen E, Ninomiya N, Nishiyori H, Noma S, Noma S, Noazaki T, Ogishima S, Ohkura N, Ohimiya H, Ohno H, Ohshima M, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov DA, Pain A, Passier R, Patrikakis M, Persson H, Piazza S, Prendergast JGD, Rackham OJL, Ramilowski JA,

Rashid M, Ravasi T, Rizzu P, Roncador M, Roy S, Rye MB, Saijyo E, Sajantila A, Saka A, Sakaguchi S, Sakai M, Sato H, Savvi S, Saxena A, Schneider C, Schultes EA, Schulze-Tanzil GG, Schwegmann A, Sengstag T, Sheng G, Shimoji H, Shimoni Y, Shin JW, Simon C, Sugiyama D, Sugiyama T, Suzuki M, Suzuki N, Swoboda RK, 't Hoen PAC, Tagami M, Takahashi N, Takai J, Tanaka H, Tatsukawa H, Tatum Z, Thompson M, Toyodo H, Toyoda T, Valen E, van de Wetering M, van den Berg LM, Verado R, Vijayan D, Vorontsov IE, Wasserman WW, Watanabe S, Wells CA, Winteringham LN, Wolvetang E, Wood EJ, Yamaguchi Y, Yamamoto M, Yoneda M, Yonekura Y, Yoshida S, Zabierowski SE, Zhang PG, Zhao X, Zucchelli S, Summers KM, Suzuki H, Daub CO, Kawai J, Heutink P, Hide W, Freeman TC, Lenhard B, Bajic VB, Taylor MS, Makeev VJ, Sandelin A, Hume DA, Carninci P, Hayashizaki Y (2014) A promoter-level mammalian expression atlas. Nature 507:462–470. doi: 10.1038/nature13182

21. Jiang C, Pugh BF (2009) Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10:161–172. doi: 10.1038/nrg2522

22. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. doi: 10.1038/nature11247

23. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E,

Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR (2012) Landscape of transcription in human cells. Nature 489:101–108. doi: doi:10.1038/nature11233

24. Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak R, Chen Y, Leonetti ME, Fields AP, Park CY, Corn JE, Weissman JS Next-generation libraries for CRISPR interference and activation. In preparation; anticipated publication 2016; will be made available on bioRxiv as soon as possible.

25. Bassik MC, Kampmann M, Lebbink RJ, Wang S, Hein MY, Poser I, Weibezahn J, Horlbeck MA, Chen S, Mann M, Hyman AA, Leproust EM, Mcmanus MT, Weissman JS (2013) A Systematic Mammalian Genetic Interaction Map Reveals Pathways Underlying Ricin Susceptibility. Cell 152:909–922. doi: 10.1016/j.cell.2013.01.030

26. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, Chen DD, Schupp PG, Vinjamur DS, Garcia SP, Luc S, Kurita R, Nakamura Y, Fujiwara Y, Maeda T, Yuan G-C, Zhang F, Orkin SH, Bauer DE (2015) BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature. doi: 10.1038/nature15521

27. Jonkers I, Lis JT (2015) Getting up to speed with transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol 16:167–177. doi: 10.1038/nrm3953

28. Denisov DA, Shpigelman ES, Trifonov EN (1997) Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. Gene 205:145–149.

29. Naftelberg S, Schor IE, Ast G, Kornblihtt AR (2015) Regulation of alternative splicing through coupling with transcription and chromatin structure. Annu Rev Biochem 84:165–198. doi: 10.1146/annurev-biochem-060614-034242

30. Segal E, Widom J (2009) What controls nucleosome positions? Trends Genet 25:335–343. doi: 10.1016/j.tig.2009.06.002

31. Tilgner H, Guigó R (2010) From chromatin to splicing: RNA-processing as a total artwork. Epigenetics 5:180–184.

32. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA (2012) The accessible chromatin landscape of the human genome. Nature 489:75–82. doi: 10.1038/nature11232

33. Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell 162:108–119. doi: 10.1016/j.cell.2015.05.048

34. Lowary PT, Widom J (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J Mol Biol 276:19–42. doi: 10.1006/jmbi.1997.1494

35. Hinz JM, Laughery MF, Wyrick JJ (2015) Nucleosomes Inhibit Cas9 Endonuclease Activity in Vitro. Biochemistry 54:7063–7066. doi: 10.1021/acs.biochem.5b01108

36. Li G, Widom J (2004) Nucleosomes facilitate their own invasion. Nat Struct Mol Biol 11:763–769. doi: 10.1038/nsmb801

37. Li G, Levitus M, Bustamante C, Widom J (2005) Rapid spontaneous accessibility of nucleosomal DNA. Nature structural & molecular …

38. Luger K, Hansen JC (2005) Nucleosome and chromatin fiber dynamics. Curr Opin Struct Biol 15:188–196. doi: 10.1016/j.sbi.2005.03.006

39. Choy JS, Lee T-H (2012) Structural dynamics of nucleosomes at single-molecule resolution. Trends in Biochemical Sciences 37:425–435. doi: 10.1016/j.tibs.2012.06.006

40. Tomschik M, van Holde K, Zlatanova J (2009) Nucleosome dynamics as studied by single-pair fluorescence resonance energy transfer: a reevaluation. J Fluoresc 19:53–62. doi: 10.1007/s10895-008-0379-1

41. Polach KJ, Widom J (1995) Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. J Mol Biol 254:130–149. doi: 10.1006/jmbi.1995.0606

42. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 507:62–67. doi: 10.1038/nature13011

43. Vasudevan D, Chua E, Davey CA (2010) Crystal structures of nucleosome core particles containing the “601”strong positioning sequence. J. Mol. Biol.

44. Anders C, Niewoehner O, Duerst A, Jinek M (2014) Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature 513:569–573. doi:

10.1038/nature13579

45. Flaus A, Owen-Hughes T (2011) Mechanisms for ATP-dependent chromatin remodelling: the means to the end. FEBS J 278:3579–3595. doi: 10.1111/j.1742-4658.2011.08281.x

46. Hargreaves DC, Crabtree GR (2011) ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. Cell Res 21:396–420. doi: 10.1038/cr.2011.32

47. Patel A, McKnight JN, Genzor P, Bowman GD (2011) Identification of residues in chromodomain helicase DNA-binding protein 1 (Chd1) required for coupling ATP hydrolysis to nucleosome sliding. J Biol Chem 286:43984–43993. doi: 10.1074/jbc.M111.282970

48. Patel A, Chakravarthy S, Morrone S, Nodelman IM, McKnight JN, Bowman GD (2013) Decoupling nucleosome recognition from DNA binding dramatically alters the properties of the Chd1 chromatin remodeler. Nucleic Acids Research 41:1637–1648. doi: 10.1093/nar/gks1440

49. O'Geen H, Henry IM, Bhakta MS, Meckler JF, Segal DJ (2015) A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. Nucleic Acids Research 43:3389–3404. doi: 10.1093/nar/gkv137

50. Naumann S, Reutzel D, Speicher M, Decker HJ (2001) Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. Leuk Res 25:313–322.

51. Kampmann M, Bassik MC, Weissman JS (2013) Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. Proc Natl Acad Sci U S A 110:E2317–E2326. doi: 10.1073/pnas.1307002110

52. Dyer PN, Edayathumangalam RS, White CL, Bao Y, Chakravarthy S, Muthurajan UM, Luger K (2004) Reconstitution of nucleosome core particles from recombinant histones and DNA. Meth Enzymol 375:23–44.

53. Luger K, Rechsteiner TJ, Richmond TJ (1999) Preparation of nucleosome core particle from recombinant histones. Meth Enzymol 304:3–19.

54. Wittmeyer J, Saha A, Cairns B (2004) DNA translocation and nucleosome remodeling assays by the RSC chromatin remodeling complex. Meth Enzymol 377:322–343. doi: 10.1016/S0076-6879(03)77020-7

55. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612. doi: 10.1002/jcc.20084

Chapter 4

Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation

Abstract

We recently found that nucleosomes directly block access of CRISPR/Cas9 to DNA (Horlbeck et al., 2016). Here, we build on this observation with a comprehensive algorithm that incorporates chromatin, position, and sequence features to accurately predict highly effective single guide RNAs (sgRNAs) for targeting nuclease-dead Cas9-mediated transcriptional repression (CRISPRi) and activation (CRISPRa). We use this algorithm to design next-generation genome-scale CRISPRi and CRISPRa libraries targeting human and mouse genomes. A CRISPRi screen for essential genes in K562 cells demonstrates that the large majority of sgRNAs are highly active. We also find CRISPRi does not exhibit any detectable non-specific toxicity recently observed with CRISPR nuclease approaches. Precision-recall analysis shows that we detect over 90% of essential genes with minimal false positives using a compact 5 sgRNA/gene library. Our results establish CRISPRi and CRISPRa as premier tools for loss- or gain-of-function studies and provide a general strategy for identifying Cas9 target sites.

Introduction

Highly multiplexed pooled genetic screening methodologies have emerged as powerful and broadly accessible tools for systematically profiling gene function at the scale of mammalian genomes (Paddison et al., 2004). Recently, a number of pooled screening platforms have been developed that utilize the bacterial CRISPR (Clustered Regularly Interspaced Palindromic Repeats)-associated nuclease Cas9 paired with libraries of single-guide RNAs (sgRNAs) to disrupt targeted genes (reviewed in (Shalem et al., 2015)). We and others have developed tools based on nuclease-dead Cas9 (dCas9)(Qi et al., 2013) to programmably interfere with (CRISPRi) or activate (CRISPRa) transcription (Gilbert et al., 2014; 2013; Konermann et al., 2015; Maeder et al., 2013; Perez-Pinera et al., 2013; Tanenbaum et al., 2014), and used these to systematically manipulate gene expression at genome scale (Gilbert et al., 2014; Konermann et al., 2015). Together, these screening platforms represent a powerful toolkit for unbiased forward gain-of-function and loss-of-function genetic screens in mammalian cells.

A key step in implementing CRISPR genetic screens is selecting sgRNAs that mediate high Cas9 activity. We and others recently found that nucleosomes provide a direct and profound impediment to Cas9 access to DNA (Hinz et al., 2015; Horlbeck et al., 2016; Isaac et al., 2016), an observation we expected to be particularly important for applications such as CRISPRi and CRISPRa, which require sustained binding of dCas9 to DNA. We found that nucleosome occupancy was predictive of Cas9 activity complementary to and independent of previously described sgRNA sequence features (Chari et al., 2015; Doench et al., 2014; Xu et al., 2015), adding an additional dimension to the set of parameters expected to influence Cas9 activity. These observations, along with the strong nucleosome-dependent phasing observed downstream of the FANTOM consortium-annotated transcription start site (TSS) (FANTOM Consortium and

the RIKEN PMI and CLST (DGT) et al., 2014; Horlbeck et al., 2016), suggested that a quantitative model incorporating all of these features could greatly enhance our ability to predict highly active sgRNAs for CRISPRi and CRISPRa.

To test this, we developed a comprehensive machine learning pipeline trained on data collected from 30 CRISPRi and 9 CRISPRa screens. We found that the resulting models were highly predictive of sgRNA efficacy and strongly weighted nucleosome positioning and specific sequence features. We used these models to design and generate CRISPRi and CRISPRa version 2 (v2) libraries, targeting human and mouse genomes, which are greatly enriched for sgRNAs with high predicted activity. These libraries include several additional improvements, including the option to screen with either 10 sgRNAs per gene or a compact half-library containing the top 5 predicted sgRNAs for each gene. To benchmark this new algorithm, we validated the human CRISPRi v2 (hCRISPRi-v2) library with a screen designed to identify genes essential for robust cell growth. In this experiment, essential genes represent a large class of expected positive controls. We identified over 2,100 essential genes with high statistical confidence, significantly improving upon our CRISPRi v1 library (Gilbert et al., 2014), and precision-recall analysis showed increased discrimination of gold standard essential and non-essential genes with both 10 sgRNA/gene and 5 sgRNA/gene hCRISPRi-v2 libraries (Hart et al., 2014). A large majority of the hCRISPRi-v2 sgRNAs targeting known essential genes produced robust growth phenotypes, a key advance over previous CRISPRi libraries (Evers et al., 2016), and our algorithm can accurately predict sgRNA activity for data from screens performed with an independently designed library and in different cell types. Furthermore, we observed that CRISPRi lacked any detectable non-specific toxicity associated with genomic DNA breaks and repair, enabling sensitive detection of genes with subtle growth phenotypes. We also conducted a screen for

genes that modify growth rates upon overexpression with our hCRISPRa-v2 and found this identified 60% more genes, with greater enrichment for previously established classes of hit genes, than our version 1 CRISPRa screen. Our results suggest that the CRISPRi and CRISPRa v2 libraries have numerous favorable properties relative to alternate approaches as a resource for targeted or genome-scale loss-of-function and gain-of-function studies in mammalian cells.

Results

An integrated machine learning approach predicts highly active sgRNAs for CRISPRi and CRISPRa

We sought to improve upon our first generation CRISPRi and CRISPRa libraries by taking a comprehensive approach that incorporated nucleosome positioning, sequence features, refinement of our original sgRNA design rules, and other potentially informative factors. In order to quantitatively model the contribution of these features to CRISPRi activity, we turned to our recently described CRISPRi activity dataset (Horlbeck et al., 2016) in which we integrated data from 30 CRISPRi screens to select 1,539 high-confidence hit genes, and normalized the phenotypes for sgRNAs targeting each gene to the strongest sgRNAs for that gene, resulting in "activity scores" for 18,380 sgRNAs. We used this set as training data for elastic net linear regression (Figure 1A) (Hui Zou, 2005). As many of the features included in the model were nonlinear with activity, we first adapted each feature set according to its relationship with activity. Categorical and non-linear parameters were binned prior to linear regression. Because the relationship between CRISPRi activity and target site distance from the TSS was highly periodic and asymmetric, as we had recently shown (Horlbeck et al., 2016), we fit sgRNA positioning features using support vector regression (SVR) to predict a continuous function for any target site (Supplementary Figure 1). An important improvement was the use of FANTOM consortium annotations instead of Ensembl/GENCODE to define the TSS (Cunningham et al., 2015; FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014; Harrow et al., 2012) (Table S2), a finding also recently reported by Radzisheuskaya and colleagues (Radzisheuskaya et al., 2016).

We first evaluated the performance of this algorithm using five-fold cross-validation. By performing regression on a training set of only 80% of the genes in the CRISPRi activity dataset, we found that the model was highly predictive of activity in the test set comprising the remaining 20% of the genes, with a receiver operating characteristic area under the curve (ROC-AUC) of 0.80 (Figure 1B). Importantly, this high predictive value was consistent across randomly selected training and test sets. As the sgRNAs in this dataset were pre-selected using our CRISPRi v1 design rules, we also tested our ability to predict the performance of sgRNAs in our previously published data set in which we tiled every target site around the TSS of 49 genes known to modulate resistance or sensitivity to the toxin ricin (Gilbert et al., 2014) and obtained an ROC-AUC of 0.91 (Figure 1B), indicating that we could identify active sgRNAs in the genome with high accuracy.

We next analyzed which features contributed most to CRISPRi activity in the predictive model (Figure 1—figure supplement 2). Overall, the predicted scores were most influenced by the position relative to the TSS, including both distance from the TSS and avoidance of canonical nucleosome-occupied regions (Figure 1C and Figure 1—figure supplement 1). In particular, the nucleosome-deprived region immediately downstream of the TSS yielded the strongest predicted activity by SVR, likely due in part to the contribution of dCas9 directly interfering with early transcriptional initiation or elongation (Gilbert et al., 2013; Qi et al., 2013). Sequence features also represented a large contribution to the model, and salient relationships included the disfavoring of guanine directly downstream of the protospacer adjacent motif (PAM) which recapitulated previous findings (Doench et al., 2014; Xu et al., 2015). Additional parameters that contributed to the prediction, included sgRNA secondary structure as predicted by ViennaRNA (Doench et al., 2016; Lorenz et al., 2011a), sgRNA protospacer length, and

chromatin accessibility features not accounted for by the nucleosome positioning relationship. The contribution of each individual parameter was also remarkably consistent across 80%/20% divisions of the training dataset, suggesting that the model was capturing underlying biological signal rather than overfitting the data.

Having established the robustness and accuracy of this approach, we used a version of our sgRNA predictions to design a CRISPRi genome-scale library targeting the human protein-coding transcriptome (hCRISPRi-v2) (Table S3; an *in silico* library design based on the final version of our predictions, hCRISPRi-v2.1, is also available in Table S3). While the predicted scores for sgRNAs in our v1 library were broadly distributed, many sgRNAs of higher predicted activity were available in the genome, and by picking the top 10 predicted sgRNAs per gene we expected that we could greatly enrich the library for guides of high activity (Figure 1D). In constructing this library, we also incorporated empirical information for highly active sgRNAs where available, revised off-target filtering, and implemented changes to the sgRNA expression vector to facilitate the processing of screen samples for sequencing (see Methods). In addition, we cloned the library as separate pools for the top 5 and next-best 5 predicted sgRNAs per gene to facilitate screens where a smaller library may be advantageous, and further divided the pools into 7 thematic sublibraries based on our previous divisions of shRNA libraries (Kampmann et al., 2015) (Figure 1E).

We then used the same approach to design a next-generation CRISPRa library. Due to the requirement for CRISPRa targeting to be upstream of the TSS for maximal activity (Gilbert et al., 2014), we collected an activity score dataset of 2,898 sgRNAs from 9 CRISPRa screens (Y.C. and M.K., personal communication) to train an independent predictive model (Figure 2A). While the input set was significantly smaller than for CRISPRi, the resulting linear regression

still had good predictive value (ROC-AUC 0.70; Figure 2B) and generally shared features of the CRISPRi model (Figure 2C and Figure 2—figure supplement 1-2). We observed that periodic relationship between distance to the TSS and sgRNA activity was less pronounced for CRISPRa than for CRISPRi (Figure 2—figure supplement 1; Figure 1—figure supplement 1), a difference we attributed to the reduced dynamic range in the nucleosome-depleted region around the TSS, and the smaller number and relatively lower expression of the genes targeted in the CRISPRa training dataset. We used the top 10 predicted sgRNAs for each gene to construct a next-generation library, which significantly increased the predicted activity of the library over our v1 designs (Figure 2D). The hCRISPRa-v2 library was partitioned into 14 sublibraries as described for the v2 CRISPRi library above (Figure 2E). Importantly, while several strategies have been described for CRISPR-mediated activation (Chavez et al., 2015; Gilbert et al., 2013; Hilton et al., 2015; Konermann et al., 2015; Maeder et al., 2013; Perez-Pinera et al., 2013; Tanenbaum et al., 2014; Zalatan et al., 2015), a recent comparison of these strategies observed that sgRNA activity generally correlated across all approaches (Tuttle et al., 2016). Our CRISPRa-v2 libraries are thus likely to serve as a valuable resource for effectively targeting most activator systems.

Finally, we applied the CRISPRi and CRISPRa models to predict highly active sgRNAs targeting the mouse protein-coding transcriptome and generated corresponding genome-scale libraries (mCRISRPi-v2 and mCRISPRa-v2) (Figures 1E, 2E). All four library designs are included as Tables S3-6, sgRNA prediction and library design scripts are available online (see Methods), and the cloned lentiviral libraries are available on Addgene.

The large majority of sgRNAs in the hCRISPRi-v2 library are effective

A central goal in developing the hCRISPRi-v2 library was to enrich for highly active sgRNAs, which would improve statistical confidence in hits and enable high sensitivity even with a compact 5 sgRNA/gene library. Underscoring the importance of this goal, Evers et al. generated a small-scale CRISPRi library targeting predetermined sets of essential and non-essential genes and conducted screens for growth phenotypes in the bladder carcinoma cell line RT-112, and found that ~50% of the CRISPRi sgRNAs targeting essential genes were inactive (Evers et al., 2016), similar to rates observed using our CRISPRi v1 library. To test whether our next-generation sgRNA prediction algorithms were able to identify these inactive sgRNAs, we evaluated the predicted activity of the sgRNAs targeting known essential genes in this screen. Despite the screen being performed with an independently designed library, using an sgRNA constant region we have found to underperform our current design(Chen et al., 2013), and in a cell type not evaluated in any of our training or test datasets, predicted activity correlated well with the sgRNA growth phenotypes observed in the screen (Figure 3A; Pearson $R = -0.58$, $P < 10^{-37}$). Over 40% of the sgRNAs in this library had a predicted activity score less than 0.4, a regime in which the vast majority (over 87%) of sgRNAs were inactive, as defined by z-score > -2 relative to non-essential gene-targeting sgRNAs, enabling *a priori* elimination of 60% of the inactive sgRNAs by simply applying this threshold. By contrast, of the sgRNAs in that library with predicted activity scores ≥0.6, 77% are active. Use of the improved sgRNA constant region would be expected to further increase this fraction (Chen et al., 2013; Dang et al., 2015).

In order to validate our hCRISPRi-v2 library design and directly compare its performance to our published v1 library screens (Gilbert et al., 2014), we conducted a screen for genes essential for robust growth in the chronic myeloid leukemia cell line K562. We calculated the growth phenotype (γ) for each sgRNA (Bassik et al., 2013; Kampmann et al., 2013) and

averaged these values across two screen replicates (Figure 3—figure supplement 1 and Table S7). We found that the hCRISPRi-v2 sgRNA growth phenotypes targeting the Evers et al. essential gene sets correlated with predicted activity as above (Figure 3A; Pearson R = -0.42, $P < 10^{-21}$), and therefore designing the libraries based on the top predicted scores selected for highly active sgRNAs. To quantify the fraction of active sgRNAs in our genome-scale libraries, we performed sgRNA-level ROC analysis, ranking sgRNAs by growth phenotype γ and classifying them as true or false positives if they targeted essential or non-essential genes, respectively (Evers et al., 2016). This analysis showed that hCRISPRi-v2 was greatly enriched for active sgRNAs (Figure 3B), and in particular the top 5 sgRNA/gene library contained 80% active sgRNAs at 5% false positives, comparable to the pilot nuclease library tested by Evers et al. This improvement was due to the significant reduction in the number of inactive sgRNAs rather than any difference in the noise as assessed by the background distribution of non-essential gene-targeting sgRNAs (Figure 3—figure supplement 1B). In some instances, as with the known essential gene *VCP*, the difference in sgRNA phenotypes between v1 and v2 libraries was likely attributable to the transition from Ensembl to FANTOM as the TSS annotation source (Figure 3—figure supplement 1C)(Cunningham et al., 2015; FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014; Harrow et al., 2012). Taken together, the above observations indicate the lower fraction of effective sgRNAs in previous libraries was a result of the algorithm and the TSS annotation used rather than any intrinsic limitation of CRISPRi.

The hCRISPRi-v2 library robustly identifies essential genes in precision-recall analysis

We next sought to evaluate whether the enrichment for active sgRNAs in hCRISPRi-v2 over CRISPRi v1 resulted in improved accuracy and confidence for calling hit genes. We

analyzed the screens with a consistent pipeline, scoring genes both by assigning a phenotype based on the mean of the top 3 sgRNAs targeting the gene (by absolute value) and by calculating the Mann-Whitney p-value of all 10 sgRNAs compared to the negative control sgRNAs. We visualized these gene scores as a volcano plot, with the phenotype effect size on the x-axis and p-value on the y-axis (Figure 3C and Table S8). Many hit genes exhibited much stronger p-values, including a substantial fraction that reached the optimal p-value obtainable in the Mann-Whitney test, indicating that the phenotypes of all targeting sgRNAs for these genes were more pronounced than any of the 3,790 non-targeting control sgRNAs. We also modeled noise and off-target effects in the system by generating a large set of "negative control genes" comprised of randomly selected sets of 10 non-targeting sgRNAs and scoring these genes as we did for true genes. In order to classify genes as hits in the screen, we used a score that integrated effect size and statistical confidence and applied the same threshold to both v1 and v2 screens. While in both screens fewer than 0.21% of these negative control genes scored as hits by these criteria, representing a ~2% empirically estimated false discovery rate overall, we could confidently identify 2,150 essential gene hits in the v2 screen while our v1 screen identified 1,408 essential genes (Figure 3C).

In order to assess whether the stronger sgRNA- and gene-level growth phenotype γ scores produced by the hCRISPRi-v2 library resulted in improved discrimination of essential genes, we turned to precision-recall analysis of large "gold standard" essential and non-essential gene sets introduced by Hart and colleagues(Hart et al., 2014). We ranked genes by their growth phenotype (γ) score and calculated at each phenotype threshold the trade-off between the recall of true essential genes and the avoidance of false positive non-essential genes, termed precision. In this analysis, the hCRISPRi-v2 library recalled over 91.2% of the gold standard essential

genes at 95% precision compared to 81.5% in CRISPRi v1 (Figure 3D). We also found that the precision-recall of the top 5 sgRNA/gene half library was essentially identical to the full 10 sgRNA per gene library. Therefore, the 5 sgRNA/gene CRISPRi library represents a compelling tool for applications such as cell sorting-based screens(Liberali et al., 2015), or for *in vivo* screens where cell engraftment and library representation may represent a limiting factor (Braun et al., 2016; Chen et al., 2015).

Finally, we wanted to benchmark our hCRISPRi-v2 library against other recent genome-scale growth screens performed with nuclease-active Cas9 and novel second-generation libraries (Doench et al., 2016; Hart et al., 2015; Wang et al., 2015). Although these screens were performed in different labs and generally in different cell lines, precision-recall analysis offers a useful metric to compare these screens in an unbiased fashion (Hart et al., 2014). We found that our CRISPRi v2 library showed comparable or in many cases much greater discrimination (Figure 3—figure supplement 2). One published CRISPR nuclease screen was conducted in K562 (Wang et al., 2015) with a ~10 sgRNA/gene library, allowing for more direct comparison albeit still tempered by differences between labs and screening protocols. This screen recalled somewhat fewer (78.7% vs 90.8%) essential genes at 95% precision than our v2 library screen (Figure 3D). Together, these results indicate that our hCRISPRi-v2 library has a low false negative rate with few false positives, and the enrichment for highly active sgRNAs enables robust detection of phenotypes even with a compact library.

CRISPRi does not induce non-specific toxicity at amplified genomic loci

We were also intrigued by the observation by Wang and colleagues of K562-specific essentiality of many genes neighboring the BCR-ABL translocation, which they demonstrated to

be mediated by non-specific toxicity of CRISPR-induced double-stranded breaks in the amplified locus (Wang et al., 2015; Wu et al., 1995). This toxicity appears to be pervasive as similar effects have been observed across a range of cancer cell lines (Aguirre et al., 2016; Munoz et al., 2016). To test whether this toxicity could be caused by CRISPRi as well, we used our hCRISPRi-v2 screen as a representative dataset. When we compared the phenotypes of CRISPRi sgRNAs to CRISPR nuclease at the *BCR* amplicon, with all phenotypes standardized to the distribution of negative controls to facilitate comparison, we found that sgRNAs targeting *BCR* were strongly depleted in both screens, as expected based on the critical role of the BCR-ABL fusion in this cancer cell line(Naumann et al., 2001), but few other CRISPRi sgRNAs in the region elicited growth defects (Figure 3E). We also found that CRISPR nuclease sgRNAs targeting the non-essential gene set were generally depleted relative to negative controls or chromosome Y-targeting sgRNAs, which should have no targets in the female-derived K562 cell line (Klein et al., 1976), suggesting that in a CRISPR screen Cas9 nuclease activity can lead to measurable toxicity not related to the function of individual genes but instead due to the formation of on-target DNA double strand breaks, even with alleles present only at 2-3 copies (Naumann et al., 2001). CRISPRi did not exhibit this generic toxicity at non-essential genes, allowing for detection of genes with subtle phenotypes relative to negative controls. Importantly, however, the vast majority of sgRNAs targeting essential genes showed clear separation from the non-essential gene distribution (Figure 3E), demonstrating the high degree of sensitivity for detecting loss-of-function phenotypes with both CRISPR nuclease and CRISPRi screens.

The hCRISPRa-v2 library identifies more genes that modify robust growth rates upon overexpression

Finally, we sought to validate our hCRISPRa-v2 library design by conducting a screen for growth phenotypes in K562 cells expressing SunTag-VP64 constructs (Gilbert et al., 2014; Tanenbaum et al., 2014). We conducted two replicate growth screens (Figure 4—figure supplement 1A and Table S9) and analyzed the screen as with the hCRISPRi-v2 screen above to directly compare the results to our published CRISPRa screens (Gilbert et al., 2014). Our hCRISPRa-v2 screen identified 540 genes to modify robust growth rates upon overexpression, 257 more than our previous CRISPRa screen (Figure 4A and Table S10). Beyond these additional hits, the v1 and v2 screens showed good agreement (Figure 4—figure supplement 1B), and the top categories in DAVID analysis of the v1 screen (Huang et al., 2009), enrichment for transcription factor genes (in particular homeobox and forkhead box transcription factors), received ~3-fold greater enrichment scores in the hCRISPRa-v2 hits (Figure 4—figure supplement 1C), indicating the strong biological coherence of the additional genes. Analysis of the sgRNA growth phenotype (γ) scores for genes that were hits in both v1 and v2 screens showed that a greater fraction of sgRNAs were highly active (69.3% in hCRISPRa-v2 with 5 sgRNAs/gene versus 45.1% in CRISPRa v1; Figure 4C and Figure 4—figure supplement 1D), further validating improvements in the library design. In addition, as with our CRISPRi results, several genes identified in the hCRISPRi-v2 screen but not in v1, including hematopoietically-expressed homeobox *HHEX* and forkhead box C1 *FOXC1*, could be attributed to the use of the CAGE-based FANTOM5 TSS annotation (Figure 4D and Figure 4—figure supplement 1E). Finally, we compared the growth phenotypes from hCRISPRi-v2 to hCRISPRa-v2 and found that the two methods identified non-overlapping sets of genes that modify robust growth (Figure 4—figure supplement 1F), consistent with our previous results and highlighting the complementary information provided by these two approaches.

Discussion

Establishing design rules for effective reagents is critical to the implementation of genome-scale screening technologies. Our previous work established genome-scale CRISPRi and CRISPRa libraries as robust, specific, and complementary tools for dissecting biological pathways in human cells (Gilbert et al., 2014). Here, we significantly improve upon this technology by developing a comprehensive predictive model to accurately identify highly active sgRNAs. This model includes both features specific for CRISPRi and CRISPRa, such as positioning relative to the TSS, as well as features like nucleosome occupancy that we expect to be generally important for most Cas9-mediated applications (Horlbeck et al., 2016), and was able to accurately predict sgRNA activity in screen performed in a cell type it had not previously evaluated (Evers et al., 2016). We used this prediction algorithm to design four new genome-scale libraries targeting human and mouse genomes. These libraries are available on Addgene and *in silico* to facilitate design of focused libraries or targeted experiments (Tables S3-6).

By performing CRISPRi and CRISPRa screens for genes that modify robust growth, we validate our sgRNA predictions and find that our hCRISPRi-v2 and hCRISPRa-v2 libraries represent a significant advance over our previous work in the fraction of highly active sgRNAs, the number of hits detected, and the statistical confidence of these hits. For CRISPRi, these improvements result in the accurate discrimination of essential genes by precision-recall analysis even with a compact 5 sgRNA/gene library. We believe that the greatly improved sgRNA prediction and lack of non-specific toxicity due to nuclease activity, combined with our previous findings that CRISPRi enables inducible, reversible, and homogenous manipulation of gene expression (Gilbert et al., 2014; Mandegar et al., 2016; Qi et al., 2013), make CRISPRi a state-of-the-art approach for loss-of-function studies.

