Title
Effects of translation inhibitors and compromised guide RNAs in eukaryotic cells: indirectly and directly impinging on transcription

Author
Santos, Daniel

Publication Date
2019

Peer reviewed|Thesis/dissertation

Effects of translation inhibitors and compromised guide RNAs in eukaryotic cells: indirectly and directly impinging on transcription

by
Daniel Alan Santos

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Jonathan S. Weissman, Ph.D.
A16A9953185F4C9...

Jonathan S. Weissman, Ph.D.
Chair

DocuSigned by:
Carol A. Gross, Ph.D.

Carol A. Gross, Ph.D.

DocuSigned by:
David A. Agard
328308DD7E5546A...

David A. Agard

Committee Members

Acknowledgments

My gratitude for all of the nudges, big and small, that led me to and kept me on the path toward this dissertation is deeper than words can adequately express. But here goes nothing:

First and foremost, I thank my thesis advisor, Jonathan Weissman, for taking me on as a student and providing the environment, resources, and guidance to help me grow as a scientist. It has been a true privilege to observe and learn from Jonathan's approach to scientific inquiry, both from the standpoint of uncovering biological truths, as well as advancing new technologies. On a more personal note, despite being almost comically over-booked at all times, Jonathan always managed to find time for me when it counted most. Whether helping me to navigate an abrupt transition between projects or helping me to think strategically about which skills would best prepare me for life after UCSF, Jonathan was there to offer his time and counsel, both of which have proven invaluable. Getting to know Jonathan outside of lab has been just as enjoyable, and I will sincerely miss the celebrations at Magnolia, impromptu scotch round-tables, State of the Lab addresses, late nights out at conferences abroad, and chairlift chats at Squaw.

I have also benefitted enormously from Jonathan's ability to maintain a healthy lab full of brilliant, generous, fascinating, and entertaining individuals. In particular, I give thanks to my rotation mentors Gene-Wei Li and Liz Boydston for introducing me to the lab and showing me the ropes, among other things too numerous to list here; to my Bay 4 compatriots Britt Adamson, Calvin Jan, Nicki Schirle Oakdale, Ezgi Hacisuleyman, Manu Leonetti, Han Li, and Albert Xu for fellowship and taking turns hitting the -80° alarm mute button; to Alex Fields for convincing me to learn Python and introducing me to pain au chocolat; to Jeff Hussmann, Michelle Chan, Jeff Quinn, Joseph Replogle, and Tom Norman for periodic book club get-togethers and/or Spikeball

tournaments; to Kamena Kostova and Sandra Torres for science, dumplings, and turtle maintenance; to Marco Hein for all the lunchtime chats; to Katerina Popova for being a wonderful rotation student, and ultimately, a wonderful bay mate; to Kelsey Hickey and Matt Shurtleff for always knowing precisely when a beer is sorely needed, and crucially, where to find a cold one; to James Nuñez for helping me with flow cytometry emergencies no fewer than 350 times; to Jin Chen and Matt Jones for teaching me about learning; to all of the great technicians over the years, especially Christina Liem and Zach Cogan for their help with tissue culture; to Ben Tu, a JSW lab alum, for fruitful collaborations; to Manny DeVera for being a more efficient version of MacGuyver; to Christopher Reiger and Joan Kanter for administrative support and colorful commentary; and to my other fellow graduate students Josh Dunn, Chris Williams, Edwin Rodriguez, Yi-Chang Liu, Meghan Zubradt, and Max Horlbeck for lighting the way. I especially want to thank Marco Jost for graciously allowing me to work with him on the allelic series project (along with all of the coaching required to get me off of YPD and on to RPMI), and Reuben Saunders, both for being a comrade on said project and for filling the shoes vacated by Britt as my next-door neighbor in lab—no easy task.

Part of the recipe that has made the Weissman lab such a special place for me is the uniquely collegial, culturally laid-back yet scientifically intense environment at UCSF. I am truly grateful to the other members of my thesis committee, Carol Gross and David Agard, for their unwavering support and guidance; to the instructors of the Tetrad program curriculum; to Rachel Mozesson, Toni Hurley, Danny Dam, and Billy Luh for Tetrad administrative support; and to Holly Ingraham for teaching me to write a decent grant application. I also want to thank the labs of Joe DeRisi and Hana El-Samad for being fantastic neighbors; Eric Chow and Derek Bogdanoff for

sequencing support at the UCSF Center for Advanced Technology; and Kari Herrington for microscopy support at the UCSF Nikon Imaging center. I am thankful for my classmates' companionship (and occasional commiseration) throughout the years, particularly Leeanne Goodrich, Melanie Silvis, Anne Pipathsouk, Han Tran, Alex Long, Kevin Colón, Diana Summers, Frances Hundley, and Jordan Tsai.

I wouldn't be writing this if not for all of the positive influences I've had throughout my early life. I thank my biology teacher Roy Benavidez for kick-starting my fascination for molecular biology by making DNA visible to the naked eye; and Asim Dasgupta for being an engaging virology professor and providing my first exposure to bench research in his lab. I thank all of my former colleagues at Favrille and Sapphire Energy for keeping me excited about a career in science (may both organizations rest in peace), particularly Chris Yohn and Yan Poon, whose mentorship and friendship continue to this day.

It's not common to abandon steady employment in one's late-20s to return to 5-to-? years of school, but the love and encouragement from friends and family allowed me to trust myself in following my gut. After all these years, friendships stretching back to Lake Tahoe, through UCLA, and down to San Diego endure, and keep me sane. (You all know who you are). My parents, Bob and Verlie Santos, always did everything they could to emphasize academics and support me in all of my pursuits even when it was not easy to do so, and for that I am eternally grateful. My brother Phil was my earliest partner in crime, and the afternoons we spent digging up bugs at Ponderosa Park no doubt sparked my curiosity about the natural world. These days I value our debates about science, politics, and life in general, which keeps our relationship closer and my thinking sharper.

Finally, none of this would have been possible without the continual love, support, and encouragement from my partner. Serena Ngo, you challenge and inspire me, and I love you more than you know.

Chapter 2 is reprinted largely as it appears in:

Santos,D.A., Shi,L., Tu,B.P. and Weissman,J.S. (2019) Cycloheximide can distort measurements of mRNA levels and translation efficiency. *Nucleic Acids Research*, **47**, 10.1093/nar/gkz205.

DAS and JSW primarily conceived, designed, and interpreted the experiments, and wrote the manuscript. LS harvested all samples for the YMC experiment. BPT assisted with design and supervised execution of the YMC experiment.

Chapter 3 is reprinted as a manuscript in preparation:

Jost,M.*, Santos,D.A.*, Saunders,R.A., Horlbeck,M.A., Hawkins,J.S., Scaria,S.M., Norman,T.M., Hussmann,J.A., Liem,C.R., Gross,C.A., and Weissman,J.S. Titrating gene expression with allelic series of CRISPR guide RNAs.

MJ conducted the large-scale growth screen, supervised the constant region and perturb-seq experiments, implemented the linear machine learning model, analyzed data, conceived experiments, and wrote the manuscript. DAS conducted the GFP and constant region screens, implemented the deep learning model, designed and conducted the compact library screen, analyzed data, conceived experiments, and wrote the manuscript. RAS designed the constant region library and conducted a pilot screen, designed and conducted the perturb-seq experiment, analyzed data, conceived experiments, and edited the manuscript. MAH assisted with the large-scale growth screen and with JSH designed the large-scale library. SMS evaluated modified constant region activities by RT-qPCR. JAH and TMN assisted with data analysis. CRL assisted with library cloning and screens. CAG supervised the generation of the large-scale library. JSW conceived and supervised experiments, and wrote the manuscript.

* MJ and DAS contributed equally to this work.

Effects of translation inhibitors and compromised guide RNAs in eukaryotic cells: indirectly and directly impinging on transcription

Daniel A. Santos

Abstract

The flow of genetic information through transcription and translation—known as the central dogma of biology—enables all aspects of life, from the response of a solitary yeast cell to an environmental stimulus, to the intricate choreography guiding development of a single-celled zygote into a complex organism with functionally distinct tissues. Recent technological advances have provided tools to observe and/or perturb molecular processes underlying the central dogma with unprecedented resolution and precision. In this dissertation, I describe two cases where such tools were used to study and manipulate gene expression in eukaryotic cells.

First, I show that in budding yeast, under nutrient limiting conditions, the commonly used translation inhibitor cycloheximide induces rapid transcriptional upregulation of hundreds of genes involved in ribosome biogenesis. This generates a large pool of mRNAs that cannot be translated due to the presence of the inhibitor, leading to the appearance of strong translational regulation. This work provides a novel mechanistic interpretation for perplexing reports in the translation field, and hopefully serves to guide experimental design moving forward.

Second, I describe the development of allelic series of systematically compromised sgRNAs to titrate expression of human genes with CRISPR interference. Large-scale measurements of compromised sgRNA activities enable identification of empirically validated intermediate-activity sgRNAs and determination of the factors governing sgRNA activity using deep learning, facilitating

construction of a compact sgRNA library to titrate expression of ~2,400 essential genes and a genome-wide *in silico* library. Staging cells along a continuum of essential gene expression levels using sgRNA series combined with rich single-cell RNA-seq readout reveals expression threshold-specific responses and gene-specific expression-to-phenotype relationships, thus highlighting such reagents as a general tool to titrate the expression of human genes, with potential applications ranging from tuning of biochemical pathways to identification of suppressors for diseases of dysregulated gene expression.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

During my time in graduate school, when someone would ask me what our lab works on, it was always difficult to offer a succinct response. While this partly stemmed from my tendency toward long-windedness, the reality is that the Weissman lab is continually evolving and expanding into new scientific frontiers. I like the way Jonathan framed this once: we are both explorers and engineers; we can see that there are interesting canyons in the distance, but in order to get there we first need to figure out how to build a suspension bridge. Although the lab has never strayed too far from a core interest in protein homeostasis, this explorer philosophy has taken the lab through various stages, shining spotlights in different corners of biology. I joined the lab during a transition of sorts, a few years after the development of ribosome profiling for monitoring translation in cells (1, 2), and right as the potential for CRISPR-based systems to precisely edit and perturb mammalian genomes was more fully being realized (3). This timing turned out to be fortuitous, ultimately allowing me to work on two distinct projects in areas spanning a decade of the lab's interests.

The first project was spawned from my rotation, when I attempted to leverage the principles uncovered by Gene-Wei Li *et al*. around proportional synthesis of multi-subunit complexes (4) to discover novel protein-protein interactions, and perhaps more interestingly, novel "moonlighting" functions of super-stoichiometric proteins. This effort entailed a great deal of ribosome profiling to determine protein synthesis rates, and for a subset of those samples we included RNA-seq to enable calculation of translation efficiencies (or, the mean number of ribosomes per mRNA molecule for a given gene, in a given condition). We were surprised to discover that for hundreds of functionally related genes, translation efficiencies appeared to be dynamically changing by 10-fold or more throughout several growth conditions including the yeast metabolic cycle (5), representing a potentially pervasive and unusual regulatory strategy. After spending much time attempting to build

a complicated translation reporter system to uncover the mechanism of this regulation, I realized that the degree to which translation appeared to be repressed was a function of the translation inhibitor concentration used to treat yeast cells, and follow-up experiments clearly showed how the apparent translational repression was merely a transcriptional artifact of using this particular drug. The disappointment of losing the trail of this illusory regulatory mechanism, however, quickly gave way to the realization that I now had an opportunity to write a cautionary tale that would be important for the translation community (6), as well as freedom to work on new things.

As it happens, while my translation project was winding down, Marco Jost had a new project (or two) ramping up, and consequently less time to devote to his ongoing exploration of the effects of systematically compromised single guide RNAs (sgRNAs) in CRISPRi (7) systems. Our lab and others previously established that mismatches in sgRNAs have differential effects on Cas9 (or its nuclease-dead counterpart, dCas9) activity (8–11), and Marco had set out both to learn the relationships between specific mismatches and their resulting activities, and to utilize this knowledge to build a tool enabling precise tuning of gene expression levels. When I came on board, Marco had laid the groundwork for using computational approaches to predict how much a given mutation would compromise sgRNA function, and this served as my first foray into the worlds of machine learning and mammalian cell biology. The CRISPR screening approaches we used in this project had already become bread-and-butter methodologies for Weissman lab, and I was lucky to learn them in-house. Beyond providing an opportunity to gain technical know-how, the collaborative and intellectual aspects of this project were immensely gratifying, especially considering the potential applications this tool has for basic biological as well as biomedical studies. I look forward to seeing where this bridge takes us.

References

1. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R.S. and Weissman,J.S. (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, **324**, 218–223.

2. Ingolia,N.T., Lareau,L.F. and Weissman,J.S. (2011) Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell*, **147**, 789–802.

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

4. Li,G.-W., Burkhardt,D., Gross,C. and Weissman,J.S. (2014) Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell*, **157**, 624–635.

5. Tu,B.P., Kudlicki,A., Rowicka,M. and McKnight,S.L. (2005) Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes. *Science*, **310**, 1152–1158.

6. Santos,D.A., Shi,L., Tu,B.P. and Weissman,J.S. Cycloheximide can distort measurements of mRNA levels and translation efficiency. *Nucleic Acids Res*, **47**, 10.1093/nar/gkz205.

7. 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.

8. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, **337**, 816–821.

9. Sternberg,S.H., Redding,S., Jinek,M., Greene,E.C. and Doudna,J.A. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, **507**, 62–67.

10. Szczelkun,M.D., Tikhomirova,M.S., Sinkunas,T., Gasiunas,G., Karvelis,T., Pschera,P., Siksnys,V. and Seidel,R. (2014) Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *PNAS*, **111**, 9798–9803.

11. 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. *Cell*, **159**, 647–661.

12. Nishimasu,H., Ran,F.A., Hsu,P.D., Konermann,S., Shehata,S.I., Dohmae,N., Ishitani,R., Zhang,F. and Nureki,O. (2014) Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell*, **156**, 935–949.

Chapter 2

Cycloheximide can distort measurements of mRNA levels and translation efficiency

Introduction

The development of ribosome profiling has broadly enabled genome-wide analyses of active translation *in vivo* (1). The technique is based on deep sequencing of ribosome-protected mRNA fragments, or ribosome footprints, providing a quantitative snapshot of ribosome positions along mRNAs at single nucleotide resolution. When combined with traditional RNA-seq, it is possible to determine the ratio of ribosome footprints to the number of mRNA molecules for a given gene, thereby providing a measure of translation efficiency (TE) for that gene. In contrast to absolute measures of translation rates provided by ribosome profiling, TE measurements require accurate quantitation of both translation rates and mRNA abundances. TE is partially governed by intrinsic features of an mRNA such as its sequence and structure (2), but TE can also vary dynamically as a regulatory strategy. Two notable examples in budding yeast are the transcription factors Hac1 and Gcn4, whose mRNAs rapidly transition from low to high TE states when cells encounter specific stresses. Interestingly, the mechanisms driving the TE switch for these factors are completely distinct: *HAC1* mRNA is spliced by the ER-resident protein Ire1 to remove inhibitory secondary structures during the unfolded protein response (3), and *GCN4* utilizes a series of short upstream open reading frames (uORFs) in its 5′ untranslated region (UTR) that sequester ribosomes away from the main ORF in the absence of stress-induced eIF2α phosphorylation (4).

While these well-studied examples highlight the complexities of translational regulation, other instances of dynamic TE switching remain poorly understood. We initially set out to identify and characterize TE regulation during the yeast metabolic cycle (YMC), a process involving synchronized growth with well-established, coordinated gene expression changes. The YMC is initiated by culturing *Saccharomyces cerevisiae* in a chemostat with limiting glucose to control growth

rate and maintain constant culture density. Under these conditions, cells grow and divide in sync, and a redox cycle is established with periodic bursts of respiration resulting in abrupt decreases in dissolved oxygen (5). Gene expression profiling has identified three distinct phases of gene expression as part of this cycle—termed the Reductive Building (RB), Reductive Charging (RC), and Oxidative (OX) phases—with greater than 50% of the overall transcriptome exhibiting variable expression (6).

Using ribosome profiling and RNA-seq, we observed that the majority of genes involved in ribosome biogenesis (commonly referred to as *ribi* genes (7)) appear to dynamically shift their TE state between the RC and OX phases of the metabolic cycle. Moreover, analysis of previous ribosome profiling studies suggests similarly large TE changes for ribi genes during amino acid starvation (8) and meiosis (9), initially suggesting a shared mechanism. However, these experiments were all conducted by treating cells with the translation elongation inhibitor cycloheximide (CHX) prior to harvesting, and upon repeating the amino acid starvation experiment without CHX, we show that TEs of ribi genes remain unchanged. We additionally demonstrate that CHX causes rapid accumulation of ribi transcripts in a TORC1-dependent manner, and the magnitude of the response to CHX depends on strain genotype and the choice of drug vehicle. Thus, CHX distorts measures of TE by causing rapid transcriptional changes which creates a new pool of untranslated mRNAs. These results underscore the caution that needs to be taken when using CHX as an experimental tool.

Materials and Methods

Yeast strains and media

For the Yeast Metabolic Cycle, cells were continuously cultured in a minimal medium consisting of 5 g/l $(NH_4)_2SO_4$, 2 g/l KH_2PO_4, 0.5 g/l $MgSO_4 \cdot 7H_2O$, 0.1 g/l $CaCl_2 \cdot 2H_2O$, 0.02 g/l $FeSO_4 \cdot 7H_2O$, 0.01 g/l $ZnSO_4 \cdot 7H_2O$, 0.005 g/l $CuSO_4 \cdot 5H_2O$, 0.001 g/l $MnCl_2 \cdot 4H_2O$, 1 g/l yeast extract, 10 g/l glucose, 0.5 ml/l 70% (vol/vol) H_2SO_4, and 0.5 ml/l Antifoam 204 (Sigma). For batch growth experiments, cells were grown in Synthetic Defined (SD) medium (Difco yeast nitrogen base and 2% glucose supplemented with amino acids [RDHILKMFTWYV], uracil, and adenine; or without supplementation for starvation experiments). The prototrophic *S. cerevisiae* strain CEN.PK was used for the YMC experiment. Unless otherwise stated, all other experiments utilized the strain BY4741, or derivatives thereof. Genomic knock-ins and knock-outs were generated using standard techniques (10). The *NOP2* coding sequence was replaced with the homologous sequence from *Kluyveromyces lactis* using the *URA3* pop-in/pop-out method, resulting in a marker-less strain with endogenous *NOP2* regulatory sequences. Plasmids carrying the promoter-gene hybrid reporters were generated using Gibson assembly (11), and the resulting expression cassettes were PCR-amplified and inserted at the dispensable *YHRCdelta14* LTR locus (12). Ribi transcription factors were C-terminally tagged with EGFP, and histone H2B was C-terminally tagged with mRuby2 (13). Strains, plasmids, and oligonucleotides used in this study are listed in Supplementary Tables S2.1-S2.3.

Yeast Metabolic Cycle

A continuous culture of CEN.PK was established in a chemostat as previously described (14). Metabolic cycles were monitored using a dO_2 probe, and 16 time points were chosen spanning a single cycle. At each time point, 20 ml of culture was mixed with 40 µl of a 50 mg/ml cycloheximide stock (in 100% ethanol), for a 100 µg/ml final concentration. The culture was shaken for 2 min, then centrifuged for 2 min at 4,000 × *g*. The cell pellet was then re-suspended in ice-cold lysis buffer (20 mM Tris pH 8, 140 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 100 µg/ml cycloheximide, 1% Triton X-100, and 0.025 U/µl Turbo DNase) and dripped into liquid nitrogen (ℓN_2). Frozen droplets were stored at -80 °C until further processing

Amino acid starvation

Cells were grown overnight at 30 °C to saturation in SD, and then diluted to OD_{600} < 0.1 in fresh SD medium. Cultures were incubated with vigorous shaking until the OD reached 0.4-0.6, and then half of the culture was centrifuged for 3 min at 3200 × *g*. The pellet was re-suspended in prewarmed SD medium lacking amino acids, and cells were shaken for an additional 20 min. CHX was then added at a final concentration of 100 µg/ml (for the matched replete sample, CHX was added to the remaining half of the culture that was not starved), and cultures were shaken for an additional 2 min before centrifuging for 3 min at 3200 × *g*. Pellets were re-suspended in ice-cold lysis buffer and frozen drop-wise in ℓN_2 as described for the YMC.

For experiments without CHX pretreatment, 20 min-starved (or replete) cultures were transferred to a vacuum filtration apparatus and cells were collected on a 0.45 µm cellulose nitrate

membrane (Whatman). Cells were then quickly scraped from the membrane with a metal spatula and immediately plunged into ℓN_2. Frozen cells were stored at -80 °C until further processing.

Ribosome profiling and RNA-seq

Cells plus lysis buffer were cryogenically pulverized in a SPEX 6870 Freezer/Mill for 1 min at 15 cycles per min (for vacuum-filtered samples, frozen droplets of lysis buffer were added to the cells). The lysate powder was thawed and immediately clarified by two sequential centrifugation steps at 4 °C: first for 5 min at 3,000 × *g*, and then for 10 min at 20,000 × *g*.