While the libraries described here target genes annotated as protein-coding, CRISPRi and CRISPRa have been shown to be effective for repressing and activating transcription of non-coding genes as well (Gilbert et al., 2014; Luo et al., 2016; Zhao et al., 2014). Our v2 sgRNA prediction algorithms may enable design of libraries to systematically manipulate expression of these transcripts as well. Furthermore, combining CRISPRi and CRISPRa with methods for robustly expressing multiple sgRNAs (Kabadi et al., 2014; Wong et al., 2016; Zalatan et al., 2015) will allow for simultaneous control of several genes, facilitating dissection of cellular pathways and systematic mapping of mammalian genetic interactions (Bassik et al., 2013; Costanzo et al., 2010; Schuldiner et al., 2005). Broadly, increased quantitative understanding of the factors dictating (d)Cas9 activity and specificity will greatly enhance the expanding set CRISPR-mediated technologies for controlling gene expression (Hilton et al., 2015; Perez-Pinera et al., 2013; Vojta et al., 2016), imaging targeted loci (Chen et al., 2013; Shao et al., 2016), or precisely editing the genome (Komor et al., 2016; Tsai et al., 2014).

Methods

Machine learning for CRISPRi and CRISPRa sgRNA activity

Training and test datasets

The CRISPRi activity score dataset was obtained from Horlbeck et al., 2016 (Horlbeck et al., 2016). CRISPRi and CRISPRa ricin tiling data was obtained from Gilbert et al., 2014 (Gilbert et al., 2014). CRISPRa activity scores were generated as previously described for the CRISPRi activity dataset, using data from 9 published and unpublished screening datasets. Hit genes were selected using the formula $|\text{effect size Z-score} \times \log_{10} \text{p-value}| \geq 20$ in any screen, and the phenotypes for sgRNAs targeting each gene were extracted from the screen in which the gene was a hit and normalized to the mean of the top 3 sgRNAs by absolute value. All datasets are included here as Table S1.

Generating TSS annotations

In order to leverage the high accuracy of the FANTOM TSS annotations but remain compatible with the comprehensive, systematic, and established Ensembl transcript models, a hybrid approach was taken. First, the full set of protein-coding genes and transcripts were selected as previously described (Gilbert et al., 2014), using Ensembl release 74 (corresponding to genome assemblies hg19 for human and mm10 for mouse; RRID:SCR_002344) and the APPRIS pipeline to identify the relevant functional transcripts and establish a preliminary set of TSS annotations (Cunningham et al., 2015). FANTOM CAGE peak BED files (RRID:SCR_002678; Human: http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/TSS_human.bed.gz, accessed March 2, 2015; Mouse:

http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/TSS_classifier/TSS_mouse.bed.gz, accessed June 28, 2015 and lifted over to mm10 coordinates with UCSC liftOver tool) were then used to refine this annotation. Specifically, for each gene, the TSS was identified as the same-stranded peaks within 30kb of any Ensembl TSS that matched the gene symbol and was labeled "p1@gene" (here referred to as "primary TSS") or "p2@gene" ("secondary TSS"; included only where the peak passed the FANTOM robust threshold for TSS), and the annotation support was labeled as "CAGE, matched peaks." If multiple matches were found the closest TSS to a known Ensembl TSS was used. If no match was found, the gene could then be matched with any same-stranded CAGE peak within 500bp labeled as "p1" or "p2," and annotation support was considered "CAGE, primary peaks." In the above cases, primary and secondary TSSs were targeted separately (i.e. by 10 sgRNAs each) if they were farther than 1kb apart, or together as one "P1P2" TSS. Where no matched or un-matched primary peaks were found, TSS annotations could be refined by any robust peaks or permissive peaks within 200bp of the annotation (labeled "CAGE, robust peak" or "CAGE, permissive peak") or simply use the Ensembl/APPRIS annotation where no CAGE support was available ("Annotation"). This combined annotation is included as Table S2.

Calculating features

Position relative to the primary and secondary TSSs was calculated from the genomic coordinate of the 3'G of the PAM for each sgRNA to the upstream and downstream edge of each TSS range. Sequence features of the sgRNA and target sites were determined using custom scripts written in Python 2.7 (RRID:SCR_008394) with the Biopython module (RRID:SCR_007173)(Cock et al., 2009). The hCRISPRi-v2.1 algorithm included a change in

these scripts to fix how certain sequence homopolymers were scored. All other libraries were generated with this improvement incorporated, and all analysis in this paper was performed with the v2.1 algorithm. RNA folding metrics were calculated using the ViennaRNA package (RRID:SCR_008550; version 2.2.5) with default parameters (Lorenz et al., 2011b). Chromatin features at the target site were calculated as described previously (Horlbeck et al., 2016), averaging the signal at each base of the target site including the PAM. Custom Python scripts with the module bxpython (v0.5.0, https://github.com/bxlab/bx-python) to extract the processed continuous signal from the following BigWig files obtained from the ENCODE consortium: MNase-seq ENCFF000VNN (Michael Snyder lab, Stanford University), DNase-seq ENCFF000SVY (Gregory Crawford lab, Duke University), and FAIRE-seq ENCFF000TLU (Jason Lieb lab, University of North Carolina)(ENCODE Project Consortium, 2012). Beyond the nucleosome positioning information incorporated in the sgRNA positioning learning models, no chromatin data was used for predicting sgRNA activity for the mouse genome.

Machine learning

Training and test activity score sets were first divided into 80%/20% or 67%/33% sets of genes for CRISPRi or CRISPRa, respectively, as described in the results section. The training set parameters were then transformed according to their distribution as depicted in Figure 1A. For binning parameters, a fixed width was chosen for each feature and applied over the range of values, with the upper-most and lower-most bins collapsed with the neighboring bins if the number of data points at each edge were sparse. Each feature was then split into individual parameters for each bin and sgRNAs were assigned a 1 for the bin if the value fell within the bin or 0 if not. For linearizing sgRNA positioning parameters with continuous curves, sgRNA

positions were fit to the activity score (individually for the distance to each TSS coordinate) using SVR with a radial basis function kernel and hyperparameters C and gamma determined using a grid search approach cross-validated within the training set. The fit score at each position was then used as the transformed linear parameter. Binary parameters were assigned a 1 if true or a 0 if false. All linearized parameters were z-standardized and fit with elastic net linear regression, with the l1/l2 ratio set by cross-validated grid search. All machine learning and downstream analysis was performed with custom Python scripts and the scikit-learn suite, version 0.15.0 (RRID:SCR_002577) (Pedregosa et al., 2011).

Design of next-generation CRISPRi and CRISPRa libraries

Prediction of sgRNA activity scores

All sequences within -25 and +500bp (for CRISPRi) or -550 and -25bp (for CRISPRa) of the upstream or downstream edge of the primary or secondary TSS and containing 19bp followed by an NGG PAM were extracted as potential sgRNAs for downstream scoring of predicted activity. All sequences were prepended with a 5' G to enable robust transcription from the U6 promoter, whether or not this base was present in the genomic sequence. Parameters were calculated for all sgRNAs as above, and transformed and scored using the CRISPRi or CRISPRa regression model from an arbitrarily chosen training set (test set ROC-AUC corresponding to these sets reported in results section).

Off-target scoring

Prediction of sgRNA off-target effects was performed using weighted Bowtie (v1.0.0, RRID:SCR_005476 (Ben Langmead et al., 2009)) alignment largely as previously described

(Gilbert et al., 2014) with several adjustments. The "--tryhard" flag was added to the Bowtie command to increase sensitivity for mismatched sgRNA target sites. The hg19 and mm10 genomes used for alignment were masked at mitochondrial sequences and pseudoautosomal sequence to eliminate "false positive" multiple alignments. Most importantly, as CRISPRi and CRISPRa have maximal effects proximal to the TSS, potential off-target alignments in these regions now were prioritized by creating a reference sequence corresponding to 1kb windows around each TSS as defined above, along with the 5' end of every Ensembl transcript annotation. Reference sequences were generated using bedtools (v2.17.0, RRID:SCR_006646 (Quinlan and Hall, 2010)). In order to pass at the strictest threshold, sgRNAs were required to have no more than 1 alignment (the sgRNA target site itself) with "mismatch score" (Gilbert et al., 2014) less than 31 proximal to the TSS and under 21 in the genome. (For hCRISPRi-v2, 96.6% of sgRNAs incorporated passed at this threshold.) In cases of difficult to target genes or close gene families, sgRNAs were allowed at relaxed thresholds. In descending order, these were: 1 alignment under 31 proximal to the TSS (no genomic threshold), 1 alignment under 21 in the genome, 2 alignments under 31 proximal to the TSS, and 3 alignments under 31 proximal to the TSS.

sgRNA selection

sgRNAs were chosen for inclusion into the genome-scale libraries based on predicted activity scores, empirical activity scores where available, off-target filtering, overlap with other sgRNAs already selected, and sequences with no restriction sites for enzymes used in cloning or sequencing sample processing (BstXI, BlpI, and SbfI). Empirical activity scores for CRISPRi/a v1 sgRNAs were generated as for the training sets above at a lower discriminant threshold of 7, and the corresponding sgRNAs were standardized to 19bp with a 5' G preprended as above and

subjected to the same revised off-target scoring procedure. For each TSS targeted by the library, up to 2 sgRNAs with the strongest empirical evidence were included first if the empirical activity score was at least 0.75, the sgRNA was less that 5kb from the revised v2 TSS, the sgRNA passed the most stringent off-target filter, the sgRNA plus flanking cloning sequences did not contain extra restriction sites, and the sgRNA target site was at least 3bp shifted from any previously selected target sequence. Once 2 empirically validated v1 sgRNAs were included, further sgRNAs fitting these criteria were not included but their predicted activity scores were increased by 0.2 to reflect the balance of information from the algorithm and empirical activity. Predicted sgRNAs were then sorted by best predicted activity score and included if the sgRNA passed the most stringent off-target filter, the sgRNA plus flanking cloning sequences did not contain extra restriction sites, the sgRNA target site was at least 3bp shifted from any previously selected target sequence, and no more than 10 sgRNAs had been selected for the TSS. If fewer than 10 sgRNAs were selected by this algorithm, off-target stringency was iteratively relaxed as above and selection was continued to attain 10 sgRNAs. If 10 sgRNAs passing the most relaxed threshold could not be identified, the TSS was not targeted by the library.

sgRNA on-target and off-target prediction algorithms, library design scripts, and associated files are available at https://github.com/mhorlbeck/CRISPRiaDesign.

Negative controls

For each library, the frequency of each DNA base at each position along the sgRNA protospacer sequence was calculated. Random sgRNA protospacer sequences weighted by these base frequencies were then generated to mirror the composition of the targeting sgRNAs. These

were then filtered for sgRNAs with 0 alignments with a mismatch score less than 31 proximal to the TSS and 0 alignments under 21 in the genome as above.

Library cloning

Protospacer sequences were appended with cloning sequences and then unique PCR adapters corresponding to the designated sublibrary. The half libraries were determined from the first 5 and second 5 sgRNAs selected into the library for each TSS according the the algorithm above. Genes were then partitioned into one of the 13 sublibraries defined by Kampmann et al., 2015 (Kampmann et al., 2015), compressed into the indicated 7 groupings. Each sgRNA was ordered as two oligonucleotide sequences to produce a narrower distribution of sgRNA representation. Overall, oligo sequences were 84bp and had the following format:

5’ pcr adapter BstXI protospacer BlpI 3’ pcr adapter

==========CCACCTTGTTG**GNNNNNNNNNNNNNNNNN**GTTTAAGAGC

TAAGCTG==========

?|||||||||||||||||

Genomic

sequence: ……NNNNNNNNNNNNNNNNNNN***NGG***……

Oligonucleotides were synthesized by Agilent Technologies (RRID:SCR_013575; Santa Clara, CA) on 244K oligo arrays, and cloned into the sgRNA expression vector as previously described (Gilbert et al., 2014). The library sgRNA expression vector “pCRISPRia-v2” was identical to the

CRISPRi/a v1 plasmid (pU6-sgRNA EF1Alpha-puro-T2A-BFP, Addgene #60955) with the addition of two SbfI restriction sites used for sequencing sample processing.

Genome-scale CRISPRi and CRISPRa screen for essential genes

The screens for genes required for robust growth were conducted essentially as previously described(Gilbert et al., 2014). Briefly, plasmid libraries were packaged into lentivirus in HEK293T cells (RRID:CVCL_0063) and infected into a previously established polyclonal K562 cell line stably expressing dCas9-KRAB grown in 3L spinner flasks (Bellco, Vineland, NJ). After two days, infected cells were selected with 0.75μg/mL puromycin (Tocris, Bristol, UK) for two days, allowed to recover for one day, and then cultured at a minimum of $750x10^6$ cells in 1.5L standard media (RPMI-1640 with 10% Fetal Bovine Serum and 1x supplemental glutamine, penicillin, and streptomycin) from "T0" to "endpoint," determined by ~10 cell doublings after T0. CRISPRi screen cells were mock-treated with 0.1% DMSO (Sigma-Aldrich, St. Louis, MO) but otherwise left untreated. Screens were performed as independent replicates starting from the infection step. The K562 dCas9-KRAB and SunTag-VP64 cell lines were obtain from (Gilbert et al., 2014) and had been constructed from K562 cells originally obtained from ATCC (RRID:CVCL_0004). Cytogenetic profiling by array comparative genomic hybridization (not shown) closely matched previous characterizations of the K562 cell line (Naumann et al., 2001). All cell lines tested negative for mycoplasma contamination (MycoAlert Kit, Lonza, Basel, Switzerland) in regular screenings.

Frozen samples of $250x10^6$ cells collected at T0 and endpoint were processed as previously described (Gilbert et al., 2014), with the substitution of an SbfI restriction digest (SbfI-HF, New England Biolabs, Ipswich, MA) in place of the MfeI digest in the genomic DNA

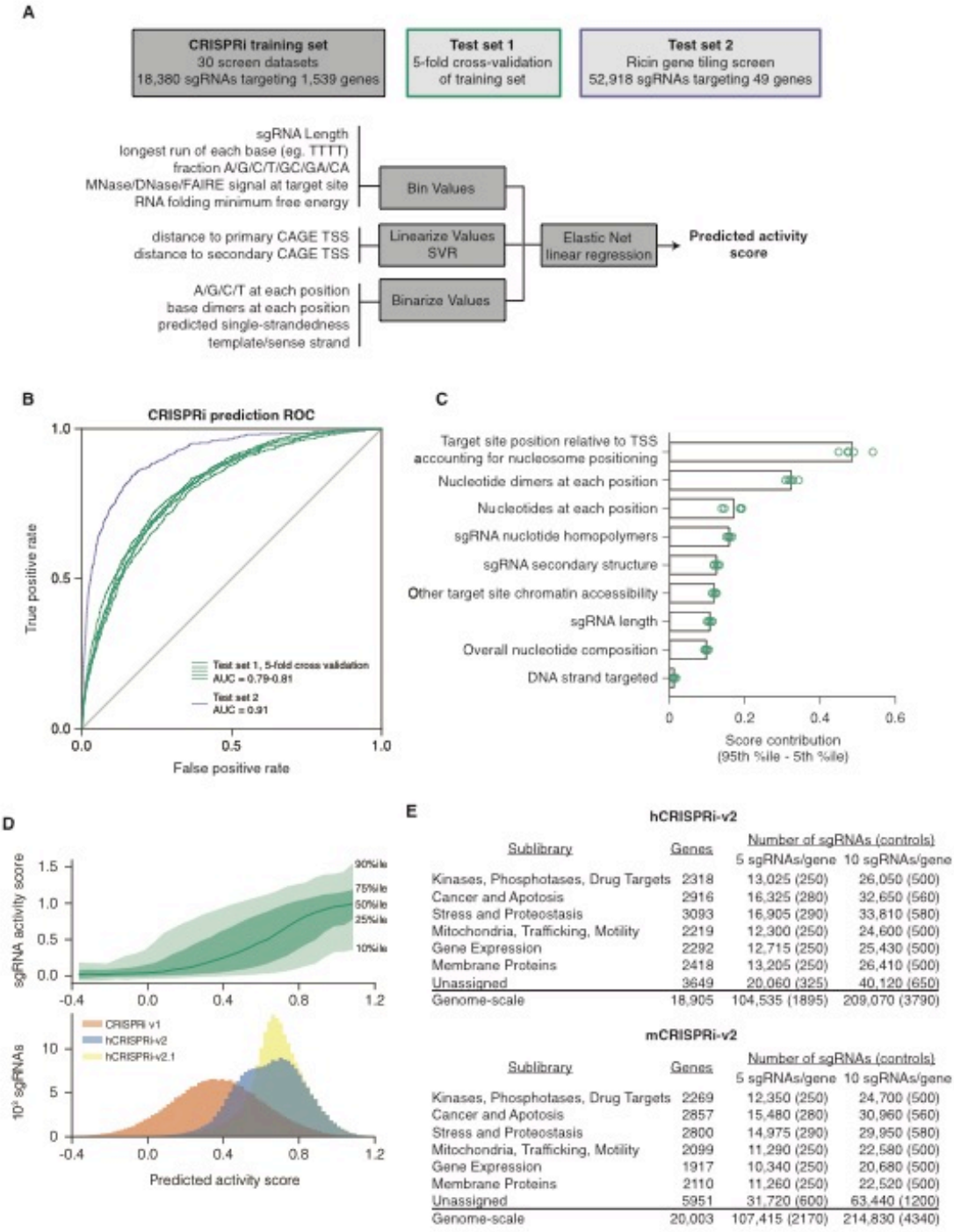
fragmentation and enrichment step. The sgRNA-encoding regions were sequenced on an Illumina HiSeq-4000 using custom primers. Sequencing reads were aligned to the expected library sequences using Bowtie (v1.0.0, (Ben Langmead et al., 2009)) and read counts were processed using custom Python scripts (available at https://github.com/mhorlbeck/ScreenProcessing) based on previously established shRNA screen analysis pipelines (Bassik et al., 2013; Kampmann et al., 2013). sgRNAs represented with fewer than 50 sequencing reads in both T0 and Endpoint were excluded from analysis. sgRNA growth phenotypes (γ) were calculated by normalizing sgRNA $\log_2$ enrichment from T0 to endpoint samples and normalizing by the number of cell doublings in this time period. CRISPRi v1 screen data from Gilbert et al., 2014 was re-analyzed using this pipeline, and the hCRISPRi/a-v2 5 sgRNA/gene libraries were evaluated by analyzing the sgRNA read counts corresponding to only the 5 sgRNA/gene sublibraries. Gene ontology analysis was conducted using DAVID 6.7 (Huang et al., 2009) with default search categories and with background lists representing the genes targeted by the CRISPRa v1 or hCRISPRa-v2 libraries where appropriate. For Figure 4B and Figure 4—figure supplement 4C, "shared hit" genes were 70 genes that scored as strong anti-growth hits (phenotype z-score x $-\log_{10}$ p-value $\leq$ -10) in both CRISPRa v1 and hCRISPRa-v2.

Acknowledgements

We would like to thank Dr. Manuel Leonetti, Ben Barsi-Rhyne, Dr. Jonathan Friedman, and Dr. Jodi Nunnari for generously sharing unpublished screening data for determination of sgRNA activity. We would also like to thank Dr. Xuebing Wu and members of the Weissman lab, particulary Dr. Joshua Dunn and Manny DeVera, for helpful discussions and assistance. We thank Dr. Laurakay Bruhn, Dr. Daniel Ryan, Dr. Luke Fairbairn, and Dr. Peter Tsang of Agilent Technologies for their assistance on the design and synthesis of oligonucleotide pools. MAH, LAG, JEV, BA, YC, APF, and JSW were supported by the Howard Hughes Medical Institutes and the National Institutes of Heath (P50 GM102706, U01 CA168370, R01 DA036858). LAG was supported by the Leukemia and Lymphoma Society. RAP, CYP, and JEC were supported by the Li Ka Shing Foundation. MK was supported by NIH/NIGMS DP2 GM119139.

Figures

Figure 1



hCRISPRi-v2

Sublibrary	Genes	Number of sgRNAs (controls)	
		5 sgRNAs/gene	10 sgRNAs/gene
Kinases, Phosphotases, Drug Targets	2318	13,025 (250)	26,050 (500)
Cancer and Apotosis	2916	16,325 (280)	32,650 (560)
Stress and Proteostasis	3093	16,905 (290)	33,810 (580)
Mitochondria, Trafficking, Motility	2219	12,300 (250)	24,600 (500)
Gene Expression	2292	12,715 (250)	25,430 (500)
Membrane Proteins	2418	13,205 (250)	26,410 (500)
Unassigned	3649	20,060 (325)	40,120 (650)
Genome-scale	18,905	104,535 (1895)	209,070 (3790)

mCRISPRi-v2

Sublibrary	Genes	Number of sgRNAs (controls)	
		5 sgRNAs/gene	10 sgRNAs/gene
Kinases, Phosphotases, Drug Targets	2269	12,350 (250)	24,700 (500)
Cancer and Apotosis	2857	15,480 (280)	30,960 (560)
Stress and Proteostasis	2800	14,975 (290)	29,950 (580)
Mitochondria, Trafficking, Motility	2099	11,290 (250)	22,580 (500)
Gene Expression	1917	10,340 (250)	20,680 (500)
Membrane Proteins	2110	11,260 (250)	22,520 (500)
Unassigned	5951	31,720 (600)	63,440 (1200)
Genome-scale	20,003	107,415 (2170)	214,830 (4340)

Figure 1. A machine learning approach for identifying highly active sgRNAs for CRISPRi. (A) Schematic of machine learning strategy and datasets. 808 features were calculated for each sgRNA, linearized as indicated, and z-standardized. A linear regression model was then generated using these features to fit to the activity scores of the CRISPRi training set (Horlbeck et al., 2016). 20% of the genes in the training set were reserved to test the predictive value of the resulting model. For display in Figures 1B, C, and Figure 1—figure supplement 2, five non-overlapping 20% datasets were randomly selected and training was performed on the corresponding 80% sets. An orthogonal dataset, based on tiling of every possible sgRNA within 10kb of the TSS of 49 genes known to modulate sensitivity to ricin (Bassik et al., 2013; Gilbert et al., 2014), was also used to assess the predictive value of this model. (B) ROC analysis of the ability of the machine learning approach in (A) to predict highly active sgRNAs. For test set 1, sgRNAs with an activity score greater than 0.75 were considered highly active. For test set 2, sgRNAs with a phenotype greater than 0.75 of the maximum phenotype for each gene were considered highly active. (C) Relative contribution of feature categories to the sgRNA predicted scores. The individual weighting of each feature assigned by the linear regression model (see Figure 2—figure supplement 2) was grouped by the indicated categories, and the summed weights for each sgRNA within the 20% test datasets was calculated. The scores of the 95th and 5th percentile sgRNAs were subtracted to compute the overall contribution of the feature category to the distribution of predicted activity scores. Bars indicate the mean of the contributions from five 20% datasets (green circles). Target site position includes both the distance to the TSS and the periodic relationship as fit by SVR (Figure 1—figure supplement 1). (D) Distribution of predicted activity scores in next-generation CRISPRi libraries. (Top) Predicted CRISPRi activity correlates with empirical activity scores. For the 80%/20% division

used to predict sgRNAs for the hCRISPRi-v2.1 library, predicted scores for the 20% test set were plotted against the empirical activity score. Activity score percentiles are from all sgRNAs within 0.25 of the indicated activity score. Predicted activity was highly correlated with activity, with a Pearson R of 0.56 ($P < 10^{-296}$). (Bottom) Distribution of predicted activity scores for CRISPRi v1, hCRISPRi-v2, and hCRISPRi-v2.1, as calculated by the hCRISPRi-v2.1 regression model. (E) Composition of hCRISPRi-v2 and mCRISPRi-v2 sublibraries.

Supplementary Figure 1

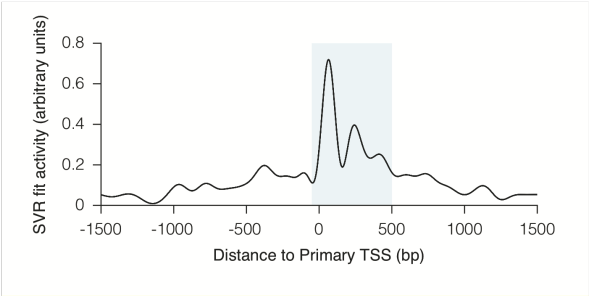


Figure 1—figure supplement 1. Relationship between CRISPRi activity and sgRNA position relative to the TSS as predicted by SVR.

An SVR model with radial basis function kernel was trained on an 80% division of the CRISPRi activity score dataset, using the position of the sgRNA relative to the upstream end of the primary FANTOM TSS for each gene as the sole feature. Hyperparameter values for SVR were selected automatically using cross-validation within the training set. To display the relationship between sgRNA position and CRISPRi activity fit by this model, predicted scores were generated for each position within a 3kb window around the TSS. The resulting curve recapitulated the previously observed periodic relationship shown to be out-of-phase with nucleosome positioning (Horlbeck et al., 2016), and use of this SVR model within the general machine learning approach enabled regression against this highly complex relationship. The shaded area indicates the region relative to the TSS where predicted activity scores for all sgRNAs were calculated for potential inclusion in construction of the hCRISPRi-v2(.1) and mCRISPRi-v2 library designs.

Figure 1--figure supplement 2

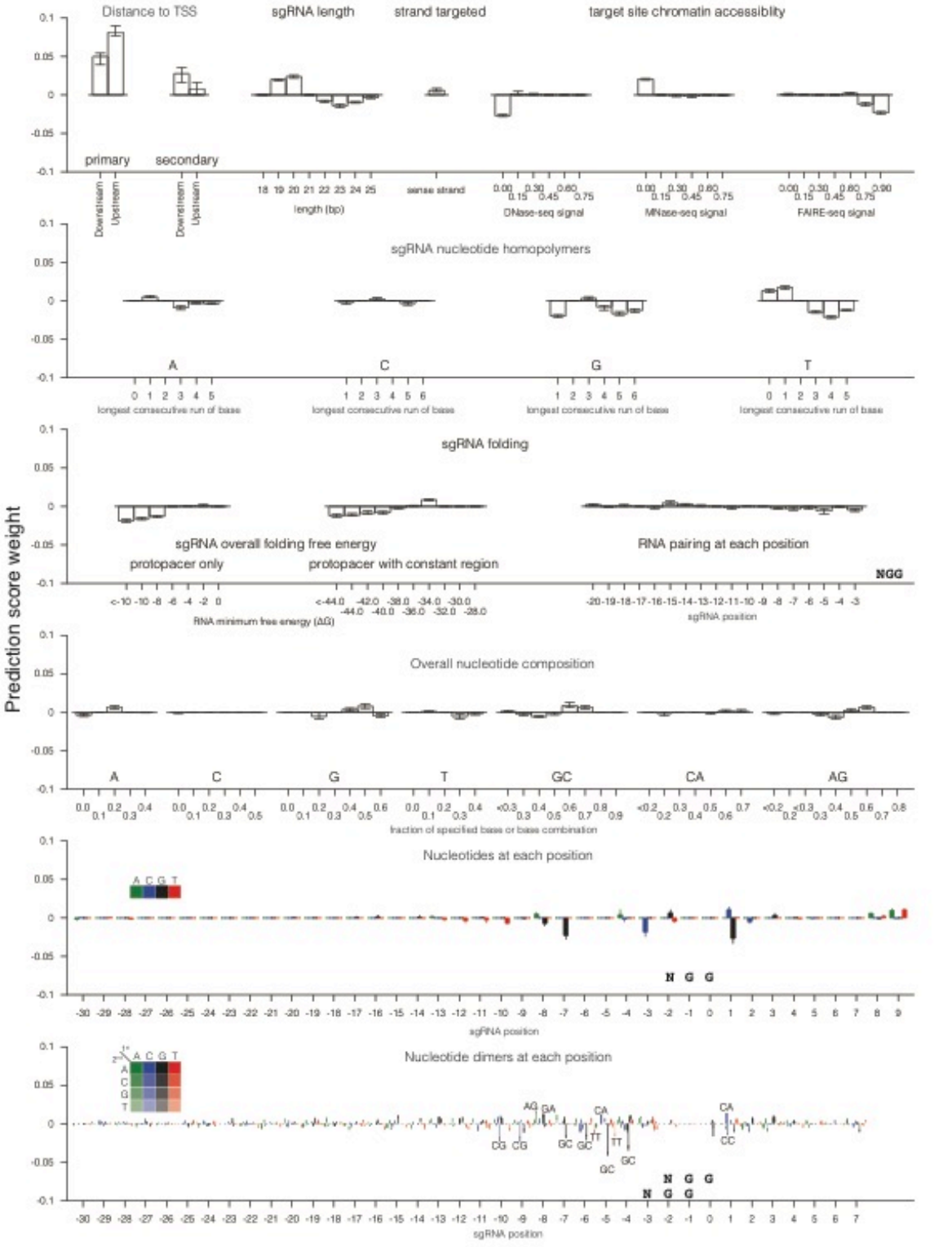
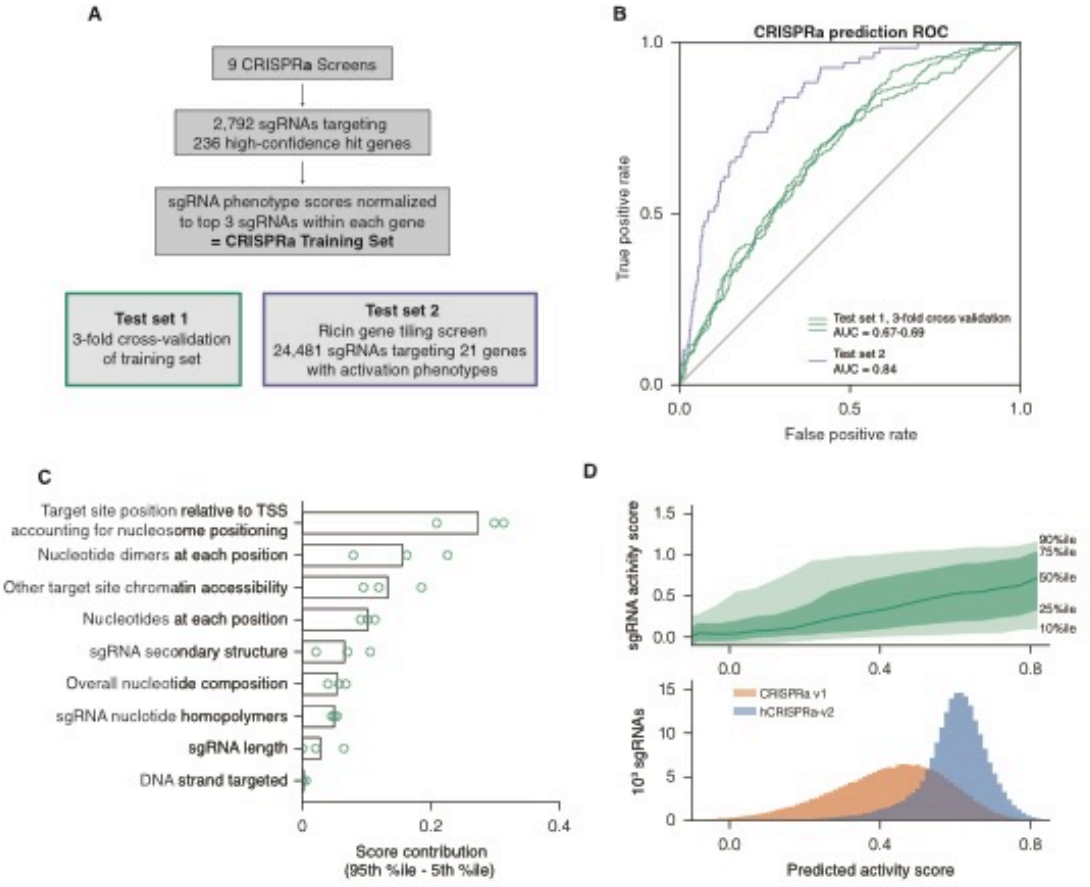


Figure 1—figure supplement 2. Individual sgRNA feature contributions to predicted CRISPRi activity.

Linear regression coefficients for each model trained according to the 80%/20% divisions displayed in Figure 1 are displayed, with bars indicating the mean of the five divisions and error bars indicating minimum and maximum feature coefficients. As each feature was z-standardized before linear regression, coefficients are directly comparable. For binned feature categories, x-axis values represent the minimum value of the bin (inclusive) unless otherwise indicated.

Figure 2



E

hCRISPRa-v2

Sublibrary	Genes	Number of sgRNAs (controls)	
		5 sgRNAs/gene	10 sgRNAs/gene
Kinases, Phosphotases, Drug Targets	2320	13,030 (250)	26,060 (500)
Cancer and Apotosis	2921	16,335 (280)	32,670 (560)
Stress and Proteostasis	3093	16,905 (290)	33,810 (580)
Mitochondria, Trafficking, Motility	2220	12,285 (250)	24,570 (500)
Gene Expression	2288	12,695 (250)	25,390 (500)
Membrane Proteins	2405	13,145 (250)	26,290 (500)
Unassigned	3668	20,145 (325)	40,290 (650)
Genome-scale	18,915	104,540 (1895)	209,080 (3790)

mCRISPRa-v2

Sublibrary	Genes	Number of sgRNAs (controls)	
		5 sgRNAs/gene	10 sgRNAs/gene
Kinases, Phosphotases, Drug Targets	2268	12,340 (250)	24,680 (500)
Cancer and Apotosis	2856	15,475 (280)	30,950 (560)
Stress and Proteostasis	2796	14,950 (290)	29,900 (580)
Mitochondria, Trafficking, Motility	2099	11,290 (250)	22,580 (500)
Gene Expression	1915	10,330 (250)	20,660 (500)
Membrane Proteins	2104	11,225 (250)	22,450 (500)
Unassigned	5901	31,495 (600)	62,990 (1200)
Genome-scale	19,939	107,105 (2170)	214,210 (4340)

Figure 2. A machine learning approach for identifying highly active sgRNAs for CRISPRa. (A) Schematic of CRISPRa datasets. CRISPRa activity scores were generated from screen data and subjected to 3-fold cross-validation due to the smaller sample size. Ricin tiling data was limited to 21 genes that were previously shown to modulate sensitivity to ricin upon CRISPRa overexpression (Gilbert et al., 2014). (B) ROC analysis of machine learning approach using CRISPRa datasets, conducted as in Figure 1B. (C) Relative contribution of feature categories for CRISPRa, calculated as in Figure 1C. (D) Distribution of predicted activity scores in next-generation CRISPRa libraries. (Top) Predicted CRISPRa activity correlates with empirical activity scores. For the 67%/33% division used to predict sgRNAs for the hCRISPRa-v2 library, predicted scores for the 33% test set were plotted against the empirical activity score. Activity score percentiles are from all sgRNAs within 0.25 of the indicated activity score. Predicted activity was highly correlated with activity, with a Pearson R of 0.41 ($P < 10^{-38}$). (Bottom) Distribution of predicted activity scores for CRISPRa v1 and hCRISPRa-v2 as calculated by the hCRISPRa-v2 regression model. (E) Composition of hCRISPRa-v2 and mCRISPRa-v2 sublibraries.

Supplementary Figure 3

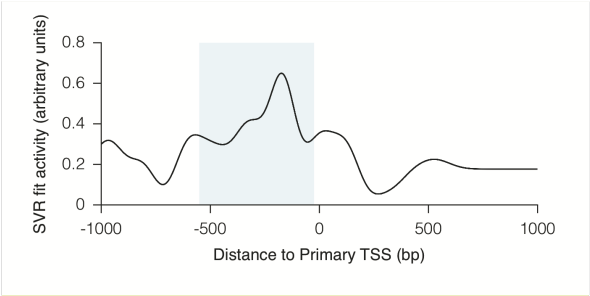


Figure 2—figure supplement 1. Relationship between CRISPRa activity and sgRNA position relative to the TSS as predicted by SVR.

An SVR model was trained on a 67% division of the CRISPRa dataset using sgRNA position relative to the downstream end of the primary FANTOM TSS for each gene. Analysis was conducted as described for Figure 1—figure supplement 1. The shaded area indicates the region relative to the TSS where predicted activity scores for all sgRNAs were calculated for potential inclusion in construction of the hCRISPRa-v2 and mCRISPRa-v2 library designs.

Figure 2--figure supplement 2

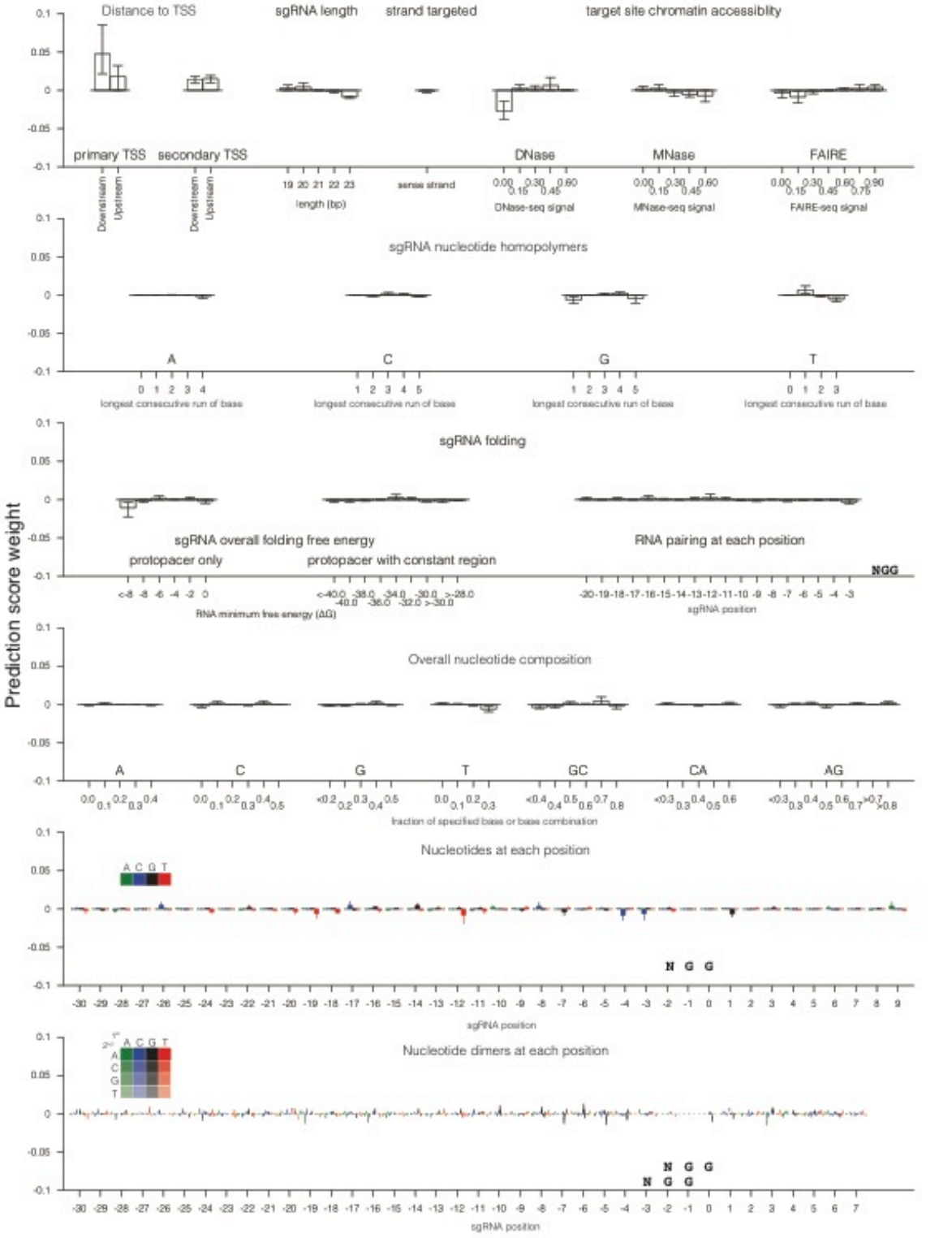


Figure 2—figure supplement 2. Individual sgRNA feature contributions to predicted CRISPRa activity.

Linear regression coefficients for each model trained according to the 67%/33% divisions displayed in Figure 2 are displayed, with bars indicating the mean of the three divisions and error bars indicating minimum and maximum feature coefficients. As each feature was z-standardized before linear regression, coefficients are directly comparable. For binned feature categories, x-axis values represent the minimum value of the bin (inclusive) unless otherwise indicated.

Figure 3

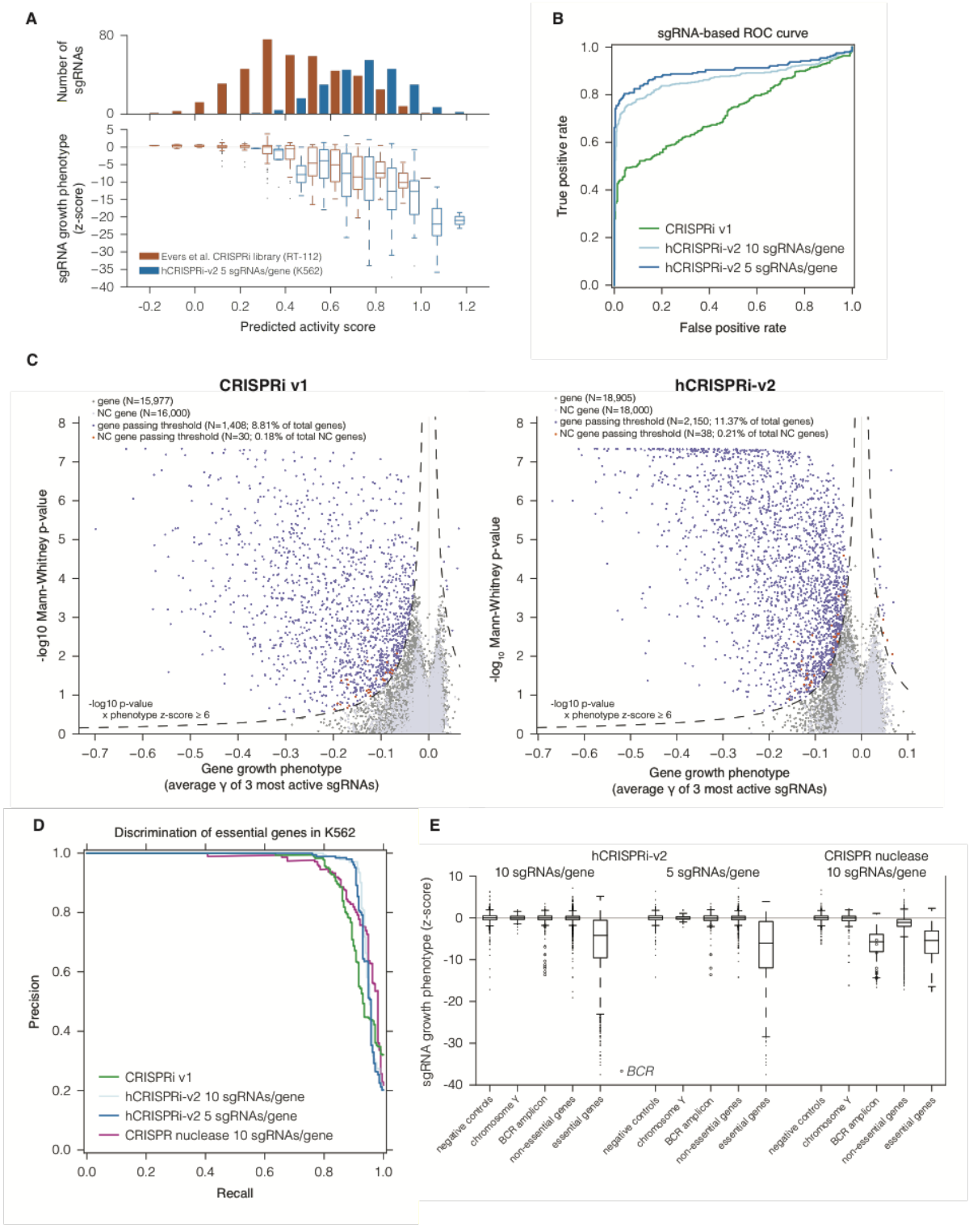

A
Number of sgRNAs
sgRNA growth phenotype (z-score)
Evers et al. CRISPRi library (RT-112)
hCRISPRi-v2 5 sgRNAs/gene (K562)
Predicted activity score
B
sgRNA-based ROC curve
True positive rate
False positive rate
CRISPRi v1
hCRISPRi-v2 10 sgRNAs/gene
hCRISPRi-v2 5 sgRNAs/gene
C
CRISPRi v1
hCRISPRi-v2
gene (N=15,977)
NC gene (N=16,000)
gene passing threshold (N=1,408; 8.81% of total genes)
NC gene passing threshold (N=30; 0.18% of total NC genes)
gene (N=18,905)
NC gene (N=18,000)
gene passing threshold (N=2,150; 11.37% of total genes)
NC gene passing threshold (N=38; 0.21% of total NC genes)
-log10 Mann-Whitney p-value
-log10 p-value x phenotype z-score ≥ 6
Gene growth phenotype (average γ of 3 most active sgRNAs)
D
Discrimination of essential genes in K562
Precision
Recall
CRISPRi v1
hCRISPRi-v2 10 sgRNAs/gene
hCRISPRi-v2 5 sgRNAs/gene
CRISPR nuclease 10 sgRNAs/gene
E
hCRISPRi-v2
10 sgRNAs/gene
5 sgRNAs/gene
CRISPR nuclease
10 sgRNAs/gene
sgRNA growth phenotype (z-score)
BCR
negative controls
chromosome Y
BCR amplicon
non-essential genes
essential genes

Figure 3. hCRISPRi-v2 outperforms CRISPRi v1 in screens for essential genes in K562.

(A) Distribution and predicted scores for sgRNAs targeting essential genes. (Top) Predicted activity scores for sgRNAs from Evers et al., 2016 or hCRISPRi-v2 targeting essential genes as defined by Evers et al. (Evers et al., 2016), binned in increments of 0.1. (Bottom) sgRNA growth phenotypes of the sgRNAs in the above bins, z-standardized to the distribution of sgRNAs targeting Evers et al. non-essential genes. (B) ROC analysis of sgRNAs from CRISPRi v1 or hCRISPRi-v2 targeting essential and non-essential genes. sgRNAs were ranked by γ, and considered true or false positives if they targeted essential or non-essential genes, respectively, as defined by Evers et al. (C) Volcano plots of gene phenotypes and p-values for growth screens performed with CRISPRi v1 (Gilbert et al., 2014) and hCRISPRi-v2. For each screen, genes phenotypes were calculated by averaging the growth phenotype (γ) of the 3 sgRNAs with the strongest γ by absolute value, and gene p-values were calculated by performing the Mann-Whitney test comparing all sgRNAs targeting the gene to the full set of negative control sgRNAs. For genes with multiple TSSs targeted, sgRNAs were grouped by TSS and the TSS with the lowest p-value was used for downstream analysis. A comparable number of negative control (NC) genes were generated by randomly sampling 10 non-targeting sgRNAs (with replacement) and analyzed as true genes. Empirically derived thresholds (dashed lines) were calculated as shown, using the NC gene distribution to derive the background standard deviation for z-score. (D) Precision-recall analysis of essential gene screens performed in K562. Statistical precision and recall of essential and non-essential gene sets (Hart et al., 2014) were calculated for genes ranked by growth phenotype in K562. For both CRISPRi and CRISPR nuclease screens(Wang et al., 2015), gene-level phenotypes were calculated as the average $\log_2$ fold-change of all sgRNAs targeting the gene (termed CRISPR Scores in ref. (Wang et al., 2015)). (E)

Boxplots of CRISPRi and CRISPR nuclease sgRNA phenotypes for several gene sets. sgRNA γ scores (CRISPRi) or $\log_2$ enrichments (nuclease) were z-standardized to the corresponding negative control set. Boxplots display the distribution of the negative control sgRNAs or sgRNAs targeting genes on Y chromosome (excluding pseudo-autosomal genes), within the *BCR* amplicon, or in the gold standard essential sets used in (D). Individual phenotypes for sgRNAs targeting *BCR* are overlaid with the corresponding boxplot.

Figure 3--figure supplement 1

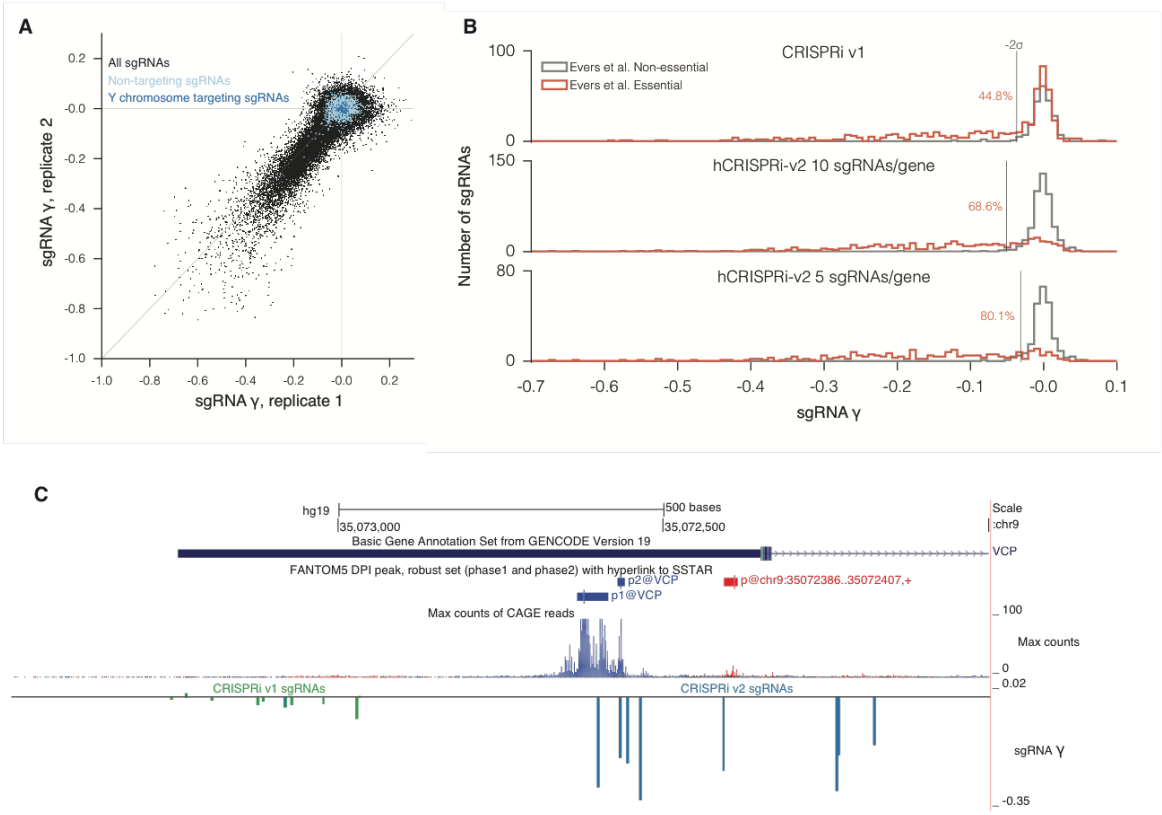


Figure 3—figure supplement 1. sgRNA phenotypes from CRISPRi v1 and hCRISPRi-v2 growth screens.

(A) hCRISPRi-v2 sgRNA phenotypes correlate between screen replicates. sgRNA γ scores were calculated by computing $\log_2$ enrichments of read counts between screen start and endpoint samples and normalizing by estimated cell doublings over the course of the screen. Phenotypes for non-targeting sgRNAs and sgRNAs targeting the Y chromosome generally do not correlate between screen replicates; Spearman R=0.08 ($P < 10^{-7}$) and R = 0.05 (P = 0.5), respectively. (B) Histograms of growth phenotypes (γ) for sgRNAs used in the analysis in Figure 3B. Percentages indicate number of essential-targeting sgRNAs with negative γ more than two standard deviations from the mean of the non-essential-targeting sgRNAs. (C) UCSC Genome Browser tracks depicting Ensembl and FANTOM annotations for example gene *VCP*. CRISPRi v1 sgRNAs were chosen to be -50bp to +300bp relative to the 5' end of the *VCP* transcript model. hCRISPRi-v2 sgRNAs were the top predicted sgRNAs chosen from all sites -25bp to +500bp relative to the "p1@VCP" and "p2@VCP" FANTOM TSS annotation. FANTOM annotations were generated from CAGE sequencing of over 800 human cell types and tissues, summarized by the maximum CAGE sequencing counts track.

Figure 3--figure supplement 2

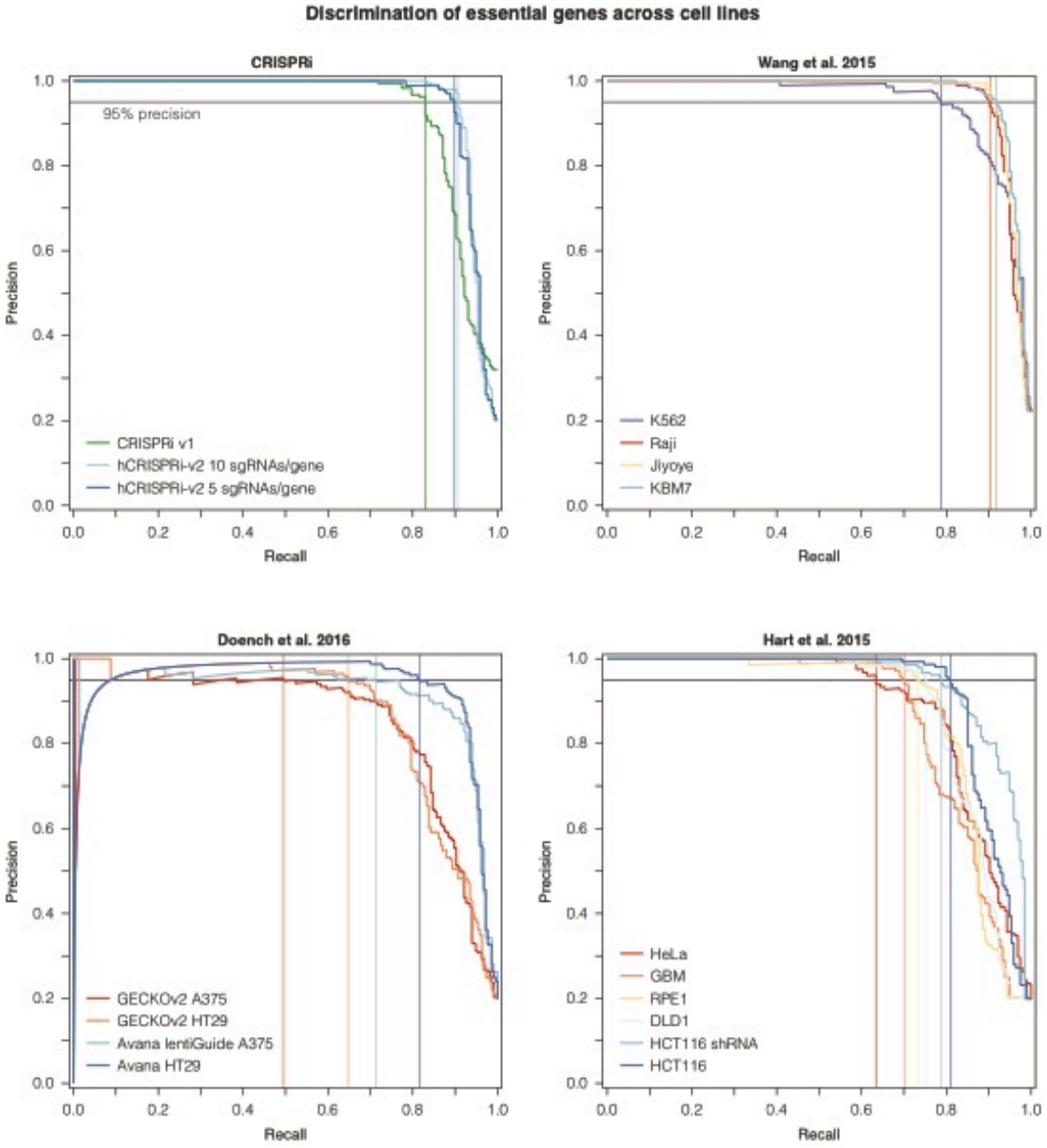


Figure 3—figure supplement 2. Precision-recall analysis of second-generation CRISPR nuclease essential gene screens.

Analysis was conducted as in Figure 3D. CRISPRi, Wang et al., and Doench et al. datasets were ranked according to the average $\log_2$ fold-change of all sgRNAs targeting a given gene. For Hart et al., genes were ranked according to their published Bayes Factor scores as sgRNA-level data was unavailable (Doench et al., 2016; Hart et al., 2015; Wang et al., 2015).

Figure 4

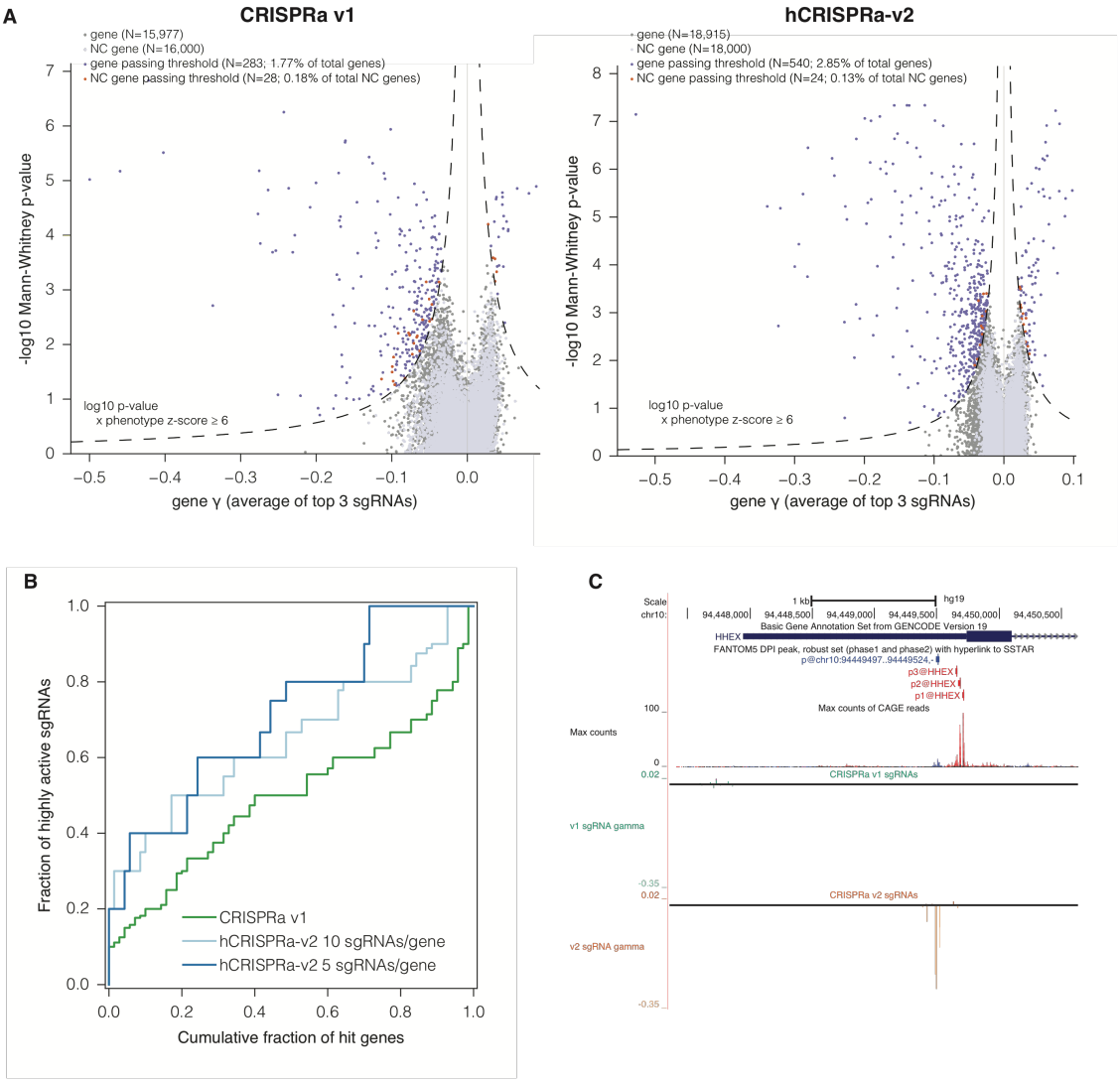

A
CRISPRa v1
gene (N=15,977)
NC gene (N=16,000)
gene passing threshold (N=283; 1.77% of total genes)
NC gene passing threshold (N=28; 0.18% of total NC genes)
-log10 Mann-Whitney p-value
log10 p-value
x phenotype z-score ≥ 6
gene γ (average of top 3 sgRNAs)
hCRISPRa-v2
gene (N=18,915)
NC gene (N=18,000)
gene passing threshold (N=540; 2.85% of total genes)
NC gene passing threshold (N=24; 0.13% of total NC genes)
B
Fraction of highly active sgRNAs
CRISPRa v1
hCRISPRa-v2 10 sgRNAs/gene
hCRISPRa-v2 5 sgRNAs/gene
Cumulative fraction of hit genes
C
Scale
chr10:
1 kb
hg19
Basic Gene Annotation Set from GENCODE Version 19
HHEX
FANTOM5 DPI peak, robust set (phase1 and phase2) with hyperlink to SSTAR
p3@HHEX
p2@HHEX
p1@HHEX
Max counts of CAGE reads
Max counts
CRISPRa v1 sgRNAs
v1 sgRNA gamma
CRISPRa v2 sgRNAs
v2 sgRNA gamma

Figure 4. hCRISPRa-v2 outperforms CRISPRa v1 in screens for genes that modify growth rates upon overexpression.

(A) Volcano plots of gene phenotypes and p-values for growth screens performed with CRISPRa v1 (Gilbert et al., 2014) and hCRISPRa-v2, presented as in Figure 3C. (B) Cumulative distributions of fraction of highly active sgRNAs targeting strong hit genes shared between CRISPRa v1 and hCRISPRa-v2 screens. Highly active sgRNAs for CRISPRa were defined as those with negative γ scores more than two standard deviations from the mean of non-targeting control sgRNAs (see Figure 4—figure supplement NNN). (C) UCSC Genome Browser tracks depicting TSS annotations and CRISPRa growth phenotypes for example gene *HHEX*.