Ribosome footprinting and library generation were carried out essentially as described in (15). RNA fragments from ~26-34 nt were selected following RNase I digestion and PAGE separation. Barcode sequences were included on 3′ cloning linkers, and samples with unique barcodes were pooled together post-ligation when possible. A dual rRNA depletion strategy was employed, first with Ribo-Zero Gold for Yeast (Illumina), and then with biotinylated antisense oligos against rRNA species that co-migrate with ribosome footprints as described in (9).

For RNA-seq, RNA was first purified from clarified lysates using TRIzol (Invitrogen) and then rRNA was removed using Ribo-Zero Gold for Yeast. rRNA-depleted RNA was used to generate TruSeq Stranded libraries (Illumina) per the manufacturer's protocol. Ribosome profiling and RNA-seq libraries were sequenced on an Illumina HiSeq 4000 in single read 50-base mode. Each set of matched ribosome profiling and RNA-seq data is derived from a single biological sample.

Sequencing Data Analysis

For ribosome profiling libraries generated in this study, linker sequences were removed from sequencing reads and samples were de-multiplexed using FASTX-clipper and -barcode splitter, respectively (http://hannonlab.cshl.edu/fastx_toolkit/). Unique molecular identifiers and sample barcodes were then removed from reads using a custom Python script. Bowtie v1.1.2 (http://bowtie-bio.sourceforge.net/) was used to filter out reads aligning to rRNAs and tRNAs, and all surviving reads were aligned to the *S. cerevisiae* genome using tophat v2.1.1 (https://ccb.jhu.edu/software/tophat/). Counts per gene and normalized counts per gene (in reads per kilobase per million mapped reads, or RPKM) were calculated using the plastid cs program (16), with counts assigned to ribosomal P-sites determined by plastid's psite program. Regions of the yeast genome that could not be uniquely mapped from a 26-base read with 2 mismatches were identified using plastid's crossmap program, and these regions, along with the first 30 and last 5 codons of each coding sequence (CDS), were masked from RPKM calculations (for some analyses, cytoplasmic ribosomal protein abundances were separately quantified without crossmap masks, since most of these genes have nearly identical paralogs in *S. cerevisiae*). RNA-seq data were processed in the same manner, except reads first had to be reverse-complemented due to the TruSeq Stranded chemistry, and counts were assigned to the 5′-most aligned base. For viewing alignments, wiggle files were generated from genome alignments using plastid's make_wiggle program, and data were visualized in the IGV browser (http://software.broadinstitute.org/software/igv/).

For genes to be included in downstream analyses, they were required to have at least 128 mRNA counts (or 32 mRNA counts for the meiosis data from (9)) and at least 1 footprint count in the CDS. Genes listed as dubious ORFs in the Saccharomyces Genome Database (SGD,

https://www.yeastgenome.org/) were not considered for analysis. Translation efficiencies were calculated for each gene by dividing the CDS footprint RPKM by the CDS mRNA RPKM, and fold changes in mRNA abundance and TE were normalized to set the median fold change of all genes to 1. Ribosome biogenesis factors were defined as non-ribosomal proteins annotated as "ribosomal small subunit biogenesis" or "ribosomal large subunit biogenesis" in the SGD Go Slim database. Data were analyzed and plotted using Pandas, Matplotlib, and Seaborn python libraries.

RT-qPCR

Amino acid starvation was carried out as described for ribosome profiling libraries, except cells were mock-treated with vehicle (ethanol or DMSO) in the no-CHX samples. Cells were harvested prior to adding CHX, and then 3, 6, and 9 min post-CHX by centrifuging 1 ml for 20 s at 11,000 × *g*. After aspirating the supernatant, cell pellets were immediately frozen in ℓN_2, and RNA was purified by phenol-chloroform extraction and ethanol precipitation (17). Residual genomic DNA was degraded using a TURBO DNA-free kit (Ambion). Oligo(dT) primers were annealed to mRNA, and cDNA was generated using AMV reverse transcriptase (Promega). *NOP2* and *ACT1* (nucleolar protein 2, ribi gene; actin, housekeeping reference) cDNAs were quantified on a Roche LightCycler 480 instrument using GoTaq (Promega) PCR reactions containing SYBR Green (Life Technologies), and relative abundances were calculated using the $2^{-\Delta\Delta CT}$ method (18). The starvation plus rapamycin experiment was conducted as described above, except cells were incubated for 20 min in starvation medium containing 200 nM rapamycin (Sigma) or vehicle (DMSO) prior to adding CHX. All qPCR experiments were conducted with technical triplicates.

Time Lapse Fluorescence Microscopy

Cells containing RFP-tagged histones and GFP-tagged ribi transcription factors were grown to OD_{600} 0.5-0.7 in SD medium and loaded into a CellASIC ONYX multi-chamber microfluidic plate for haploid yeast (Millipore). The growth chamber was perfused with SD medium at 10.8 kPa for 15 min, and then the media source was switched to SD lacking amino acids. Following 15 min of starvation the media was again switched to SD lacking amino acids plus 100 μg/ml CHX. Variations of the above scheme were also used to assess (i) the response to 200 nM rapamycin in the starvation medium, (ii) the response when CHX is not added after 20 min of starvation, and (iii) the baseline response to switching between replete media sources. For each strain/condition, two fields of view were imaged every 5 min on an inverted Nikon Ti microscope with a 40x objective (brightfield, GFP, and mCherry channels). The plate was maintained at 30 °C during image acquisition. For rapamycin experiments without microfluidics, cells were transferred to glass-bottomed wells coated with Concanavalin A and allowed to settle for 5 min. Wells were then washed twice with the appropriate medium to remove non-adhered cells, and images were acquired as described above.

Image Analysis

The ratio of nuclear to cytoplasmic GFP-tagged transcription factors was determined using CellProfiler 3.1.5 (http://cellprofiler.org/). Nuclei were initially detected based on HTB2-mRuby2 signal, and then cell boundaries surrounding nuclei were defined based on GFP signal using the Propagation method with Robust Background thresholding. Cytoplasm was defined as the region of the cell not occupied by the nucleus. For each cell, the mean GFP intensities in the nucleus and

cytoplasm were normalized to the mean GFP intensity of the entire cell. Nuclear localization was calculated as the normalized nuclear intensity divided by the normalized cytoplasmic intensity. A minimum of 200 cells were analyzed per strain/condition in the microfluidic experiments; 28-112 cells in glass-bottom wells.

Results

Ribi transcripts are under apparent dynamic translational regulation during the YMC, meiosis, and amino acid starvation

To obtain a global view of translational regulation during the different phases of the YMC, we performed ribosome profiling and RNA-seq at multiple YMC time points. Cells were grown in a chemostat, removed at defined time points, immediately treated with CHX to arrest translation, and subsequently processed to generate ribosome profiling and mRNA libraries suitable for analysis by next generation sequencing (Figure 2.1A). In total, 16 time points spanning a single ~4.5-hour cycle were analyzed (Figure 2.1B, Supplementary Figure S2.1A). At each time point, TEs were calculated for each protein-coding gene by dividing the gene's normalized ribosome profiling counts by its normalized RNA-seq counts (footprint RPKM / mRNA RPKM). We then quantified the range of TEs exhibited over the span of the YMC for each gene by dividing its minimum TE in the time course by its maximum TE. By this metric the median TE change for all genes during the YMC is 2.2-fold, with genes in the 99th percentile changing TE by at least 20-fold (Figure 2.1C, darker shade).

The same analysis was applied to previously published ribosome profiling and RNA-seq data acquired along a meiosis time course in yeast (9). In this context the median change in TE is 3.9-fold (Figure 2.1C, lighter shade); however, both distributions have long tails extending into the range of 100-fold or more. To investigate whether any of the strongly translationally regulated genes are shared between these distinct biological processes, we selected genes with TE change of 8-fold or greater from each experiment and analyzed the overlap of these gene sets. Remarkably, nearly half of the genes under strong translational regulation in the YMC are also strongly regulated in meiosis (Supplementary Figure S2.1B). Moreover, the overlapping gene set is highly enriched for ribi factors (p=6×10^{-24}) supporting the possibility of a common regulatory mechanism.

In addition to displaying a wide range of TEs in the YMC and meiotic time courses, the TEs of individual ribi genes are also highly correlated throughout each time course (Supplementary Figures S2.1C-D). By comparing TE and mRNA levels between high and low ribi gene translation states, a similar pattern emerges for both time courses in which decreased TE coincides with decreased mRNA content for the ribi genes (Figure 2.1D-E). Moreover, a previously published ribosome profiling study showed similar decreases in TEs and mRNA levels of ribi genes during amino acid starvation (8). Taken at face value, these observations suggest that a large group of genes involved in the energetically demanding ribi pathways are coordinately regulated in their mRNA abundances and translation efficiencies.

Observed decrease in ribi TE during amino acid starvation is CHX-dependent

To understand the mechanistic basis of this effect we chose to focus on the amino acid starvation because, of the three transitions where we observed coordinated changes in the measured

TE of ribi genes, this is the most experimentally facile. We were able to replicate the apparent change in TE during amino acid starvation using a slightly modified protocol in order to maintain consistency with the YMC harvesting scheme (Figure 2.2A, top). Cells were switched from minimal glucose medium with amino acids to minimal glucose medium without amino acids, and after 20 minutes of amino acid starvation we observed marked reductions in TEs and mRNA abundances of ribi genes as well as a pronounced increase in GCN4 TE (Figure 2.2B), which is a hallmark of starvation-induced eIF2α phosphorylation.

Up to this point all of the experiments showing changes in ribi gene TEs were conducted with CHX pretreatment, which has been well documented to impact ribosome profiling measurements by causing accumulation of ribosomes near translation start sites (8, 19), and skewing the distribution of ribosome positions in a codon-dependent manner (20, 21). However, since we took appropriate precautions such as masking footprint reads from the beginning of ORFs, such effects should have a minimal impact on the measure of the average ribosome density used to determine the overall rate of translation of a message; a key component of TE measurements. Nonetheless, we wanted to ensure that the TE measurements were not affected by CHX. We therefore used an alternative harvesting protocol that does not require CHX treatment, but instead relies on rapid filtration and freezing to arrest translation (Figure 2.2A, bottom). Ribosome profiling and RNA-seq were then carried out using the standard protocol.

Remarkably, we found that the change in TEs was essentially abolished when cells were not treated with CHX (Figure 2.2C). Only two ribi genes, *JIP5* and *RIO1*, still exhibited large decreases in TE. *JIP5*'s proximity to an upstream gene makes it difficult to assess whether the transcript architecture changes in response to starvation; on the other hand, it is clear that the *RIO1* transcript

is significantly extended on the 5′ end in starved cells, which incorporates a uORF that appears to sequester ribosomes away from the canonical ORF (Supplementary Figure S2.2A). This "long undecoded transcript isoform", or LUTI, represents a pervasive regulatory mechanism that was recently described in meiosis (22–24).

In order to determine how CHX pretreatment might lead to an apparent low TE for the ribi genes, we first compared normalized counts per gene in starvation experiments with and without CHX (Figure 2.2D, Supplementary Figure S2.2B). For the vast majority of non-ribi genes, CHX treatment did not have a substantial effect on measured gene expression in any condition. However, the ribi genes experienced significant CHX-dependent changes in both footprint and mRNA abundances, and this effect was greatly exaggerated in starved cells. Somewhat paradoxically, pretreatment with CHX leads to increased mRNA and decreased footprint counts for ribi transcripts, which conspire to make TE measurements much lower in starvation experiments that include a CHX pretreatment step.

CHX causes rapid induction of ribi transcripts in starved cells

A simple explanation for the difference in mRNA abundances with and without CHX pretreatment would be that CHX induces transcription of ribi genes. To test this possibility, we treated starved cultures with CHX or vehicle and measured the mRNA abundance of a typical ribi gene, *NOP2*, using RT-qPCR over a 9-minute time course. Surprisingly, *NOP2* mRNA abundance increased in both the CHX- and vehicle-treated samples, although the magnitude of the increase was much larger with CHX (Figure 2.3A, orange lines). In this case the vehicle was ethanol, which is the drug manufacturer's recommended solvent. The working solution was made at a 500X

concentration, therefore the quantity of added ethanol in the culture after treatment was 0.2% by volume. We next tested whether the response could be replicated with larger or smaller amounts of ethanol alone. Indeed, treatments of 0.4% and 0.1% ethanol proportionally scaled the *NOP2* mRNA increase relative to 0.2% ethanol (Supplementary Figure S2.3A).

Since CHX is also routinely dissolved in dimethyl sulfoxide (DMSO), we repeated the same experiment with CHX in DMSO or with DMSO alone. While *NOP2* mRNA abundance did increase following several minutes of drug treatment, the magnitude of the increase was 3- to 4-fold smaller compared to treatment with CHX in ethanol (Figure 2.3A, purple lines). This discrepancy does not appear to be due to gross differences in the activity of CHX in each solvent, since both formulations arrested growth with equal potency (Supplementary Figure S2.3B). Interestingly, the response to ethanol is strain-specific, as a prototrophic strain treated with CHX in ethanol exhibits a similar *NOP2* mRNA increase as our auxotrophic lab strain treated with CHX in DMSO (Supplementary Figure S2.3C). Therefore, in yeast starvation experiments that include CHX pretreatment—especially if the strain is auxotrophic and CHX is dissolved in ethanol—cells accumulate ribi transcripts from the time CHX is added to the culture until cells are frozen in liquid nitrogen.

We reasoned that if the apparent TE decrease in starved cells is due to transcription of new ribi messages following CHX treatment, then a ribi promoter should be necessary and sufficient for the effect. Since most ribi genes are essential in yeast, we started by replacing the endogenous *NOP2* coding sequence with a homologous sequence from *K. lactis* which contains enough nucleotide differences to be readily distinguished from the *S. cerevisiae* sequence using deep sequencing. We then introduced different *NOP2* reporter constructs elsewhere in the genome, allowing us to

unambiguously quantify their expression via ribosome profiling and RNA-seq, and thus determine the influence of promoters, UTRs, and coding sequences on TE (Figure 2.3B, Supplementary Figure S2.3D). As expected, with CHX pretreatment, a fully wild-type *NOP2* gene exhibited decreases in TE and mRNA levels following amino acid starvation (Figure 2.3C). However, when *NOP2* was instead transcribed from the non-ribi *CCW12* promoter, the decrease in mRNA abundance was severely attenuated and the change in TE was eliminated entirely. Finally, an exogenous GFP sequence flanked by non-ribi *ADH1* UTRs transcribed from the *NOP2* promoter exhibited decreases in mRNA and TE similar to those observed for wild-type *NOP2*. Collectively, these reporter experiments demonstrate that a ribi promoter is both necessary and sufficient to recapitulate the low TEs observed in amino acid-starved cells when CHX pretreatment is used.

CHX-induced ribi transcription requires TORC1 signaling

We next sought to identify the factors mediating the transcriptional response of ribi genes to CHX. The Target Of Rapamycin (TOR) protein kinase emerged as a candidate given its central role in regulating ribosome biogenesis in response to nutrient availability (25). In budding yeast TOR is a member of two protein complexes, TORC1 and TORC2, the former of which promotes growth when conditions are favorable, and is inactivated by the macrolide antibiotic rapamycin (26). Upon nutrient starvation or rapamycin treatment, reduced TORC1 signaling leads to inhibition of rRNA and ribosomal protein gene transcription, as well as global attenuation of translation initiation (27). Since ribi gene transcription was reactivated in the presence of CHX despite poor nutrient conditions, we speculated that the TORC1 pathway might be involved. To test this possibility, we measured changes in *NOP2* mRNA abundance induced by CHX in the presence and absence of

rapamycin. After shifting cells to starvation medium containing rapamycin or vehicle and incubating for 20 minutes, CHX was added and *NOP2* mRNA abundance was monitored over time by qPCR. *NOP2* transcript levels increased by nearly 8-fold in vehicle-treated cells but remained unchanged in rapamycin-treated cells (Figure 2.4A), demonstrating that TORC1 signaling is required for CHX-mediated ribi transcription.

Expression of ribi genes is specifically regulated by at least three transcription factors—Dot6, Tod6, and Stb3—which bind two conserved sequence motifs in ribi promoters to repress transcription (28). These factors are phosphorylated by the kinase Sch9, itself a direct target of TORC1 (29), which is thought to mediate their nuclear localization and/or chromatin binding affinity (30). To assess whether CHX treatment affects the sub-cellular localization of these factors, we C-terminally tagged each with EGFP and monitored their localization in live cells. Using a microfluidic device, we were able to rapidly switch between different media and image the same cells over time. In replete medium, all three transcription factors are evenly distributed throughout cells, and within 5 minutes of amino acid starvation they accumulate in the nucleus. Strikingly, within 5 minutes of treating starved cells with CHX, all three transcription factors begin to exit the nucleus (Figure 2.4B-C), and after 15 minutes nuclear localization is reversed by 35-50% (Tod6, Dot6) or up to 90% (Stb3). Rapamycin did not prevent these factors from leaving the nucleus upon CHX exposure using this experimental setup (Supplementary Figure S2.4A), however it is well-documented that small molecules can be absorbed by the polymer that constitutes the channels of this microfluidic device (31–33). After repeating the rapamycin/CHX treatment regimen in a flask and imaging in glass-bottom wells, we observed significantly more nuclear localization post-CHX for all three transcription factors in rapamycin-treated cells (Supplementary Figure S2.4B). Taken

together, these data suggest that CHX works through TORC1 and Sch9 to relieve transcriptional repression of ribi genes, despite poor nutrient availability.

Discussion

Ribosome profiling provides an unprecedented view into the absolute rate of translation, and paired with mRNA measurements, translation efficiency. However, such measurements depend on the ability to accurately freeze the translational state of a cell which can be challenging due to the rapid kinetics of translation initiation and elongation. To overcome this challenge, translation elongation inhibitors such as CHX are often used to "lock" ribosomes in place prior to harvesting. Such treatment comes with well-described caveats such as accumulation of ribosome density near translation start sites, and "smearing" of ribosome density in gene bodies. However, with proper precautions, these effects typically have minimal impact on the overall rate of translation as measured by the average ribosome density across the body of an mRNA. The present work adds the prospect of rapid CHX-induced transcription as a distinct and potentially more pervasive artifact when one is attempting to measure TE.

Here we show that what appears to be strong translational regulation of a large group of ribi genes during amino acid starvation is in fact an experimental artifact caused by CHX pretreatment. Even brief drug exposure (5-6 min) leads to a substantial increase in ribi mRNAs, and due to the presence of CHX, these new messages cannot be translated. This together with a decrease in ribosome-protected footprints of ribi genes leads to artificially low measured translation efficiencies. Using *NOP2* as a representative ribi gene, we find that the increased ribi mRNA content in CHX-treated samples is due to rapid accumulation of transcripts upon drug exposure. Remarkably, the

solvent used to make a CHX working solution has a large effect on the response to the drug: while CHX dissolved in DMSO leads to a slight increase in mRNA after 9 minutes, CHX dissolved in ethanol multiplies the increase by as much as 4-fold. Even more surprising is the observation that as little as 0.1% ethanol by itself is enough to transiently increase *NOP2* levels, but only in an auxotrophic lab strain. This heightened sensitivity to ethanol could be the result of abnormal utilization of glucose by auxotrophs, which may respond more readily to otherwise unfavorable carbon sources such as ethanol and increase expression of pro-growth genes (34). Since the NOP2 increase from CHX in ethanol is much greater than the sum of the individual increases induced by CHX in DMSO or ethanol alone, CHX and ethanol may influence ribi mRNA abundance through separate pathways.

The CHX-dependent decrease in ribi gene ribosome-protected footprints is surprising in light of prior observations that CHX inhibits mRNA decay (35), therefore one might assume *a priori* that CHX treatment would actually increase relative footprint abundances in cases when transcription is shut off (e.g. ribi genes during starvation) and old mRNAs are turned over. Notably, ribi transcripts are highly enriched in certain codons for aspartate, glutamate, and lysine (Supplementary Figure S2.2C), and all genes enriched in these codons—ribi or otherwise—tend to have fewer footprint counts in CHX-treated samples (Supplementary Figure S2.2D). Although this effect of CHX on footprint counts remains enigmatic, it is possible that when charged tRNA pools are depleted (as is the case during amino acid starvation), in the presence of CHX, ribosomes engaged in translating these codons are more likely to encounter collisions that trigger ribosome quality control pathways, resulting in clearance of these ribosomes from the mRNA (36).

When conditions are unfavorable for growth, ribi gene expression is shut off by at least three transcriptional repressors. Using time-lapse fluorescent microscopy, we show that adding CHX to starved cells causes these factors to exit the nucleus within 5 minutes. These factors are phosphorylated by the TORC1 substrate Sch9 to regulate their activity, and Sch9 phosphorylation is known to increase in response to CHX (29). Our observations are consistent with a model in which CHX activates TORC1 signaling in starved cells, overriding the starvation response and causing nuclear export of the ribi gene transcriptional repressors. Transcription of the ribi genes thus resumes, but in the presence of CHX the new mRNAs are not translated, leading to what appears to be strong translational repression (Figure 2.5). Although we cannot formally rule out a CHX-driven change in ribi mRNA stability, the most parsimonious explanation given our observations is that the accumulation of mRNAs is due to increased transcription. In support of this hypothesis, we show that when CHX pretreatment is used to harvest cells, a ribi promoter sequence is necessary and sufficient to impart "low" TE on a reporter gene following amino acid starvation, independent of UTR sequences. We also show that CHX causes rapid re-partitioning of ribi transcription factors within the cell, but blocking TORC1 signaling with rapamycin prior to adding CHX suppresses ribi transcriptional repressor egress from the nucleus and abolishes accumulation of a ribi transcript in starved cells. It is worth noting that ribi mRNA abundances still decrease following amino acid starvation in a strain with all three transcriptional repressor genes deleted, so additional factors are likely involved (Supplementary Figure S2.4C). It is also important to note that genes encoding ribosomal protein (RP) subunits are not subject to these effects, as their overall TE fold changes in the YMC and meiosis are comparatively low (Supplementary Figure S2.1B), and CHX pretreatment did not lead to increases in their measured mRNA abundances or decreases in measured ribosome

footprints (data not shown). Interestingly, studies in the fission yeast *Schizosaccharomyces pombe* have shown CHX-dependent changes in RP gene TEs during nitrogen starvation and histidine biosynthesis inhibition, suggesting a similar CHX-induced mechanism operating on a distinct regulon (37, 38). Ribi and RP genes are regulated by different sets of transcription factors in *S. cerevisiae* (39–41), and ribi gene expression peaks just before that of the RPs during the YMC, supporting the idea that decoupling ribi from RP gene regulation allows ribosome assembly machinery to accumulate before subunits are produced (42).

Although the precise mechanism of CHX-induced Sch9 phosphorylation remains unknown, a plausible explanation is that an increase in cytoplasmic concentrations of free amino acids following global translation arrest mimics a shift to nutrient-rich conditions, leading to TORC1 activation (43, 44). Therefore, it is possible that other stress conditions resulting in Sch9 dephosphorylation could be subject to translation inhibitor artifacts. These conditions include carbon/nitrogen/phosphate starvation, osmotic stress, redox stress, and heat shock (29); and potentially affected genes in the TOR regulon are numerous. Indeed, cells in two distinct growth conditions—the YMC and meiosis, both involving carbon and/or nitrogen limitation—had similarly low TE of ribi genes when CHX pretreatment was used in the ribosome profiling protocol. CHX is an invaluable tool, but it is important to be aware of the complex interplay between drug formulation, culture conditions, and cell genotype that can lead to unexpected results. We hope that this work will aid future ribosome profiling experimental design by highlighting additional pitfalls one might encounter as a result of treating cells with CHX, as well as ways to mitigate them.

Data Availability

The raw and processed sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus under accession number GSE125038.

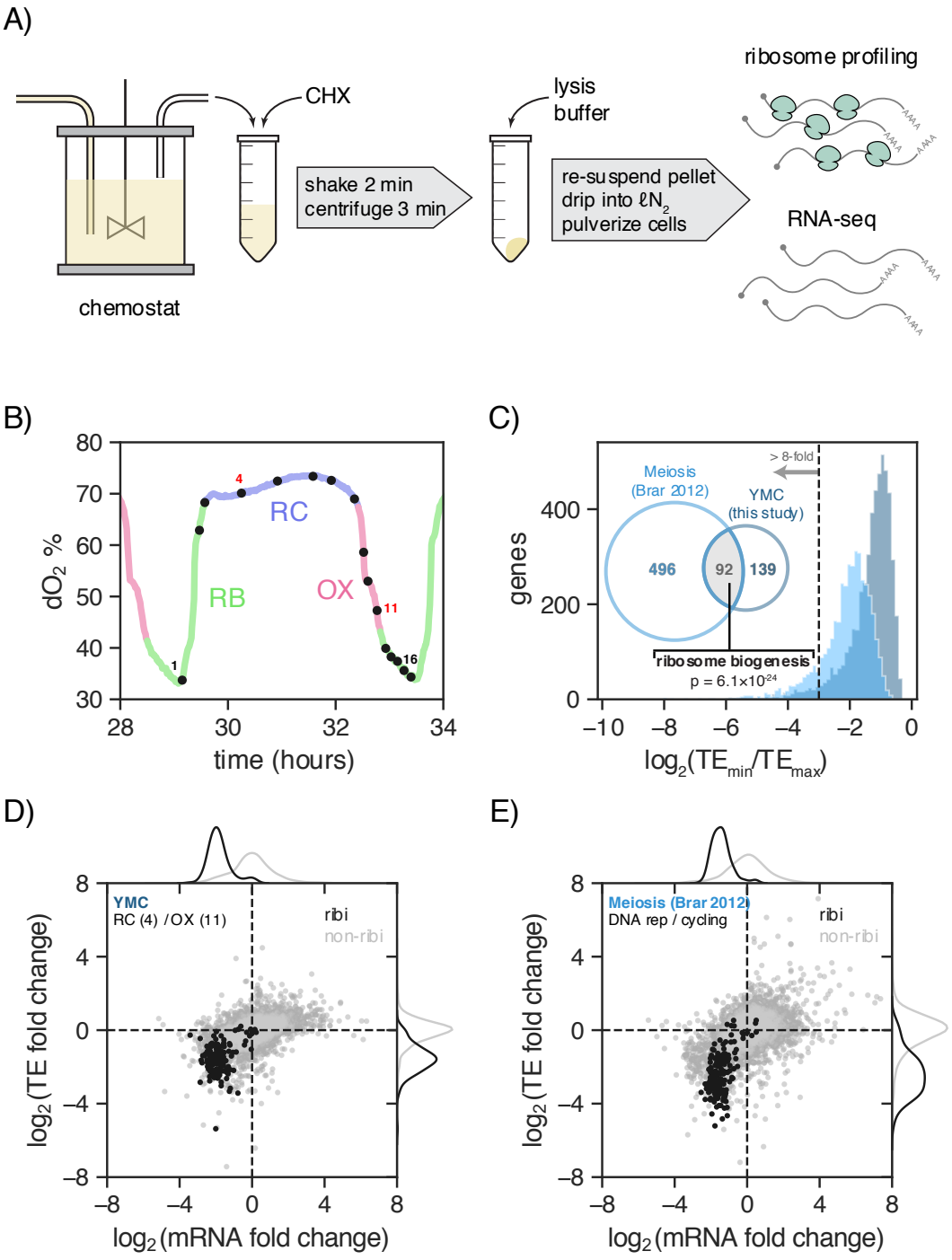


Figure 2.1. Apparent translational control of ribi genes in the Yeast Metabolic Cycle and meiosis. **(A)** Harvesting scheme from the YMC chemostat for ribosome profiling and RNA-seq. **(B)** Periodic fluctuations in dissolved oxygen for the span of the YMC during which samples were taken. Samples are numbered 1 through 16; samples 4 and 11 (red) were used to compare TE fold changes in panel **D**. RB, reductive building; RC, reductive charging; OX, oxidative phase. **(C)** Histograms of TE fold change (min/max) for all genes in the YMC and meiosis (9). Venn diagram shows the number of

genes in each experiment exceeding 8-fold TE change, with the intersection highly enriched for ribosome biogenesis genes (SGD GO Process finder). **(D)** TE fold change vs mRNA fold change for all ribi (black) and non-ribi (grey) genes between the RC and OX phases (YMC time points 4 and 11, respectively). Kernel density estimates of the distributions are plotted in the margins. **(E)** TE fold change vs mRNA fold change, as in panel **D**, between the cycling vegetative and DNA replication time points of meiosis (9).

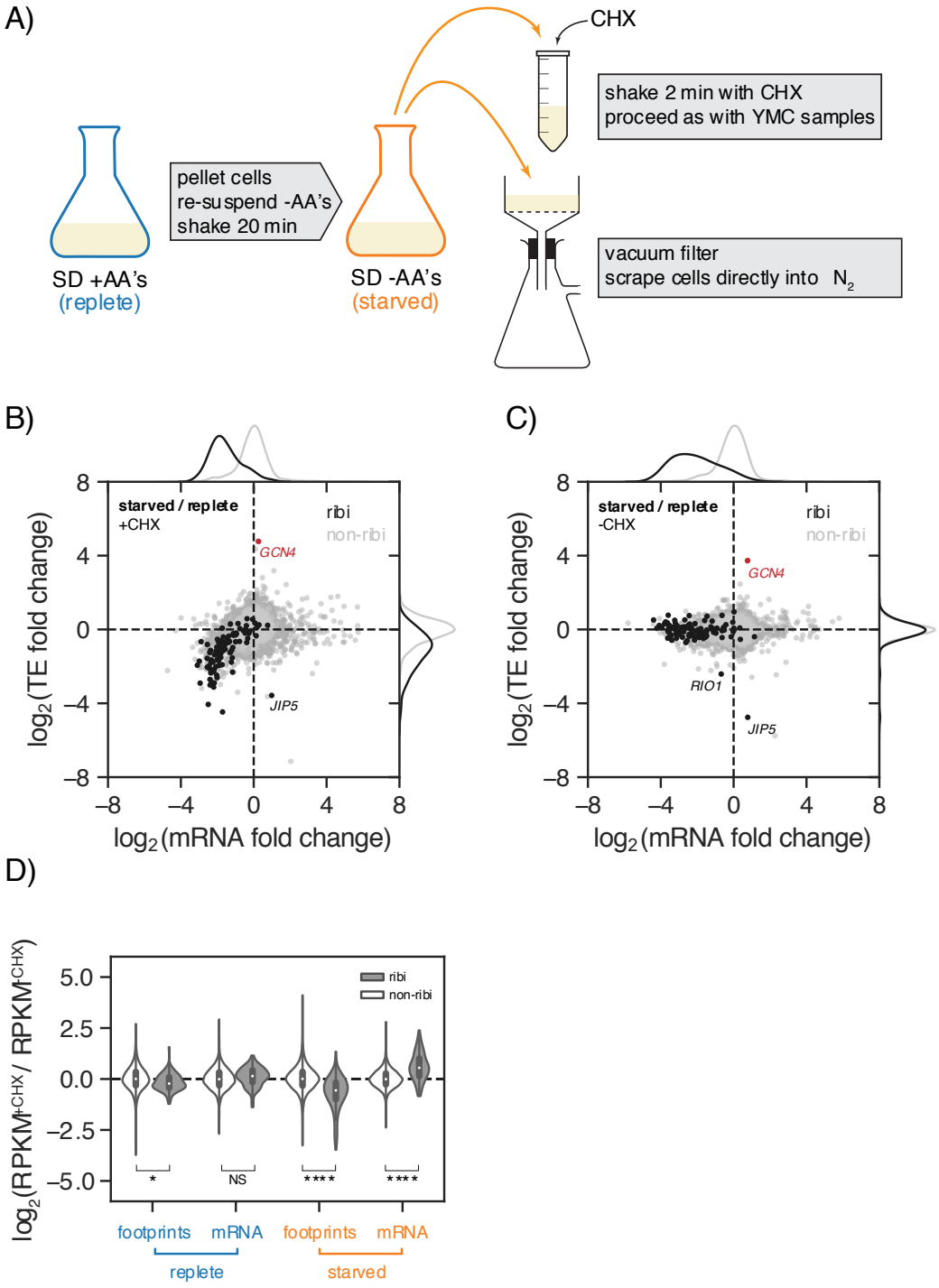


Figure 2.2. Effect of CHX on ribi gene TEs during amino acid starvation. **(A)** Amino acid starvation harvesting scheme with CHX (top) or without CHX (bottom). **(B)** TE fold change vs mRNA fold change in starved vs replete cells, with CHX pretreatment. As in the YMC and meiosis, ribi genes decrease along both dimensions. *GCN4* translational activation is a hallmark of eIF2α

phosphorylation in starved cells. **(C)** Without CHX pretreatment, ribi TEs are unchanged except for *RIO1* and *JIP5*. **(D)** Violin plot of normalized count ratios with and without CHX pretreatment. The plot is separated by treatment (replete vs. starved), measured quantity (footprints vs. mRNA), and gene set (ribi vs. non-ribi). In starved cells, CHX pretreatment causes a decrease in footprints of ribi genes and an increase in mRNA levels of ribi genes, whereas non-ribi genes are relatively unaffected. This is also observed in replete cells but to a lesser degree. * $p<10^{-2}$, **** $p<10^{-25}$, NS not significant; two-sided t-test.

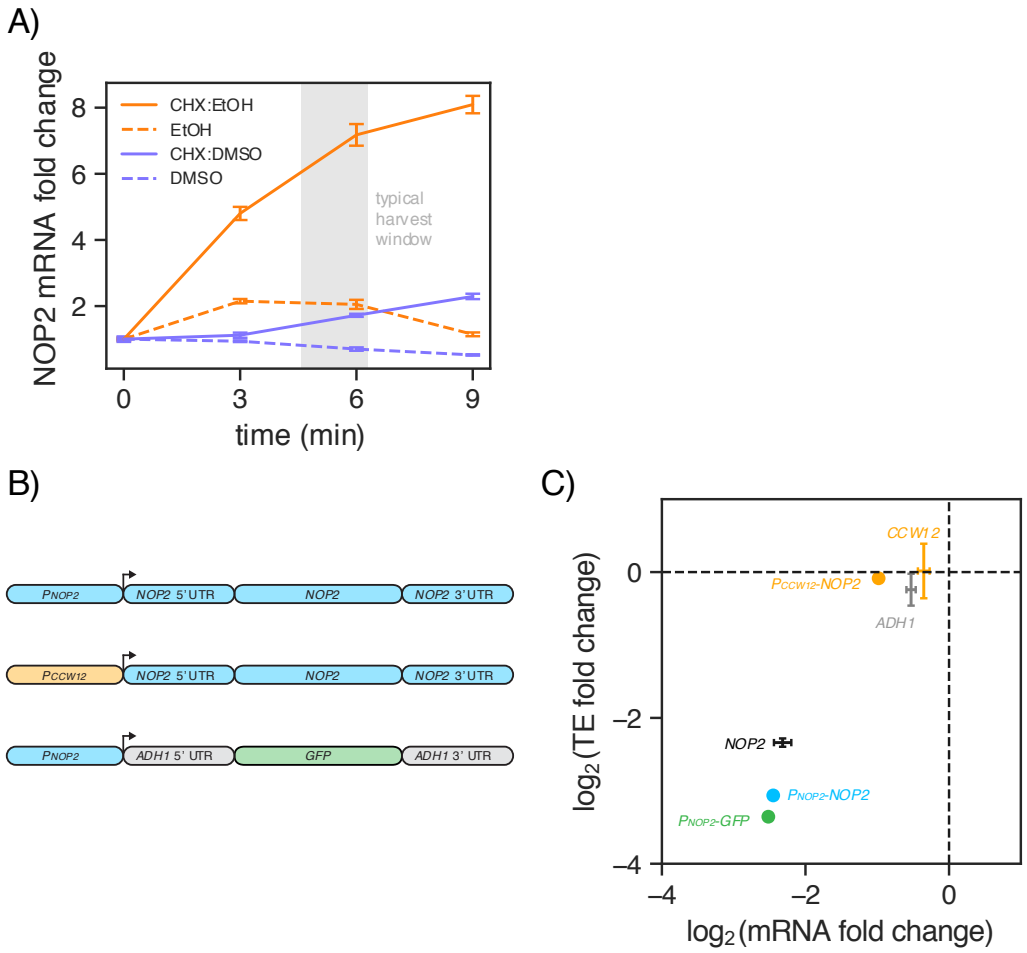


Figure 2.3. CHX influences the TE of a ribi gene via transcription. **(A)** CHX or vehicle was added to starved cells at t=0 and relative abundance of a ribi mRNA (*NOP2*) was monitored over time. *NOP2* transcripts accumulate faster and to a greater degree when CHX is dissolved in ethanol as opposed to DMSO. The shaded area indicates the typical amount of elapsed time between CHX pretreatment and freezing of cells for ribosome profiling experiments in this study. Error bars represent the standard deviation of 3 technical replicates from a single biological sample. **(B)** Reporter constructs with various promoter, coding sequence, and UTR elements were integrated into the dispensable *YHRCdelta14* locus. Sequences from the ribi gene *NOP2* are shaded blue. **(C)** TE fold change vs. mRNA fold change for reporter strains in panel **B** following amino acid starvation. TE and mRNA only decrease when the *NOP2* promoter is driving expression. Fold changes for endogenous genes (*NOP2*, *CCW12*, *ADH1*) are represented as mean ± standard deviation across 2-4 independent experiments.

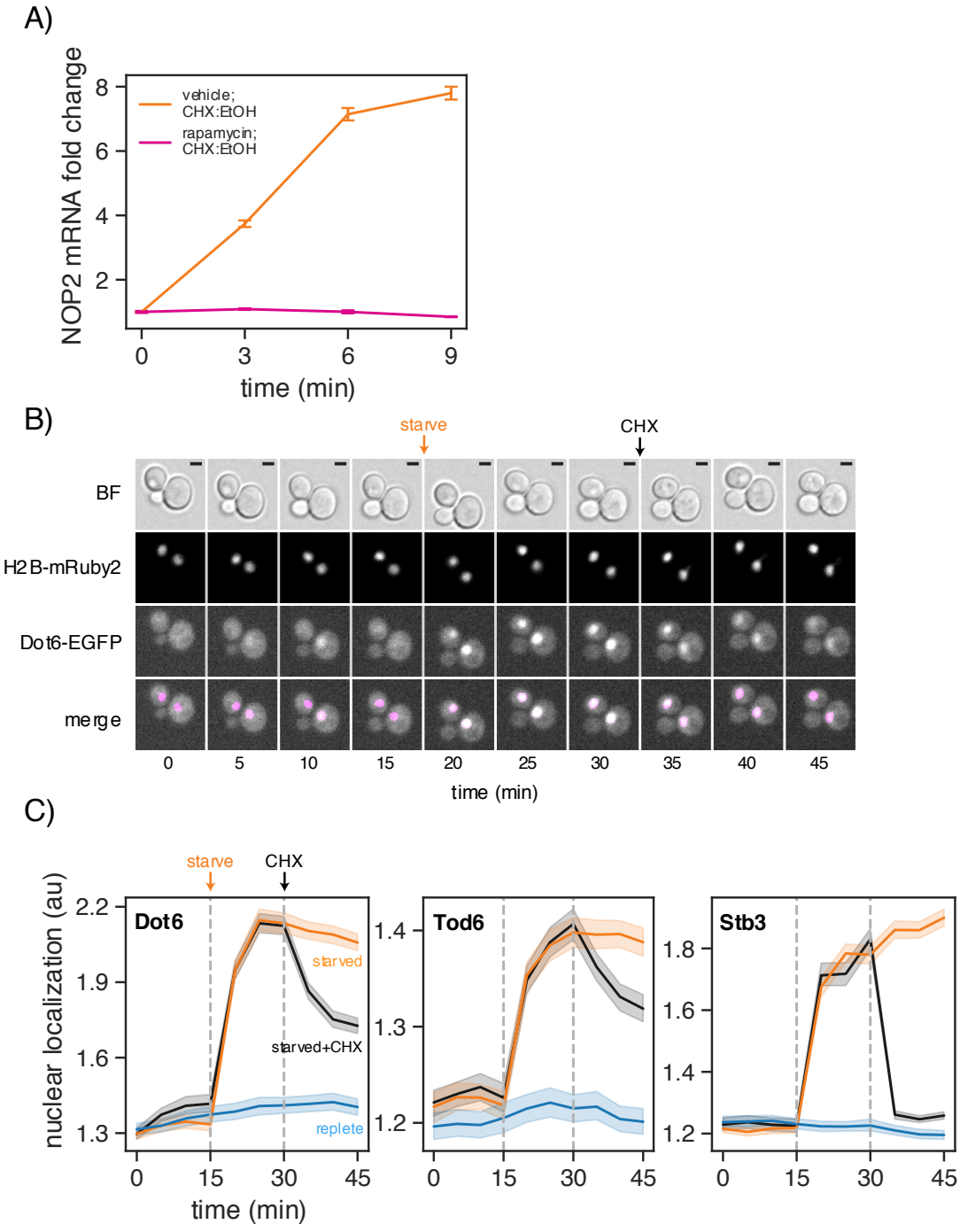


Figure 2.4. CHX-induced ribi gene transcription requires TORC1 signaling. **(A)** Cells were transferred to starvation medium and simultaneously treated with rapamycin or vehicle (DMSO). After 20 min CHX was added and *NOP2* mRNA abundance was monitored over time. Rapamycin treatment abolishes transcriptional activation. Error bars represent the standard deviation of 3 technical replicates from a single biological sample. **(B)** Dot6 and histone H2B were tagged with EGFP and mRuby2, respectively, and cells were immobilized in a CellASIC microfluidic device. Images were acquired as replete medium was flowed over the cells for 15 min, followed by 15 min of

starvation medium, and finally 15 min of starvation medium plus CHX. Dot6 rapidly localizes to the nucleus upon starvation and exits the nucleus after CHX treatment. Nuclei are false-colored magenta in the merged image. BF, brightfield with 2 μm scale bar. (C) The experiment in panel B was conducted with EGFP-tagged Dot6, Tod6, and Stb3; cells were also kept in replete medium (blue) or not treated with CHX (orange) as controls. Nuclear localization was defined as the fraction of GFP signal in the nucleus relative to the signal in the cytoplasm. All three transcription factors exit the nucleus upon CHX treatment; Stb3 responds most robustly. Localization data is plotted as the mean of all cells observed (n = 213-310) with 95% confidence interval.