Figure 4--figure supplement

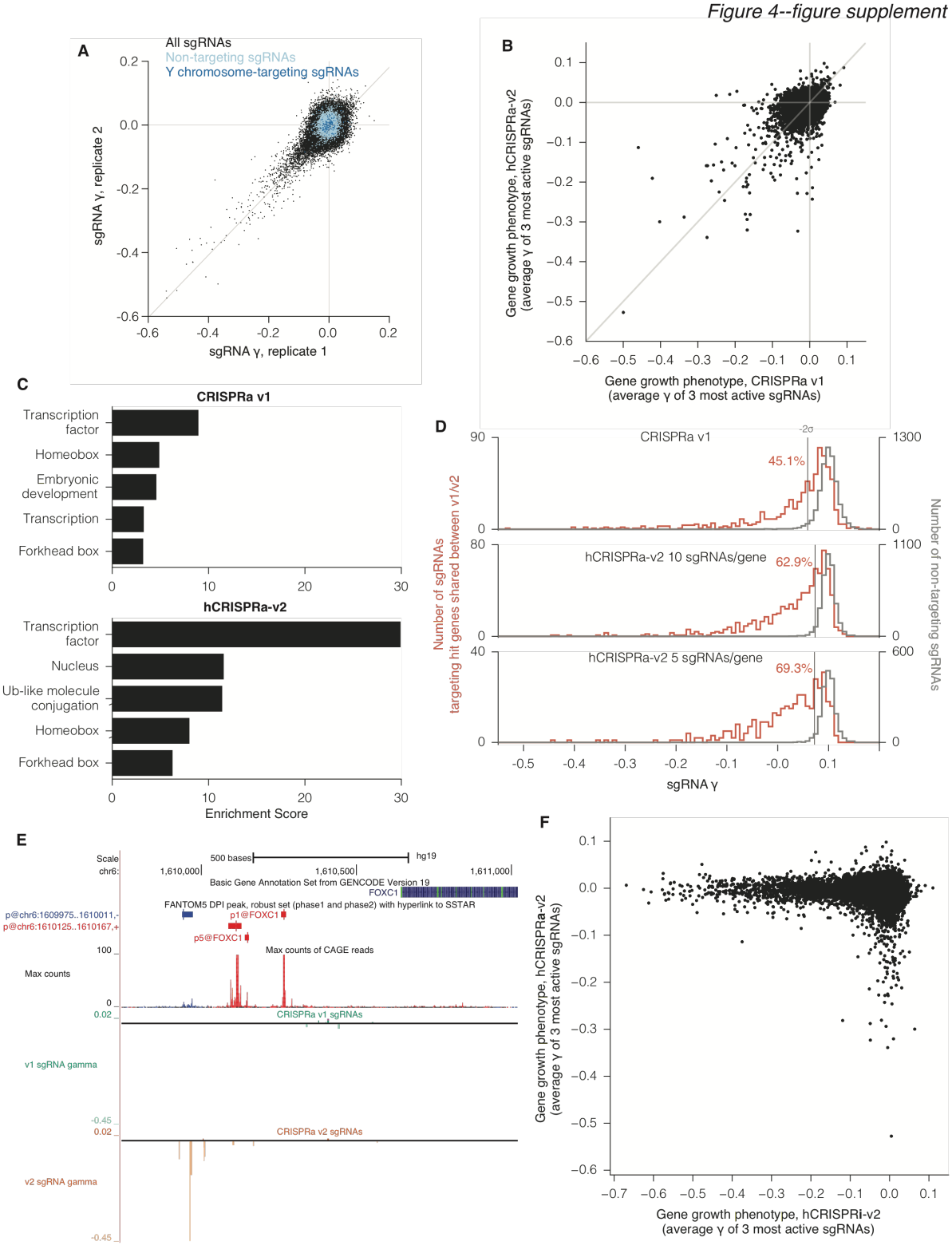


Figure 4—figure supplement 1. sgRNA phenotypes and gene category enrichment scores from CRISPRa v1 and hCRISPRa-v2 growth screens.

(A) hCRISPRa-v2 sgRNA phenotypes correlate between screen replicates, presented as in Figure 3—figure supplement 1A. Phenotypes for non-targeting sgRNAs and sgRNAs targeting the Y chromosome correlate poorly between screen replicates; Spearman R=0.10 ($P < 10^{-8}$) and R = 0.15 (P = 0.05), respectively. (B) Comparison of CRISPRa v1 and hCRISPRa-v2 gene growth phentoypes (γ). (C) DAVID enrichment scores for hit gene categories from CRISPRa v1 and hCRISPRa-v2 screens. CRISPRa v1 categories represent the top 5 categories identified. hCRISPRa-v2 categories include the top 3 identified along with homeobox and forkhead box categories. (D) Histograms of growth phenotypes (γ) for sgRNAs used in the analysis in Figure 4B. Percentages indicate number of sgRNAs targeting v1 and v2 shared hit genes with negative γ more than two standard deviations from the mean of the non-targeting control sgRNAs. (E) UCSC Genome Browser tracks depicting TSS annotations and CRISPRa growth phenotypes for example gene *FOXC1*. (F) Comparison of hCRISPRi-v2 and hCRISPRa-v2 gene growth phenotypes (γ).

Supplementary Tables

Table S1. CRISPRi and CRISPRa activity score datasets

Table S2. TSS annotations for hg19 and mm10 genomes.

Table S3. Library composition of hCRISPRi-v2 and hCRISPRi-v2.1

Table S4. Library composition of mCRISPRi-v2

Table S5. Library composition of hCRISPRa-v2

Table S6. Library composition of mCRISPRa-v2

Table S7. sgRNA read counts and growth phenotypes for hCRISPRi-v2 screens performed in K562

Table S8. Gene growth phenotypes and p-values for hCRISPRi-v2 screens performed in K562

Table S9. sgRNA read counts and growth phenotypes for hCRISPRa-v2 screens performed in K562

Table S10. Gene growth phenotypes and p-values for hCRISPRa-v2 screens performed in K562

References

Aguirre, A.J., Meyers, R.M., Weir, B.A., Vazquez, F., Zhang, C.-Z., Ben-David, U., Cook, A., Ha, G., Harrington, W.F., Doshi, M.B., et al. (2016). Genomic copy number dictates a gene-independent cell response to CRISPR-Cas9 targeting. Cancer Discov.

Bassik, M.C., Kampmann, M., Lebbink, R.J., Wang, S., Hein, M.Y., Poser, I., Weibezahn, J., Horlbeck, M.A., Chen, S., Mann, M., et al. (2013). A Systematic Mammalian Genetic Interaction Map Reveals Pathways Underlying Ricin Susceptibility. *152*, 909–922.

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

Braun, C.J., Bruno, P.M., Horlbeck, M.A., Gilbert, L.A., Weissman, J.S., and Hemann, M.T. (2016). Versatile in vivo regulation of tumor phenotypes by dCas9-mediated transcriptional perturbation. Proc Natl Acad Sci U S A *113*, E3892–E3900.

Chari, R., Mali, P., Moosburner, M., and Church, G.M. (2015). Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nature Methods *12*, 823–826.

Chavez, A., Scheiman, J., Vora, S., Pruitt, B.W., Tuttle, M., P R Iyer, E., Lin, S., Kiani, S., Guzman, C.D., Wiegand, D.J., et al. (2015). Highly efficient Cas9-mediated transcriptional programming. Nature Methods *12*, 326–328.

Chen, B., Gilbert, L.A., Cimini, B.A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S., et al. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *155*, 1479–1491.

Chen, S., Sanjana, N.E., Zheng, K., Shalem, O., Lee, K., Shi, X., Scott, D.A., Song, J., Pan, J.Q., Weissleder, R., et al. (2015). Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *160*, 1246–1260.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science *327*, 425–431.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Research *43*, D662–D669.

Dang, Y., Jia, G., Choi, J., Ma, H., Anaya, E., Ye, C., Shankar, P., and Wu, H. (2015). Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. Genome Biology *16*, 823.

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature Biotechnology *34*, 184–191.

Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J., and Root, D.E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nature Biotechnology *32*, 1262–U130.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *489*, 57–74.

Evers, B., Jastrzebski, K., Heijmans, J.P.M., Grernrum, W., Beijersbergen, R.L., and Bernards, R. (2016). CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. Nature Biotechnology.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassman, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. *507*, 462–470.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *159*, 647–661.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *154*, 442–451.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Research *22*, 1760–1774.

Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R., and Moffat, J. (2014). Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol *10*, 733–733.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *163*, 1515–1526.

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nature Biotechnology *33*, 510–517.

Hinz, J.M., Laughery, M.F., and Wyrick, J.J. (2015). Nucleosomes Inhibit Cas9 Endonuclease Activity in Vitro. Biochemistry *54*, 7063–7066.

Horlbeck, M.A., Witkowsky, L.B., Guglielmi, B., Replogle, J.M., Gilbert, L.A., Villalta, J.E., Torigoe, S.E., Tjian, R., and Weissman, J.S. (2016). Nucleosomes impede Cas9 access to DNA in vivo and in vitro. eLife *5*, e12677.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research *37*, 1–13.

Hui Zou, T.H. (2005). Regularization and variable selection via the Elastic Net.

Isaac, R.S., Jiang, F., Doudna, J.A., Lim, W.A., Narlikar, G.J., and Almeida, R. (2016). Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. eLife *5*, 1.

Kabadi, A.M., Ousterout, D.G., Hilton, I.B., and Gersbach, C.A. (2014). Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. Nucleic Acids Research *42*, e147–e147.

Kampmann, M., Bassik, M.C., and Weissman, J.S. (2013). Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. Proc Natl Acad Sci U S A *110*, E2317–E2326.

Kampmann, M., Horlbeck, M.A., Chen, Y., Tsai, J.C., Bassik, M.C., Gilbert, L.A., Villalta, J.E., Kwon, S.C., Chang, H., Kim, V.N., et al. (2015). Next-generation libraries for robust RNA interference-based genome-wide screens. Proc Natl Acad Sci U S A *112*, E3384–E3391.

Klein, E.E., Ben-Bassat, H.H., Neumann, H.H., Ralph, P.P., Zeuthen, J.J., Polliack, A.A., and Vánky, F.F. (1976). Properties of the K562 cell line, derived from a patient with chronic myeloid leukemia. Int J Cancer *18*, 421–431.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *517*, 583–588.

Liberali, P., Snijder, B., and Pelkmans, L. (2015). Single-cell and multivariate approaches in genetic perturbation screens. Nat. Rev. Genet. *16*, 18–32.

Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011a). ViennaRNA Package 2.0. Algorithms Mol Biol *6*, 26.

Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011b). ViennaRNA Package 2.0. Algorithms Mol Biol *6*, 26.

Luo, S., Lu, J.Y., Liu, L., Yin, Y., Chen, C., Han, X., Wu, B., Xu, R., Liu, W., Yan, P., et al. (2016). Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. Stem Cell 1–17.

Maeder, M.L., Linder, S.J., Cascio, V.M., Fu, Y., Ho, Q.H., and Joung, J.K. (2013). CRISPR RNA-guided activation of endogenous human genes. Nature Methods *10*, 977–979.

Mandegar, M.A., Huebsch, N., Frolov, E.B., Shin, E., Truong, A., Olvera, M.P., Chan, A.H., Miyaoka, Y., Holmes, K., Spencer, C.I., et al. (2016). CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. Cell Stem Cell.

Munoz, D.M., Cassiani, P.J., Li, L., Billy, E., Korn, J.M., Jones, M.D., Golji, J., Ruddy, D.A., Yu, K., McAllister, G., et al. (2016). CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. Cancer Discov.

Naumann, S., Reutzel, D., Speicher, M., and Decker, H.J. (2001). Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. Leuk. Res. *25*, 313–322.

Paddison, P.J., Silva, J.M., Conklin, D.S., Schlabach, M., Li, M., Aruleba, S., Balija, V., O'Shaughnessy, A., Gnoj, L., Scobie, K., et al. (2004). A resource for large-scale RNA-interference-based screens in mammals. *428*, 427–431.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python.

The Journal of Machine Learning Research *12*, 2825–2830.

Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., et al. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nature Methods *10*, 973–976.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *152*, 1173–1183.

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

Radzisheuskaya, A., Shlyueva, D., Müller, I., and Helin, K. (2016). Optimizing sgRNA position markedly improves the efficiency of CRISPR/dCas9-mediated transcriptional repression. Nucleic Acids Research gkw583.

Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J.F., et al. (2005). Exploration of the Function and Organization of the Yeast Early Secretory Pathway through an Epistatic Miniarray Profile. *123*, 13–13.

Shalem, O., Sanjana, N.E., and Zhang, F. (2015). High-throughput functional genomics using CRISPR–Cas9. Nat. Rev. Genet. *16*, 299–311.

Shao, S., Zhang, W., Hu, H., Xue, B., Qin, J., Sun, C., Sun, Y., Wei, W., and Sun, Y. (2016). Long-term dual-color tracking of genomic loci by modified sgRNAs of the CRISPR/Cas9

system. Nucleic Acids Research gkw066.

Tanenbaum, M.E., Gilbert, L.A., Qi, L.S., Weissman, J.S., and Vale, R.D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *159*, 635–646.

Tsai, S.Q., Wyvekens, N., Khayter, C., Foden, J.A., Thapar, V., Reyon, D., Goodwin, M.J., Aryee, M.J., and Joung, J.K. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nature Biotechnology *32*, 569–576.

Tuttle, M., Pruitt, B.W., Ben Ewen-Campen, Chari, R., Ter-Ovanesyan, D., Haque, S.J., Cecchi, R.J., Kowal, E.J.K., Buchthal, J., Housden, B.E., et al. (2016). Comparison of Cas9 activators in multiple species. Nature Methods 1–7.

Vojta, A., Dobrinić, P., Tadić, V., Bočkor, L., Korać, P., Julg, B., Klasić, M., and Zoldoš, V. (2016). Repurposing the CRISPR-Cas9 system for targeted DNA methylation. Nucleic Acids Research gkw159.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096–1101.

Wong, A.S.L., Choi, G.C.G., Cui, C.H., Pregernig, G., Milani, P., Adam, M., Perli, S.D., Kazer, S.W., Gaillard, A., Hermann, M., et al. (2016). Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. Proceedings of the National Academy of Sciences *113*, 2544–2549.

Wu, S.Q., Voelkerding, K.V., Sabatini, L., Chen, X.R., Huang, J., and Meisner, L.F. (1995).

Extensive amplification of bcr/abl fusion genes clustered on three marker chromosomes in human leukemic cell line K-562. Leukemia *9*, 858–862.

Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C.A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J.S., et al. (2015). Sequence determinants of improved CRISPR sgRNA design. Genome Research *25*, 1147–1157.

Zalatan, J.G., Lee, M.E., Almeida, R., Gilbert, L.A., Whitehead, E.H., La Russa, M., Tsai, J.C., Weissman, J.S., Dueber, J.E., Qi, L.S., et al. (2015). Engineering Complex Synthetic Transcriptional Programs with CRISPR RNA Scaffolds. *160*, 339–350.

Zhao, Y., Dai, Z., Liang, Y., Yin, M., Ma, K., He, M., Ouyang, H., and Teng, C.-B. (2014). Sequence-specific inhibition of microRNA via CRISPR/CRISPRi system. Sci. Rep. *4*, –.

Chapter 5

CRISPRi-based genome-scale identification of functional long non-coding RNA loci in human cells

Abstract

The human genome produces thousands of long non-coding RNAs (lncRNAs) – transcripts >200 nucleotides long that do not encode proteins. While critical roles in normal biology and disease have been revealed for a subset of lncRNAs, the function of the vast majority remains untested. Here, we developed a CRISPR interference (CRISPRi) platform targeting 16,401 lncRNA loci in 7 diverse cell lines including 6 transformed cell lines and human induced pluripotent stem cells (iPSCs). Large-scale screening identified 499 lncRNA loci required for robust cellular growth, of which 89% showed growth modifying function exclusively in one cell type. We further found that lncRNA knockdown can perturb complex transcriptional networks in a cell type-specific manner. These data underscore the functional importance and cell type-specificity of many lncRNAs.

Introduction

Sequencing efforts have revealed that the human genome produces tens of thousands of long non-coding RNAs (lncRNAs), transcripts over 200 nucleotides that are often spliced and polyadenylated but have no apparent protein coding potential (*1*-*3*). Certain lncRNAs play critical roles in cellular function, development, and disease (*4*, *5*). However, of the very large set of lncRNAs – many of which are differentially expressed in tissues and disease states – only a very small fraction have established biological functions, and even fewer are known to function in fundamental aspects of cell biology such as cell proliferation. Currently, it is not possible to predict which lncRNAs are functional, let alone what function they perform. Thus, a large-scale, systematic approach to evaluating the function of the vast population of lncRNAs is critical to understanding the roles that these non-coding transcripts play in cell biology.

A central limitation to systematic efforts to evaluate lncRNA function has been the lack of highly specific, scalable tools for inhibiting lncRNA gene activity (*6*). Gene deletion studies conducted in mice, flies, and human cells have yielded important biological insights about lncRNAs, but this approach is difficult to scale up (*7*-*10*). CRISPR/Cas9 nuclease approaches based on introduction of indels – while both scalable and useful for targeted loss of function studies of protein coding genes by altering the coding frame – are not well suited for the study of lncRNA gene function, as small deletions do not generally disrupt their biological activity (*11*-*13*). Nonetheless, larger Cas9-mediated genetic deletions can be effective at eliminating lncRNA genes (*6*, *14*-*17*). Screens based on RNA interference (RNAi) have been valuable (*18*, *19*) despite challenges with off-target effects (*20*). However, many lncRNAs localize to the nucleus, where RNAi exhibits variable knockdown efficiency (*21*).

We previously developed CRISPRi, a technology which can repress transcription of any gene via the targeted recruitment of the nuclease-dead dCas9-KRAB repressor fusion protein to the transcriptional start site (TSS) by a single guide RNA (sgRNA) (*22*-*24*). As CRISPRi acts only within a small window (1kb) around the targeted TSS (*23*), and as dCas9 occludes only 23bp of the targeted DNA strand (*25*), CRISPRi allows for precise perturbation of any lncRNA gene. By catalyzing repressive chromatin modifications around the TSS and serving as a transcriptional roadblock, CRISPRi tests a broad range of lncRNA gene functions including the production of *cis*- and *trans*-acting RNA transcripts (*4*), *cis*-mediated regulation related to lncRNA transcription itself (*26*-*29*), and enhancer-like function of some lncRNA loci (*14*, *15*, *30*). The repressive chromatin modification H3K9me3 catalyzed by CRISPRi is highly specific, with little to no off-target effects due to either spurious dCas9 binding or unintended silencing of distal regulatory elements, as measured by ChIP-seq or RNA-seq (*22*, *31*-*34*) see also Figure 4C below). To enhance CRISPRi for large-scale screening, we have improved upon the design of CRISPRi sgRNA libraries to optimize on-target activity while further minimizing-off target effects, enabling highly sensitive detection of essential coding genes (*35*).

Here, we developed CRISPRi libraries targeting 16,401 lncRNA loci (with 10 sgRNAs per TSS), and conducted screens for genes that are required for robust growth in 7 human cell types—6 transformed cell lines and induced pluripotent stem cells (iPSCs)(*36*). These large-scale screens, coupled with extensive validation studies, greatly increased the number of lncRNA genes known to have biological function and revealed lncRNA function to be highly cell type-specific. Our studies thus help elucidate the biology contained within the lncRNA genome, and provide a tool for both large-scale and targeted investigations of lncRNA function.

Results

CRISPRi screens identify lncRNA loci that modify cell growth

We first designed an sgRNA library to enable genome-scale CRISPRi screening of lncRNA gene function. We generated a comprehensive lncRNA gene set by merging three major non-coding transcriptome annotations (*37*-*39*), prioritized ~1/3 of these genes based on expression in any of a panel of cancer and non-transformed cell lines (Table S1), and designed 10 sgRNAs targeting each lncRNA transcription start site (TSS) using the hCRISPRi-v2.1 algorithm (*35*) (Figure 1A and S1). The cell lines represent a broad range of cell types studied by the ENCODE project (*40*), including a chronic myeloid leukemia cell line (K562), the cervical cancer line HeLa, a glioblastoma line (U87), and two mammary adenocarcinoma lines (MCF7 and MDA-MB-231). We also chose an iPSC line that inducibly expresses CRISPRi components (*33*, *41*). The library, termed "CRiNCL" for CRISPRi Non-Coding Library, is available as pooled lentiviral plasmid libraries on Addgene and *in silico* as Table S2.

We used this library to conduct screens for lncRNA loci that increase or decrease cell growth in each of 7 cell lines. We infected the full lentiviral library or targeted sublibraries (Figure S2A) into each cell line engineered to express dCas9-KRAB (*22*, *23*, *33*, *42*), selected for infected cells by puromycin selection, and cultured for between 12 and 20 days, measuring sgRNA enrichment by Illumina sequencing (Figure 1B and Table S3). The fraction of cells infected with the sgRNA library remained stable over the course of the screen (*23*), indicating that CRISPRi targeting of lncRNA loci does not exhibit non-specific toxicity (Figure S2B). To facilitate comparisons between screens conducted for different durations and in cell lines with different growth rates, we normalized sgRNA enrichment by total cell doublings to obtain the

quantitative growth phenotype γ, which reflects the positive or negative impact on cell growth caused by knockdown of a given gene (*43*) (Figure 1B).

Analysis of biological replicates revealed that the γ for targeting sgRNAs showed strong and reproducible phenotypes (Pearson r = 0.34-0.90) while non-targeting control sgRNAs were tightly distributed around 0 (Figure 1C and S2C, Table S3). We averaged replicate sgRNA phenotypes and used these to score lncRNA genes (*23*, *35*), calculating gene phenotypes from the mean of the top 3 sgRNAs targeting the gene and Mann-Whitney p-values from all 10 sgRNAs compared to non-targeting control sgRNAs (Figure 1D and S3A, Table S4). Within each screen, we also randomly sampled non-targeting sgRNA phenotypes to generate "negative control genes" and analyzed them as with lncRNA genes (see Methods), enabling us to estimate an empirical false discovery rate for each screen as well as the combined screen dataset (Figure S2D). LncRNA genes were considered to be hits if their combined phenotype effect size and p-value (referred to here as "screen score") exceeded a consistent threshold applied to each screen corresponding to an empirical false discovery rate of 5% (Figure S3C). Overall, we found between 28 and 438 lncRNA loci hits in each cell line (Figure 1E and S3A, Table S4).

We observed that for 169 of these lncRNA hits, the TSS of the non-coding gene was within 1kb of the TSS of a coding gene previously found to be essential in a CRISPRi screen (*23*), making it difficult to determine whether the observed phenotypes were due to knockdown of the target lncRNA or direct inhibition of the neighboring coding gene (Figure S3B). We thus removed these hits from the total set of hit genes for downstream analyses (Figure 1E, S3A, S3D), resulting in 169 "neighbor hits" and 499 "lncRNA hits," 299 of which are distal from any protein coding gene (~90% of which would not measurably impact growth upon knockdown). The 1kb threshold was chosen based on the maximum distance at which CRISPRi is effective as

revealed by analysis of dense sgRNA tiling and genome-scale screens (Figure S4) (*23*); increasing this threshold to 10kb classifies only an additional 19 genes as neighbor hits (Figure S3D).

A larger fraction of lncRNAs hits were observed in the iPSC screen, suggesting that this cell line is either more susceptible to growth perturbations or that iPSCs were differentiating to other cell types with lower growth rates. We therefore investigated iPSC differentiation in a secondary fluorescence activated cell sorting (FACS)-based screen by assessing loss of pluripotency as indicated by decreased *POU5F1/OCT4* expression. CRISPRi targeting of only 9 lncRNA loci reduced *POU5F1/OCT4* expression (Figure S5, Tables S5-6), suggesting that the majority of lncRNA hits identified in iPSCs primarily affect cell growth. To confirm that the increased fraction of lncRNA hits in iPSCs was not due to technical differences in CRISPRi function between cell lines, we performed a CRISPRi screen for protein-coding genes required for cell growth in iPSCs (Figure S6A, Table S7). These results corresponded well with our previously published K562 growth screen (*35*) in both the number of genes found to have function and in the ability to specifically identify known essential genes (Figure S6B-C) (*44*). Taken together, our screens identified 499 lncRNA genes that modify cell growth and have no essential coding gene neighbors, representing a large set of unstudied non-protein-coding genes serving important functions in cell biology.

lncRNA CRISPRi phenotypes are reproducible with robust knockdown

Extensive validation studies argue for the low false-positive and -negative rates of our studies. First, we individually cloned the top two sgRNAs targeting 65 representative lncRNA hit loci, 41 of which were hits in only one cell line. We tested whether the observed phenotypes

from the screens were reproducible using internally-controlled growth assays, in which the fraction of cells infected with an sgRNA were measured over time by flow cytometry. We monitored the growth effects of sgRNAs in the cell lines in which they exhibited a phenotype in the screen, as well as several sgRNAs in cell lines where they showed no effect, and found that the individual sgRNA growth phenotypes (γ) correlated well with the screen γ (Pearson r = 0.72, Figure 2A). This confirmed both that lncRNA knockdown phenotypes were reproducible and that the difference in lncRNA phenotype between cell lines was not due to technical differences between genome-scale screens. Analyzing these phenotypes over time further revealed distinct kinetics of cell depletion mediated by lncRNA knockdown (Figure 2B). For 12 lncRNA hits, we measured the levels of knockdown by qPCR and found over 70-95% knockdown for most of the targeted transcripts (14/14 sgRNAs in U87; 10/16 sgRNAs in MCF7) despite the effect of cellular depletion (Figure S7A).

In four cell lines, knockdown of lncRNA *PVT1* had a pro-growth phenotype. As *PVT1* had previously been characterized as a proto-oncogene (*45*) and pro-growth phenotypes in cancer cell lines are uncommon (*23*, *46*), we validated the pro-growth phenotype (Figure 2C and S7A) and investigated this complex locus further by conducting a CRISPRi screen in U87 cells with an sgRNA library tiling every possible site along the locus (17,469 sgRNAs). We found that only sgRNAs within 1kb of the most upstream TSSs, which is distal to any mapped enhancers, caused a consistent pro-growth phenotype (Figure 2D and S7B, Table S8). Within this TSS region, the majority of sgRNAs promoted cell growth, and knockdown of the major isoform was confirmed by qPCR (Figure S7A). sgRNAs outside of this 1 kb window around the TSS, which would not be expected to affect transcription of the major *PVT1* isoform (*23*), showed no consistent impact

on growth, arguing that the observed pro-growth phenotype is mediated by transcriptional interference.

Repression of lncRNA loci elicits lncRNA-specific transcriptome responses

To better understand the consequences of lncRNA CRISPRi, we performed RNA-seq following CRISPRi knockdown of 42 lncRNA hits in 3 cell types. 32 of these lncRNA loci were hits in only one cell type. Selected lncRNA loci did not have essential coding gene neighbors, and 2 or more sgRNAs per gene were tested individually. Distinct sgRNAs targeting the same lncRNA TSS resulted in highly correlated transcriptome responses (mean Pearson r = 0.980; Figure 2E) that were generally proximal to each other in hierarchical clustering analysis (Figure S8A-D). By contrast, pairs of sgRNAs targeting different hit lncRNA loci with the same phenotype direction had transcriptome responses that were more dissimilar (mean Pearson r = 0.942, Mann-Whitney p-value compared to same-gene pairs = $6.4x10^{-08}$;), suggesting distinct molecular mechanisms of the lncRNAs despite having similar phenotypes (Figure 2E).

RNA-seq analysis of differential gene expression also revealed several clusters of co-expressed genes, suggesting that growth modifier lncRNA loci regulate critical pathways (Figure S8A-D and Table S9). For instance, 2 lncRNA knockdowns that caused increased growth in U87 cells clustered by upregulation of translation genes ($p = 3.2x10^{-37}$), while other pro-growth sgRNAs showed correlated changes in expression of DNA replication ($p = 2.0x10^{-10}$) and post-transcriptional regulation ($p = 3.0x10^{-08}$). Clusters enriched for genes in the p53 pathway (e.g. *ATF3*) were upregulated by many anti-growth sgRNAs in both U87 and HeLa cells. Interestingly, K562 cells showed clusters of genes enriched for platelet degranulation ($p = 1.6x10^{-05}$) and response to decreasing oxygen levels ($p = 5.0x10^{-05}$). The median magnitude of

log_2 fold changes for differentially expressed genes in U87, HeLa, and K562 were 0.67, 0.86, and 1.17, respectively (Figure S8E), with several genes exhibiting > 2 fold up- or down-regulation consistently across many samples (Figure S8F). These results indicate that different lncRNAs can regulate distinct biological pathways that affect cell growth and proliferation.

Analysis of the chromosomal location of differentially expressed genes did not reveal a global trend toward transcriptional changes on the targeted chromosome (Figure S9). We did however find that knockdown of 14 lncRNA loci resulted in local transcriptional changes within a 20 gene window (Figure S10), suggesting that certain lncRNAs may preferentially act locally.

CRISPRi robustly inhibits lncRNA transcription

The fraction of growth modifier lncRNA loci identified in our screens (1-8% per cell line) was less than the fraction of essential protein-coding genes in previous reports (10-11%) (*35*, *46*). We therefore wanted to assess whether lncRNA genes that did not appear as a hit in any screen were true negatives or simply a result of ineffective repression by CRISPRi. To this end, using all 10 sgRNAs per gene, we measured the knockdown of five arbitrarily selected lncRNA genes that had no observed phenotype in any cells and were expressed in both K562 and U87 cells (Figure 2F and S7C). Of these 100 knockdown measurements, 61 showed over 90% repression of the targeted lncRNA. Furthermore, with the exception of *LOC100506710* in U87 cells, all lncRNAs were repressed by at least 90% by at least three different sgRNAs. For all sgRNAs, lncRNA knockdown efficiency correlated with their predicted CRISPRi activity, and the efficiency of knockdown was highly correlated between K562 and U87 cells (Pearson r = 0.78; Figure 2G). Based on these findings, with the exception of cases where a small amount of residual transcript is sufficient for lncRNA function, we infer that the majority of lncRNA loci

that did not appear as a screen hit produce transcripts that are not essential for robust growth of the cell line screened.

Growth modifier lncRNA function is highly cell type-specific

We next determined the number of lncRNA hits that were unique to a specific cell type or common to any combination of two or more of the cell types screened. The vast majority (89.4%) of lncRNA hits were unique to only one cell type, with none being a hit in 5 or more cell types (Figure 3A-C). Even when we restricted this analysis to the 1,329 lncRNAs expressed in all 7 cell types, 82.6% of the lncRNA hits modified growth in only one cell type (Figure 3B). Analysis of cell type-specificity scores based on the Jensen-Shannon distance, which quantifies how closely a given distribution resembles "perfect" specificity (*37*), revealed that the specificity of lncRNA screen scores was far greater than the specificity of lncRNA expression, for lncRNA hits (Figure 3D). Therefore, differential expression patterns alone are not sufficient to predict functional lncRNAs. Cross comparison of screen score distributions for lncRNAs that scored as hits in each cell type revealed that the threshold used for calling hits did not account for the cell type specificity (Figure 3E, S11D-E). Furthermore, cross-comparison of screen scores between replicates does not support technical variation as the source of the apparent cell type-specific function (Figure 3F and S11F).

In contrast to the sparse cell type overlap of lncRNA hits, analysis of published protein coding screens across similar numbers of cell types (*46*, *47*) revealed that the majority (54.8% in (*47*), 67.3% in (*46*)) of essential protein coding genes are hits in 2 or more cell types, with 20.4% and 30.8% being essential to all cell types screened in (*47*) and (*46*), respectively (Figure 3C, S11A-B). In addition, "neighbor hits" (lncRNA loci that are within 1kb of an essential protein

coding gene), were more likely to modify growth in multiple cell types, suggesting that CRISPRi targeted to these loci represses the adjacent essential coding gene, at least in some cases (Figure 3C, S11C,E).