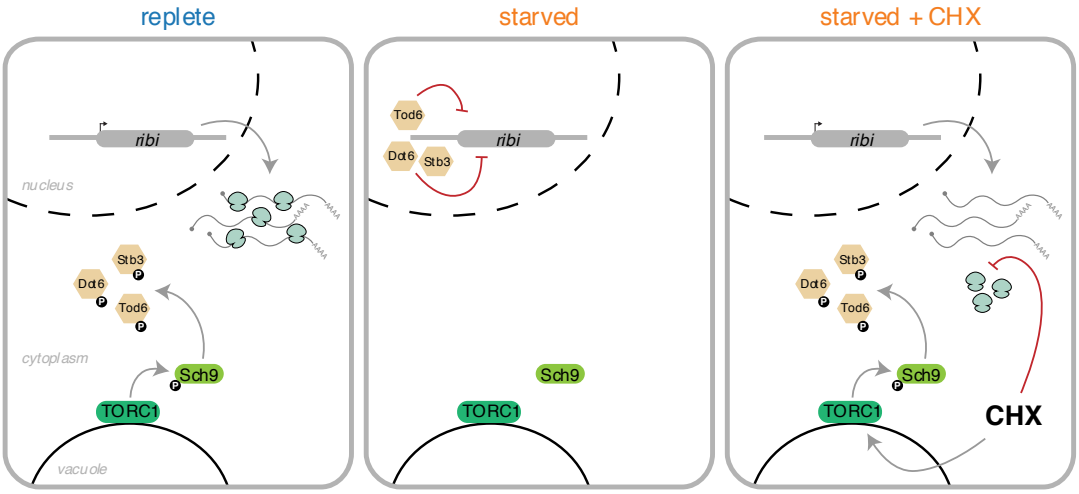
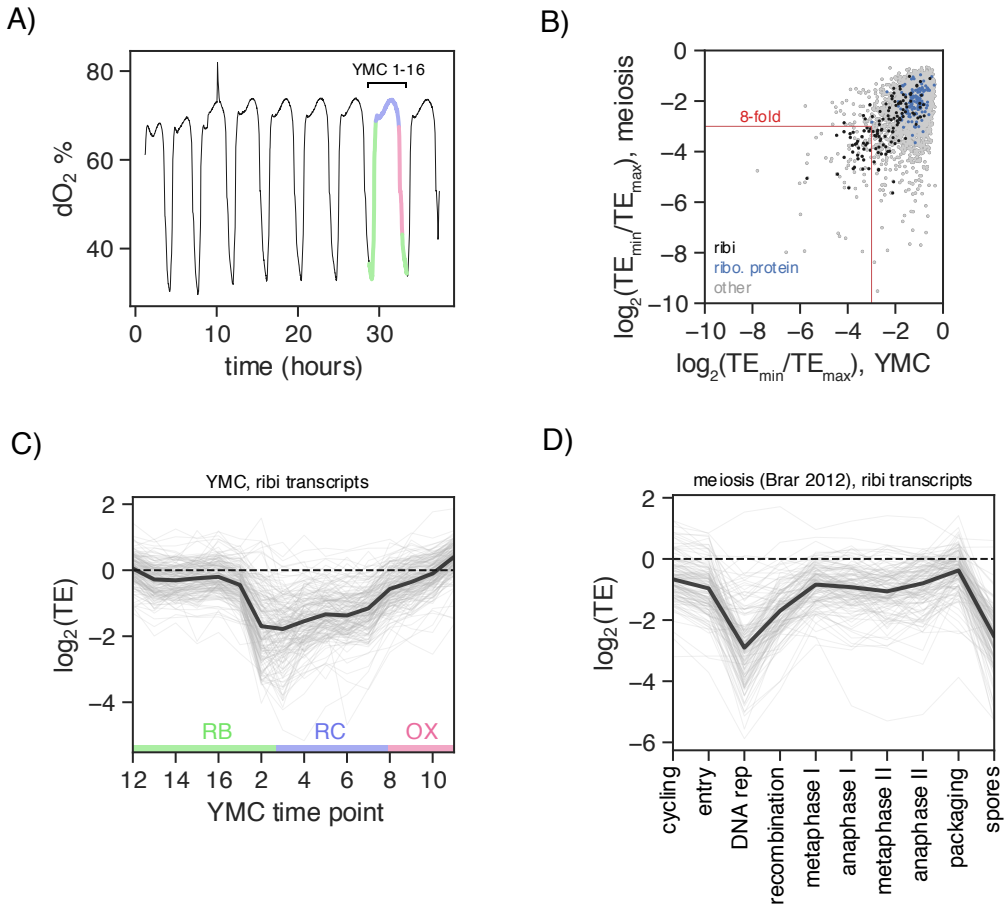
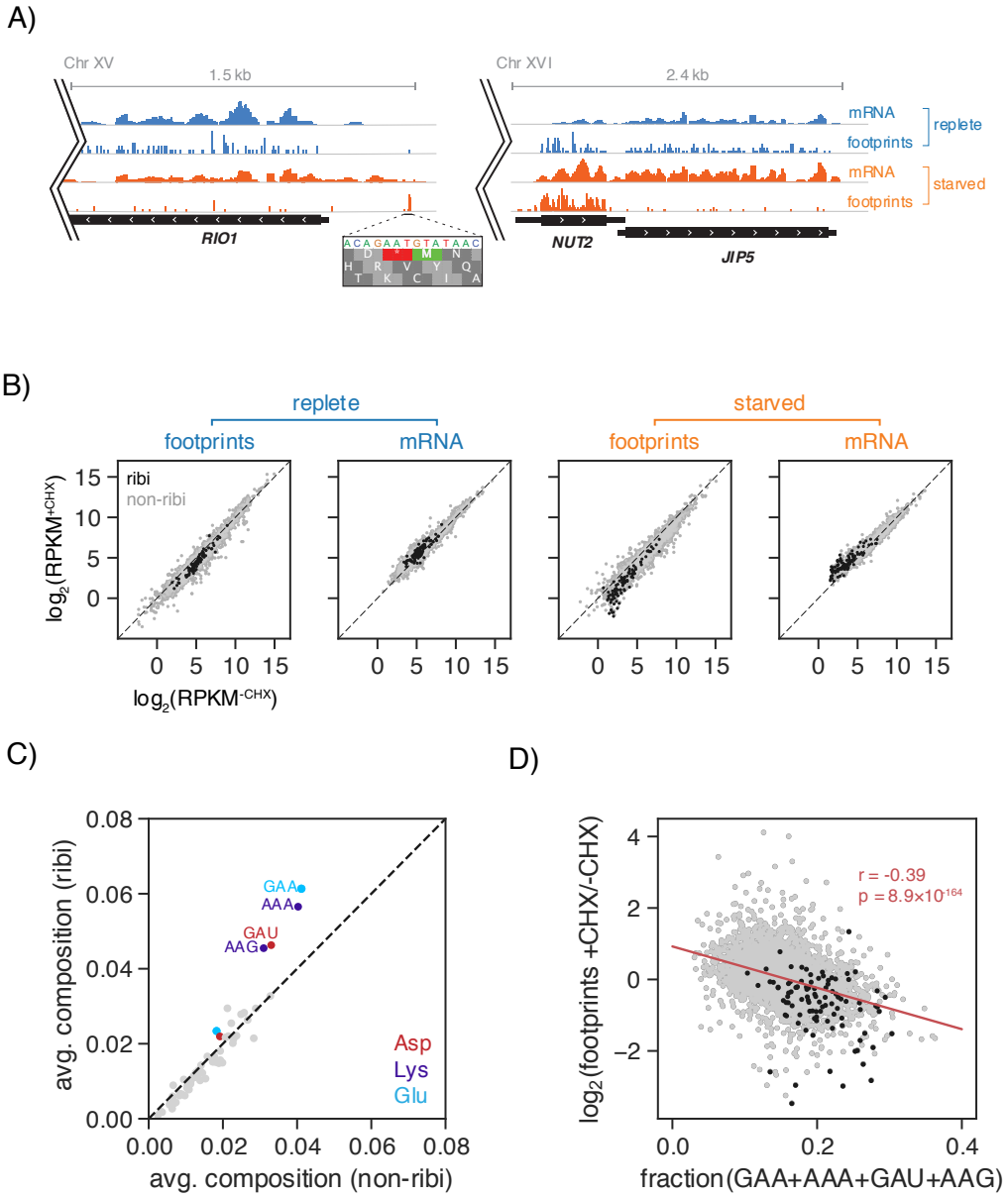


Figure 2.5. Proposed model of the influence of CHX on ribi gene TEs. In replete conditions TORC1 signaling is active, leading to inhibitory phosphorylation of ribi gene transcriptional repressors and active transcription/translation of ribi genes. In starved cells, TORC1 signaling decreases, thereby allowing ribi transcriptional repression. Upon adding CHX to starved cells, TORC1 signaling resumes and new ribi transcripts are produced. In the presence of CHX, however, these mRNAs cannot be translated and apparent TE is low.

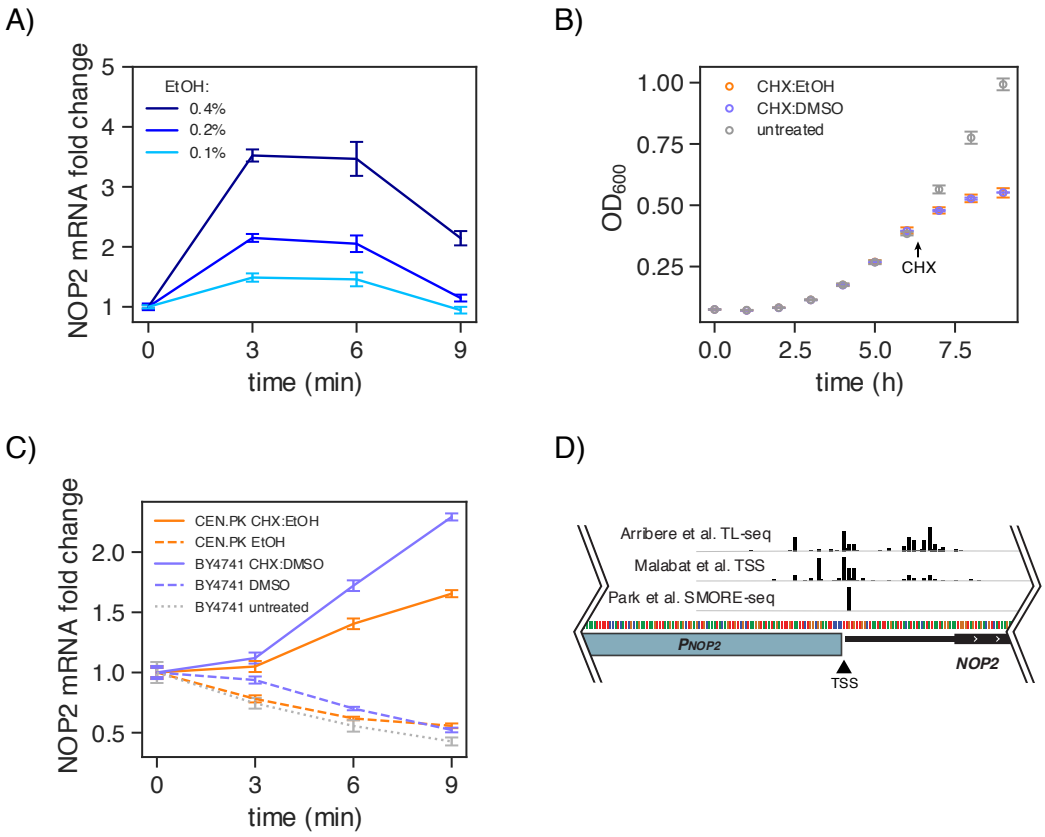


Supplementary Figure S2.1. **(A)** Overview of metabolic cycles observed over ~40 hours. Decreases in dissolved oxygen correspond to the OX phase of the YMC. Time points 1-16 are highlighted; detail shown in Figure 2.1B. **(B)** Scatter plot of the TE range (min/max) exhibited by each gene in meiosis vs the YMC. The set of genes with 8-fold or more TE change in both time courses (red box) is highly enriched for ribi factors, but not protein subunits of the cytoplasmic ribosome. **(C)** TEs of ribi genes over the YMC time course. Each gene is shown as a thin grey line; the median of all ribi genes is shown as a thick black line. In general, TEs appear to be low during the RB to RC transition, and high at the end of OX. **(D)** TEs of ribi genes through meiosis, from (9). TEs appear to be low during DNA replication and in spores.

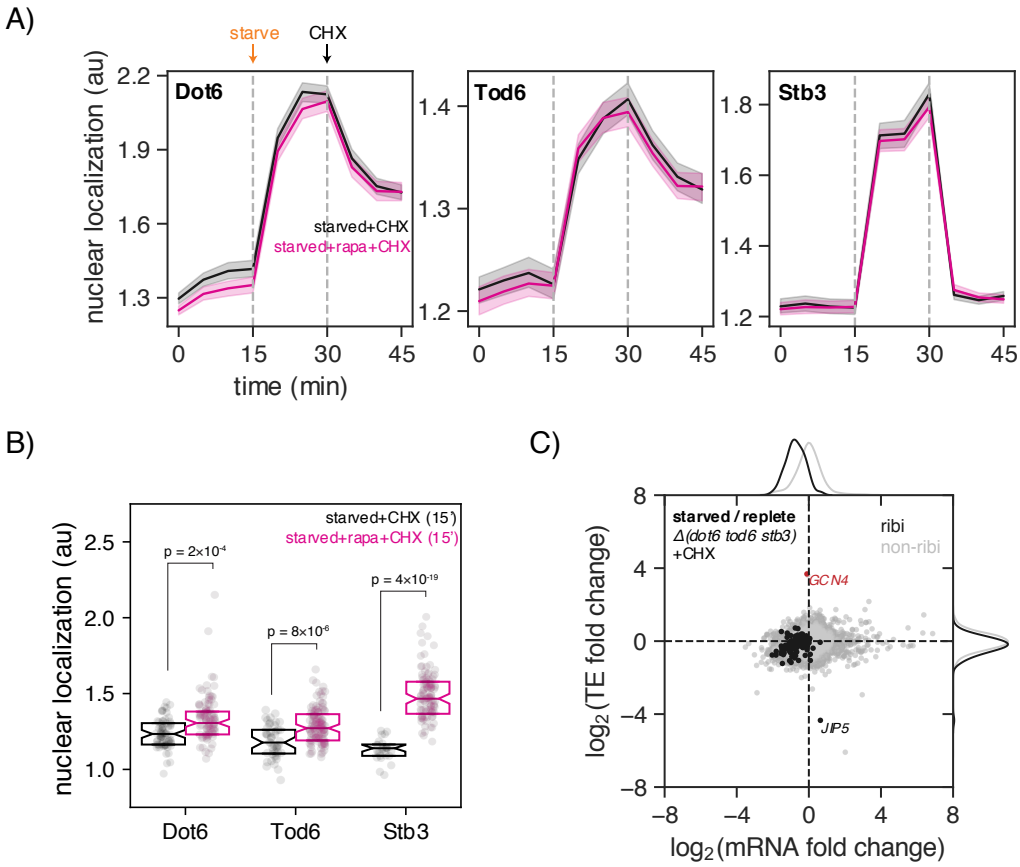


Supplementary Figure S2.2. **(A)** mRNA and footprint counts mapped to the *RIO1* and *JIP5* loci. In starved cells the *RIO1* transcript is extended on the 5′ end (minus strand; 5′ end is to the right). The longer isoform incorporates a short ORF consisting of a start codon immediately followed by a stop codon, and a large peak of ribosome density is observed here in starved cells. Footprint density is greatly reduced at *JIP5* in starved cells, however the proximity of *NUT2* makes it difficult to assess whether the transcript architecture changes. Y-axis scales are consistent for mRNA and footprint tracks, respectively, at each locus. Footprint counts are mapped to ribosomal P-sites; RNA-seq counts are evenly apportioned across the length of each sequencing read. **(B)** Scatter plots of normalized counts (RPKM) in amino acid starvation experiments with- vs. without-CHX pretreatment. Ribi gene counts fall off-diagonal in starved cells. **(C)** Average codon composition of ribi vs non-ribi genes. Certain Asp, Lys, and Glu codons are enriched in ribi genes. **(D)** Genes

enriched in GAA, AAA, GAU, and AAG codons have lower footprint counts in CHX-treated vs -untreated cells. While ribi genes (black) are enriched in these codons, the correlation holds for all genes.



Supplementary Figure S2.3. **(A)** Ethanol was added to starved cells at t=0 and *NOP2* mRNA abundance was monitored over time. Percentages reflect the final concentration (vol/vol) of added ethanol in the culture. Error bars represent the standard deviation of 3 technical replicates from a single biological sample. **(B)** BY4741 was grown in SD in a 30 °C shaking incubator and OD was monitored over time. CHX dissolved in ethanol or DMSO was added immediately following the 6-hour time point. CHX inhibits growth regardless of the solvent. Error bars represent the standard deviation of 3 biological replicates. **(C)** The prototrophic strain CEN.PK has an attenuated response to CHX in ethanol compared to the same treatment in the auxotrophic strain BY4741 (compare to Figure 2.3A, solid orange line). The change in *NOP2* mRNA is of similar magnitude to BY4741 treated with CHX in DMSO. Both strains have nearly identical responses to the respective vehicle treatments, which resemble an untreated control. Error bars represent the standard deviation of 3 technical replicates from a single biological sample. **(D)** The *NOP2* promoter was defined as the 285 bp region from the boundary of the upstream gene (*GCD10*) to the consensus *NOP2* transcript start site (45-47). Blue box, *NOP2* promoter; thin black box, 5′ UTR from (48); thick black box, coding sequence.



Supplementary Figure S2.4. The experiment in Figure 2.4C was repeated with rapamycin added to the starvation medium. In a microfluidic plate, rapamycin does not alter the localization of these transcription factors following CHX treatment. **(B)** Repeat of the above experiment in a flask. Cells were pelleted and re-suspended in starvation medium with (pink) or without (black) 200 nM rapamycin. After 15 min CHX was added, and after an additional 15 min cells were imaged in glass-bottom wells. All three transcription factors show increased nuclear localization in the presence of rapamycin. *P*-values were calculated using a two-sided t-test. **(C)** Ribi mRNAs still decrease following 20 min of amino acid starvation in a *dot6 tod6 stb3* triple deletion strain, though to a lesser degree than the WT strain (compare to Figure 2.2B)

Table S2.1. Strains used in this study.

NAME	DESCRIPTION	MAT	GENOTYPE	PARENT	SOURCE
yJW1201	BY4741	a	*his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*	S288C	Brachmann et al. (1998). PMID: 9483801
yJW1857	CEN.PKa	a	WT	sporulated from CEN.PK122	van Dijken et al. (2000). PMID: 10862876
yJW1860	*K.lac NOP2*	a	*nop2::klNOP2 his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*	BY4741	this study
yJW1861	*K.lac NOP2;* P_{CCW12}*-NOP2*	a	*nop2::klNOP2 yhrcdelta14::*P_{CCW12}*-NOP2::kanMX6 his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*	yJW1860	this study
yJW1862	*K.lac NOP2;* P_{NOP2}*-NOP2*	a	*nop2::klNOP2 yhrcdelta14::*P_{NOP2}*-NOP2::kanMX6 his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*	yJW1860	this study
yJW1863	P_{NOP2}*-5'UTR*$_{ADH1}$*-EGFP-3'UTR*$_{ADH1}$	a	*yhrcdelta14::*P_{NOP2}*-5'UTR*$_{ADH1}$*-EGFP-3'UTR*$_{ADH1}$*::kanMX6 his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*	BY4741	this study
yJW1864	*Δ(dot6 tod6 stb3)*	a	*dot6::LEU2 tod6::URA3 stb3::kanMX6 his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*	BY4741	this study
yJW1865	Dot6-EGFP; H2B-mRuby2	a	*dot6::DOT6-EGFP::kanMX6 htb2::HTB2-mRuby2::SpHIS5 his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*	BY4741	this study
yJW1866	Tod6-EGFP; H2B-mRuby2	a	*tod6::TOD6-EGFP::kanMX6 htb2::HTB2-mRuby2::SpHIS5 his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*	BY4741	this study
yJW1867	Stb3-EGFP; H2B-mRuby2	a	*stb3::STB3-EGFP::kanMX6 htb2::HTB2-mRuby2::SpHIS5 his3-Δ1 leu2-Δ0 met15-Δ0 ura3-Δ0*	BY4741	this study

Table S2.2. Plasmids used in this study.

NAME	DESCRIPTION	PURPOSE	SOURCE
pFA6a-link-yoEGFP-Kan	EGFP-Kan	Tag ribi transcription factors with EGFP	Addgene 44900
pFA6a-link-yomRuby2-SpHis5	mRuby2-His	Tag H2B with mRuby2	Addgene 44858
pJW1746	pRS306_YHRCdelta14UP-P_{CCW12}-NOP2-KanR-YHRCdelta14DOWN	PCR template to generate yJW1861	this study
pJW1747	pRS306_YHRCdelta14UP-P_{NOP2}-NOP2-KanR-YHRCdelta14DOWN	PCR template to generate yJW1862	this study
pJW1748	pRS306_YHRCdelta14UP-P_{NOP2}-5'UTR_{ADH1}-EGFP-3'UTR_{ADH1}-KanR-YHRCdelta14DOWN	PCR template to generate yJW1863	this study

Table S2.3. Oligonucleotides used in this study.

NAME	SEQUENCE	PURPOSE	DESCRIPTION
oAF66	/5rApp/NNNNNATCGAGATCGGAAGAGCAC ACGTCTGAACTC/3ddC/	ribosome profiling	3' cloning linker (ATCG)
oAF67	/5rApp/NNNNNTAGCAGATCGGAAGAGCAC ACGTCTGAACTC/3ddC/	ribosome profiling	3' cloning linker (TAGC)
oAF68	/5rApp/NNNNNCGATAGATCGGAAGAGCAC ACGTCTGAACTC/3ddC/	ribosome profiling	3' cloning linker (CGAT)
oAF69	/5rApp/NNNNNGCTAAGATCGGAAGAGCAC ACGTCTGAACTC/3ddC/	ribosome profiling	3' cloning linker (GCTA)
oAF70	/5rApp/NNNNNAGTCAGATCGGAAGAGCAC ACGTCTGAACTC/3ddC/	ribosome profiling	3' cloning linker (AGTC)
oAF71	/5rApp/NNNNNGACTAGATCGGAAGAGCAC ACGTCTGAACTC/3ddC/	ribosome profiling	3' cloning linker (GACT)
oAF72	/5rApp/NNNNNCTGAAGATCGGAAGAGCAC ACGTCTGAACTC/3ddC/	ribosome profiling	3' cloning linker (CTGA)
oAF73	/5rApp/NNNNNTCAGAGATCGGAAGAGCAC ACGTCTGAACTC/3ddC/	ribosome profiling	3' cloning linker (TCAG)
oAF74	/5Phos/AGATCGGAAGAGCGTCGTGTAGGG AAAGAG/iSp18/CTGGAGTTCAGACGTGTG	ribosome profiling	RT
oAF75	AATGATACGGCGACCACCGAGATCTACACT CTTTCCCTACACGACGCTC	ribosome profiling	PCR (universal)
oAF76	CAAGCAGAAGACGGCATACGAGATTACAAG GTGACTGGAGTTCAGACGTGTGCTC	ribosome profiling	PCR (index 12)
oAF77	CAAGCAGAAGACGGCATACGAGATATTGG CGTGACTGGAGTTCAGACGTGTGCTC	ribosome profiling	PCR (index 6)
oAF78	CAAGCAGAAGACGGCATACGAGATGGAAC TGTGACTGGAGTTCAGACGTGTGCTC	ribosome profiling	PCR (index 14)
oAF79	CAAGCAGAAGACGGCATACGAGATAAGCTA GTGACTGGAGTTCAGACGTGTGCTC	ribosome profiling	PCR (index 10)
oAF80	CAAGCAGAAGACGGCATACGAGATGCCTA AGTGACTGGAGTTCAGACGTGTGCTC	ribosome profiling	PCR (index 3)
oAF81	CAAGCAGAAGACGGCATACGAGATCGTGA TGTGACTGGAGTTCAGACGTGTGCTC	ribosome profiling	PCR (index 1)

NAME	SEQUENCE	PURPOSE	DESCRIPTION
oAF82	CAAGCAGAAGACGGCATACGAGATCCACT CGTGACTGGAGTTCAGACGTGTGCTC	ribosome profiling	PCR (index 23)
oAF83	CAAGCAGAAGACGGCATACGAGATTTGAC TGTGACTGGAGTTCAGACGTGTGCTC	ribosome profiling	PCR (index 13)
oDAS 220	CGTCTGGATTGGTGGTTCTATC	qPCR	ACT1 (For)
oDAS 221	GGACCACTTTCGTCGTATTCTT	qPCR	ACT1 (Rev)
oDAS 880	TCTATCCGCCATTGATTCTGTT	qPCR	NOP2 (For)
oDAS 881	CTATGACAGCTTCGTCCTCTTC	qPCR	NOP2 (Rev)

References

1. Ingolia,N.T., Hussmann,J.A. and Weissman,J.S. (2018) Ribosome Profiling: Global Views of Translation. *Cold Spring Harb. Perspect. Biol.*, 10.1101/cshperspect.a032698.

2. Gingold,H. and Pilpel,Y. (2011) Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.*, **7**, 481.

3. Rüegsegger,U., Leber,J.H. and Walter,P. (2001) Block of HAC1 mRNA Translation by Long-Range Base Pairing Is Released by Cytoplasmic Splicing upon Induction of the Unfolded Protein Response. *Cell*, **107**, 103–114.

4. Mueller,P.P. and Hinnebusch,A.G. (1986) Multiple upstream AUG codons mediate translational control of GCN4. *Cell*, **45**, 201–207.

5. Meyenburg,H.K. von (1969) Energetics of the budding cycle of Saccharomyces cerevisiae during glucose limited aerobic growth. *Arch. Für Mikrobiol.*, **66**, 289–303.

6. Tu,B.P. (2005) Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes. *Science*, **310**, 1152–1158.

7. Jorgensen,P., Rupeš,I., Sharom,J.R., Schneper,L., Broach,J.R. and Tyers,M. (2004) A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes Dev.*, **18**, 2491–2505.

8. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.

9. Brar,G.A., Yassour,M., Friedman,N., Regev,A., Ingolia,N.T. and Weissman,J.S. (2012) High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. *Science*, **335**, 552–557.

10. Rothstein,R. (1991) [19] Targeting, disruption, replacement, and allele rescue: Integrative DNA transformation in yeast. In Methods in Enzymology, Guide to Yeast Genetics and Molecular Biology. Academic Press, Vol. 194, pp. 281–301.

11. Gibson,D.G., Young,L., Chuang,R.-Y., Venter,J.C., Iii,C.A.H. and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–345.

12. Flagfeldt,D.B., Siewers,V., Huang,L. and Nielsen,J. (2009) Characterization of chromosomal integration sites for heterologous gene expression in Saccharomyces cerevisiae. *Yeast*, **26**, 545–551.

13. Lee,S., Lim,W.A. and Thorn,K.S. (2013) Improved Blue, Green, and Red Fluorescent Protein Tagging Vectors for S. cerevisiae. *PLOS ONE*, **8**, e67902.

14. Shi,L. and Tu,B.P. (2013) Acetyl-CoA induces transcription of the key G1 cyclin CLN3 to promote entry into the cell division cycle in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci.*, **110**, 7318–7323.

15. MGlincy,N.J. and Ingolia,N.T. (2017) Transcriptome-wide measurement of translation by ribosome profiling. *Methods*, **126**, 112–129.

16. Dunn,J.G. and Weissman,J.S. (2016) Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics*, **17**, 958.

17. Ares,M. (2012) Isolation of Total RNA from Yeast Cell Cultures. *Cold Spring Harb. Protoc.*, **2012**, pdb.prot071456.

18. Livak,K.J. and Schmittgen,T.D. (2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method. *Methods*, **25**, 402–408.

19. Gerashchenko,M.V. and Gladyshev,V.N. (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.*, **42**, e134–e134.

20. Ingolia,N.T., Lareau,L.F. and Weissman,J.S. (2011) Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell*, **147**, 789–802.

21. Hussmann,J.A., Patchett,S., Johnson,A., Sawyer,S. and Press,W.H. (2015) Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLOS Genet.*, **11**, e1005732.

22. Chen,J., Tresenrider,A., Chia,M., McSwiggen,D.T., Spedale,G., Jorgensen,V., Liao,H., van Werven,F.J. and Ünal,E. (2017) Kinetochore inactivation by expression of a repressive mRNA. *eLife*, **6**, e27417.

23. Chia,M., Tresenrider,A., Chen,J., Spedale,G., Jorgensen,V., Ünal,E. and van Werven,F.J. (2017) Transcription of a 5' extended mRNA isoform directs dynamic chromatin changes and interference of a downstream promoter. *eLife*, **6**, e27420.

24. Cheng,Z., Otto,G.M., Powers,E.N., Keskin,A., Mertins,P., Carr,S.A., Jovanovic,M. and Brar,G.A. (2018) Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. *Cell*, **172**, 910-923.e16.

25. Powers,T. and Walter,P. (1999) Regulation of ribosome biogenesis by the rapamycin-sensitive TOR-signaling pathway in Saccharomyces cerevisiae. *Mol. Biol. Cell*, **10**, 987–1000.

26. Loewith,R., Jacinto,E., Wullschleger,S., Lorberg,A., Crespo,J.L., Bonenfant,D., Oppliger,W., Jenoe,P. and Hall,M.N. (2002) Two TOR Complexes, Only One of which Is Rapamycin Sensitive, Have Distinct Roles in Cell Growth Control. *Mol. Cell*, **10**, 457–468.

27. Barbet,N.C., Schneider,U., Helliwell,S.B., Stansfield,I., Tuite,M.F. and Hall,M.N. (1996) TOR controls translation initiation and early G1 progression in yeast. *Mol. Biol. Cell*, **7**, 25–42.

28. Lippman,S.I. and Broach,J.R. (2009) Protein kinase A and TORC1 activate genes for ribosomal biogenesis by inactivating repressors encoded by Dot6 and its homolog Tod6. *Proc. Natl. Acad. Sci.*, **106**, 19928–19933.

29. Urban,J., Soulard,A., Huber,A., Lippman,S., Mukhopadhyay,D., Deloche,O., Wanke,V., Anrather,D., Ammerer,G., Riezman,H., et al. (2007) Sch9 Is a Major Target of TORC1 in Saccharomyces cerevisiae. *Mol. Cell*, **26**, 663–674.

30. Huber,A., French,S.L., Tekotte,H., Yerlikaya,S., Stahl,M., Perepelkina,M.P., Tyers,M., Rougemont,J., Beyer,A.L. and Loewith,R. (2011) Sch9 regulates ribosome biogenesis via Stb3, Dot6 and Tod6 and the histone deacetylase complex RPD3L. *EMBO J.*, **30**, 3052–3064.

31. W. Toepke,M. and J. Beebe,D. (2006) PDMS absorption of small molecules and consequences in microfluidic applications. *Lab. Chip*, **6**, 1484–1486.

32. Gomez-Sjoberg,R., Leyrat,A.A., Houseman,B.T., Shokat,K. and Quake,S.R. (2010) Biocompatibility and Reduced Drug Absorption of Sol–Gel-Treated Poly(dimethyl siloxane) for Microfluidic Cell Culture Applications. *Anal. Chem.*, **82**, 8954–8960.

33. Wang,J.D., Douville,N.J., Takayama,S. and ElSayed,M. (2012) Quantitative Analysis of Molecular Absorption into PDMS Microfluidic Channels. *Ann. Biomed. Eng.*, **40**, 1862–1873.

34. Boer,V.M., Amini,S. and Botstein,D. (2008) Influence of genotype and nutrition on survival and metabolism of starving yeast. *Proc. Natl. Acad. Sci.*, **105**, 6930–6935.

35. Beelman,C.A. and Parker,R. (1994) Differential effects of translational inhibition in cis and in trans on the decay of the unstable yeast MFA2 mRNA. *J. Biol. Chem.*, **269**, 9687–9692.

36. Brandman,O. and Hegde,R.S. (2016) Ribosome-associated protein quality control. *Nat. Struct. Mol. Biol.*, **23**, 7–15.

37. Duncan,C.D.S. and Mata,J. (2017) Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe. *Sci. Rep.*, **7**, 10331–10331.

38. Duncan,C.D.S., Rodríguez-López,M., Ruis,P., Bähler,J. and Mata,J. (2018) General amino acid control in fission yeast is regulated by a nonconserved transcription factor, with functions analogous to Gcn4/Atf4. *Proc. Natl. Acad. Sci.*, **115**, E1829–E1838.

39. Wade,J.T., Hall,D.B. and Struhl,K. (2004) The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes. *Nature*, **432**, 1054–1058.

40. Rudra,D., Zhao,Y. and Warner,J.R. (2005) Central role of Ifh1p–Fhl1p interaction in the synthesis of yeast ribosomal proteins. *EMBO J.*, **24**, 533–542.

41. Zhao,Y., McIntosh,K.B., Rudra,D., Schawalder,S., Shore,D. and Warner,J.R. (2006) Fine-Structure Analysis of Ribosomal Protein Gene Transcription. *Mol. Cell. Biol.*, **26**, 4853–4862.

42. Kuang,Z., Cai,L., Zhang,X., Ji,H., Tu,B.P. and Boeke,J.D. (2014) High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast. *Nat. Struct. Mol. Biol.*, **21**, 854–863.

43. Beugnet,A., Tee,A.R., Taylor,P.M. and Proud,C.G. (2003) Regulation of targets of mTOR (mammalian target of rapamycin) signalling by intracellular amino acid availability. *Biochem. J.*, **372**, 555–566.

44. Binda,M., Péli-Gulli,M.-P., Bonfils,G., Panchaud,N., Urban,J., Sturgill,T.W., Loewith,R. and De Virgilio,C. (2009) The Vam6 GEF Controls TORC1 by Activating the EGO Complex. *Mol. Cell*, *35*, 563–573.

45. Arribere,J.A. and Gilbert,W.V. (2013) Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res.*, **23**, 977–987.

46. Malabat,C., Feuerbach,F., Ma,L., Saveanu,C. and Jacquier,A. (2015) Quality control of transcription start site selection by nonsense-mediated-mRNA decay. *eLife*, **4**.

47. Park,D., Morris,A.R., Battenhouse,A. and Iyer,V.R. (2014) Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic Acids Res.*, **42**, 3736–3749.

48. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349

Chapter 3

Titrating gene expression with allelic series of CRISPR guide RNAs

Introduction

The complexity of biological processes, from stress responses to morphogenesis arises from not only the set of expressed genes but also quantitative differences in their expression levels. As a classical example, some genes are haploinsufficient and thus are sensitive to a 50% decrease in expression, whereas for other genes disease only occurs upon far stronger depletion. The ability to systematically control gene expression levels and explore their relationships to biological phenotypes would have immediate practical and conceptual implications. For example, rescue of disease-causing mutations by chemical treatment or gene therapy ultimately requires restoring gene expression to functionally sufficient levels, and *a priori* these sufficiency levels are unclear. Vice versa, a cornerstone of cancer drug development is to inhibit generally essential functions to sufficient magnitude to ensure toxicity to rapidly proliferating cancer cells while sparing healthy cells, however the location of each therapeutic window is target-specific. More fundamentally, gene expression levels are determined by opposing evolutionary forces including the cost of protein synthesis and the need for robustness against random fluctuations, but these tradeoffs have not been comprehensively explored. Tools to precisely control gene expression level would transform efforts in these and additional areas, but despite dramatic advances in functional genomics the current set of tools has been primarily optimized for complete knockout or knockdown or massive overexpression and does not afford such control.

The discovery and development of artificial transcription factors, such as TALEs or the CRISPR-based effectors underlying CRISPR interference (CRISPRi) and activation (CRISPRa), has brought such tools within reach. CRISPR systems in particular have attracted considerable attention as the targeting to a locus of interest through sequence complementarity to an associated single guide

RNA (sgRNA) affords uniquely high programmability. Investigations of CRISPR targeting mechanisms have established that both activity and binding of Cas9 or its nuclease-dead variants (dCas9) can be modulated by introducing mismatches into the sgRNA targeting region, modifying the sgRNA constant region, and other approaches (1–5). Here, we report a systematic approach to control dCas9 effector binding through series of modified sgRNAs, or allelic series of sgRNAs, as a general method to titrate gene expression. We describe both a compact sgRNA library to titrate the expression of essential genes, empirically validated through pooled screens, and a genome-wide *in silico* library derived from deep learning analysis of the empirical data. Rich single-cell RNA-seq phenotypes recorded at different expression levels of essential genes reveal gene-specific expression-to-phenotype relationships and expression level-dependent cell responses and highlight the utility of such modified sgRNAs in staging cells along a continuum of expression levels.

Materials and Methods

Reagents and cell lines

K562 and Jurkat cells were grown in RPMI 1640 medium (Gibco) with 25 mM HEPES, 2 mM L-glutamine, 2 g/L $NaHCO_3$ supplemented with 10% (v/v) standard fetal bovine serum (FBS, HyClone or VWR), 100 units/mL penicillin, 100 µg/mL streptomycin, and 2 mM L-glutamine (Gibco). HEK293T and HeLa cells were grown in Dulbecco's modified eagle medium (DMEM, Gibco) with 25 mM D-glucose, 3.7 g/L $NaHCO_3$, 4 mM L-glutamine and supplemented with 10% (v/v) FBS, 100 units/mL penicillin, 100 µg/mL streptomycin, and 2 mM L-glutamine. K562 and HeLa cells are derived from female patients/donors. Jurkat cells are derived from a male patient. HEK293T are derived from a female fetus. All cell lines were grown at 37 °C in the presence of 5%

CO_2. All cell lines were periodically tested for Mycoplasma contamination using the MycoAlert Plus Mycoplasma detection kit (Lonza).

DNA transfections and virus production

Lentivirus was generated by transfecting HEK239T cells with four packaging plasmids (for expression of VSV-G, Gag/Pol, Rev, and Tat, respectively) as well as the transfer plasmid using TransIT-LT1 Transfection Reagent (Mirus Bio). Viral supernatant was harvested two days after transfection and filtered through 0.44 μm PVDF filters and/or frozen prior to transduction.

Cloning of individual guide RNAs

Individual perfectly matched or mismatched sgRNAs were cloned essentially as described previously (4). Briefly, two complementary oligonucleotides (Integrated DNA Technologies), containing the targeting region as well as overhangs matching those left by restriction digest of the backbone with BstXI and BlpI, were annealed and ligated into a version of pU6-sgCXCR4-2 (marked with a puromycin resistance cassette and mCherry, Addgene #46917 (6)), modified to include a BlpI site, digested with BstXI (NEB or Thermo Fisher Scientific) and BlpI (NEB) or Bpu1102I (Thermo Fisher Scientific). The ligation product was transformed into Stellar chemically competent *E. coli* cells (Takara Bio) and plasmid was prepared following standard protocols.

Individual evaluation of sgRNA phenotypes for GFP knockdown

For individual evaluation of GFP knockdown phenotypes, sgRNAs were individually cloned as described above. The sgRNA expression vectors were individually packaged into lentivirus and

transduced into target cells at MOI < 1 (15 – 40% infected cells) by centrifugation at 1,000 × *g* and 33 °C for 0.5-2 h. To measure GFP knockdown, sgRNA expression vectors were transduced into GFP^+ K562 CRISPRi cells (6) and GFP levels were recorded 10 d after transduction by flow cytometry using a FACSCelesta flow cytometer (BD Biosciences), gating for sgRNA-expressing cells ($mCherry^+$). Experiments were performed in duplicate from the transduction step.

Design of large-scale mismatched sgRNA library

To generate the list of targeting sgRNAs for the large-scale mismatched sgRNA library, hit genes from a growth screen performed in K562 cells with the CRISPRi v2 library (7) were selected by calculating a discriminant score (phenotype z-score × $-\log_{10}$(Mann-Whitney *P*)). Discriminant scores for negative control genes (randomly sampled groups of 10 non-targeting sgRNAs) were calculated as well, and hit genes were selected above a threshold such that 5% of the hits would be negative control genes (i.e. an estimated empirical 5% FDR). This procedure resulted in the selection of 2,477 genes. Of these genes, 28 genes for which the second strongest sgRNA by absolute value had a positive growth phenotype were filtered out as these were likely to be scored as hits solely due to a single sgRNA. For the remaining 2,449 genes, the two sgRNAs with the strongest growth phenotype were selected, for a total of 4,898 perfectly matched sgRNAs.

For each of these sgRNAs, a set of 23 variant sgRNAs with mismatches was designed: 5 with a single randomly chosen mismatch within 7 bases of the PAM, 5 with a single randomly chosen mismatch 8-12 bases from the PAM, and 3 with a single randomly chosen mismatch 13-19 bases from the PAM (the first base of the targeting region was never selected for this purpose as it is an

invariant G in all sgRNAs to enable transcription from the U6 promoter). The remaining 10 variants had 2 randomly chosen mismatches selected from positions –1 to –19.