Cell type-specific lncRNAs elicit highly divergent phenotypes

We sought to better understand the cell type-specific function of specific lncRNAs. We focused on *LINC00263*, which despite being expressed in all 7 cell lines screened, had a much stronger negative growth phenotype in U87 than in any other cell line (Figure S12A). The abundance of *LINC00263* transcript in a given cell line was also poorly correlated with the corresponding screen phenotype (Pearson r = 0.266). Validating these screen results, in internally controlled growth assays, two distinct sgRNAs to the TSS of *LINC00263* reduced the propagation of only U87 cells and not K562, MCF7 or HeLa cells (Figure 4A). H3K9me3 is a chromatin modification that is a result of local dCas9-KRAB activity (*31*), and in both U87 and HeLa cells with *LINC00263* CRISPRi targeting, ChIP-seq analysis demonstrated equal enrichment of H3K9me3 specifically at the *LINC00263* promoter for two independent sgRNAs (Figure 4B,C, S12B,C). However, despite such evidence of equivalent and specific CRISPRi targeting, U87 and HeLa cells had substantially different transcriptome changes after *LINC00263* knockdown. While U87 cells upregulated genes related to ER stress (e.g. *ATF4*, *CHAC1*; GO term $p = 4.51x10^{-09}$) and apoptosis (e.g. *DDIT3*, *SOD2*; GO term $p = 3.39x10^{-08}$), only *LINC00263* itself was differentially expressed in HeLa cells (adj. $p < 0.05$; Figure 4D). In K562 cells, these same 2 sgRNAs also produced very little transcriptional change (Figure S12D). Of note, in all three cell lines, knockdown efficiency of *LINC00263* was equivalent (Figure 4D, S12D). Consistent with our observations for *LINC00263*, knockdown of *PVT1* and *LINC00909,*

which were hits in U87 but not in HeLa, produced many more differentially expressed genes in U87 (Figure S12E). By contrast, depletion of *LINC00680,* which was a hit in both U87 and HeLa cells, resulted in comparable numbers of differentially expressed genes in U87 and HeLa cells (Figure S12E). Our results suggest that the specificity of lncRNA function is not due to differences in CRISPRi activity, but is related to differences in transcriptional networks across cell types.

We then targeted the *LINC00263* lncRNA transcript with antisense oligonucleotides (ASOs) that degrade RNA via an RNaseH-based mechanism. In both U87 and HeLa cells, ASOs reduced *LINC00263* transcript levels by 85-95% (Figure 4E). However, *LINC00263* ASOs decreased proliferation in U87 cells but not in HeLa cells (Figure 4F,G). The magnitude of proliferation decrease was also comparable to CRISPRi (Figure S12F,G), further supporting the cell type-specific function of this lncRNA. ASO knockdown of three other U87 lncRNA hits also reduced cell proliferation (Figure S12H,I), providing additional evidence for the functional contribution of the lncRNA molecule in these examples.

Machine learning identifies features predictive of growth modifier lncRNAs

Using data from our genome-scale screens, we sought to identify properties of the lncRNA hits that can distinguish them from non-hit lncRNAs. 18 classes of genomic data such as enhancer maps, expression levels, chromosomal looping data, conservation, and copy number variation from ENCODE (*40*), FANTOM (*48*), Vista (*49*), and other sources (*50-52*) were compared with all lncRNA loci screened in this study. 8 of these properties (expression, Pol2/CTCF looping by ChIA-PET, enhancers and super enhancers from (*51*), copy number variation) were cell type-dependent. Generalized linear models were constructed to assess which

genomic properties are predictive of lncRNA function (see Methods). Expression levels within each cell line, lncRNA gene body within 1kb of a mapped FANTOM Enhancer, lncRNA gene body within 5kb of a cancer-associated single nucleotide polymorphism (SNP) (*50*), and the number of exons were significant predictors of lncRNAs hits ($p < 0.01$) in repeated 10-fold cross validation (Figure 5, Table S10). 99.6% of lncRNA genes that were screened but not apparently expressed were not called as hits (Figure 5C). Whether the 11 growth modifier hits of such "non-expressed" lncRNA loci represent non-lncRNA mediated effects, inaccurate quantitation of the transcript levels, or effects mediated by lncRNAs acting at low expression remains to be determined. In support of the latter possibility, *HOTTIP* has been reported to function despite being expressed at ~0.3 copies per cell (*53*). Nonetheless, many highly expressed lncRNAs were not hits (*e.g.* 154 non-hit lncRNAs were detected at FPKM > 100), and the accuracy for predicting lncRNA hits was greater for a model using all variables as compared to a model that relied only on expression levels (Figure 5B).

Compared to non-hit lncRNAs, hit lncRNA gene bodies were 1.66 times more likely to be within 1kb of a mapped enhancer (Figure 5D). This represented 127 of the lncRNA hit loci identified in our screens. However, the FANTOM enhancer annotations used for our analyses were derived from hundreds of different cell types, and thus only a fraction of these enhancers are active in any given cell type in our screen (*48*, *49*). Hit loci were also 1.4 times more likely to be within 5kb of a cancer-associated SNP (Figure 5E). That our hits were enriched for multi-exonic lncRNAs is consistent with the concept that lncRNA splicing can be an aspect of lncRNA function (*26*) (Figure 5F). However, the explanatory power of exon number was relatively low, and our screen did identify several single exon hits such as *NEAT1*. However, no genomic property analyzed, alone or in aggregate, fully predicted growth modifier lncRNAs in a given

cell type, underscoring the importance of performing loss-of-function screens for defining sets of functional genes.

Discussion

By employing CRISPRi for systematic, large-scale screens for lncRNA function in multiple cell lines, we identified 499 lncRNA loci that are required for robust cell growth. This work increases considerably the number of known functional lncRNAs and revealed that the large majority (89%) of lncRNA genes identified modified growth in just one cell type. Studies of the protein-coding genome with similar large-scale screening efforts showed that an essential gene in one cell type is highly likely to be essential in the other cell types tested (*46*, *47*). In contrast to protein coding genes, of the 1,329 lncRNA genes expressed in all of the seven different cell lines tested, not one lncRNA gene was required for robust cell growth in all cell types, with the large majority of lncRNA gene hits being specific to just one cell line. Our results thus reveal a critical role of cellular context in determining lncRNA function.

Several clues to this specificity of lncRNA function emerge from our analyses. First, although cell type-specific expression of lncRNAs was the strongest predictor of lncRNA hits in our machine learning model (Figure 5A,C), it did not fully explain this functional specificity (Figure 3, 5B). For example, RNA-seq analysis points to *LINC00263* playing a role in a complex transcriptional network required for U87 cells, but that despite being expressed in other cell types, *LINC00263* appears dispensable for the normal expression of nearly all genes in these other cells (Figure 4D and S12D,E). Taking advantage of the scale of our dataset, we have also begun to discover genomic features that predict growth modifying function. Our finding that enhancer proximity and chromosome contacts correlate with lncRNA function suggests that higher-order chromatin structure can play a role in such specificity of lncRNA function (*28*, *29*) (*30*). The extent to which cell type-specific function of enhancer-templated lncRNAs results from repression of the transcript itself or its genomic locus remains an important open question.

In any case, the association of lncRNA function with higher order chromatin structure is consistent with the emerging view that chromosomal looping between lncRNA promoters and target genes differs between cell types (*54*) and is critical to lncRNA function (*55*). Finally, our finding that genomic regions containing growth modifier lncRNAs are enriched for cancer risk SNPs suggests that these lncRNAs may contribute to the pathogenesis of cancer.

Regardless of the mechanism(s) of the observed cell type-specificity of lncRNAs, this finding has implications for understanding the biological roles of lncRNAs. LncRNAs appear to have originated much later than protein coding genes, consistent with their not playing generic housekeeping roles (*3*, *56*). Our study, which focused on lncRNAs required for robust cell growth, underestimates the true number of functional lncRNAs in these cell types, as lncRNAs have been shown to regulate more evolutionarily complex cellular decisions such as cell fate (*7*, *19*, *57*, *58*), cancer metastasis (*59*, *60*), and perhaps neuronal function (*61*). The CRISPRi tools developed here can now be applied to the study of such higher order cellular processes, where lncRNAs might exhibit even greater richness of function. Finally, the exquisite cell type-specificity of lncRNA gene function has clear implications for targeted therapy.

Acknowledgments

We thank the members of the Lim and Weissman labs, particularly Alex Fields, Joshua Dunn, Manny DeVera, Miao Cui, and David Wu for helpful discussions and assistance. We thank Annie Truong for assistance with iPS cell culturing and Nathan Salomonis for iPSC RNA-seq data. We also thank Eric Chow and Derek Bogdanoff of the UCSF Center for Advanced Technology for sequencing assistance, and Laurakay Bruhn, Daniel Ryan, Luke Fairbairn, and Peter Tsang of Agilent Technologies for their assistance on the design and synthesis of oligonucleotide pools.

This project was supported by NIH 1R01NS091544-01A1, VA 5I01 BX000252-07, NIH SPORE DRP, the Shurl and Kay Curci Foundation, the LoGlio Foundation, and the Hana Jabsheh Initiative (to D.A.L.). S.J.L. is supported by NIH F30 NS092319-01. M.A.H., J.E.V., M.Y.C., Y.C., L.A.G., and J.S.W. were supported by the Howard Hughes Medical Institutes and the National Institutes of Health (P50 GM102706, U01 CA168370, R01 DA036858). S.W.C. and H.Y.C. are supported by National Institutes of Health (R35-CA209919, P50-HG007735). B.R.C and M.A.M are supported by the Gladstone Institutes and National Institutes of Health (U01HL100406, P01HL089707, R01HL130533). L.A.G. is supported by the NIH/NCI Pathway to Independence Award (K99CA204602). Oligonucleotide pools were provided courtesy of the Innovative Genomics Initiative. Sequencing data are deposited in GSE85011.

Materials and Methods

lncRNA CRISPRi library design

lncRNA target selection

LncRNA annotations were retrieved from Ensembl build 75 (using the biotypes lincRNA, antisense, 3 prime overlapping ncRNA, processed transcript, sense intronic, sense overlapping) (*39*), the Broad human lincRNA catalog (*37*), the MiTranscriptome (*38*) , and a set of human brain specific lncRNAs (*42*). Annotations were merged using the cuffmerge command in Cufflinks v2.2.1 (*62*).

LncRNAs that were transcribed in at least one of the 7 cell lines in this study were identified by quantifying the expression of lncRNAs using RNA-seq data. RNA-seq data were obtained from ENCODE and other sources: HEK293T (GSE56010), HeLa (GSE30567, GSE33480, GSE23316), K562 (GSE30567, GSE33480, GSE23316), MCF7 (GSE30567, GSE33480), MDAMB231 (GSE73526, GSE45732), iPS (clone PCBC15hsi2012040401 (*63*)), HFF (GSE69906). RNA-seq was performed in-house for U87 cells, using the illumina TruSeq Stranded mRNA kit. Reads were quality trimmed using seqtk v1.0 and aligned to the human genome (GRCh37) with tophat v2.0.10, using the merged transcriptome reference as the transcriptome index, the prefilter-multihits flag, and strand specific flag when appropriate. Transcript abundance estimation was performed using Cufflinks v2.2.1. For each gene, the median FPKM value of the replicate samples were obtained. For each cell line, a minimum expression threshold was set between 0.25-0.50 FPKM in order to screen as many genes as possible, given cell culture scale limitations. 21,578 lncRNAs were identified.

Generating lncRNA TSS annotations

From these 21,578 transcripts passing the expression filter, an initial set of 17,740 TSSs were obtained from transcripts belonging to the same gene and with 5' ends within 100bp of each other. These TSS annotations were further refined using the FANTOM cap analysis of gene expression (CAGE)-based TSS annotations as previously described (*35*) with adjustments. LncRNA TSS annotations could not be directly matched to FANTOM "p1@gene" CAGE peaks, and instead were matched to any same-stranded CAGE peak within 400bp labeled as "p1" or "p2," and annotation support was labeled as "CAGE, primary peaks." If no primary peaks were found, annotations could instead be refined by robust or permissive peaks within 200bp of the starting annotation, and were labeled as "CAGE, robust peak" or "CAGE, permissive peak," respectively. Where no CAGE peaks were found (due to the cell type-specific nature of lncRNA expression only 30% of TSS annotations were refined with CAGE peaks), the TSS as determined by the annotation sets above was used and labeled "Annotation." 66 of the original TSSs were assigned the same start site by this method, reducing the total number targeted to 17,674. This annotation is included as Table S1. As detailed below, a further 692 TSSs could not be uniquely targeted, reducing the total TSSs to 16,982. Finally, to avoid redundant information from different TSSs located in close proximity, TSSs within 100bp of each other were assigned to a single gene ID (designated LHnnn in Tables S1-6,8) for a total of 16,401 distinct lncRNA target loci.

sgRNA selection

All potential sgRNAs within 25bp upstream and 500bp downstream of the refined lncRNA TSS annotations were scored for predicted activity using the hCRISPRi-v2.1 algorithm, scored for

off-target sites near TSSs and in the genome using weighted Bowtie v1.0.0 (*64*), and filtered for required restriction sites (BstXI, BlpI, and SbfI) and overlap with higher-ranking sgRNAs as previously described (*35*). For 692 TSSs, 10 sgRNAs passing all filters could not be found and were discarded. 87.5% of sgRNAs accepted into the library passed the highest off-target stringency threshold, while 10.9% passed at the second-highest stringency. Non-targeting control sgRNAs were generated randomly weighted by the per-base nucleotide frequencies of the targeting sgRNAs in the library, and filtered for no target sites in the genome. sgRNAs targeting lncRNA genes specifically expressed in a subset of cell types were assigned to the appropriate hierarchical sublibrary (along with a proportional number of non-targeting controls) to enable screening only the desired gene sets (Figure S2A). Sublibraries were designed as the intersection of genes expressed in the cell lines indicated in Figure S2A, and then the full set of genes for a given cell line could be generated by combining sublibraries as follows:

iPSC = Common + (iPSC, HFF) + iPSC

HFF (not screened in this study) = Common + (iPSC, HFF) + iPSC

U87 = Common + Cancer common + (U87, HEK293T) + U87

HEK293T = Common + Cancer common + (U87, HEK293T) + HEK293T

K562* = Common + Cancer common + (K562, HeLa, MCF7) + (K562, HeLa) + K562

HeLa = Common + Cancer common + (K562, HeLa, MCF7) + (K562, HeLa) + HeLa

MCF7/MDA-MB-231 = Common + Cancer common + (K562, HeLa, MCF7) + MCF7

*all 13 sublibraries were screened in K562s to validate the cell line expression sublibrary strategy; see Figure 5C

Oligonucleotide pools were designed with flanking cloning and PCR sites as described, synthesized by Agilent Technologies (Santa Clara, CA), and cloned into the library sgRNA expression vector pCRISPRia-v2 (*23*, *35*).

For the *PVT1* tiling library, all possible sgRNAs from 25bp upstream of the *PVT1* locus to 25bp downstream that passed the second-highest off-target stringency filter and restriction site filters were included. Non-targeting sgRNAs matching this design in base composition and off-target stringency were included, and oligonucleotide pools were synthesized and cloned as above.

CRISPRi screens

Growth screens

Several cell lines expressing dCas9-KRAB were obtained from previous publications: HEK293T(*22*), HeLa(*22*), K562 (*23*), iPSCs (WTC-CRISPRi Gen IC)(*33*), and U87(*42*). MCF7 and MDA-MB-231 were generated for this study by infecting lentivirus expressing dCas9-KRAB-BFP (Addgene #46911; (*22*)) and sorting for single cell clones stably expressing high BFP. Replicates for these cell lines were performed in different clones. All cell lines except iPSCs were infected in duplicate with sgRNA the sublibraries described above or the *PVT1* tiling library, packaged with TransIT-LT1 (Mirus, Madison, WI) transfection in HEK293T cells (not expressing dCas9-KRAB), at an initial infection rate of 30-50% (300-500x coverage of the library). Cells were cultured for two days following infection, treated for two days with 0.75-1.00 μg/mL puromycin, allowed to recover for one day, and then cultured at a minimum coverage of 1000x for 12 days (K562, HEK293T, U87, HeLa) or 20 days (MCF7, MDA-MB-231) starting from this "T0." K562 cells were passaged daily, while adherent cells were split on

alternate days. iPSCs were infected at ~15% infection (with double the starting cell population to yield 300x coverage), grown for 3 days, selected with 1.5 μg/mL puromycin for 9 days, and allowed to recover for 3 days. iPSCs were then divided into two independent replicates and treated daily with 2μM doxycycline starting from this T0 and for the following 18 days (with primary endpoint at 12 days used here unless otherwise specified). Cells with a minimum of 1000x library coverage were harvested the day following puromycin recovery (T0) and at the endpoint, and processed for sequencing on Illumina HiSeq 2500 or 4000 as previously described(*23*, *35*).

Sequencing reads were aligned to the expected CRiNCL library sequences, counted, and quantified using the ScreenProcessing pipeline (https://github.com/mhorlbeck/ScreenProcessing; (*35*)) Negative control genes were generated by randomly sampling (with replacement) ~10 non-targeting sgRNAs per negative control gene to match the true gene TSSs targeted by the library, and then scoring the negative control genes for effect size and Mann-Whitney p-value as was done for the true genes. Genes (LHnnn) with multiple TSSs were collapsed to a single score by selecting the one with the lowest Mann-Whitney p-value.

In order to call hit genes from screens, we defined a "screen score" incorporating both the effect size and the p-values of genes in each screen. The screen score was calculated as | γ z-score from negative control gene distribution | x $-\log_{10}$ p-value, and for all screens a threshold of greater than or equal to 7 was applied to call hits.

Neighbor hits were classified by first calculating the distance between the each lncRNA TSS and the TSS of the closest protein coding gene (TSS-pc distance). LncRNAs whose TSS-pc distance was less than 1000 bp, and whose neighboring protein coding gene was 1) scored as essential in our previous screen in K562 cells (*23*), 2) expressed in the cell type in consideration, and 3) had the same phenotype direction as the lncRNA, were then classified as neighbor hits. Hits that did not meet these criteria were left as lncRNA hits.

OCT4 FACS-based screen

iPSCs were harvested 9 days post-doxycycline addition from the growth screen samples above and fixed with 4% paraformaldehyde for 10 minutes at room temperature followed by PBS wash. 9 days was chosen to balance the longer duration of continuous target gene knockdown with dropout of cells containing sgRNAs conferring a negative growth phenotype. Cells were permeablized with 0.5% saponin (Sigma) in PBS with 4% FBS and 2mM EDTA, stained with 1:100 mouse monoclonal α-*OCT3/4* antibody (sc-5279, Santa Cruz Biotechnology), washed with permeablization buffer, and stained with 1:200 Goat α-Mouse IgG-488 (A11029, Invitrogen) (*33*). Cells in the top and bottom 30% of *OCT4* signal as measured on the FITC-A channel were sorted for purity on a FACS AriaII custom. Sorted cells were then harvested for de-crosslinked genomic DNA using QIAamp DNA Formalin Fixed Paraffin Embedded Tissue kit (QIAGEN) following manufacturer's instructions but omitting paraffin removal steps, using one column per million sorted cells. Genomic DNA was directly amplified for Illumina sequencing using Q5 DNA polymerase (New England Biolabs) and sequenced on a HiSeq 4000. Sequencing data was analyzed as described above.

A screen for protein-coding genes required for robust growth in iPSCs was performed as with the iPSC lncRNA screen, with an endpoint at 14 days post-doxycycline addition. The screen was performed using the hCRISPRi-v2 H1 Drug Targets, Kinases, and Phosphatases sublibrary with 5 sgRNAs/gene (*35*), and was analyzed as above.

sgRNA Validation

sgRNAs for individual validation were cloned by annealing oligo pairs containing the sgRNA protospacer and flanking BstXI and BlpI cloning sites and ligating the resulting fragment into the sgRNA expression vector pU6-sgRNA EF1Alpha-puro-T2A-BFP (Addgene #60955). Internally controlled growth assays were performed by infecting cells with sgRNA lentiviruses at MOI < 1.0 and measuring the sgRNA+ fraction by BFP using flow cytometry on an LSRII (BD). Experiments were performed in biological triplicates from the infection step.

RT-qPCR

Cells were puromycin selected for 4 d (1 μg/mL) and subjected to 1 d recovery. K562 cells were infected for 48 hr, followed by 2 d puromycin treatment (3 μg/mL) and 2 d recovery. RNA was harvested with TRIzol and purified using the Direct-zol MiniPrep RNA purification kits (Zymo Research) with the on-column DNAse digestion step. cDNA were prepared with Transcriptor First Strand cDNA Synthesis Kit (Roche) using the oligo-dT protocol, and RT-qPCR was performed using LightCycler 480 SYBR Green I Master Mix (Roche) on a LightCycler 480 instrument (Roche). Experiments were performed in biological triplicates from the infection step. sgRNA protospacer sequences and RT-qPCR primers are listed in Table S11.

RNA-seq sample preparation and data analysis following CRISPRi

Cells were infected with sgRNA lentiviruses for 48 hr, followed by 4 d puromycin treatment (1 μg/mL) and 1 d recovery. K562 cells were infected for 48 hr, followed by 2 d puromycin treatment (3 μg/mL) and 2 d recovery. RNA was harvested using TRIzol and purified using the Direct-zol MiniPrep RNA purification kits (Zymo Research) with the on-column DNAse digestion step. RNA integrity was confirmed using the Agilent 2200 RNA ScreenTape. RNA-seq libraries were generated using TruSeq HT Stranded mRNA kit according to manufacturer's protocol (illumina). cDNA was validated using the Agilent 2200 DNA 1000 ScreenTape, Qubit 2.0 Fluorometer (Life Technologies), and ddPCR (Bio-Rad). Cluster generation and sequencing was performed on a HiSeq 4000, using the single end 50 read protocol. Reads were aligned to the human genome (GRCh37) using the spliced read aligner HISAT2 v2.0.3 (*65*) against an index containing SNP and transcript information (*genome_snp_tran*). Quantification of Ensembl build 75 genes was carried out with featureCounts (*66*) using only uniquely mapped reads.

Library complexity was calculated by counting the number of genes with greater than 2 reads, and knockdown efficiency was calculated by normalizing gene Transcripts per Million (TPM) for the experimental samples with the mean TPM of the control knockdown samples. Samples with fewer than 11,000 genes detected and weaker than 40% lncRNA knockdown were filtered. Pairwise Pearson correlations between RNA-seq samples were obtained using the sets of genes exhibiting significant variation within each cell type using the likelihood ratio test in DESeq2 (*67*) with an adjusted p value threshold of 0.001. Differential expression analysis for individual lncRNA knockdowns was performed using the Wald test in DESeq2 with an adjusted p value threshold of 0.05, using unique sgRNAs against the same lncRNA TSS as biological replicates.

For hierarchical clustering of co-expressed genes across multiple samples, we first grouped cells by cell type and then by the direction of phenotype of the sgRNA. Within each subgroup, we obtained the set of variable genes using the likelihood ratio test in DESeq2 (*67*) with an adjusted p value threshold of 0.001. These genes were then used for complete linkage hierarchical clustering using Pearson correlation coefficients as the distance metric. Sequencing data are deposited in GSE85011.

ChIP-seq sample preparation and data analysis

Cells were infected with sgRNA lentiviruses for 48 hr, followed by 4 d puromycin treatment (1 ug/mL) and 1 d recovery. Genome-wide histone modifications were determined by ChIP against H3K9me3 (Abcam ab8898) on 5 million cells as described in (*68*). Cells were cross-linked by adding 37% formaldehyde to a final concentration of 1% into culture medium and gently shaking for 10 min at room temperature. Reaction was quenched with glycine, and cells were then washed twice with ice-cold PBS containing protease inhibitors (1mM PMSF, 1X Roche cOmplete EDTA-free cocktail). Cells were scraped off of the plate using a cell lifter and pelleted for 5 min at 2,000 rpm at 4°C. Pellet was snap-frozen in liquid nitrogen and stored at −80°C. Pellet was then thawed and resuspended in Cell Lysis Buffer (5 mM PIPES pH 8, 85 mM KCl, freshly added 1% IGEPAL) with protease inhibitors (Pierce Halt Protease Inhibitor Cocktail). Cells were then homogenized using a type B glass dounce homogenizer, pelleted, and resuspended in Nuclei Lysis Buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS). Chromatin was sonicated in Diagenode TPX tubes using the Diagenode Bioruptor for 20 cycles and DNA was ranged from 150–700 bps as determined by gel electrophoresis. Debris was pelleted and discarded, and an aliquot was removed for Input DNA sequencing from the

sonicated chromatin within the supernatant. Sonicated chromatin was then diluted 5-fold in IP Dilution Buffer (50 mM Tris–HCl pH 7.4, 150 mM NaCl, 1% IGEPAL, 0.25% deoxycholic acid, 1 mM EDTA pH 8) with protease inhibitors and pre-cleared with Life Technologies Protein G Dynabeads for 2 hr at 4°C. 5 µg of antibody was added per million cells, and samples were incubated overnight at 4°C. Antibody-bound chromatin was then collected using Life Technologies Protein G Dynabeads and washed twice with IP Dilution Buffer, twice with IP Wash Buffer 2 (100 mM Tris–HCl pH 9, 500 mM LiCl, 1% IGEPAL, 1% deoxycholic acid), and once with IP Wash Buffer 3 (100 mM Tris–HCl pH 9, 500 mM LiCl, 150 mM NaCl, 1% IGEPAL, 1% deoxycholic acid). Precipitated chromatin was then eluted for 30 min at 65°C with Elution Buffer (1% SDS, 50 mM $NaHCO_3$). ChIP and Input DNA crosslinks were reversed by adding 5 M NaCl and heating at 65°C overnight. The following day, 10 mg/ml RNase A was added to precipitated chromatin, and chromatin was incubated for 30 min at 37°C. DNA was then recovered using Agencourt AMPure XP Beads and quantified using the Life Technologies Qubit Fluorometer.

ChIP DNA was then used for library preparation using the Kapa HyperPlus library preparation kit. 100ng of ChIP DNA was used for end repair and A-tailing. Illumina adapters were then ligated to the end-repair products. The library was amplified for 6 cycles before post-amplification cleanup using SPRI beads. Libraries were then quantified with the Life Technologies Qubit Fluorometer, and library size was confirmed using Agilent TapeStation 2200. ChIP-seq libraries were sequenced on a HiSeq 4000, 50 read single end.

Reads were aligned to the human genome (GRCh37) using bowtie v2.2.8 (*69*). Enrichment at promoter regions, which were defined as +/- 1kb of each TSS and generated from Ensembl GRCh37 build 75, were quantified using featureCounts v1.5.0-p2 (*66*). Signal was visualized using deepTools2 bamCoverage (*70*), normalizing reads to 1x sequencing depth. Differential H3K9me3 enrichment was analyzed using DESeq2 (*67*) , treating distinct sgRNAs against the same lncRNA TSS as replicate samples. Sequencing data are deposited in GSE85011.

Antisense oligonucleotide knockdown and proliferation assay

Antisense locked nucleic acid gapmers were designed against *LINC00263* using the Exiqon web server. Cells were transfected with the specified ASOs including negative control "A" at a final concentration of 50nM using Lipofectamine RNAimax Reagent (Invitrogen) according to the manufacturer's instructions. After 48 hours of transfection cells were seeded in duplicate. In order to maintain gene depletion, cells were transfected for a second time 7 days after the first transfection. Cell counting was performed every 2 days using Countess Automated Cell Counter (Invitrogen).

Flow cytometry for cell cycle analysis

Cells were transfected with the specified ASOs as described above. After 72 hours of transfection cells were pulsed with 33 μM bromodeoxyuridine (BrdU) for 20 min, and afterwards fixed in 70% ethanol. Subsequently cells were stained with primary anti-BrdU antibody (Clone B44; BD Biosciences) for 1 h, followed by 1 h incubation with Alexa Fluor 488 anti-mouse IgG (Invitrogen). DNA was counterstained by 0.1 mg/ml propidium iodide supplemented with RNase

for 1 h at 37°C. Analysis was performed on a FACSCalibur using CellQuest software (BD). Quantification and analysis of cell-cycle profiles were obtained using FlowJo (Tree Star, Inc).

Machine learning of lncRNA properties

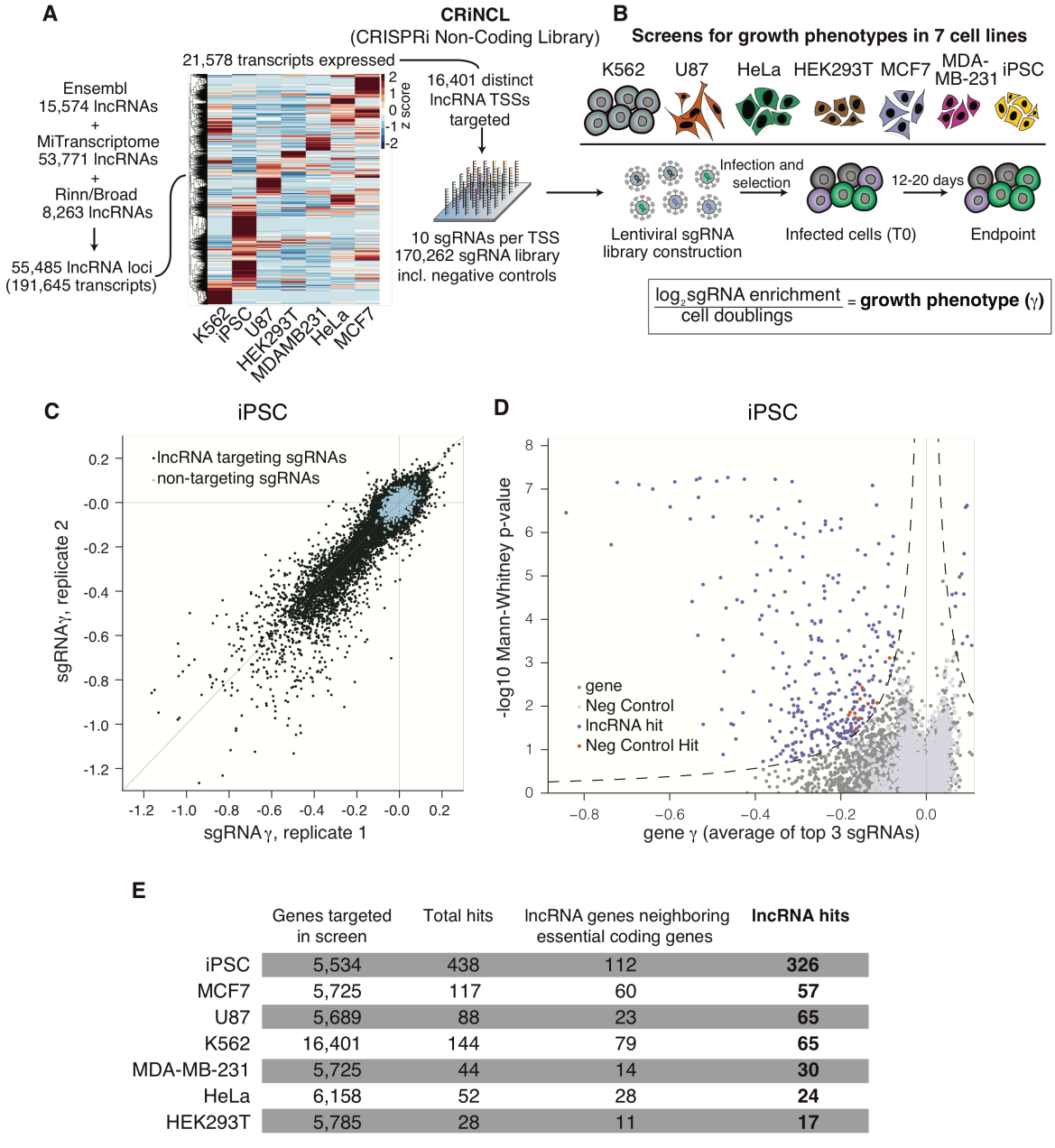
Genomic features were obtained from multiple sources. RNA-seq data were the same as used above for sgRNA library generation. Enhancer maps were obtained from the Fantom 5 Transcribed Enhancer Atlas (*48*), and VISTA Enhancer Browser set of experimentally confirmed human enhancers (*49*). Cell type-specific enhancer and super-enhancer maps for HeLa, U87, K562, and MCF7 cells were obtained from (*51*). LncRNA loci were considered near a (super)enhancer if it overlapped with or was within 1kb of a mapped enhancer. Cancer associated SNPs from the NHGRI GWAS Catalog were obtained from (*50*) and noted if any were within 5 kb of a lncRNA locus. Cell type-specific copy number variation data for HEK293T, HeLa, K562, U87, and MCF7 cells were obtained from ENCODE (GSE40698) and intersected with lncRNA loci. ChIA-pet data for HeLa, K562, and MCF7 cell lines were obtained from ENCODE (GSE39495). LncRNA loci that were overlapped completely by a Pol2 or CTCF loop with a score of at least 400 were noted. LncRNAs with mouse orthologs were identified using Slncky (*52*).