To assess the off-target potential of mismatched sgRNAs, we extended our previous strategy to estimate sgRNA off-target effects (4, 7). Briefly, for each target in the genome, a FASTQ entry was created for the 23 bases of the target including the PAM, with the accompanying empirical Phred score indicating an estimate of the anticipated importance of a mismatch in that base position. Bowtie (http://bowtie-bio.sourceforge.net) was then used to align each designed sgRNA back to the genome, parameterized so that sgRNAs were considered to mutually align if and only if: (i) no more than 3 mismatches existed in the PAM-proximal 12 bases and the PAM, and (ii) the summed Phred score of all mismatched positions across the 23 bases was less than a threshold. This was done iteratively with decreasing thresholds, and any sgRNAs which aligned successfully to no other site in the genome at a particular threshold were then deemed to have a specificity at said threshold. The compiled sgRNA sequences were then filtered for sgRNAs containing BstXI, BlpI, and SbfI sites, which are used during library cloning and sequencing library preparation, and 2,500 negative controls (randomly generated to match the base composition of our hCRISPRi-v2 library) were added. Sequences of sgRNAs and descriptions of mismatches are listed in Table S3.1.

Pooled cloning of mismatched sgRNA libraries

Pooled sgRNA libraries were cloned largely as described previously (4, 8, 9). Briefly, oligonucleotide pools containing the desired elements with flanking restriction sites and PCR adapters were obtained from Agilent Technologies. The oligonucleotide pools were amplified by 15 cycles of PCR using Phusion polymerase (NEB). The PCR product was digested with BstXI and

Bpu1102I (Thermo Fisher Scientific), purified, and ligated into BstXI/Bpu1102I-digested pCRISPRia-v2 at 16 °C for 16 h. The ligation product was purified by isopropanol precipitation and then transformed into MegaX DH10B electrocompetent cells (Thermo Fisher Scientific) by electroporation using the Gene Pulser Xcell system (Bio-Rad), transforming ~100 ng purified ligation product per 100 µL cells. The cells were allowed to recover in 3-6 mL SOC medium for 2 h. At that point, a small 1-5 µL aliquot was removed and plated in three serial dilutions on LB plates with selective antibiotic (carbenicillin). The remainder of the culture was inoculated into 0.5 to 1 L LB supplemented with 100 µg/mL carbenicillin, grown at 37 °C with shaking at 220 rpm for 16 h and harvested by centrifugation. Colonies on the plates were counted to confirm a transformation efficiency greater than 100-fold over the number of elements (>100x coverage). The pooled sgRNA plasmid library was extracted from the cells by GigaPrep (Qiagen or Zymo Research). Even coverage of library elements was confirmed by sequencing a small aliquot on a HiSeq 4000 (Illumina).

Large-scale mismatched sgRNA screen and sequencing library preparation

Large-scale screens were conducted similarly to previously described screens (4, 7). The large-scale library was transduced in duplicate into K562 CRISPRi and Jurkat CRISPRi cells at MOI <1 (percentage of transduced cells 2 days after transduction: 20-40%) by centrifugation at 1,000 × *g* and 33 °C for 2 h. Replicates were maintained separately in 0.5 L to 1 L of RPMI-1640 in 1 L spinner flasks for the course of the screen. 2 days after transduction, the cells were selected with puromycin for 2 days (K562: 2 days of 1 µg/mL; Jurkat: 1 day of 1 µg/mL and 1 day of 0.5 µg/mL), at which point transduced cells accounted for 80-95% of the population, as measured by flow cytometry using an LSR-II flow cytometer (BD Biosciences). Cells were allowed to recover for 1 day in the

absence of puromycin. At this point, t_0 samples with a 3,000x library coverage (400×10^6 cells) were harvested and the remaining cells were cultured further. The cells were maintained in spinner flasks by daily dilution to 0.5×10^6 cells mL^{-1} at an average coverage of greater than 2,000 cells per sgRNA with daily measurements of cell numbers and viability on an Accuri bench-top flow cytometer (BD BioSciences) for 11 days, at which point endpoint samples were harvested by centrifugation with 3,000x library coverage.

Genomic DNA was isolated from frozen cell samples and the sgRNA-encoding region was enriched, amplified, and processed for sequencing essentially as described previously (7). Briefly, genomic DNA was isolated using a NucleoSpin Blood XL kit (Macherey-Nagel), using 1 column per 100×10^6 cells. The isolated genomic DNA was digested with 400 U SbfI-HF (NEB) per mg DNA at 37 °C for 16 h. To isolate the ~500 bp fragment containing the sgRNA expression cassette liberated by this digest, size separation was performed using large-scale gel electrophoresis with 0.8% agarose gels. The region containing DNA between 200 and 800 bp of size was excised and DNA was purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). The isolated DNA was quantified using a QuBit Fluorometer (Thermo Fisher Scientific) and then amplified by 23 cycles of PCR using Phusion polymerase (NEB) and appending Illumina adapter and unique sample indices in the process. Each DNA sample was divided into 5-50 individual 100 μL reactions, each with 500 ng DNA as input. To ensure base diversity during sequencing, the samples were divided into two sets, with all samples for a given replicate always being assigned to the same set. The two sets had the Illumina adapters appended in opposite orientations, such that samples in set A were sequenced from the 5′ end of the sgRNA sequence in the first 20 cycles of sequencing and samples in set B were sequenced from the 3′ end of the sgRNA sequence in the next 20 cycles of sequencing. With updates

to Illumina chemistry and software, this strategy is no longer required to ensure high sequencing quality, and all samples are amplified in the same orientation. Following the PCR, all reactions for a given DNA sample were combined and a small aliquot (100-300 μL) was purified using AMPure XP beads (Beckman-Coulter) with a two-sided selection (0.65x followed by 1x). Sequencing libraries from all samples were combined and sequencing was performed on a HiSeq 4000 (Illumina) using single-read 50 runs and with two custom sequencing primers (oCRISPRi_seq_V5 and oCRISPRi_seq_V4_3′, Table S3.11). For samples that were amplified in the same orientation, only a single custom sequencing primer was added (oCRISPRi_seq_V5), and the samples were supplemented with a 5% PhiX spike-in.

Sequencing reads were aligned to the library sequences, counted, and quantified using the Python-based ScreenProcessing pipeline (https://github.com/mhorlbeck/ScreenProcessing). Calculation of phenotypes was performed as described previously (4). Untreated growth phenotypes (γ) were derived by calculating the $\log_2$ change in enrichment of an sgRNA in the endpoint and t_0 samples, subtracting the equivalent median value for all non-targeting sgRNAs, and dividing by the number of doublings of the population (4, 9). Read counts and phenotypes for individual sgRNAs are available in Table S3.2 and Table S3.3, respectively. To calculate relative activities, phenotypes of mismatched sgRNAs were divided by those for the corresponding perfectly matched sgRNA. Relative activities were filtered for series in which the perfectly matched sgRNA had a growth phenotype greater than 5 z-scores outside the distribution of negative control sgRNAs for all further analysis. Relative activities from both cell lines were averaged if the series passed the z-score filter in both. All analyses were performed in Python 2.7 using a combination of Numpy (v1.16.1), Pandas (v0.24.2), Scipy (v1.1.0), and Seaborn (v0.9.0).

Design and pooled cloning of constant region variants library

The sequences in the library of modified constant regions were derived from the sgRNA (F+E) optimized sequence (10) modified to include a BlpI site (4). Each modified constant region was paired with 36 sgRNA targeting sequences (3 sgRNAs targeting each of 10 essential genes and six non-targeting negative control sgRNAs). The cloning strategy (described below) allowed the mutation of most positions in the sgRNA constant region. A variety of modifications were made, including substitutions of all single bases not in the BlpI restriction site (which is used for cloning), double substitutions including all substitutions at base-paired position pairs not in the BlpI site, and a variety of triple, quadruple, and sextuple substitutions, including base-pair-preserving substitutions at adjacent base-pairs.

The library was ordered and cloned in two parts. One part consisted of ~100 modifications to the eight bases upstream of the BlpI restriction site. Constant region variants with mutations in this section were paired with each of the 36 targeting sequences, ordered as a pooled oligonucleotide library (Twist Biosciences), and cloned into pCRISPRia-v2 as described above. The second part consisted of ~900 modifications to the 71 bases downstream of the BlpI restriction site. This part was cloned in two steps. First, all 36 targeting sequences were individually cloned into pCRISPRia-v2 as described above. The vectors were then pooled at an equimolar ratio and digested with BlpI (NEB) and XhoI (NEB). The modified constant region variants were ordered as a pooled oligonucleotide library (Twist Biosciences), PCR amplified with Phusion polymerase (NEB), digested with BlpI (NEB) and XhoI (NEB), and ligated into the digested vector pool, in a manner identical to previously published protocols and as described above, except for the different restriction enzymes.

Compact mismatched sgRNA library and constant region library screens

Screens with the compact mismatched sgRNA library and the constant region library were conducted largely as described above, with smaller modifications during the screening procedure and an updated sequencing library preparation protocol. Briefly, the libraries were transduced in duplicate into K562 CRISPRi (both libraries) or HeLa CRISPRi cells (compact mismatched sgRNA library) as described above. K562 replicates were maintained separately in 0.15 to 0.3 L of RPMI-1640 in 0.3 L spinner flasks for the course of the screen. HeLa replicates were maintained in sets of ten 15-cm plates. Cells were selected with puromycin as described above (K562: 1 day of 0.75 μg/mL and 1 day of 0.85 μg/mL; HeLa: 2 days of 0.8 μg/mL and 1 day of 1 μg/mL). The remainder of the screen was carried out at >1,000x library coverage (K562 compact mismatched sgRNA library: >2,000x; HeLa compact mismatched sgRNA library: >1,000x; K562 constant region library: >2,000x). Multiple samples were harvested after 4 to 8 days of growth.

Genomic DNA was isolated from frozen cell samples as described above. The subsequent sequencing library preparation was simplified to omit the enrichment step by gel extraction. In particular, following the genomic DNA extraction, DNA was quantified by absorbance at 260 nm using a NanoDrop One spectrophotometer (Thermo Fisher Scientific) and then directly amplified by 22-23 cycles of PCR using NEBNext Ultra II Q5 PCR MasterMix (NEB), appending Illumina adapter and unique sample indices in the process. Each DNA sample was divided into 50-200 individual 200 μL reactions, each with 10 μg DNA as input. All samples were amplified using the same strategy and in the same orientation. The PCR products were purified as described above and sequencing libraries from all samples were combined. For the compact mismatched library screens, sequencing was performed on a HiSeq 4000 (Illumina) using single-read 50 runs with a 5% PhiX

spike-in and a custom sequencing primer (oCRISPRi_seq_V5, Table S3.11). For the constant region screens, the downstream PCR primer was adapted to allow for amplification of the entire constant region and to append a standard Illumina read 2 primer binding site. Sequencing was then performed in the same manner including oCRISPRi_seq_v5 primer and a 5% PhiX spike-in, but using paired-read 150 runs.

Sequencing reads were processed as described above. Read counts and phenotypes for individual sgRNAs are available in Tables S3.5-S3.6 (constant region screen) and Tables S3.9-S3.10 (compact mismatched sgRNA library screen).

Generation and evaluation of individual constant region variants by RT-qPCR

Constant region variants were evaluated in the background of a constant region with an additional base pair substitution in the first stem loop (fourth base pair changed from AT to GC (11)). Ten constant region variants with average relative activities between 0.2 and 0.8 from the screen and carrying substitutions after the BlpI site were selected (Table S3.11). Cloning of individual constant regions was performed essentially as the cloning of sgRNA targeting regions, described above, except that the BlpI and XhoI restriction sites were used for cloning (the XhoI site is immediately downstream of the constant region) and that cloning was performed with a variant of pCRISPRia-v2 (marked with a puromycin resistance cassette and BFP, Addgene #84832 (7)). For each of the ten constant region variants as well as the constant region carrying only the stem loop substitution, two different targeting regions against *DPH2* were then cloned as described above (Table S3.13). These 22 vectors as well as a vector with a non-targeting negative control sgRNA (Table S3.13) were individually packaged into lentivirus and transduced into K562 CRISPRi cells at

MOI < 1 (10 – 50% infected cells) by centrifugation at 1000 × *g* and 33 °C for 2 h. Cells were allowed to recover for 2 days and then selected to purity with puromycin (1.5 – 3 μg/mL), as assessed by measuring the fraction of BFP^+ cells by flow cytometry on an LSR-II (BD Biosciences), allowed to recover for 1 day, and harvested in aliquots of $0.5 – 2 \times 10^6$ cells for RNA extraction. RNA was extracted using the RNeasy Mini kit (Qiagen) with on-column DNase digestion (Qiagen) and reverse-transcribed using SuperScript II Reverse Transcriptase (Thermo Fisher Scientific) with oligo(dT) primers in the presence of RNaseOUT Recombinant Ribonuclease Inhibitor (Thermo Fisher Scientific). Quantitative PCR (qPCR) reactions were performed in 22 μL reactions by adding 20 μL master mix containing 1.1X Colorless GoTaq Reaction Buffer (Promega), 0.7 mM $MgCl_2$, dNTPs (0.2 mM each), primers (0.75 μM each), and 0.1X SYBR Green with GoTaq DNA polymerase (Promega) to 2 μL cDNA or water. Reactions were run on a LightCycler 480 Instrument (Roche). For each cDNA sample, reactions were set up with qPCR primers against *DPH2* and *ACTB* (sequences listed in Table S3.11). Experiments were performed in technical triplicates.

Machine learning

In order to establish a subset of highly active sgRNAs with which to train a machine learning model, we filtered for perfectly matched guides with a growth phenotype greater than 10 z-scores outside the distribution of negative control sgRNAs in the K562 and/or Jurkat pooled screens (K562 $\gamma < -0.21$; Jurkat $\gamma < -0.35$). All singly mismatched variants derived from sgRNAs passing the filter were then included, and relative activities were calculated as described previously, averaging the replicate measurements for each sgRNA. In cases where a perfectly matched sgRNA passed the filter in the K562 and Jurkat screen, the average relative activity across both cell types was calculated for

each mismatched variant; otherwise the relative activities for only one cell type were considered. This filtering scheme resulted in 26,248 mismatched sgRNAs comprising 2,034 series targeting 1,292 genes, with approximately 40% of relative activity values averaged from K562 and Jurkat cells.

For each sgRNA, a set of features was defined based on the sequences of the genomic target and the mismatched guide. First, the genomic sequence extending from 22 bases 5′ of the beginning of the PAM to 1 base 3′ of the end of the PAM (26 bases in all) is binarized into a 2D array of shape (4, 26), with 0s and 1s indicating the absence or presence of a particular nucleotide at each position, respectively. Next, a similar array is constructed representing the mismatch imparted by the sgRNA, with an additional potential mismatch at the 5′ terminus of the sgRNA (position –20), which invariably begins with G in our libraries due to the mU6 promoter. Thus, the mismatched sequence array is identical to the genomic sequence array except for 1 or 2 positions. Finally, the arrays are stacked into a 3D volume of shape (4, 26, 2), which serves as the feature set for a particular sgRNA.

The training set of sgRNAs was established by randomly selecting 80% of guide series, with the remaining 20% set aside for model validation. A convolutional neural network (CNN) regression model was then designed using Keras (https://keras.io/) with a TensorFlow backend engine, consisting of two sequential convolution layers, a max pooling layer, a flattening layer, and finally a three-layer fully connected network terminating in a single neuron. Additional regularization was achieved by adding dropout layers after the pooling step and between each fully connected layer. To penalize the model for ignoring under-represented sgRNA classes (e.g. those with intermediate relative activity), training sgRNAs were binned according to relative activity, and sample weights inversely proportional to the population in each bin were assigned. Hyperparameters were optimized using a randomized grid search with 3-fold cross-validation with the training set as

input. Parameters included the size, shape, stride, and number of convolution filters, the pooling strategy, the number of neurons and layers in the dense network, the extent of dropout applied at each regularization step, the activation functions in each layer, the loss function, and the model optimizer. Ultimately, 20 CNN models with identical starting parameters were individually trained for 8 epochs in batches of 32 sgRNAs. Performance was assessed by computing the average prediction of the 20-model ensemble for each validation sgRNA and comparing it to the measured value.

A linear regression model was trained on the same set of sgRNAs, albeit with modified features more suited for this approach. These features include the identities of bases in and around the PAM, whether the invariant G at the 5′ end of the sgRNA is base paired, the GC content of the sgRNA, the change in GC content due to the point mutation, the location of the protospacer relative to the annotated transcription start site, the identities of the 3 RNA bases on either side of the mismatch, and the location and type of each mismatch. All features were binarized except for GC and delta GC content. In total, each guide was represented by a vector of 270 features, 228 of which describe the mismatch position and type (19 possible positions by 12 possible types). Prior to training, feature vectors were z-normalized to set the mean to 0 and variance to 1. Finally, an elastic net linear regression model was created using the scikit-learn Python package (https://scikit-learn.org), and key hyperparameters (alpha and L1 ratio) were optimized using a grid search with 3-fold cross validation during training.

Design of compact library

Genes targeted by the compact allelic series library were required to have at least one perfectly matched sgRNA with a growth phenotype greater than 2 z-scores outside the distribution of negative control sgRNAs ($\gamma < -0.04$) in a single replicate of a K562 pooled screen (this work or (7)). By this metric, 4,722 unique sgRNAs targeting 2,405 essential genes were included. Next, for each perfectly matched sgRNA, variants containing all 57 single mismatches in the targeting sequence (positions –19 to ⁻1) were generated *in silico*, and sequences with off-target binding potential in the human genome were filtered out as described for the large-scale library. Remaining variant sgRNAs were whitelisted for potential selection in subsequent steps.

For each gene being targeted, if both of the perfectly matched sgRNAs imparted growth phenotypes greater than 3 z-scores outside the distribution of negative controls ($\gamma < -0.06$) in this work's large-scale K562 screen, then one series of 4 variant guides was generated from each. Otherwise, one series of 8 variants was generated from the sgRNA with the stronger phenotype. Both perfectly matched sgRNAs were included regardless of their growth phenotype, for a total of 2 perfectly matched and 8 mismatched sgRNAs per gene.

In order to select mismatched guides, we first divided the relative activity space into 6 bins with edges at 0.1, 0.3, 0.5, 0.7, and 0.9. For each series, we attempted to select sgRNAs from each of the middle 4 bins (centers at 0.2, 0.4, 0.6, and 0.8 relative activity) as measured in this work's K562 screen. If multiple sgRNAs were available in a particular bin, they were prioritized based on distance to the center of the bin and variance between replicate measurements. If no previously measured sgRNA was available in a given bin, then the CNN model was run on all whitelisted (novel) mismatched sgRNAs belonging to that series, and sgRNAs were selected based on predicted activity

as needed. In total, the compact library was composed of 4,722 unique perfectly matched sgRNAs, 19,210 unique mismatched sgRNAs, and 1,202 non-targeting control sgRNAs. Approximately 68% of mismatched sgRNAs were evaluated in previous screens (72% single mismatches, 28% double mismatches), with the remaining 32% imputed from the CNN model (all single mismatches). Sequences of sgRNAs and descriptions of mismatches are listed in Table S3.8.

Perturb-seq

The perturb-seq experiment targeted 25 genes involved in a diverse range of essential functions (Table S3.12). For each target gene, the original sgRNAs and 4-5 mismatched sgRNAs covering the range from full relative activity to low relative activity were chosen from the large-scale screen. These 128 targeting sgRNAs as well as 10 non-targeting negative control sgRNAs were individually cloned into a modified variant of the CROP-seq vector (12, 13) as described above, except into the different vector. Lentivirus was individually packaged for each of the 138 sgRNAs and was harvested and frozen in array. To determine viral titers, each virus was individually transduced into K562 CRISPRi cells by centrifugation at 1,000 × *g* and 33 °C for 2 h, and the fraction of transduced cells was quantified as BFP^+ cells using an LSR-II flow cytometer (BD Biosciences) 48 h after transduction.

To generate transduced cells for single-cell RNA-seq analysis, virus for all 138 sgRNAs was pooled immediately before transduction and then transduced into K562 CRISPRi cells by centrifugation at 1,000 × *g* and 33 °C for 2 h. To achieve even representation at the intended time of single-cell analysis, the virus pooling was adjusted both for titer and expected growth-rate defects. 3 d after transduction, transduced (BFP^+) cells were selected using FACS on a FACSAria2 (BD

Biosciences) and then resuspended in conditioned media (RPMI formulated as described above except supplemented with 20% FBS and 20% supernatant of an exponentially growing K562 culture). 2 d after sorting, the cells were loaded onto three lanes of a Chromium Single Cell 3′ V2 chip (10x Genomics) at 1,000 cells/μL and processed according to the manufacturer's instructions. The CROP-seq sgRNA barcode was PCR amplified from the final single cell RNA-seq libraries with a primer specific to the sgRNA expression cassette (oBA503, Table S3.11) and a standard P5 primer (Table S3.11), purified on a Blue Pippin 1.5% agarose cassette (Sage Science) with size selection range 436-534 bp, and pooled with the single cell RNA-seq libraries at a ratio of 1:100. The libraries were sequenced on a HiSeq 4000.