To generate machine learning models, lncRNAs phenotypes were binarized as hit (1) or non-hit (0) and used as the response variable. Categorical variables were assigned as either 1 or 0. Missing data, e.g. super-enhancer or CNV information for cell lines for which data were not available, were assigned the value of 0. Predictor variables were then centered to the mean and z standardized. Expression levels were log transformed. To avoid confounding by nearby protein

coding genes, only lncRNAs whose TSS were > 1kb from a protein coding TSS were considered. Several classes of models were generated and tested, using the R package caret on randomly sampled training (75% of data) and testing (25% of data) sets from our screen results. Logistic regression outperformed both support vector machines (least squares, polynomial kernel, radial kernel) and random forests in accurately classifying test sets of lncRNAs as hits or non-hit. Therefore, we used logistic regression to identify significant predictors of lncRNA hits. 100 iterations of ten-fold cross validation was performed by randomly withholding 10% of the dataset and training logistic regression models using the remaining data. Those predictors that repeatedly scored as significant ($p < 0.01$) were noted as reliable.

Figures

Liu, Horlbeck et al., Figure 1



E

	Genes targeted in screen	Total hits	lncRNA genes neighboring essential coding genes	**lncRNA hits**
iPSC	5,534	438	112	**326**
MCF7	5,725	117	60	**57**
U87	5,689	88	23	**65**
K562	16,401	144	79	**65**
MDA-MB-231	5,725	44	14	**30**
HeLa	6,158	52	28	**24**
HEK293T	5,785	28	11	**17**

Figure 1. CRISPRi screens identify lncRNA genes that modify cell growth

A) Schematic of CRISPRi library design strategy. Three lncRNA annotation sets were merged, prioritized by expression in the indicated cell lines, and targeted by 10 sgRNAs per TSS using the hCRISPRi-v2.1 algorithm. Heatmap represents expression as z-score of fragments per kilobase million (FPKM) within each cell line (see Figure S1 for TPM values). B) Schematic of growth screens performed in 7 different cell lines, and formula for calculation of the growth phenotype (γ). C) Scatter plot of sgRNA phenotypes from two independent replicates of a CRISPRi screen performed in iPSCs. D) Volcano plot of gene γ and p-value. Screen replicates were averaged, and sgRNAs targeting the same gene were collapsed into a growth phenotype for each gene by the average of the 3 top scoring sgRNAs by absolute value, and assigned a p-value by the Mann-Whitney test of all 10 sgRNAs compared to the non-targeting controls. Negative control genes were randomly generated from the set of non-targeting sgRNAs, and dashed lines represents a threshold for calling hits by screen score (see Methods). Neighbor hits are not displayed for clarity (see Figure S3A,B). E) Summary table of all CRISPRi growth screens performed.

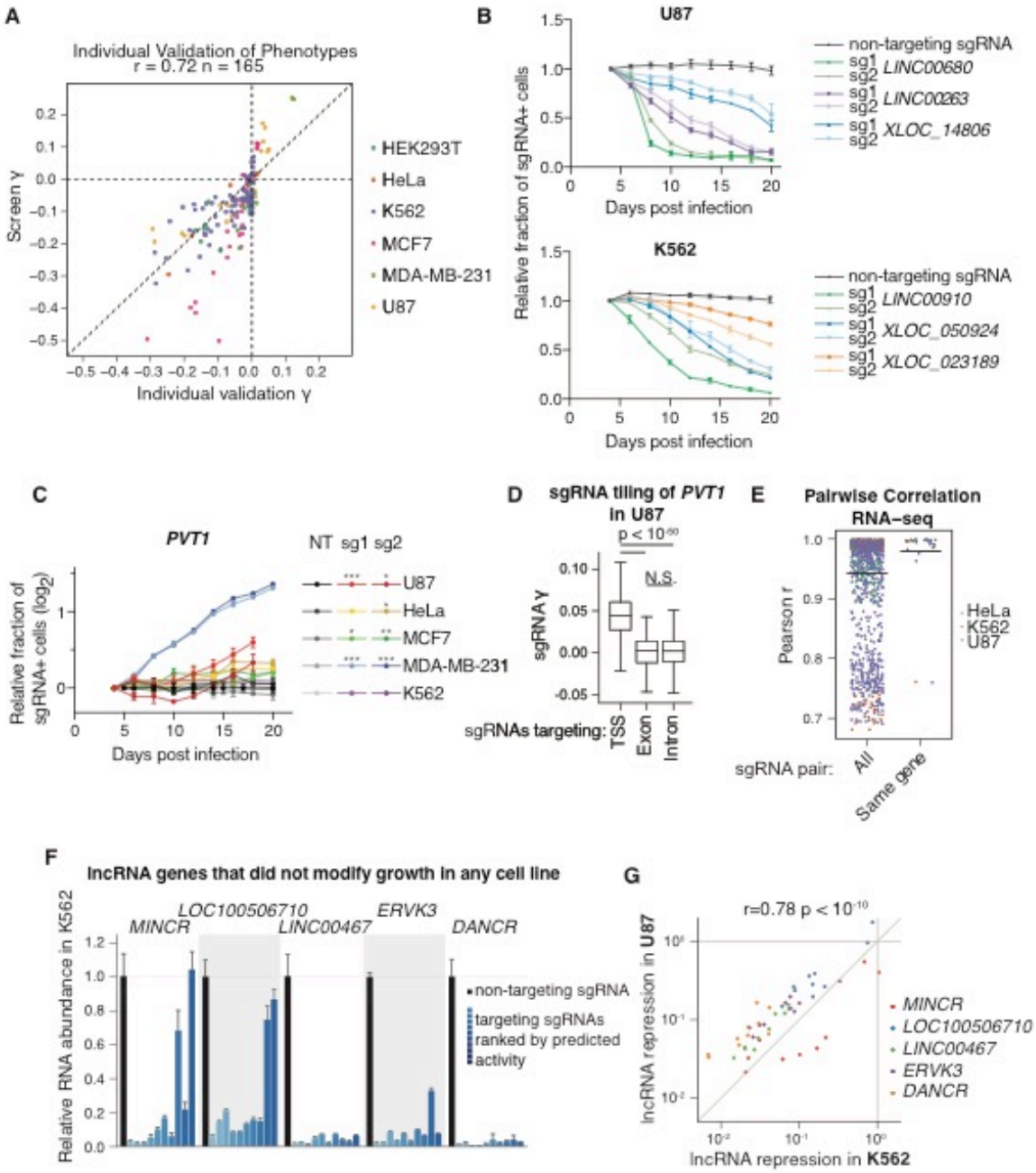

A
Individual Validation of Phenotypes
r = 0.72 n = 165
Screen γ
Individual validation γ
HEK293T
HeLa
K562
MCF7
MDA-MB-231
U87
B
U87
Relative fraction of sgRNA+ cells
Days post infection
non-targeting sgRNA
sg1 LINC00680
sg2
sg1 LINC00263
sg2
sg1 XLOC_14806
sg2
K562
Relative fraction of sgRNA+ cells
Days post infection
non-targeting sgRNA
sg1 LINC00910
sg2
sg1 XLOC_050924
sg2
sg1 XLOC_023189
sg2
C
PVT1
NT sg1 sg2
Relative fraction of sgRNA+ cells (log2)
Days post infection
U87
HeLa
MCF7
MDA-MB-231
K562
D
sgRNA tiling of PVT1 in U87
p < 10-50
N.S.
sgRNA γ
sgRNAs targeting:
TSS
Exon
Intron
E
Pairwise Correlation RNA-seq
Pearson r
HeLa
K562
U87
sgRNA pair:
All
Same gene
F
lncRNA genes that did not modify growth in any cell line
Relative RNA abundance in K562
MINCR
LOC100506710
LINC00467
ERVK3
DANCR
non-targeting sgRNA
targeting sgRNAs ranked by predicted activity
G
r=0.78 p < 10-10
lncRNA repression in U87
lncRNA repression in K562
MINCR
LOC100506710
LINC00467
ERVK3
DANCR

Figure 2. Validation of screen results shows reproducible phenotypes, correlated transcriptome responses, and robust knockdown of target transcripts

A) Individual sgRNA phenotypes from internally-controlled growth assays (B,C) compared to sgRNA phenotypes from screens. Individual growth phenotypes were calculated from relative fraction of sgRNA-containing cells at the endpoint, divided by the number of doublings from 4 days post-infection. Screen growth phenotypes represent the replicate average phenotype from the indicated cell line. B) Internally-controlled growth assays performed with sgRNAs targeting lncRNA hit genes in U87 and K562. Cells were infected with lentivirus of the sgRNA expression vector (including a BFP marker gene) and passaged for 20 days. The fraction of sgRNA-containing cells was measured as the fraction of high-BFP-expressing cells by flow cytometry, and expressed relative to the fraction at 4 days post infection. Points represent the mean and standard deviation of 3 biological replicates. C) Internally-controlled growth assays of *PVT1*-targeting sgRNAs in 5 cell lines. Assays were performed as in (B). Asterisks represent t-test p-values compared to the non-targeting (NT) sgRNA at the assay endpoint (* < 0.05, ** < 0.01, *** < 0.001). D) Boxplot of sgRNA growth phenotypes from tiling screen of *PVT1* in U87 cells. TSS represents all sgRNAs within 1kb of the *PVT1* "p1" and "p2" TSSs as annotated by FANTOM, exon represents sgRNAs targeting any *PVT1* exon annotated by Ensembl, and intron represents all other sgRNAs (see Figure S7B). sgRNA γs are the average of two replicates. E) Pairwise correlation of gene expression profiles for independent sgRNAs. Expression profiles were measured by RNA-seq and correlations were calculated from transcripts per million (TPM) of genes with significant variation of expression (see Methods). "All" represents every sgRNA pair from the same cell line with the same phenotype direction, except same-sgRNA and same-gene pairs. F) Relative RNA abundance in K562 of lncRNA genes that were not hits in any cell

line. RNA abundance for all 10 sgRNAs targeting the indicated genes in the CRiNCL library was measured by qPCR. Each bar represents the mean and standard deviation of 3 biological replicates, and is ordered by decreasing activity as predicted by the hCRISPRi-v2.1 algorithm. G) Correlation of lncRNA repression in K562 and U87. Points represent mean values from (F) and Figure S7C.

Liu, Horlbeck et al., Figure 3

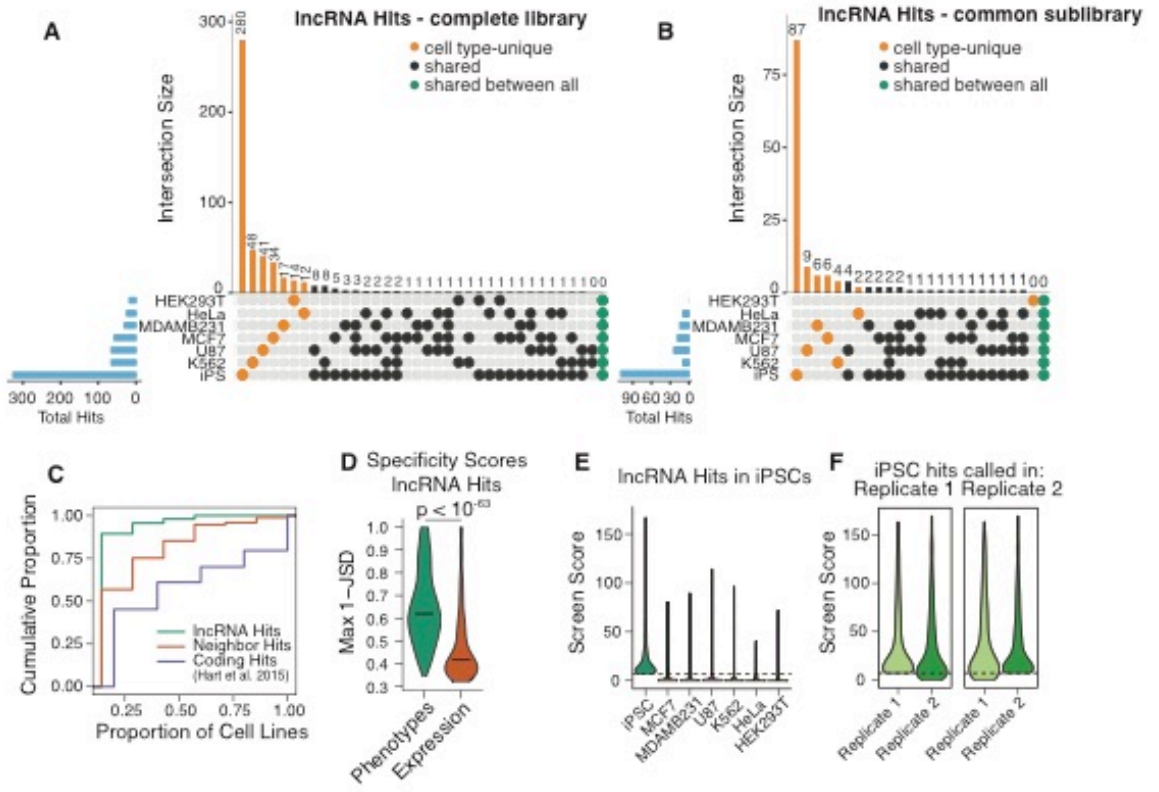


Figure 3. Growth modifier lncRNA function is highly cell type-specific

A) Numbers of lncRNA hits for each set of cell types in the complete library and (B) common sublibrary (lncRNAs that were expressed and screened in all cell types). Blue bars indicate total number of lncRNA hits in each cell type. C) Cumulative distribution function for the proportion of cell types in which each gene is a hit. Protein coding hits were obtained from Hart et al. 2015 using the authors' 5% FDR Bayes Factor threshold. D) Distributions of the maximum 1 - Jenson Shannon distance (JSD) metric of cell type-specificity for lncRNA hit screen scores and expression values. Horizontal lines – median. E) Distributions of screen scores across all cell types for lncRNAs that were hits in iPSCs. Dashed line represents screen score threshold for calling hit genes. F) Distributions of screen scores across both replicates of iPS cells, for lncRNAs that would be called as hits in replicate 1 (left) and in replicate 2 (right).

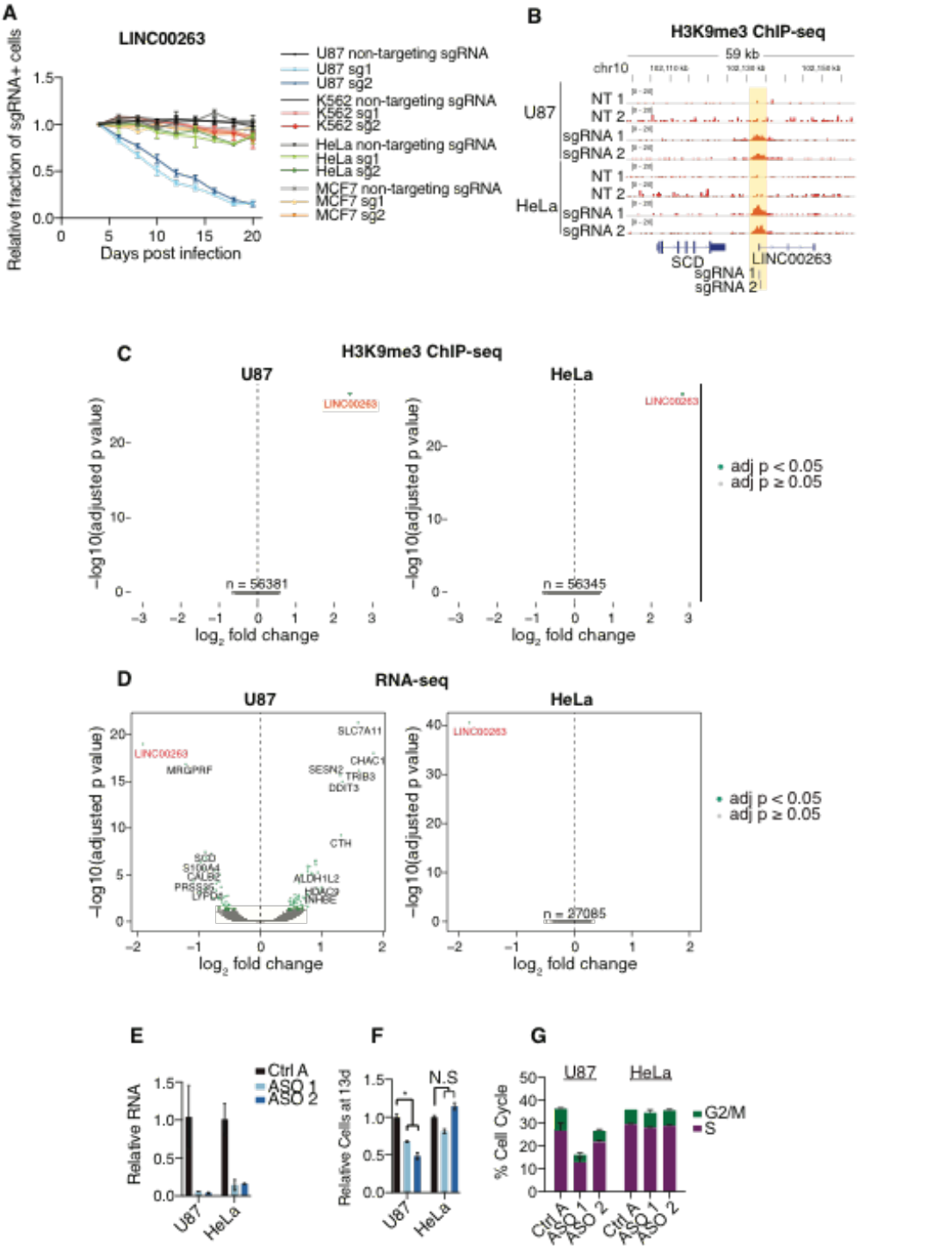

A
LINC00263
Relative fraction of sgRNA+ cells
Days post infection
U87 non-targeting sgRNA
U87 sg1
U87 sg2
K562 non-targeting sgRNA
K562 sg1
K562 sg2
HeLa non-targeting sgRNA
HeLa sg1
HeLa sg2
MCF7 non-targeting sgRNA
MCF7 sg1
MCF7 sg2
B
H3K9me3 ChIP-seq
59 kb
chr10
U87
HeLa
NT 1
NT 2
sgRNA 1
sgRNA 2
SCD
LINC00263
C
H3K9me3 ChIP-seq
U87
HeLa
−log10(adjusted p value)
log2 fold change
n = 56381
n = 56345
adj p < 0.05
adj p ≥ 0.05
D
RNA-seq
U87
HeLa
n = 27085
E
Relative RNA
Ctrl A
ASO 1
ASO 2
F
Relative Cells at 13d
N.S
G
% Cell Cycle
G2/M
S

Figure 4. Dissection of cell type-specific growth modifier lncRNA *LINC00263*

A) Internally-controlled growth assays for 2 independent sgRNAs targeting the TSS of *LINC00263* and non-targeting sgRNA in U87, K562, HeLa, and MCF7 cells. B) ChIP-seq against H3K9me3 in replicates of U87 and HeLa cells infected with non-targeting sgRNAs or *LINC00263* sgRNAs. Values represent normalized reads. C) Volcano plots for ChIP-seq samples in (B), representing genome-wide differential enrichment of H3K9me3 at promoter regions. Fold changes are between *LINC00263* sgRNAs over non-targeting sgRNAs. D) Volcano plots for RNA-seq differential expression following infection of *LINC00263* sgRNAs compared to infection of non-targeting sgRNAs. E) qPCR of ASO knockdown of *LINC00263* in U87 and HeLa cells. F) Proportion of cells at 13 days post ASO transfection, relative to control ASO. G) Percentage of cells in S or G2/M phases following ASO knockdown of *LINC00263*. * indicates p = 0.0029.

A

Variable	Cell Type Specific	Continuous or Discrete	odds	p value	% of CV where p < 0.01
Expression level	●	**C**	**1.62**	**4.52E-28**	**100**
Near FANTOM Enhancer		**D**	**1.34**	**6.44E-09**	**100**
Near cancer associated SNP		**D**	**1.10**	**1.84E-03**	**81**
Number of exons		**C**	**1.11**	**5.74E-03**	**66**
Within Pol2 ChIA-PET loop	●	D	0.83	1.27E-02	23
Near Super Enhancer	●	D	1.14	1.65E-02	23
Within CTCF ChIA-PET loop	●	D	0.83	1.81E-02	14
Near Traditional Enhancer	●	D	0.84	2.20E-02	5
Has mouse ortholog		D	1.08	4.44E-02	0
Locus is heterozygous deleted	●	D	0.92	1.98E-01	0
Is intergenic		D	1.12	2.57E-01	0
Transcript length		C	0.95	4.45E-01	0
Locus is amplified	●	D	0.96	5.12E-01	0
Is antisense		D	1.04	6.86E-01	0
Locus-Nearest coding gene dist.		C	0.98	8.12E-01	0
TSS-Nearest coding gene dist.		C	1.02	8.63E-01	0
Near Vista Enhancer		D	1.01	8.68E-01	0
Locus is homozygous deleted	●	D	0.79	9.69E-01	0

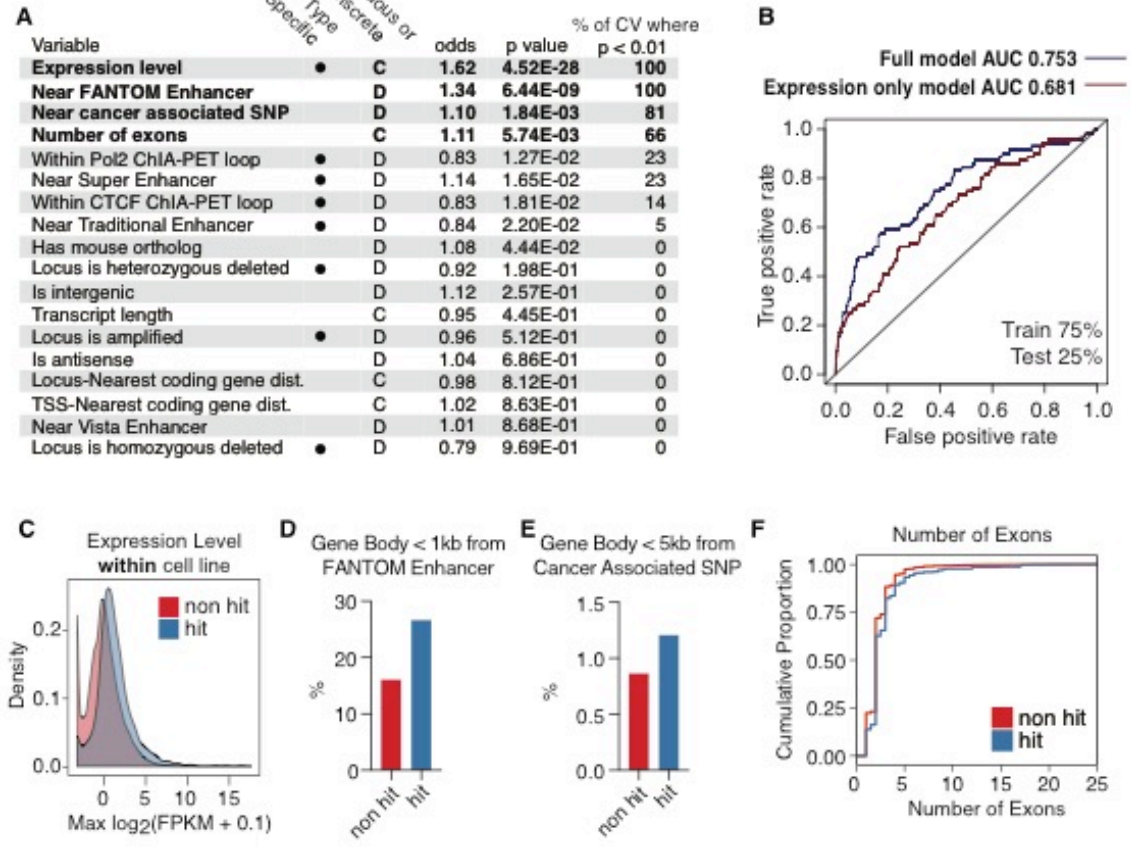


Figure 5. Machine learning identifies genomic features of growth modifier lncRNAs

A) Results from logistic regression model using 18 classes of genomic data as possible predictors of growth modifier lncRNAs. Cell type dependent variables are marked. Odds ratios represent relative impact of 1 standard deviation increase of given variable. Significant variables ($p < 0.01$) are bolded. Results of 10-fold cross validation are represented as the % of cross validation iterations where the given variable is significant. B) ROC curves for full model compared to model using only expression data. C) Density plot of expression levels for lncRNAs that scored as hits and non-hits, aggregated across all cell types. D) Percentage of non-hit (red) and hit (blue) lncRNAs whose gene bodies resided < 1 kb from an annotated FANTOM enhancer. E) Percentage of non-hit (red) and hit (blue) lncRNAs whose gene bodies resided < 5 kb from a cancer associated SNP. F) Cumulative distribution function of number of exons for non-hit (red) and hit (blue) lncRNAs transcripts.

Supplementary Figures

Liu, Horlbeck et al., Figure S1

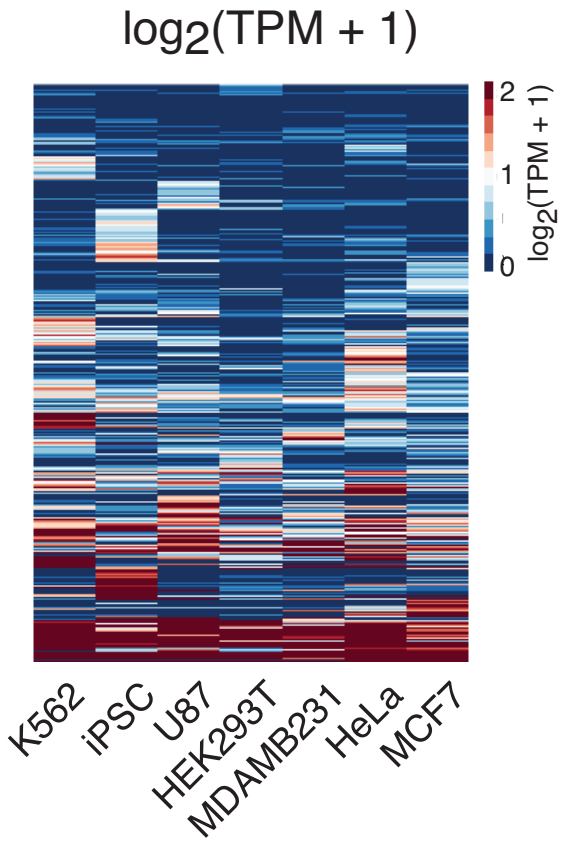


Figure S1. Expression levels of lncRNAs targeted in the CRiNCL library.

Rows correspond to those used in Figure 1A. TPM, transcripts per million.