To measure the growth rate defects conferred by each sgRNA for comparison with the transcriptional phenotypes, samples of ~500,000 transduced cells were taken from the same transduced cell population used in the perturb-seq experiment on days 2, 7, and 12 after transduction. Genomic DNA was extracted using the Nucleospin Blood kit (Macherey-Nagel) and sgRNA amplicons were prepared as described previously and above (7), albeit with no genomic DNA digestion or gel purification, and sequenced on HiSeq 4000 as described above.

Perturb-seq analysis

i. Cell barcode and UMI calling, assignment of perturbations

UMI count tables with UMI counts for all genes in each individual cell were calculated from the raw sequencing data using CellRanger 2.1.1 (10x Genomics) with default settings. Perturbation calling was performed as in (14). Briefly, reads from the specifically amplified sgRNA barcode libraries were aligned to a list of expected sgRNA barcode sequences using bowtie (flags: -v3 -q -m1).

Reads with common UMI and barcode identity were then collapsed to counts for each cell barcode, producing a list of possible perturbation identities contained by that cell. A proposed perturbation identity was identified as "confident" if it met thresholds derived by examining the distributions of reads and UMIs across all cells and candidate identities: (i) reads > 50, (ii) UMIs > 3, and (iii) coverage (reads/UMI) in the upper mode of the observed distribution across all candidate identities. As described in (15), perturbation identities were called for any cell barcode with greater than 2,000 UMIs to enable capture of cells with strong growth defects. Any cell barcode containing two or more confident identities was deemed a "multiplet", and may arise from either multiple infection or simultaneous encapsulation of more than one cell in a droplet during single-cell RNA sequencing. Cell barcodes passing the 2,000 UMI threshold and bearing a single, unambiguous perturbation barcode were included in all subsequent analyses.

ii. Expression normalization

Some portions of analysis use normalized expression data. We used a relative normalization procedure based on comparison to the gene expression observed in control cells bearing non-targeting sgRNAs, as in (14):

1. Total UMI counts for each cell barcode are normalized to have the median number of UMIs observed in control cells.
2. For each gene x, expression across all cell barcodes is z-normalized with respect to the mean (μ_x) and standard deviation (σ_x) observed in control cells:

$$x_{\text{normalized}} = \frac{x - \mu_x}{\sigma_x}$$

Following this normalization, control cells have average expression 0 (and standard deviation 1) for all genes. Negative/positive values therefore represent under/overexpression relative to control.

iii. Target gene quantification

Expression levels of genes targeted by a given sgRNA were quantified by normalizing UMI counts of the targeted gene to the total UMI count for each individual cell (Fig. S3.8). Considering raw UMI counts of the targeted gene (Fig. S3.9) or z-normalized target gene expression as described above yielded similar results. Note that the sgRNA targeting *BCR* is toxic due to knockdown of the *BCR-ABL1* fusion present in K562 cells. Knockdown was apparent both in *BCR* and *ABL1* expression, but we used *BCR* expression for further analysis as there are likely additional copies of *ABL1* that are not fused to *BCR* (and thus would not be affected by the *BCR*-targeting sgRNA) contributing to *ABL1* expression.

iv. Cell cycle analysis

Calling of cell cycle stages was performed using a similar approach to (16) and largely as described in (14). Briefly, lists of marker genes showing specific expression in different cell cycle stages from the literature were first adapted to K562 cells by restricting to those that showed highly correlated expression within our experiment. The total (log_2-normalized) expression of each set of marker genes was used to create scores for each cell cycle stage within each cell, and these scores were then z-normalized across all cells. Each cell was assigned to the cell cycle stage with the highest score.

v. Differential gene expression analysis

We took two approaches to differential expression, as described in (15). For both approaches, we only considered genes with expression greater than 0.25 UMIs per cell on average across all cells. First, for a given gene, we could assess the changes in the expression distribution of that gene induced by a given genetic perturbation by comparing to the expression distribution observed in control cells bearing non-targeting sgRNAs. We performed this comparison using a two-sample

Kolmogorov-Smirnov test and corrected for multiple hypothesis testing at an FDR of 0.001 using the Benjamini-Yekutieli procedure.

We also exploited a machine learning approach that potentially allows correlated expression patterns to be detected and that scales beyond two sample comparisons. Perturbed cells and control cells bearing non-targeting sgRNAs were each used as training data for a random forest classifier that was trained to predict which sgRNA a cell contained from its transcriptional state. As part of the training process the classifier ranks which genes have the most prognostic power in predicting sgRNA identity, which by construction will tend to vary across condition. For most further analysis, the top 100-300 genes by prognostic power were then considered.

vi. Constructing mean expression profiles

For some analyses, expression profiles were averaged across all cells with the same perturbation. In general, this was done simply by calculating the mean z-normalized expression of all genes with mean expression level of 0.25 UMI or higher across all cells in the experiment or within the specific considered subpopulation (usually all cells with sgRNAs targeting a given gene as well as all control cells with non-targeting sgRNAs).

vii. UMAP dimensionality reduction

For UMAP dimensionality reduction of all cells, the 300 genes with the highest prognostic power in distinguishing cells by targeted gene as ranked by a random forest classifier were selected. Dimensionality reduction was then performed on the z-normalized single-cell expression profiles of these 300 genes using the following parameters: n_neighbors = 40, min_dist = 0.1, metric = 'euclidean', spread = 1.0. UMAP dimensionality reduction of subpopulations containing only cells

with perturbation of a given gene or control cells was performed analogously but using the expression profiles of the 100 genes with the highest prognostic power and using n_neighbors = 15.

viii. ISR scores

Magnitude of ISR activation in individual cells was quantified as activation of the PERK (*EIF2AK3*) regulon from the gene set and activation coefficients determined in (14).

Results

Mismatched sgRNAs mediate diverse intermediate phenotypes

To comprehensively characterize the activities of mismatched sgRNAs in CRISPRi-mediated knockdown, we introduced all 57 singly mismatched variants of a GFP-targeting sgRNA (6) into GFP$^+$ K562 CRISPRi cells and measured GFP levels by flow cytometry (Fig. 3.1a). Cells harboring mismatched sgRNAs experienced knockdown levels between those of cells with the perfectly matched sgRNA (94%) and cells with a non-targeting control sgRNA (Fig. 3.1b, S3.1a, S3.1b). As expected, sgRNAs with mismatches in the PAM-proximal seed region had strongly compromised activity. By contrast, sgRNAs with mismatches in the PAM-distal region effected GFP knockdown to an extent similar to that of the unmodified sgRNA, albeit with substantial variability depending on the type of mismatch (Fig. 3.1b-c). The distributions of GFP levels with mismatched sgRNAs were largely unimodal, although the distributions were broader than with the perfectly matched sgRNA or the control sgRNA (Fig. 3.1b, S3.1b). These results suggest that series of mismatched sgRNAs can be used to titrate gene expression at the single-cell level, but that mismatched sgRNA activity is modulated by complex factors.

Rules of mismatched sgRNA activity derived from a large-scale screen

We reasoned that by measuring growth phenotypes imparted by mismatched sgRNAs in a pooled screen, we could empirically derive the factors governing the influence of mismatches on sgRNA activity. To create a library of mismatched variants for this screen, we selected 2 sgRNAs each against 2,499 genes for with growth phenotypes in K562 cells (7). For each sgRNA, we generated series of 22-23 variant sgRNAs with one or two mismatches and cloned these sgRNAs, including the original, perfectly matched sgRNA, into a pooled library comprising ~120,000 sgRNAs in 4,898 series (Fig. 3.2a, Table S3.1). We transduced K562 (chronic myelogenous leukemia) and Jurkat (acute T-cell lymphocytic leukemia) cells expressing dCas9-KRAB with the sgRNA library, harvested subpopulations at the outset of the experiment (t_0) and after 11 days of growth, and then measured the difference in relative abundance of each sgRNA between the end point and t_0 populations by next-generation sequencing to quantify how each sgRNA affects growth (γ), with a more negative value of γ indicating a stronger growth defect (4, 9) (Fig. 3.2b). Growth phenotypes of targeting sgRNAs were well-correlated in biological replicates (Fig. S3.2a-b, Tables S3.2-S3.3, Pearson r^2 [K562] = 0.82; Pearson r^2 [Jurkat] = 0.82), and growth phenotypes of perfectly matched sgRNAs in K562 cells were well-correlated with previously reported phenotypes (7) (Fig. S3.2c, Pearson r^2 = 0.86). Growth phenotypes measured in K562 and Jurkat cells were more different (Pearson r^2 = 0.37), likely reflecting cell-type specific gene essentiality (Fig. S3.2d).

Examining the phenotype patterns within sgRNA series revealed that mismatched sgRNAs mediate a range of phenotypes, spanning from that of the corresponding perfectly matched sgRNA to those of negative control sgRNAs (Fig. 3.2c). To quantify the effects of mismatches, we normalized the phenotype of each mismatched sgRNA to that of its perfectly matched sgRNA

(relative activity, Fig. 3.2b) and filtered for series in which the perfectly matched sgRNA had a strong growth phenotype (see Methods, 3,147 sgRNA series for K562 cells, 2,029 sgRNA series for Jurkat cells). The resulting relative activities derived from phenotypes measured in K562 and Jurkat cells were well-correlated (Fig. 3.2d, Pearson $r^2 = 0.71$), regardless of differences in absolute phenotype of the perfectly matched sgRNAs (Pearson $r^2 = 0.74$ for $|\gamma[\text{K562}] - \gamma[\text{Jurkat}]| > 0.2$; Pearson $r^2 = 0.70$ for $|\gamma[\text{K562}] - \gamma[\text{Jurkat}]| < 0.2$). We therefore averaged relative activities from both cell lines for further analysis (see Methods). The majority of mismatched sgRNAs were inactive (Fig. 3.2d), with a particularly high proportion of inactive sgRNAs among sgRNAs with two mismatches (Fig. S3.2e). Nonetheless, many mismatched sgRNAs exhibited intermediate activity (19,596 sgRNAs with 0.1 < relative activity < 0.9, 25.5% of sgRNAs in series passing filter).

To further understand the rules governing the impact of mismatches on sgRNA activity, we focused on sgRNAs with a single mismatch and stratified the relative activities by various properties of the mismatched sgRNAs. As expected, mismatch position was a strong determinant of activity, with sgRNAs carrying mismatches closer to the PAM having lower relative activity (Fig. 3.2e). The exact nucleotide substitution also had a strong effect, with rG:dT mismatches (A to G mutations in the sgRNA) retaining substantial activity even for mismatches close to the PAM (Fig. 3.2f). Other factors were of lower magnitude or had more context dependence. For example, sgRNAs with higher GC content retained higher activity for mismatches located 9 or more bases upstream of the PAM (positions –9 to –19), and mismatched sgRNAs with G nucleotides surrounding the site of the mismatch retained marginally higher activity for some positions (Fig. S3.2f-g). Activity of mismatched sgRNAs thus appears to be determined by general biophysical rules; a premise further supported by the high correlation of relative activities obtained in two different cell lines (Fig. 3.2d)

and the high correlation of mismatched sgRNA activities with previous *in vitro* measurements of dCas9 binding on-rates in the presence of mismatches (17) (Fig. 3.2g).

Finally, we evaluated how many sgRNA series provide access to multiple intermediate CRISPRi growth phenotypes and thereby likely enable titrating the expression of the targeted gene. With an intermediate phenotype defined as relative activity between 0.1 and 0.9, 76.1% of series contained at least 2 sgRNAs with intermediate phenotypes when only considering single mismatches, and 86.7% of series did so when also including double mismatches (Fig. S3.2h). Given that we explored ~20% of possible single mismatches and <1% of possible double mismatches, it is likely that intermediate-activity sgRNAs also exist for the remaining series. Altogether, these results suggest that systematically mismatched sgRNAs provide a general method to titrate CRISPRi activity and, consequently, target gene expression.

Controlling sgRNA activity with modified constant regions

We also explored the orthogonal approach of generating intermediate-activity sgRNAs through modifications to the sgRNA constant region, which is required for binding to Cas9. Previous work has established that such modifications have varied effects, including increases and decreases in Cas9 activity; or no measurable impact (5, 10, 11, 14, 18, 19). In these examples the mutational landscape of the constant region was only sparsely explored, and largely with the goal of preserving sgRNA activity. We therefore reasoned that saturation mutagenesis of the constant region could identify variants with intermediate activity.

To comprehensively assess the activities of modified sgRNA constant regions, we designed a library of 995 constant region variants comprising all possible single nucleotide substitutions, substitutions of base pairs for other base pairs, and combinations of these changes (see Methods),

and then determined the growth phenotypes for each constant region variant paired with 30 different targeting sequences against 10 essential genes in a pooled K562 screen (Fig. 3.3a, S3.3a; Tables S3.4-S3.6). We then calculated the relative activities for all targeting sequence:constant region pairs by normalizing phenotypes to those of the unmodified constant region, identifying 409 constant region variants that on average conferred intermediate activity (0.1-0.9, Fig. 3.3b). Ten variants selected for individual evaluation also mediated intermediate mRNA knockdown (Fig. S3.3b). Mapping the activities of constant region variants with single base substitutions onto the structure recapitulated known relationships between constant region structure and function (Fig. 3.3c). For example, mutation of bases known to mediate contacts with Cas9 (5) (e.g. the first stem loop or the nexus) generally reduced activity, whereas mutations in regions not contacted by Cas9 (e.g. the hairpin region of stem loop 2) were well-tolerated (Fig. 3.3c). Notably, several variants carrying mutations in stem loop 2 had consistently increased activities and thus could be useful tools in future applications (Fig. 3.3b-c).

We next evaluated the relative activities of constant region variants across different targeting sequences. Although the rank ordering of variant activities was largely consistent, the actual relative activities were more variable (Fig. 3.3d, S3.3c). For example, a targeting sequence against *TUBB* retained high activity with ~100 constant region variants that otherwise abolished activity, whereas a targeting sequence against *SNRPD2* lost activity with ~50 constant region variants that otherwise conferred intermediate activity (Fig. 3.3d). This heterogeneity extended to different targeting sequences against the same gene in some cases, both at the level of growth phenotype (Fig. 3.3e-g, S3.3d-e) and mRNA knockdown (Fig. S3.3b). These differences between targeting sequences could be a consequence of specific targeting sequence:constant region structural interactions or of

differences in basal sgRNA expression levels such that lowly expressed sgRNAs are more susceptible to constant region modifications. These analyses suggest that modifications to the constant region can be used to titrate sgRNA activity, but the activity of a given constant region variant with a given targeting sequence is difficult to predict. We therefore focused on sgRNAs with mismatches in the targeting region for the remainder of our work, given that these sgRNAs conferred intermediate activity in a more predictable manner.

A neural network predicts mismatched sgRNA activities with high accuracy

We next sought to leverage our large-scale data set of mismatched sgRNA activities to learn the underlying rules in a principled manner and to enable predictions of intermediate-activity sgRNAs against other genes. We reasoned that a convolutional neural network (CNN) approach would be well-suited to uncovering these rules due to the ability of CNNs to learn complex global and local dependencies on spatially-ordered features such as nucleotide sequences (20), including factors governing guide RNA activity in orthogonal CRISPR systems (21). To train a CNN model, we converted singly-mismatched variants derived from perfectly matched sgRNAs into a binarized 2D array of shape (4, 26), representing the mutated targeting sequence flanked by 2 upstream and 4 downstream genomic bases (Fig. 3.4a). This array was then stacked onto a nearly identical array representing the sequence of the respective genomic locus, yielding a 3D volume of shape (4, 26, 2) for each mismatched sgRNA that differs in 1 or 2 positions along the depth axis, depending on whether the invariant G at the 5′ terminus of the sgRNA is base paired. The sgRNA series were randomly split into a training set (80%) and a validation set (20%), yielding a final training set composed of a feature array (X) of shape (21007, 4, 26, 2) mapped to a target vector (y) of 21,007 relative activity measurements.

We constructed a model consisting of two convolution steps, a pooling step, and a 3-layer fully connected neural network (Fig. 3.4a, S3.4a) and optimized hyperparameters using a randomized grid search. We then trained 20 independent models initialized with the same parameters for 8 epochs, which minimized loss without extensive over-fitting (Fig. S3.4b). Predicted and measured relative activities were well-correlated (Pearson r^2 = 0.65), with mean predictions of the 20-model ensemble outperforming all individual models (Fig. 3.4b, S3.4c). Moreover, the distribution of correlation coefficients for individual sgRNA series was unimodal with Pearson *r* values in the 25th-75th percentile ranging from 0.77 to 0.93, indicating that the model performed comparably well for most series (Fig. 3.4c). Model accuracy varied by mismatch position and type, with the highest accuracies corresponding to mismatches in the seed region (Fig. S3.4d-e). Despite the fact that the model was trained on relative growth phenotypes, it also accurately predicted relative fluorescence values measured in the GFP experiment (Fig. 3.4d), further supporting the hypothesis that relative growth phenotypes report on biophysical attributes of specific sgRNA:DNA interactions.

To apply our model more broadly to the human genome, we generated all 57 singly-mismatched sgRNAs for the top 5 sgRNAs against each gene in the hCRISPRi-v2.1 library (7) *in silico* and predicted their relative activities using the CNN ensemble (Table S3.7). Based on the accuracy of predictions for the validation set, we estimate that for any given gene, sampling 5 sgRNAs with intermediate predicted relative activity (0.1-0.9) will yield at least one sgRNA in that activity range over 90% of the time (Fig. S3.4f-i). This resource should therefore enable generation of rationally constructed sgRNA libraries capable of titrating the knockdown of any gene(s) of interest.

Finally, we sought to further understand the features of mismatched sgRNAs that contribute most to their activity. As the contributions of individual features in a deep learning model are difficult to assess directly, we also trained an elastic net linear regression model on the same data using a curated set of features. This linear model explained less variance in relative activities than the CNN model (r^2 = 0.52, Fig S5a-b), implying that our feature set was incomplete and/or sgRNA activity is partly determined by non-linear combinations of features. Nonetheless, the relative activities predicted by the different models were well-correlated (r^2 = 0.74, Fig. S3.5c); therefore, the most important features determining sgRNA activity should correspond to those most heavily weighted in the linear model. Consistent with our earlier observations, mismatch position and type were assigned the largest absolute weights in the model, although other features such as GC content in the sgRNA and the identities of flanking bases up to 3 nucleotides away from the mismatch were heavily weighted as well (Fig. S3.5d-e). For any given position, the type of mismatch contributed differentially to the prediction, with the largest variation between types occurring in the intermediate region of the targeting sequence (Fig. S3.5f). Taken together, these data demonstrate that the activities of mismatch-containing sgRNAs are determined by multiple factors which can be gleaned using supervised machine learning approaches.

A compact mismatched sgRNA library conferring intermediate growth phenotypes

We next set out to design a compact version of our large-scale library to enable titration of essential genes with a small number of sgRNAs. We selected 2,405 genes, divided the relative activity space into six bins, and attempted to select mismatched variants from each of the center four bins (relative activities 0.1-0.9) for two sgRNA series targeting each gene. If a bin did not contain a previously measured sgRNA, we generated feature arrays for all 57 singly-mismatched sgRNAs

derived from the perfectly matched sgRNA and predicted their activities using the CNN model ensemble. After filtering for off-target binding potential, these novel sgRNAs were included as needed based on their predicted activity (Fig. 3.5a). For each gene, 2 perfectly matched and 8 mismatched sgRNAs were selected for the library, with approximately 32% of mismatched sgRNAs imputed from the CNN model (Fig. S3.6a-c).

We evaluated the relative activities of sgRNAs in the compact library using pooled growth screens in K562 and HeLa (cervical carcinoma) CRISPRi cells. Growth phenotypes were well-correlated in biological replicates from samples harvested at different time points after t_0 in both cell lines and less correlated between cell lines (Fig. S3.6d-f). The CNN model predicted imputed sgRNA activities with lower accuracy than the large-scale validation (Fig. S3.6g), although we note that imputed guides were highly enriched in PAM-distal mutations which are associated with higher model errors (Fig. S3.6b, S3.4e). Whereas the majority of mismatched sgRNAs in the large-scale screen had little to no activity, relative activities in the compact library were evenly distributed, ranging from inactive to full activity (Fig. 3.5b). Relative sgRNA activities were also reasonably well-correlated between K562 and HeLa cells (r^2 = 0.58, Fig. 3.5c), suggesting that intermediate phenotypes should be achievable for most genes in multiple cell types using this library.

Exploring essential gene loss-of-function phenotypes with sgRNA allelic series

Finally, we sought to use intermediate-activity sgRNAs to explore the relationship between gene expression levels and biological phenotype. To simultaneously measure gene expression levels and obtain rich phenotypes for a variety of sgRNA series, we used perturb-seq, an experimental strategy developed by us and others that enables matched capture of the transcriptome and the identity of an expressed sgRNA for each individual cell in pools of cells (12, 14, 22, 23) (Fig. S3.7a).