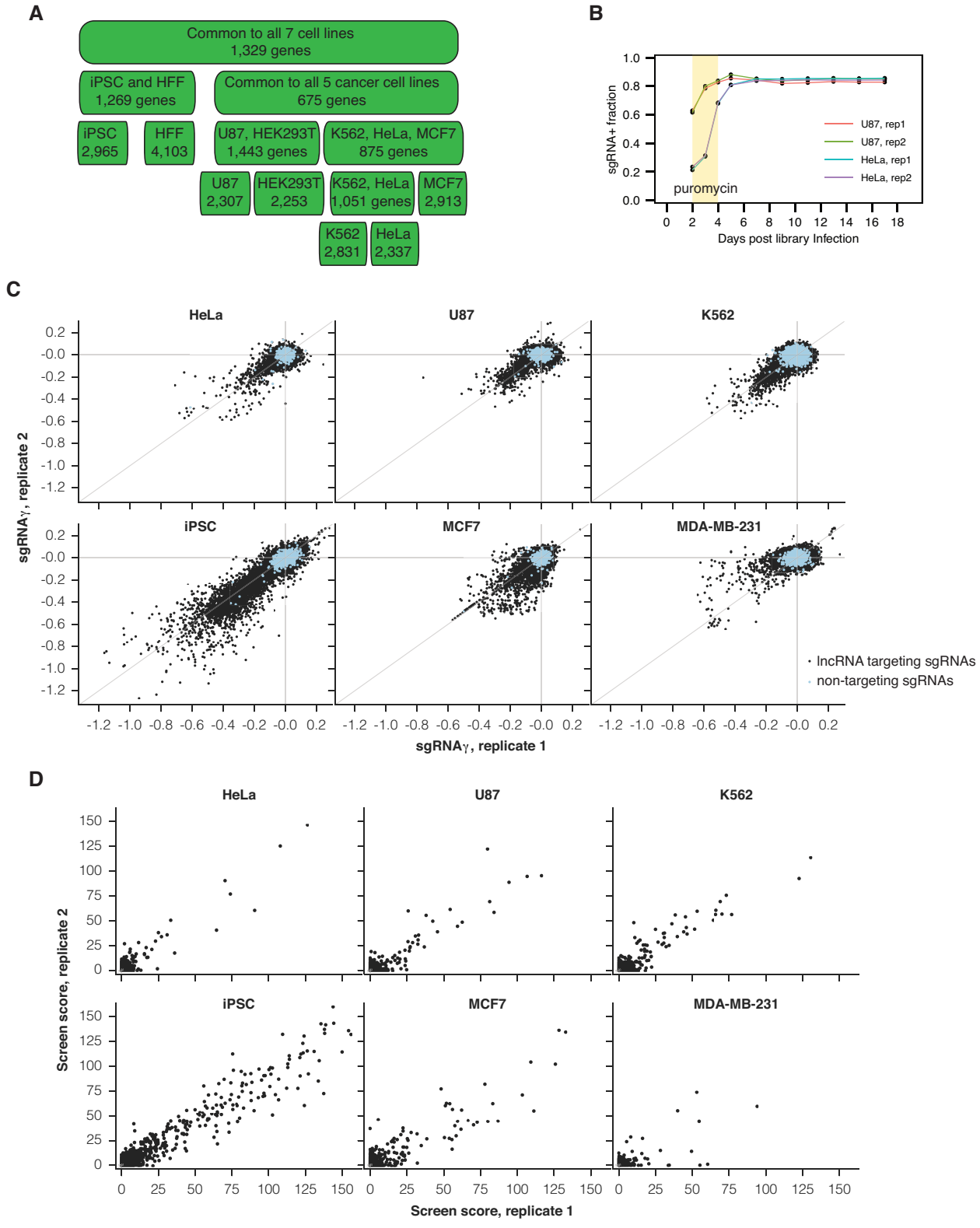
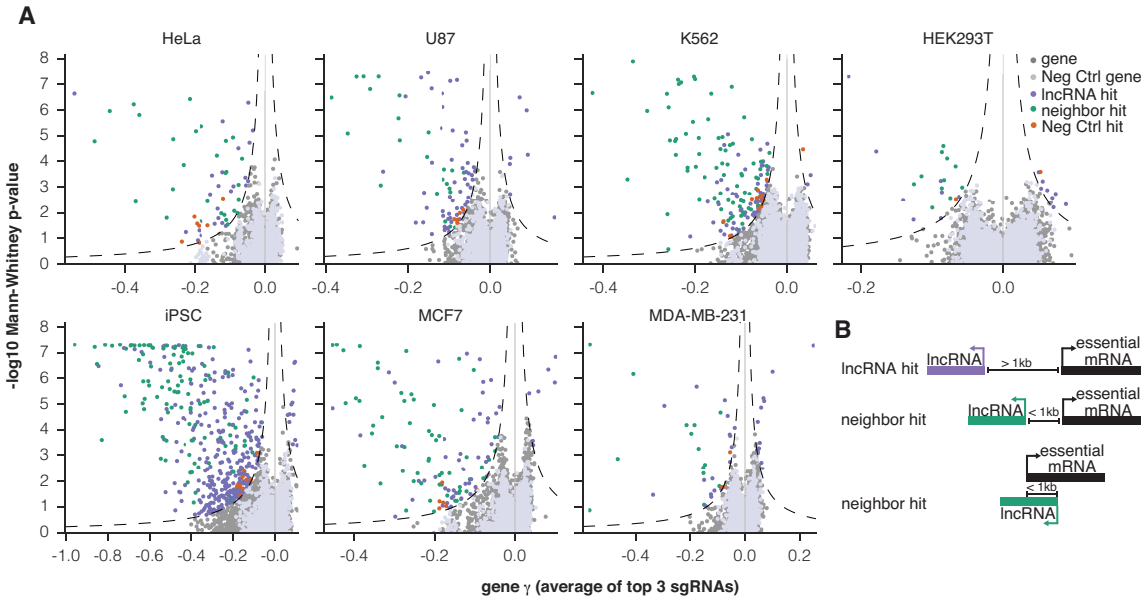
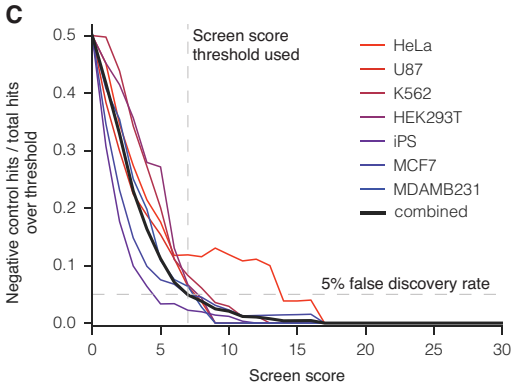

A
Common to all 7 cell lines
1,329 genes
iPSC and HFF
1,269 genes
Common to all 5 cancer cell lines
675 genes
iPSC
2,965
HFF
4,103
U87, HEK293T
1,443 genes
K562, HeLa, MCF7
875 genes
U87
2,307
HEK293T
2,253
K562, HeLa
1,051 genes
MCF7
2,913
K562
2,831
HeLa
2,337
B
sgRNA+ fraction
puromycin
U87, rep1
U87, rep2
HeLa, rep1
HeLa, rep2
Days post library Infection
C
HeLa
U87
K562
iPSC
MCF7
MDA-MB-231
sgRNAγ, replicate 2
sgRNAγ, replicate 1
lncRNA targeting sgRNAs
non-targeting sgRNAs
D
HeLa
U87
K562
iPSC
MCF7
MDA-MB-231
Screen score, replicate 2
Screen score, replicate 1

Figure S2. CRISPRi growth screens performed in seven cell lines.

A) Schematic of sublibrary divisions of the CRiNCL library. The library was divided into 13 sublibraries based on expression in 7 cell lines to facilitate library cloning and allow for targeted screens. Combinations of sublibraries were selected for screening in each cell line studied as described in Methods. B) Fraction of cells containing the sgRNA library over the course of the U87 and HeLa screens. sgRNA-containing fraction measured as the fraction of high-BFP-expressing cells by flow cytometry. C) sgRNA γ for replicate screens performed in 6 cell lines, as in Figure 1C. Only one replicate was performed for HEK293T. D) Screen scores from replicate screens for individual lncRNA loci. Screen scores were calculated as described in Methods.





D

Distance to nearest essential protein-coding gene TSS	lncRNA hits passing threshold
1.0kb	499
2.5kb	490
5.0kb	485
10.0kb	480

Figure S3. CRISPRi growth screen results and validation of thresholds used in screen analysis.

A) Volcano plots for screens performed in 7 cell lines, as in Figure 1D. Hits that were considered neighbor hits (see B) are labeled. B) Schematic of definitions of "lncRNA hit" and "neighbor hit." C) Fraction of negative control genes called as hits out of total number of hits above the indicated screen score threshold, calculated for each cell line and the combined dataset. D) Number of lncRNA hits classified after eliminating neighbor hits within the indicated distance from any essential protein-coding gene. A 1.0kb threshold was used for all analysis.

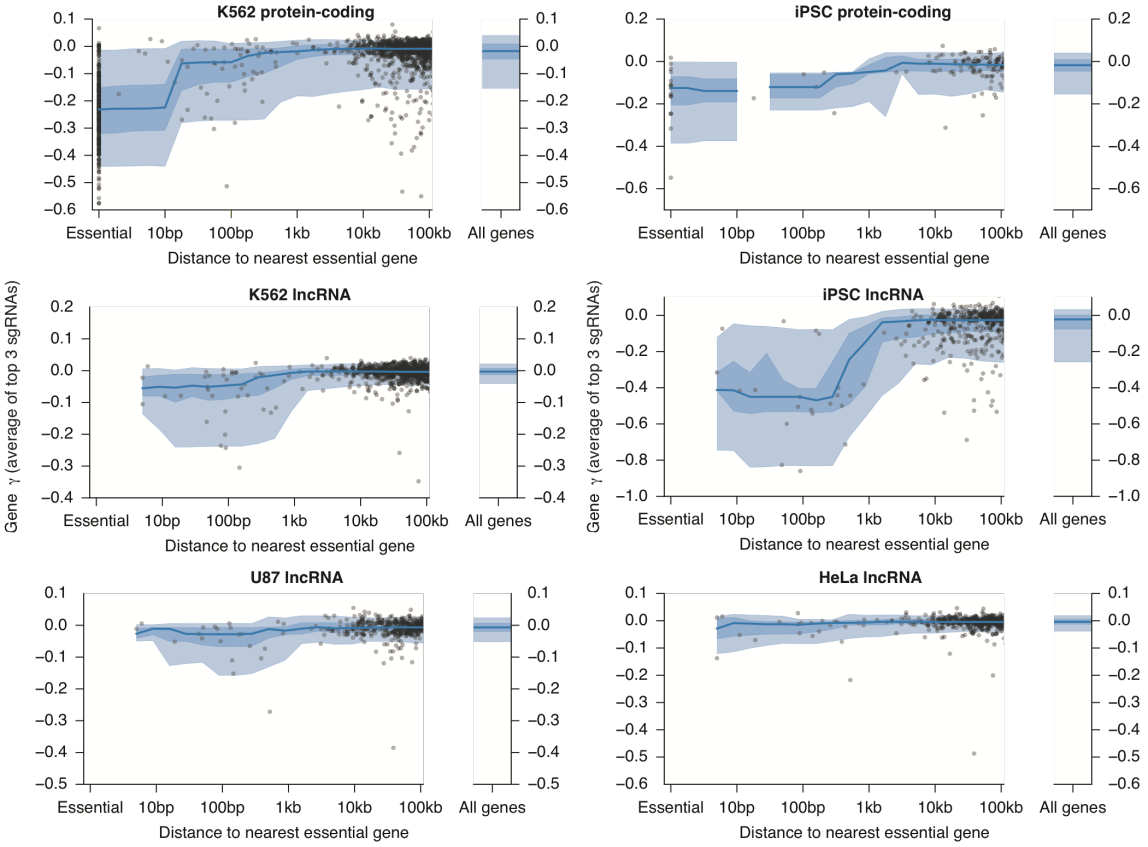

K562 protein-coding
iPSC protein-coding
K562 lncRNA
iPSC lncRNA
U87 lncRNA
HeLa lncRNA
Gene γ (average of top 3 sgRNAs)
Essential
10bp
100bp
1kb
10kb
100kb
All genes
Distance to nearest essential gene

Figure S4. CRISPRi growth phenotypes relative to gold standard essential genes.

Distribution of gene γ relative to the nearest gold standard essential protein-coding gene (*44*). Points indicate individual gene γ scores and blue shaded regions represent 5th, 25th, 50th, 75th, and 95th percentiles of all genes within 10-fold of the position. Screen data plotted are from the indicated protein-coding screens ((*35*) and Figure S4) or lncRNA screens.

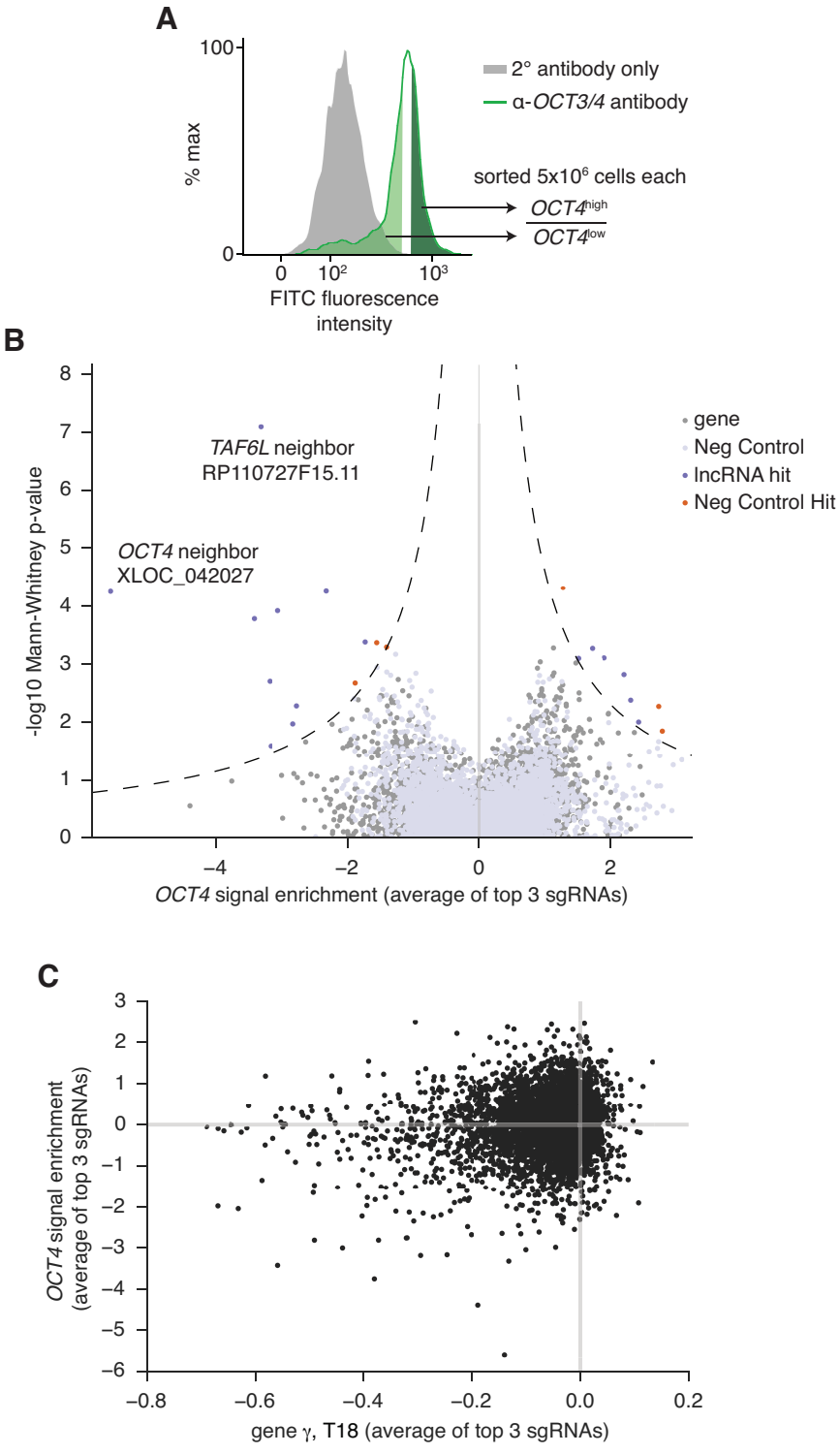

A
100
0
% max
2° antibody only
α-OCT3/4 antibody
sorted 5x10^6 cells each
OCT4^high
OCT4^low
0
10^2
10^3
FITC fluorescence intensity
B
-log10 Mann-Whitney p-value
TAF6L neighbor
RP110727F15.11
OCT4 neighbor
XLOC_042027
gene
Neg Control
lncRNA hit
Neg Control Hit
OCT4 signal enrichment (average of top 3 sgRNAs)
C
OCT4 signal enrichment (average of top 3 sgRNAs)
gene γ, T18 (average of top 3 sgRNAs)

Figure S5. A FACS-based screen for *OCT4* expression identifies genes that modify iPSC differentiation.

A) Representative FACS histogram of *OCT4* staining of iPSCs, with high and low 30% fractions highlighted. $5x10^6$ cells from each fraction were sorted and processed for Illumina sequencing. *OCT4* signal enrichment was calculated as the fraction of each sgRNA present in the high sample versus the low sample. B) Volcano plot of screen results, as in Figure 1D. C) Gene growth phenotypes from iPSC screen compared to *OCT4* signal enrichment. The results suggests that all *OCT4* screen hits also modify cell growth rates, but that most growth screen hits are not accompanied by changes in *OCT4* expression.

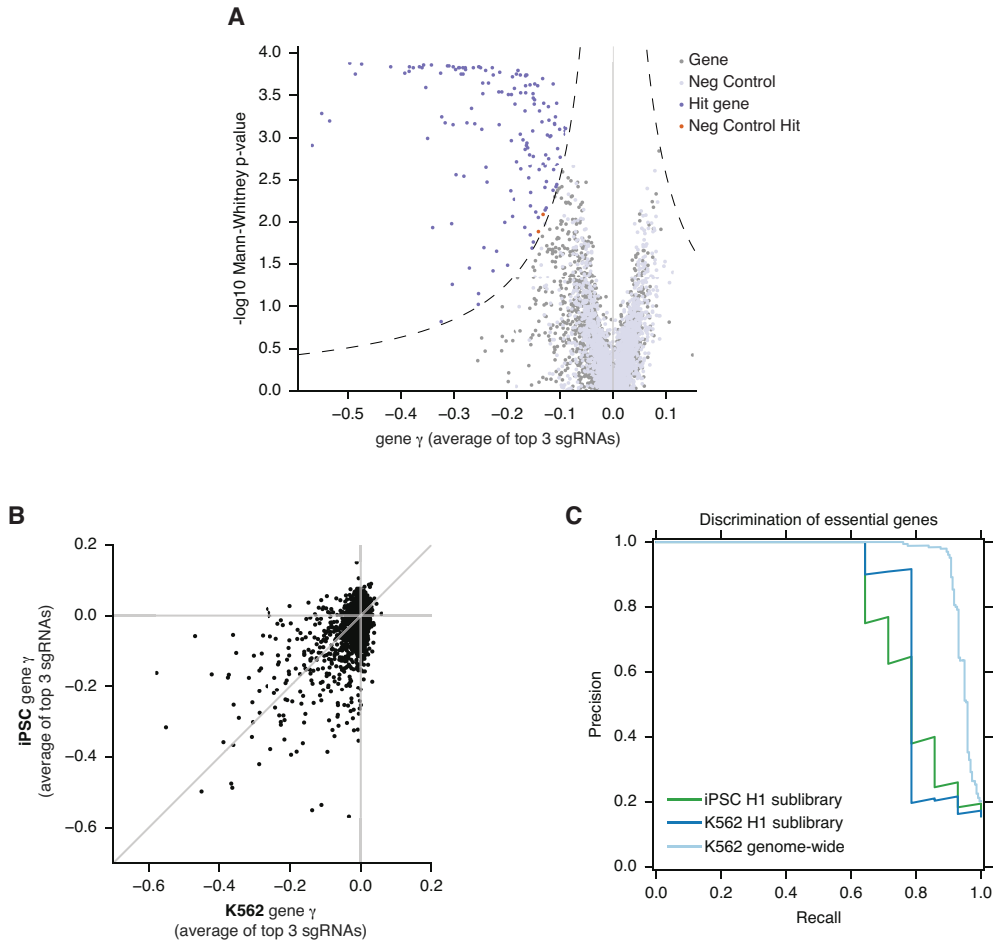

A
-log10 Mann-Whitney p-value
gene γ (average of top 3 sgRNAs)
Gene
Neg Control
Hit gene
Neg Control Hit
B
iPSC gene γ
(average of top 3 sgRNAs)
K562 gene γ
(average of top 3 sgRNAs)
C
Discrimination of essential genes
Precision
Recall
iPSC H1 sublibrary
K562 H1 sublibrary
K562 genome-wide

Figure S6. A CRISPRi screen for protein-coding genes that modify growth in iPSC.

A) Volcano plot for iPSC screen performed with the hCRISPRi-v2 H1 sublibrary, 5 sgRNAs/gene, displayed as in Figure 1D. Data are the average of two replicate screens. B) Scatter plot of gene γ from growth screens performed in K562 (*35*) and iPSC. K562 screen was performed with the genome-wide hCRISPRi-v2 library and reanalyzed here using only the H1 5 sgRNAs/gene sublibrary. C) Discrimination of gold-standard essential genes from non-essential genes (*44*) for K562 and iPSC screens, ranked by gene γ. K562 genome-wide and H1 data were analyzed using the 5 sgRNAs/gene sublibraries.

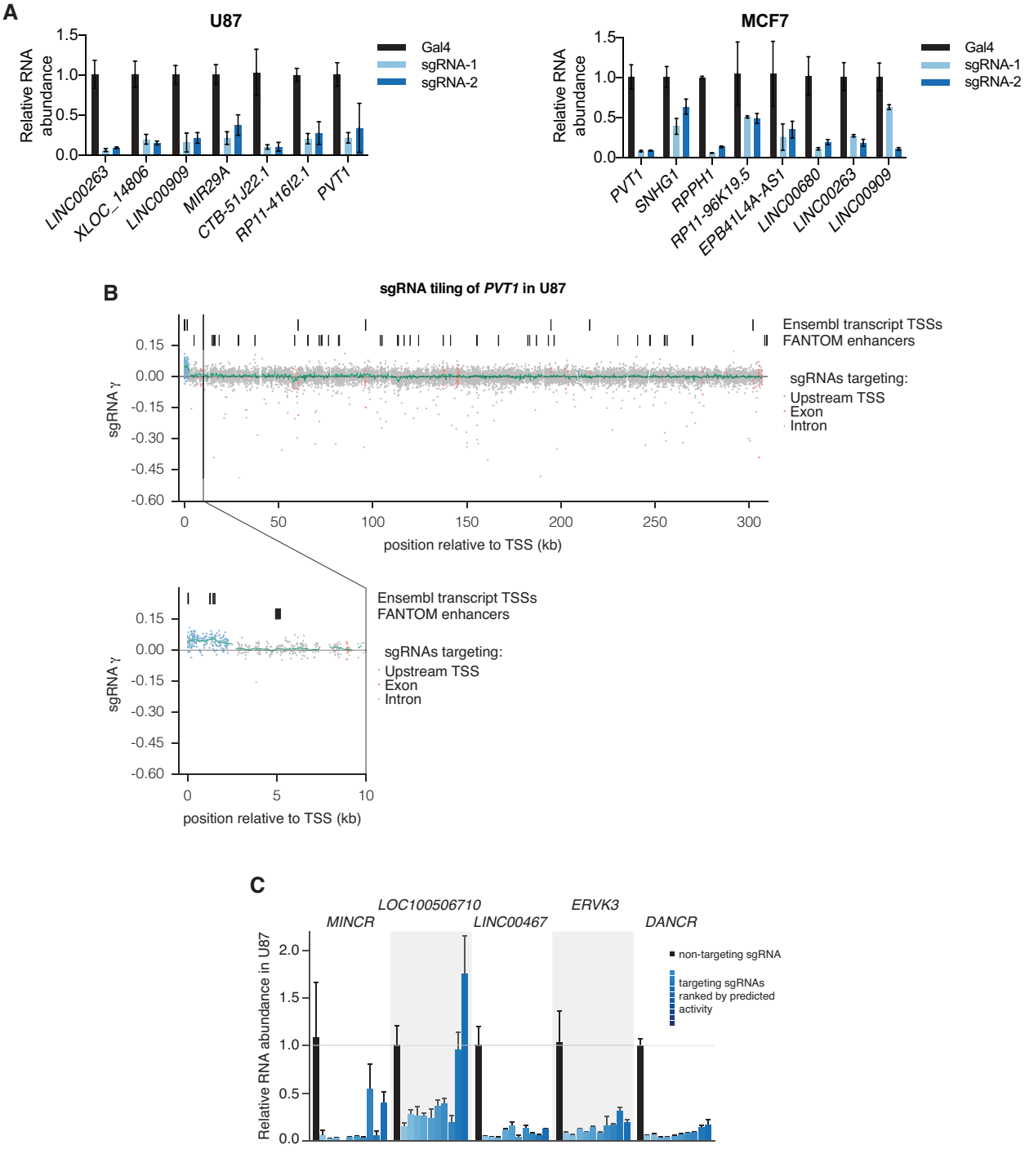

A
U87
MCF7
Relative RNA abundance
Gal4
sgRNA-1
sgRNA-2
LINC00263
XLOC_14806
LINC00909
MIR29A
CTB-51J22.1
RP11-416I2.1
PVT1
SNHG1
RPPH1
RP11-96K19.5
EPB41L4A-AS1
LINC00680
B
sgRNA tiling of PVT1 in U87
sgRNA γ
position relative to TSS (kb)
Ensembl transcript TSSs
FANTOM enhancers
sgRNAs targeting:
Upstream TSS
Exon
Intron
C
LOC100506710
ERVK3
MINCR
LINC00467
DANCR
Relative RNA abundance in U87
non-targeting sgRNA
targeting sgRNAs ranked by predicted activity

Figure S7. lncRNA CRISPRi produces robust knockdown and is specific to the TSS.

A) Relative RNA abundance of lncRNA hits upon knockdown with CRIPSRi. Bars represent mean and standard deviation of 3 biological replicates. B) sgRNA growth phenotypes from tiling screen of *PVT1* in U87 cells by position. sgRNA position was calculated as the genomic coordinate of the protospacer adjacent motif (PAM) relative to the *PVT1* p1 FANTOM TSS. sgRNA γ is the average of two replicates. TSS, exon, and intron are defined as in Figure 2D. Green line represents median phenotype of all sgRNAs within 250bp. C) Relative RNA abundance in U87 of lncRNA genes that were not hits in any cell line, as with Figure F,G.

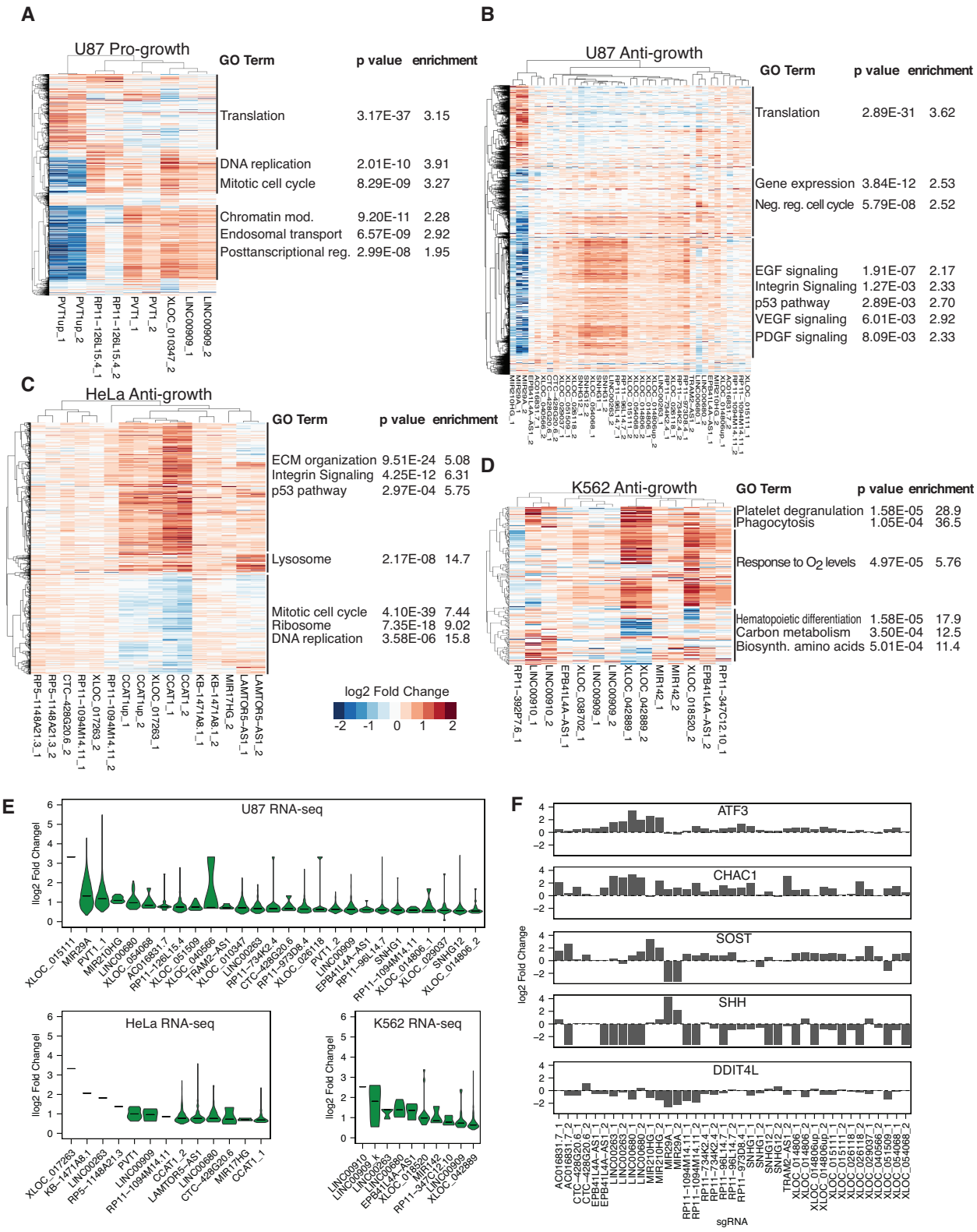

A
U87 Pro-growth
GO Term p value enrichment
Translation 3.17E-37 3.15
DNA replication 2.01E-10 3.91
Mitotic cell cycle 8.29E-09 3.27
Chromatin mod. 9.20E-11 2.28
Endosomal transport 6.57E-09 2.92
Posttanscriptional reg. 2.99E-08 1.95
PVT1up_1
PVT1up_2
RP11−126L15.4_1
RP11−126L15.4_2
PVT1_1
PVT1_2
XLOC_010347_2
LINC00909_1
LINC00909_2
B
U87 Anti-growth
GO Term p value enrichment
Translation 2.89E-31 3.62
Gene expression 3.84E-12 2.53
Neg. reg. cell cycle 5.79E-08 2.52
EGF signaling 1.91E-07 2.17
Integrin Signaling 1.27E-03 2.33
p53 pathway 2.89E-03 2.70
VEGF signaling 6.01E-03 2.92
PDGF signaling 8.09E-03 2.33
C
HeLa Anti-growth
GO Term p value enrichment
ECM organization 9.51E-24 5.08
Integrin Signaling 4.25E-12 6.31
p53 pathway 2.97E-04 5.75
Lysosome 2.17E-08 14.7
Mitotic cell cycle 4.10E-39 7.44
Ribosome 7.35E-18 9.02
DNA replication 3.58E-06 15.8
RP5−1148A21.3_1
RP5−1148A21.3_2
CTC−428G20.6_2
RP11−1094M14.11_1
RP11−1094M14.11_2
XLOC_017263_2
CCAT1up_1
CCAT1up_2
XLOC_017263_1
CCAT1_1
CCAT1_2
KB−1471A8.1_2
MIR17HG_2
LAMTOR5−AS1_1
LAMTOR5−AS1_2
log2 Fold Change
-2 -1 0 1 2
D
K562 Anti-growth
GO Term p value enrichment
Platelet degranulation 1.58E-05 28.9
Phagocytosis 1.05E-04 36.5
Response to O2 levels 4.97E-05 5.76
Hematopoietic differentiation 1.58E-05 17.9
Carbon metabolism 3.50E-04 12.5
Biosynth. amino acids 5.01E-04 11.4
RP11−392P7.6_1
LINC00910_1
LINC00910_2
EPB41L4A−AS1_1
XLOC_038702_1
LINC00909_1
LINC00909_2
XLOC_042889_1
XLOC_042889_2
MIR142_1
MIR142_2
XLOC_018520_2
EPB41L4A−AS1_2
RP11−347C12.10_1
E
U87 RNA-seq
log2 Fold Change
XLOC_015111
MIR29A
PVT1_1
MIR210HG
LINC00680
XLOC_054068
AC016831.7
RP11−126L15.4
XLOC_051509
XLOC_040566
TRAM2−AS1
XLOC_010947
LINC00263
RP11−734K2.4
CTC−428G20.6
RP11−973D8.4
XLOC_026118
PVT1_2
LINC00909
EPB41L4A−AS1
RP11−96L14.7
SNHG1
RP11−1094M14.11
XLOC_014806_1
XLOC_029037
SNHG12
XLOC_014806_2
HeLa RNA-seq
log2 Fold Change
XLOC_017263
KB−1471A8.1
LINC00263
RP5−1148A21.3
PVT1
LINC00909
RP11−1094M14.11
CCAT1_2
LAMTOR5−AS1
LINC00680
CTC−428G20.6
MIR17HG
CCAT1_1
K562 RNA-seq
log2 Fold Change
LINC00910
LINC00909_1
LINC00909_2
LINC00680
EPB41L4A−AS1
XLOC_018520
MIR142
RP11−347C12.10
LINC00909
XLOC_042889
F
ATF3
CHAC1
SOST
SHH
DDIT4L
log2 Fold Change
AC016831.7_1
AC016831.7_2
CTC−428G20.6_1
CTC−428G20.6_2
EPB41L4A−AS1_1
EPB41L4A−AS1_2
LINC00263_1
LINC00263_2
LINC00680_1
LINC00680_2
MIR210HG_1
MIR210HG_2
MIR29A_1
MIR29A_2
RP11−1094M14.11_1
RP11−1094M14.11_2
RP11−734K2.4_1
RP11−734K2.4_2
RP11−96L14.7_1
RP11−96L14.7_2
RP11−973D8.4_1
RP11−973D8.4_2
SNHG1_1
SNHG1_2
SNHG12_1
SNHG12_2
TRAM2−AS1_1
TRAM2−AS1_2
XLOC_014806_1
XLOC_014806_2
XLOC_014806up_1
XLOC_014806up_2
XLOC_015111_1
XLOC_015111_2
XLOC_026118_1
XLOC_026118_2
XLOC_029037_1
XLOC_029037_2
XLOC_040566_1
XLOC_040566_2
XLOC_054068_1
XLOC_054068_2
sgRNA

Figure S8. lncRNA CRISPRi produces co-expressed transcriptome responses.

Heatmaps of differentially expressed genes across samples segregated by (A) U87 pro-growth, (B) U87 anti-growth, (C) HeLa anti-growth, (D) K562 anti-growth. Significant gene ontology terms are indicated next to the clusters in which they are enriched, annotated with p value and enrichment. Fold changes are relative to non-targeting controls within the same cell type. E) Distributions of absolute value $\log_2$ fold changes for differentially expressed genes (adj. $p < 0.05$) for each lncRNA knockdown. F) Panel of genes consistently upregulated or downregulated across multiple CRISPRi samples in U87.

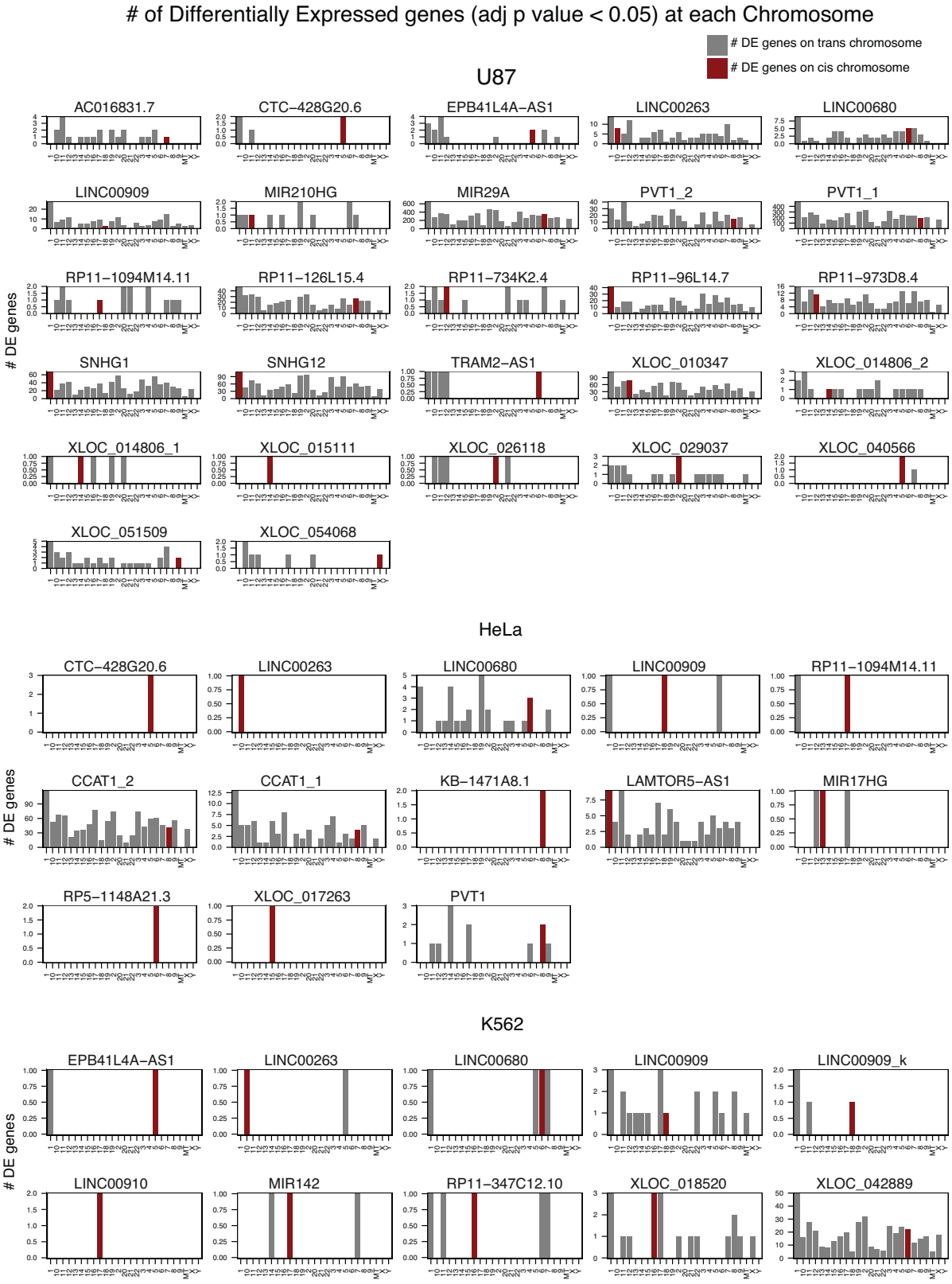

# of Differentially Expressed genes (adj p value < 0.05) at each Chromosome
DE genes on trans chromosome
DE genes on cis chromosome
U87
AC016831.7
CTC−428G20.6
EPB41L4A−AS1
LINC00263
LINC00680
LINC00909
MIR210HG
MIR29A
PVT1_2
PVT1_1
RP11−1094M14.11
RP11−126L15.4
RP11−734K2.4
RP11−96L14.7
RP11−973D8.4
SNHG1
SNHG12
TRAM2−AS1
XLOC_010347
XLOC_014806_2
XLOC_014806_1
XLOC_015111
XLOC_026118
XLOC_029037
XLOC_040566
XLOC_051509
XLOC_054068
DE genes
HeLa
CTC−428G20.6
LINC00263
LINC00680
LINC00909
RP11−1094M14.11
CCAT1_2
CCAT1_1
KB−1471A8.1
LAMTOR5−AS1
MIR17HG
RP5−1148A21.3
XLOC_017263
PVT1
DE genes
K562
EPB41L4A−AS1
LINC00263
LINC00680
LINC00909
LINC00909_k
LINC00910
MIR142
RP11−347C12.10
XLOC_018520
XLOC_042889
DE genes

Figure S9. Chromosome distribution of differentially expressed genes after each lncRNA knockdown.

Red bars indicate chromosomes harboring the lncRNAs of interest (*cis*). Gray bars represent chromosomes that do not contain the lncRNAs of interest (*trans*).

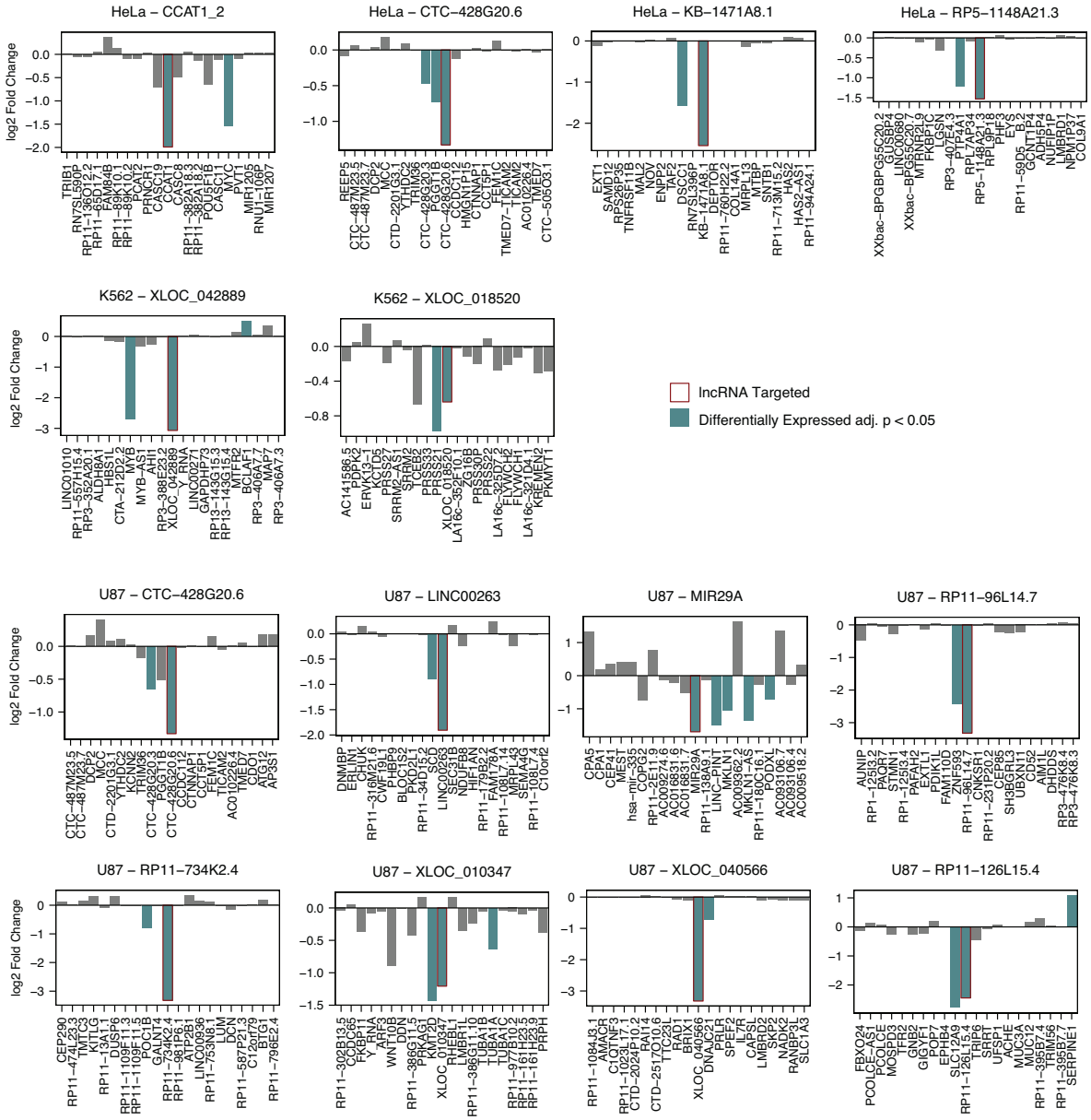

HeLa – CCAT1_2
HeLa – CTC–428G20.6
HeLa – KB–1471A8.1
HeLa – RP5–1148A21.3
K562 – XLOC_042889
K562 – XLOC_018520
lncRNA Targeted
Differentially Expressed adj. p < 0.05
U87 – CTC–428G20.6
U87 – LINC00263
U87 – MIR29A
U87 – RP11–96L14.7
U87 – RP11–734K2.4
U87 – XLOC_010347
U87 – XLOC_040566
U87 – RP11–126L15.4
log2 Fold Change

Figure S10. Local transcriptional changes within 20 gene-windows surrounding each lncRNA of interest following CRISPRi knockdown.

Red outline indicates targeted lncRNA. Blue bars indicate differentially expressed genes, among the broader set of differentially expressed genes genome-wide (DESeq2 adj. $p < 0.05$) upon knockdown.

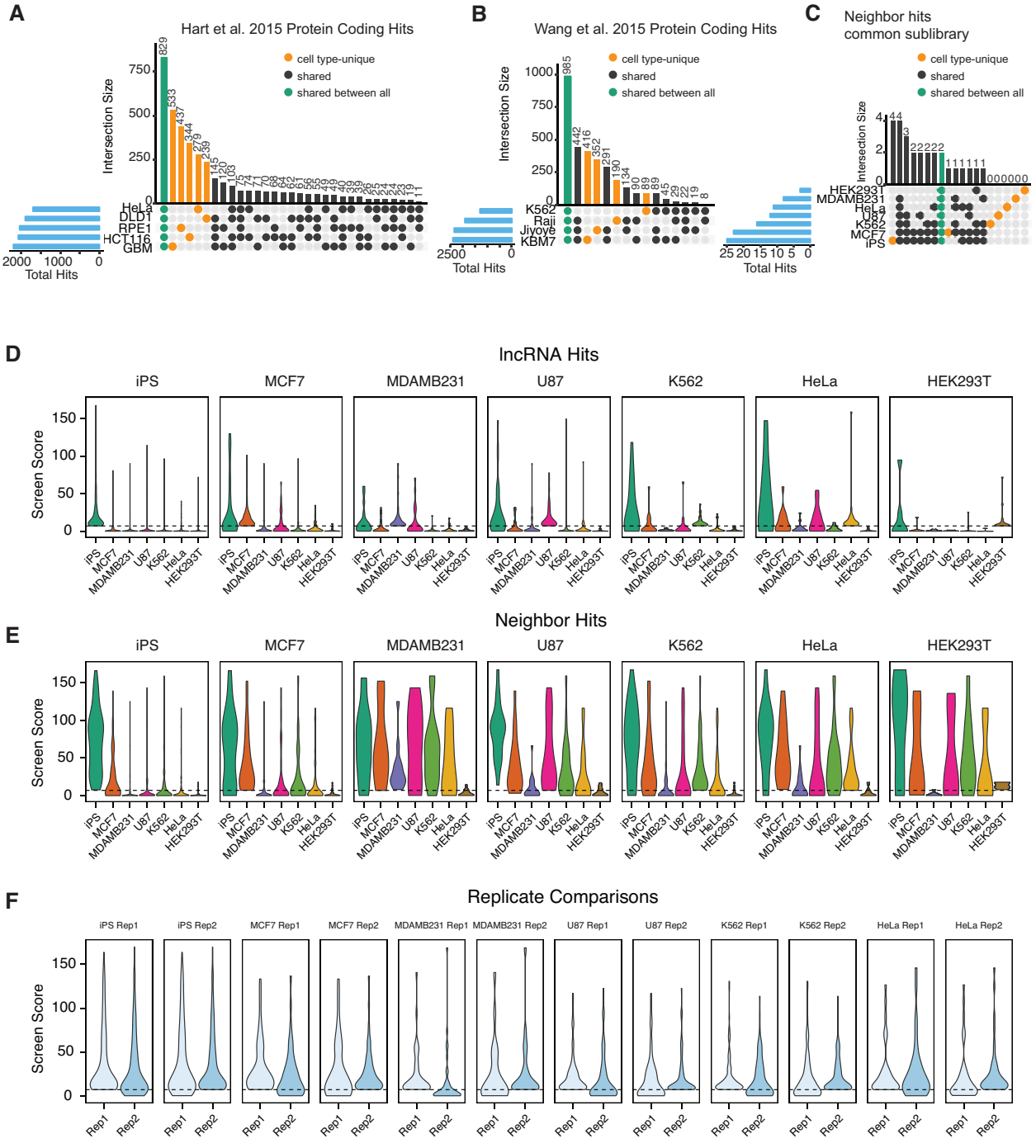

A
Hart et al. 2015 Protein Coding Hits
cell type-unique
shared
shared between all
Intersection Size
Total Hits
HeLa
DLD1
RPE1
HCT116
GBM
B
Wang et al. 2015 Protein Coding Hits
K562
Raji
Jiyoye
KBM7
C
Neighbor hits
common sublibrary
HEK293T
MDAMB231
U87
D
lncRNA Hits
Screen Score
iPS
MCF7
MDAMB231
U87
K562
HeLa
HEK293T
E
Neighbor Hits
F
Replicate Comparisons
iPS Rep1
iPS Rep2
MCF7 Rep1
MCF7 Rep2
MDAMB231 Rep1
MDAMB231 Rep2
U87 Rep1
U87 Rep2
K562 Rep1
K562 Rep2
HeLa Rep1
HeLa Rep2
Rep1
Rep2

Figure S11. lncRNA hit specificity is greater than essential protein coding gene specificity. A) Numbers of protein coding gene hits for each set of cell types screened in Hart et al. 2015 and (B) Wang et al. 2015. Wang et al. genes were considered hits if they passed a 5% false discovery threshold set by precision-recall analysis (*44*). C) Numbers of hits in our study that share promoters with essential protein coding genes (neighbor hits). Blue bars indicate total number of hits in each cell type. D) Distributions of screen scores across all cell types, for lncRNAs and (E) "neighbors" that were called hits in each given cell type. F) Distributions of screen scores across both replicates of each cell type, for lncRNAs that would be called as hits in replicate 1 (left) and in replicate 2 (right).

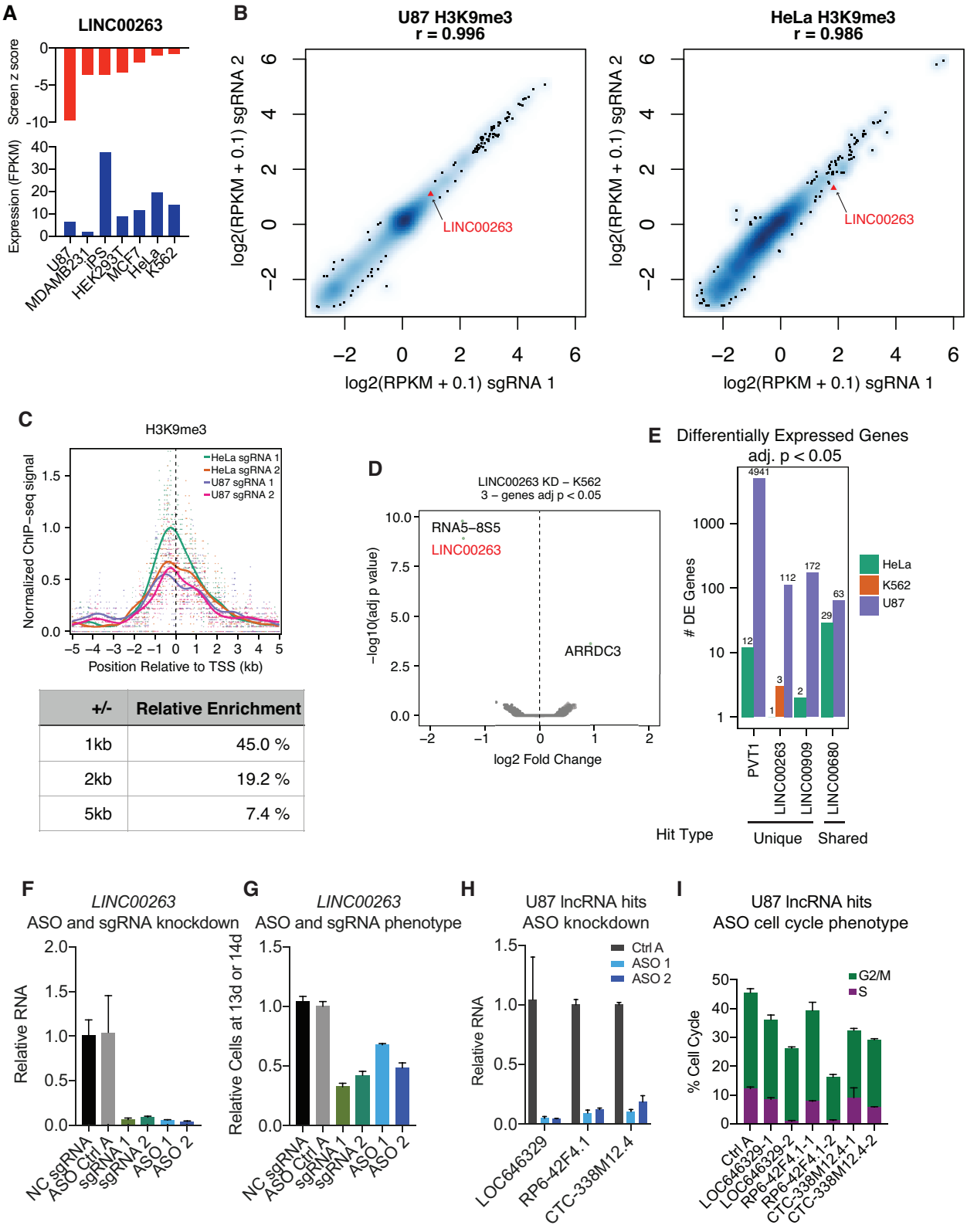

A
LINC00263
Screen z score
Expression (FPKM)
U87
MDAMB231
iPS
HEK293T
MCF7
HeLa
K562
B
U87 H3K9me3
r = 0.996
HeLa H3K9me3
r = 0.986
log2(RPKM + 0.1) sgRNA 2
log2(RPKM + 0.1) sgRNA 1
LINC00263
C
H3K9me3
HeLa sgRNA 1
HeLa sgRNA 2
U87 sgRNA 1
U87 sgRNA 2
Normalized ChIP−seq signal
Position Relative to TSS (kb)
+/-
Relative Enrichment
1kb
45.0 %
2kb
19.2 %
5kb
7.4 %
D
LINC00263 KD – K562
3 – genes adj p < 0.05
RNA5−8S5
LINC00263
ARRDC3
−log10(adj p value)
log2 Fold Change
E
Differentially Expressed Genes
adj. p < 0.05
DE Genes
HeLa
K562
U87
PVT1
LINC00263
LINC00909
LINC00680
Hit Type
Unique
Shared
F
LINC00263
ASO and sgRNA knockdown
Relative RNA
NC sgRNA
ASO Ctrl A
sgRNA 1
sgRNA 2
ASO 1
ASO 2
G
LINC00263
ASO and sgRNA phenotype
Relative Cells at 13d or 14d
H
U87 lncRNA hits
ASO knockdown
Ctrl A
ASO 1
ASO 2
Relative RNA
LOC646329
RP6-42F4.1
CTC-338M12.4
I
U87 lncRNA hits
ASO cell cycle phenotype
G2/M
S
% Cell Cycle
Ctrl A
LOC646329-1
LOC646329-2
RP6-42F4.1-1
RP6-42F4.1-2
CTC-338M12.4-1
CTC-338M12.4-2

Figure S12. Cell type-specificity of *LINC00263*.

A) Screen phenotype z scores (red) and expression values (blue) for *LINC00263* across the 7 cell types. B) Reproducibility of H3K9me3 ChIP-seq between sgRNA 1 and sgRNA 2 targeting the TSS of *LINC00263* in U87 (left) and HeLa (right) cells. C) ChIP-seq enrichment of H3K9me3 surrounding the TSS of LINC00263, comparing 2 independent sgRNAs in U87 and HeLa cells. Smoothed lines were obtained by applying a Gaussian kernel smoother against ChIP-seq coverage that had been background subtracted with coverage of H3K9me3 in cells infected with non-targeting control sgRNAs. Signal was then normalized to the peak of the highest smoothed line. Table summarizes relative enrichment of H3K9me3 at various distances beyond the TSS, obtained from the median value of the smoothed lines at each distance. D) Volcano plots for RNA-seq differential expression following infection of *LINC00263* sgRNAs compared to infection of non-targeting sgRNAs in K562 cells. E) Numbers of differentially expressed genes (DESeq2 adj $p < 0.05$) following knockdown of lncRNAs in HeLa, K562, and U87 cells. For each gene, the same sgRNAs were used across the cell types. F) qPCR comparing *LINC00263* knockdown using CRISPRi and ASO. G) Proportion of cells at 14 days post sgRNA infection, or 13 days post ASO transfection against *LINC00263*, relative to control sgRNA or control ASO, respectively. H) qPCR of ASO knockdown of additional lncRNA hits in U87 cells. I) Percentage of cells in S and G2/M phases following ASO knockdown of additional lncRNA hits in U87.

Supplementary Tables

Table S1. TSS Annotations

Table S2. CRiNCL library sgRNAs

Table S3. Growth screen sgRNA read counts and phenotypes

Table S4. Growth screen gene phenotypes and p-values

Table S5. OCT4 screen sgRNA read counts and phenotypes

Table S6. OCT4 screen gene phenotypes and p-values

Table S7. iPSC protein-coding screen sgRNA and gene phenotypes

Table S8. PVT1 tiling library sgRNAs and phenotypes

Table S9. Differential expression of genes following lncRNA CRISPRi

Table S10. Genomic Properties of lncRNAs

Table S11. Individually cloned sgRNAs and primer pairs used in this study

References

1. S. Djebali *et al.*, Landscape of transcription in human cells. *Nature*. **489**, 101–108 (2012).
2. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.*, A promoter-level mammalian expression atlas. *Nature*. **507**, 462–470 (2014).
3. I. Ulitsky, D. P. Bartel, lincRNAs: genomics, evolution, and mechanisms. *Cell*. **154**, 26–46 (2013).
4. J. L. Rinn, H. Y. Chang, Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).
5. C. P. Ponting, P. L. Oliver, W. Reik, Evolution and functions of long noncoding RNAs. *Cell*. **136**, 629–641 (2009).
6. A. R. Bassett *et al.*, Considerations when investigating lncRNA function in vivo. *eLife*. **3**, e03058 (2014).
7. M. Sauvageau *et al.*, Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife*. **2**, e01749 (2013).
8. V. H. Meller, B. P. Rattner, The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. *The EMBO Journal*. **21**, 1084–1091 (2002).
9. E. Aparicio-Prat *et al.*, DECKO: Single-oligo, dual-CRISPR deletion of genomic elements including long non-coding RNAs. *BMC Genomics*. **16**, 846 (2015).
10. T.-T. Ho *et al.*, Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. *Nucleic Acids Res.* **43**, gku1198–e17 (2014).
11. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science*. **343**, 80–84 (2014).

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

13. J. Shi *et al.*, Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat Biotechnol*. **33**, 661–667 (2015).

14. Y. Yin *et al.*, Opposing Roles for the lncRNA Haunt and Its Genomic Locus in Regulating HOXA Gene Activation during Embryonic Stem Cell Differentiation. *Cell Stem Cell*. **16**, 504–516 (2015).

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

16. A. F. Groff *et al.*, In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements. *Cell Reports*. **16**, 2178–2186 (2016).

17. S. Zhu *et al.*, Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat Biotechnol* (2016), doi:10.1038/nbt.3715.

18. M. Guttman *et al.*, lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*. **477**, 295–300 (2011).

19. N. Lin *et al.*, An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. *Molecular Cell*. **53**, 1005–1019 (2014).

20. B. Adamson, A. Smogorzewska, F. D. Sigoillot, R. W. King, S. J. Elledge, A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. *Nat Cell Biol*. **14**, 318–328 (2012).

21. Y. Zeng, B. R. Cullen, RNA interference in human cells is restricted to the cytoplasm. *RNA*. **8**, 855–860 (2002).

22. L. A. Gilbert *et al.*, CRISPR-mediated modular RNA-guided regulation of transcription in

eukaryotes. *Cell*. **154**, 442–451 (2013).

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

24. L. S. Qi *et al.*, Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*. **152**, 1173–1183 (2013).

25. H. Nishimasu *et al.*, Crystal Structure of Staphylococcus aureus Cas9. *Cell*. **162**, 1113–1126 (2015).

26. J. M. Engreitz, J. E. Haines, G. Munson, J. Chen, E. M. Perez, Neighborhood regulation by lncRNA promoters, transcription, and splicing. *bioRxiv* (2016), doi:10.1101/050948.

27. A. E. Kornienko, P. M. Guenzl, D. P. Barlow, F. M. Pauler, Gene regulation by the act of long non-coding RNA transcription. *BMC Biol.* **11**, 59 (2013).

28. U. A. Ørom *et al.*, Long noncoding RNAs with enhancer-like function in human cells. *Cell*. **143**, 46–58 (2010).

29. W. Li *et al.*, Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*. **498**, 516–520 (2013).

30. C. P. Fulco *et al.*, Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, 1–13 (2016).

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

32. A. Amabile *et al.*, Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing. *Cell*. **167**, 219–232.e14 (2016).

33. M. A. Mandegar *et al.*, CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. *Cell Stem Cell*. **18**, 541–553 (2016).

34. C. J. Braun *et al.*, Versatile in vivo regulation of tumor phenotypes by dCas9-mediated transcriptional perturbation. *Proceedings of the National Academy of Sciences*. **113**, E3892–900 (2016).

35. M. A. Horlbeck *et al.*, Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife*. **5** (2016), doi:10.7554/eLife.19760.

36. K. Takahashi *et al.*, Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Stem Cells*. **131**, 1–12 (2007).

37. M. N. Cabili *et al.*, Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*. **25**, 1915–1927 (2011).

38. M. K. Iyer *et al.*, The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics* (2015), doi:10.1038/ng.3192.

39. A. Yates *et al.*, Ensembl 2016. *Nucleic Acids Res.* **44**, D710–6 (2016).

40. T. E. P. Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. **306**, 636–640 (2004).

41. F. R. Kreitzer *et al.*, A robust method to derive functional neural crest cells from human pluripotent stem cells. *Am J Stem Cells*. **2**, 119–131 (2013).

42. S. J. Liu *et al.*, Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol*. **17**, 67 (2016).

43. M. Kampmann, M. C. Bassik, J. S. Weissman, Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. *Proceedings of the National Academy of Sciences*. **110**, E2317–26 (2013).

44. T. Hart, K. R. Brown, F. Sircoulomb, R. Rottapel, J. Moffat, Measuring error rates in

genomic perturbation screens: gold standards for human functional genomics. *Molecular Systems Biology*. **10**, 733–733 (2014).

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

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

47. T. Hart *et al.*, High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*. **163**, 1515–1526 (2015).

48. R. Andersson *et al.*, An atlas of active enhancers across human cell types and tissues. *Nature*. **507**, 455–461 (2014).

49. A. Visel, S. Minovitsky, I. Dubchak, L. A. Pennacchio, VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–92 (2007).

50. X. Yan *et al.*, Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell*. **28**, 529–540 (2015).

51. D. Hnisz *et al.*, Super-enhancers in the control of cell identity and disease. *Cell*. **155**, 934–947 (2013).

52. J. Chen *et al.*, Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol*. **17**, 19 (2016).

53. K. C. Wang *et al.*, A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*. **472**, 120–124 (2011).

54. W. Ma *et al.*, Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat Meth*. **12**, 71–78 (2015).

55. J. M. Engreitz *et al.*, The Xist lncRNA exploits three-dimensional genome architecture to

spread across the X chromosome. *Science*. **341**, 1237973–1237973 (2013).

56. A. Necsulea *et al.*, The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*. **505**, 635–640 (2014).

57. A. D. Ramos *et al.*, The long noncoding RNA Pnky regulates neuronal differentiation of embryonic and postnatal neural stem cells. *Cell Stem Cell*. **16**, 439–447 (2015).

58. M. Kretz *et al.*, Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*. **493**, 231–235 (2013).

59. T. Gutschner *et al.*, The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research*. **73**, 1180–1189 (2013).

60. R. A. Gupta *et al.*, Long non-coding RNA HOTAIR reprograms chromatinstate to promote cancer metastasis. *Nature*. **464**, 1071–1076 (2011).

61. J. A. Briggs, E. J. Wolvetang, J. S. Mattick, J. L. Rinn, G. Barry, Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution. *Neuron*. **88**, 861–877 (2015).

62. C. Trapnell *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. **28**, 511–515 (2010).

63. N. Salomonis *et al.*, Integrated Genomic Analysis of Diverse Induced Pluripotent Stem Cells from the Progenitor Cell Biology Consortium. *Stem Cell Reports*. **7**, 110–125 (2016).

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

65. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory

requirements. *Nat Meth*. **12**, 357–360 (2015).

66. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. **30**, 923–930 (2014).

67. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Genome Biol*. **11**, R106 (2010).

68. H. O'Geen, L. Echipare, P. J. Farnham, Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods Mol. Biol.* **791**, 265–286 (2011).

69. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Meth*. **9**, 357–359 (2012).

70. F. Ramírez *et al.*, deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

3/22/2017

Date