We targeted 25 essential genes involved in diverse cell biological processes (Table S3.12) and chose series of sgRNAs consisting of a perfectly matched sgRNA and 4-5 variants with intermediate growth phenotypes (138 sgRNAs total including 10 non-targeting controls, Table S3.13). We cloned each sgRNA into a modified CROP-seq vector (12, 13), transduced these vectors into K562 CRISPRi cells, and subjected the cells to single-cell RNA-seq (scRNA-seq) 5 days after transduction. In total, we captured transcriptomes for ~23,600 cells with a median number of ~31,500 transcripts per cell (Fig. S3.7b) and assigned single sgRNA identities to ~19,600 cells (83%, Fig. S3.7b) with a median number of 122 cells per sgRNA (Fig. S3.7c). We also quantified relative sgRNA abundances in the cell population after 5 and 10 days of growth to determine the growth phenotypes and relative activities conferred by each sgRNA in this vector (Fig. S3.7d-e). We used sgRNA relative activities determined after 5 days for further analyses.

We first used the scRNA-seq data to assess the expression of the gene targeted by each sgRNA series. To account for cell-to-cell variability in transcript capture efficiency, we quantified target gene UMIs as a fraction of total UMIs in a given cell (Fig. S3.8), although analyzing raw UMI counts yielded similar results (Fig. S3.9). Owing to the limited capture efficiency of scRNA-seq, for lowly expressed genes such as *CAD* and *COX11* we typically observed 0-4 UMIs per cell, with a shift to lower UMI numbers with increasing sgRNA activity (Fig. S3.8, S3.9). For genes with higher basal expression levels such as *HSPA9* and ribosomal genes, the target gene expression distributions are more clearly apparent (Fig. 3.6a, S3.8). These distributions are largely unimodal, with medians shifting downwards with increasing sgRNA activity (Fig. 3.6a). Notably, for ribosomal genes and a few other highly expressed genes, two populations with different knockdown levels are apparent (Fig. 3.6a, S3.8). These populations are present both with intermediate-activity sgRNAs and the perfectly

matched sgRNAs, suggesting that they are not a consequence of limited knockdown penetrance for intermediate-activity sgRNAs.

Beyond expression levels of the targeted gene, titration is also apparent at the level of the transcriptome phenotypes. In the simplest cases, knockdown of *POLR2H*, a core subunit of RNA polymerase II (as well as RNA polymerases I and III), *GATA1*, a central myeloid lineage transcription factor, or to a lesser extent *BCR*, which is fused to the driver oncogene *ABL1* in K562 cells, led to substantial reductions in cellular UMI counts, consistent with global inhibition of mRNA transcription (Fig. 3.6b, Fig. S3.10a). Cells with mismatched sgRNAs against these genes had intermediate UMI counts (Fig. 3.6b), with the extent of reduction in UMI counts exhibiting a linear relationship with growth phenotype (Fig. S3.10b), but non-linear relationships with target gene knockdown at least in the cases of *GATA1* and *POLR2H* (Fig. 3.6c, S3.10b). In particular, both relationships appear to be sigmoidal but with different thresholds: whereas cellular UMI counts drop rapidly once *GATA1* mRNA levels are reduced by 50%, a larger reduction of *POLR2H* mRNA levels is required to achieve a similarly sized effect. Knockdown of most other targeted genes did not perturb total UMI counts to the same extent (Fig. S3.10a) but nonetheless resulted in measurable phenotypes. Knockdown of *CAD*, for example, triggered cell cycle stalling during S-phase, as had been observed previously (14), with a higher frequency of stalling with increasing sgRNA activity (Fig. S3.10c). Finally, knockdown of most of the targeted genes induced strong transcriptional phenotypes (Fig. S3.10d). For example, knockdown of *HSPA9*, the mitochondrial Hsp70 isoform, induced the expected transcriptional signature corresponding to activation of the integrated stress response (ISR) including upregulation of *DDIT3* (CHOP), *DDIT4*, *ATF5*, and *ASNS* (asparagine synthetase), among others (24). The magnitude of this transcriptional signature increased with

increasing sgRNA activity both on the bulk population level (Fig. 3.6d) and on the single-cell level (Fig. 3.6e), although populations with intermediate-activity sgRNAs had larger cell-to-cell variation in the magnitudes of transcriptional responses. Similarly, the transcriptional responses to knockdown of other genes scaled with sgRNA activity and exhibited larger variance for intermediate-activity sgRNAs (Fig. 3.6e). These results highlight the ability of intermediate-activity sgRNAs to provide access to diverse cellular states.

We next explored the relationships between target gene expression and two metrics of phenotype, growth in bulk and transcriptional response. Within each series, the relative magnitudes of the transcriptional responses appear to be well-correlated with growth phenotype, despite substantial differences in the absolute magnitudes of the transcriptional responses with different series (Fig. 3.6f, S3.10d-f). By contrast, the relationship between magnitude of transcriptional changes and target knockdown is more varied (Fig. 3.6g, S3.10g). For *HSPA5* and *GATA1*, for example, a comparably small reduction in mRNA levels was sufficient to induce a near-maximal transcriptional response, whereas for most other genes a larger reduction was required. These relationships are similarly apparent when comparing the target gene knockdown to growth phenotype, with small reductions in *HSPA5* or *GATA1* levels triggering strong growth defects (Fig. S3.10h). These results prompt the hypothesis that expression of *GATA1* and *HSPA5* is more limiting to growth of K562 cells, with sharp transitions from growth to death once expression drops below a threshold. More broadly, these results highlight the utility of titrating gene expression with intermediate-activity sgRNAs on the single-cell level to systematically map the relationships between gene expression and phenotype.

To gain further insight into the diversity of transcriptional responses induced by depletion of essential genes, we compared the transcriptional profiles of all perturbations in this experiment. Clustering all individual perturbations according to the similarity (Pearson correlation) of their bulk transcriptomes revealed multiple groups of perturbations segregated by biological function, including a cluster of perturbations of ribosomal subunits and *POLR1D*, a subunit of the rRNA-transcribing RNA polymerase I (and of RNA polymerase III), and a cluster of perturbations that activate the integrated stress response (*HSPA9*, *HSPE1*, and *EIF2S1*/eIF2α) (Fig. S3.11a). To further visualize the space of transcriptional states, we performed dimensionality reduction on the single-cell transcriptomes using UMAP (25). The resulting projection of individual cells recapitulates the results from the clustering, as indicated for example by the close spatial proximity of cells with perturbations of *HSPA9*, *HSPE1*, and *EIF2S1* (Fig. 3.6h). Within individual series, cells project further outward in UMAP space with increasing sgRNA activity, further highlighting that target gene expression levels are titrated on the single cell level (Fig. 3.6i). We did notice that ~5% cells had mis-assigned sgRNA identities (evident e.g. within the cluster of cells with *HSPA5* knockdown). These cells had confidently assigned single perturbations and only expressed the corresponding barcode transcript, suggesting that they did not evade our doublet detection algorithm. We speculate that these cells expressed two different sgRNAs but silenced expression of one of the reporter transcripts. Given the strong trends in the results above, we concluded that this rate of mis-assignment did not substantially affect our ability to identify trends within cell populations. Together, these analyses reveal a large diversity of transcriptional responses in response to inhibiting central cell biological processes to various degrees.

Closer examination of the UMAP projection revealed more granular structure, including the grouping of a subset of cells with knockdown of *ATP5E*, a subunit of ATP synthase, with cells with ISR-activating perturbations (Fig. 3.6h). This subset of cells indeed exhibited classical features of ISR activation (Fig. S3.11b), but not all cells with *ATP5E* knockdown exhibited this phenotype (Fig. 3.6j, S11b). The frequency of ISR activation increased with lower *ATP5E* mRNA levels (Fig. 3.6j), but even at the lowest mRNA levels some cells did not exhibit ISR activation. These results suggest that depletion of ATP synthase under these conditions predisposes cells to activate the ISR, perhaps by exacerbating transient phases of mitochondrial stress, in a manner that is proportional to ATP synthase levels. More broadly, these results highlight the utility of titrating gene expression in probing cell biological phenotypes, especially in combination with rich phenotyping methods such as scRNA-seq.

Discussion

Here we describe the development of allelic series of sgRNAs to titrate gene expression in human cells, with a broad range of applications across basic and biomedical research. We highlight the utility of the approach in extracting rich phenotypes by single-cell RNA-seq along a continuum of gene expression levels, which enabled mapping of expression levels to various phenotypes and identification of expression level-dependent cell fates.

Our approach builds on *in vitro* work characterizing the effects of sgRNA modifications on (d)Cas9 binding on-rates and activity (2, 17, 26–28). The effects of mismatches in cells follow the biophysical principles established by these studies, enabling us to apply machine learning approaches to derive the underlying rules and predict series for arbitrary sgRNAs. The resulting genome-wide *in*

silico library (Table S3.7) enables titration of any expressed gene of interest. We also describe a compact 25,000-element library that is composed largely of empirically validated sgRNAs and enables titration of ~2,400 essential genes (Table S3.8), with potential applications for example in focused chemical-genetic screens or in profiling interactions between genes with extremely strong growth phenotypes. Given that target gene expression levels are largely unimodally distributed in cells with sgRNA series, these sgRNAs can be combined with both single-cell or bulk population readouts. Thus, complex phenotypes as a function of gene expression levels can be recorded by a variety of techniques tailored to the particular question, such as perturb-seq or related techniques, microscopy, bulk metabolomics or proteomics, or targeted cell biological assays, providing substantial experimental flexibility.

In perhaps the most basic application, sgRNA allelic series enable mapping expression-to-phenotype curves, with implications for both evolutionary biology and biomedical research. Indeed, using these sgRNA series to titrate essential gene expression, we found gene-specific expression-to-phenotype relationships: although all genes had a threshold expression level below which cell viability dropped rapidly, the relative locations of these thresholds varied across genes, with K562 cells being particularly sensitive to depletion of GATA1 and HSPA5 by 50% but robust to depletion of most other genes by ~80%. This variability in threshold location suggests different buffering capacities for different genes, in line with previous findings in yeast (29), the logic of which remains unclear. More comprehensive efforts to generate such dose-response curves and determine the extents to which gene expression is buffered across cell models would allow for identification of patterns for different gene sets and processes and thereby begin to reveal the underlying principles that have shaped gene expression levels. Analogous efforts to map such dose-response curves in

cancer cell types could identify specific vulnerabilities as targets for therapeutics and, vice versa, mapping these curves for cancer driver genes or genes underlying specific diseases enables defining the corresponding therapeutic windows, i.e. the required extents of inhibition or restoration, as goals for drug development.

As an additional feature, sgRNA allelic series provide access to a diversity of cell states and phenotypes including loss-of-function phenotypes that otherwise may be obscured by cell death or neomorphic behavior. Thus, our approach enables positioning cells at states of interest, for example to record chemical-gene or gene-gene interactions, or near phenotypic transitions to characterize the transcriptional trajectories. These series will also facilitate recapitulating gene expression levels of disease-relevant states such as haploinsufficiency or partial loss-of-function diseases, enabling systematic efforts to identify suppressors or modifiers as potential therapeutic targets, or modeling quantitative trait loci associated with multigenic traits in conjunction with rich phenotyping to systematically identify the mechanisms by which they interact and contribute to such traits. Finally, such sgRNA series can be equivalently used to titrate dCas9 occupancy and activity in other applications such as CRISPRa or dCas9-based epigenetic modifiers.

In some regards our allelic series approach unlocks the full potential of CRISPRi to allow for titration of single or multiple genes and thereby evaluation of dose-response relationships. This resource should be equally enabling to systematic large-scale efforts and detailed single-gene investigations in drug development, basic cell biology, and functional genomics.

Data Availability

The raw and processed single-cell sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus under accession number GSE132080. Supplementary tables containing processed screening data will be available pending publication in a peer-reviewed journal.

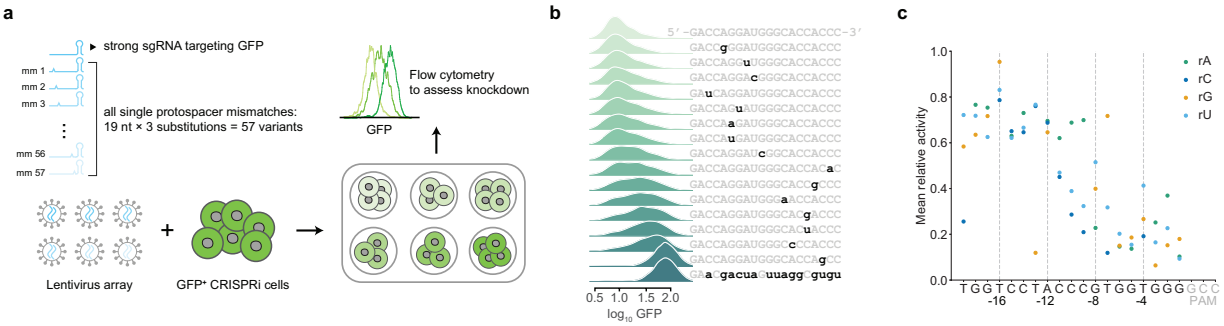


Figure 3.1. Mismatched sgRNAs titrate GFP expression at the single-cell level. **(a)** Experimental design to test knockdown conferred by all mismatched variants of a GFP-targeting sgRNA. **(b)** Distributions of GFP levels in cells with perfectly matched sgRNA (top), mismatched sgRNAs (middle), and non-targeting control sgRNA (bottom). Sequences of sgRNAs are indicated on the right. **(c)** Relative activities of all mismatched sgRNAs, defined as the ratio of percent-knockdown conferred by a mismatched sgRNA to percent-knockdown conferred by the perfectly matched sgRNA. Relative activities are displayed as the mean of two biological replicates.

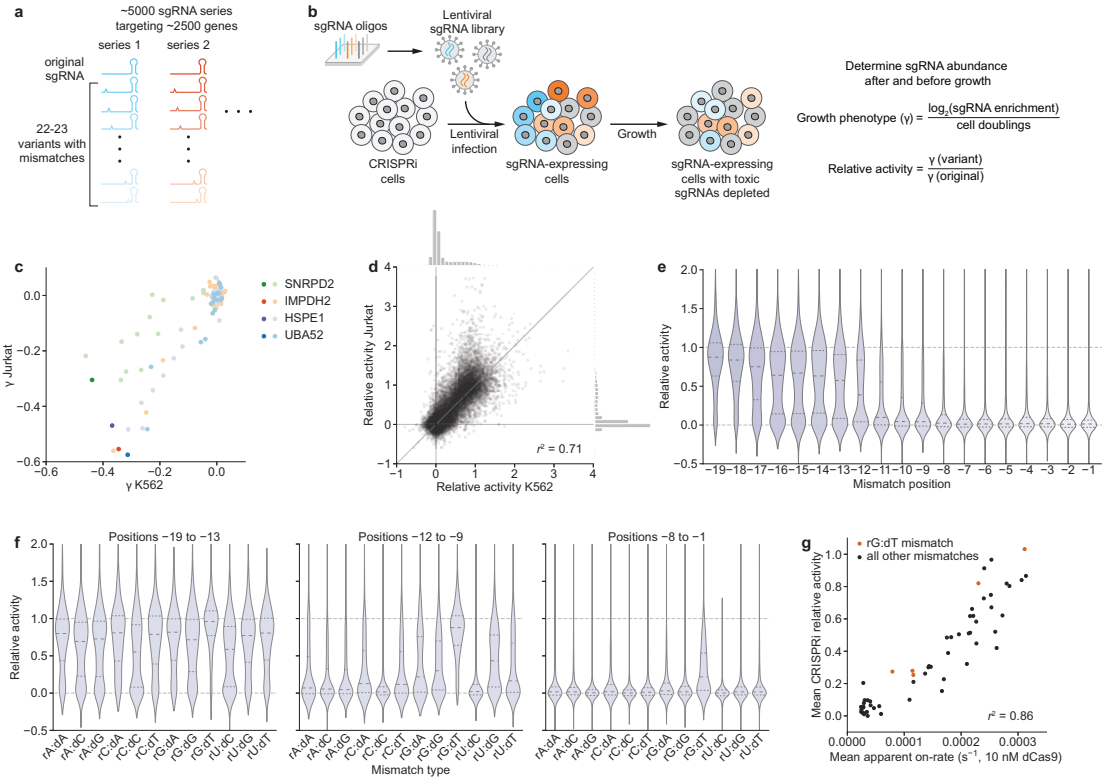


Figure 3.2. A large-scale CRISPRi screen identifies factors governing mismatched sgRNA activity. **(a)** Design of large-scale mismatched sgRNA library. **(b)** Schematic of pooled CRISPRi screen to determine activities of mismatched-sgRNAs. **(c)** Growth phenotypes (γ) in K562 and Jurkat cells for four sgRNA series, with the perfectly matched sgRNAs shown in darker colors and mismatched sgRNAs shown in corresponding lighter colors. Differences in absolute phenotypes likely reflect cell type-specific essentiality. **(d)** Comparison of mismatched sgRNA relative activities in K562 and Jurkat cells. Marginal histograms depict distributions of relative activities along the corresponding axes. **(e)** Distribution of mismatched sgRNA relative activities stratified by position of the mismatch. Position –1 is closest to the PAM. **(f)** Distribution of mismatched sgRNA relative activities stratified by type of mismatch, grouped by mismatches located in positions –19 to –13 (PAM-distal region), positions –12 to –9 (intermediate region), and positions –8 to –1 (PAM-proximal/seed region). **(g)** Comparison of mean apparent on-rates measured *in vitro* for mismatched variants of a single sgRNA (17) and mean relative activities from large-scale screen. Values are compared for identical combinations of mismatch type and mismatch position; mean relative activities were calculated by averaging relative activities for all mismatched sgRNAs with a given combination.

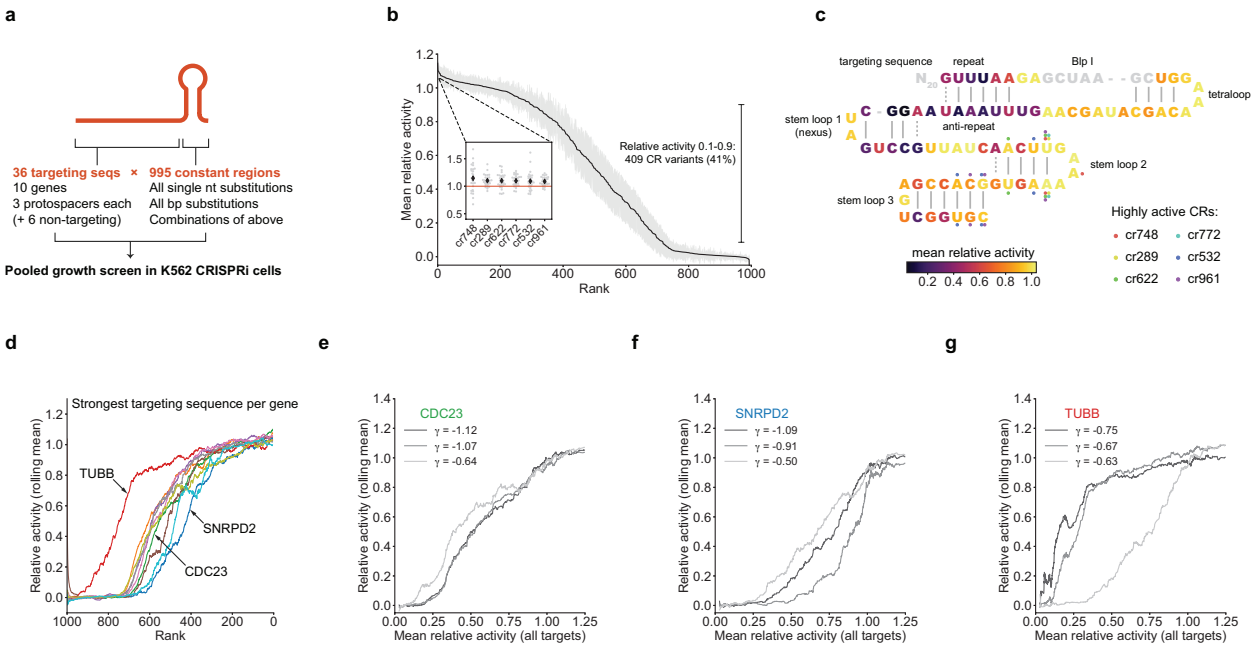


Figure 3.3. Identification and characterization of intermediate-activity constant regions. **(a)** Design of constant region variant library. **(b)** Mean relative activities of constant region variants, calculated by averaging relative activities for all targeting sequences. Grey margins denote 95% confidence interval. Inset: Focus on 6 constant region variants with higher activity than the original constant region. Black diamonds denote mean relative activity, grey dots relative activities with individual targeting sequences. **(c)** Mapping of constant region variant relative activities onto constant region structure. Each constant region base is colored by the average relative activity of the three single constant region variants carrying a single mutation at that position. Positions mutated in 6 highly active constant regions (inset in panel **b**) are indicated by colored dots. **(d)** Constant region activities by targeting sequence, plotted against ranked mean constant region activity. For each gene, the activities with the strongest targeting sequence are shown as rolling means with a window size of 50. **(e-g)** Constant region activities by targeting sequence for all three targeting sequences against the indicated genes. Growth phenotypes (γ) of each targeting sequence paired with the unmodified constant region are indicated in the legend.

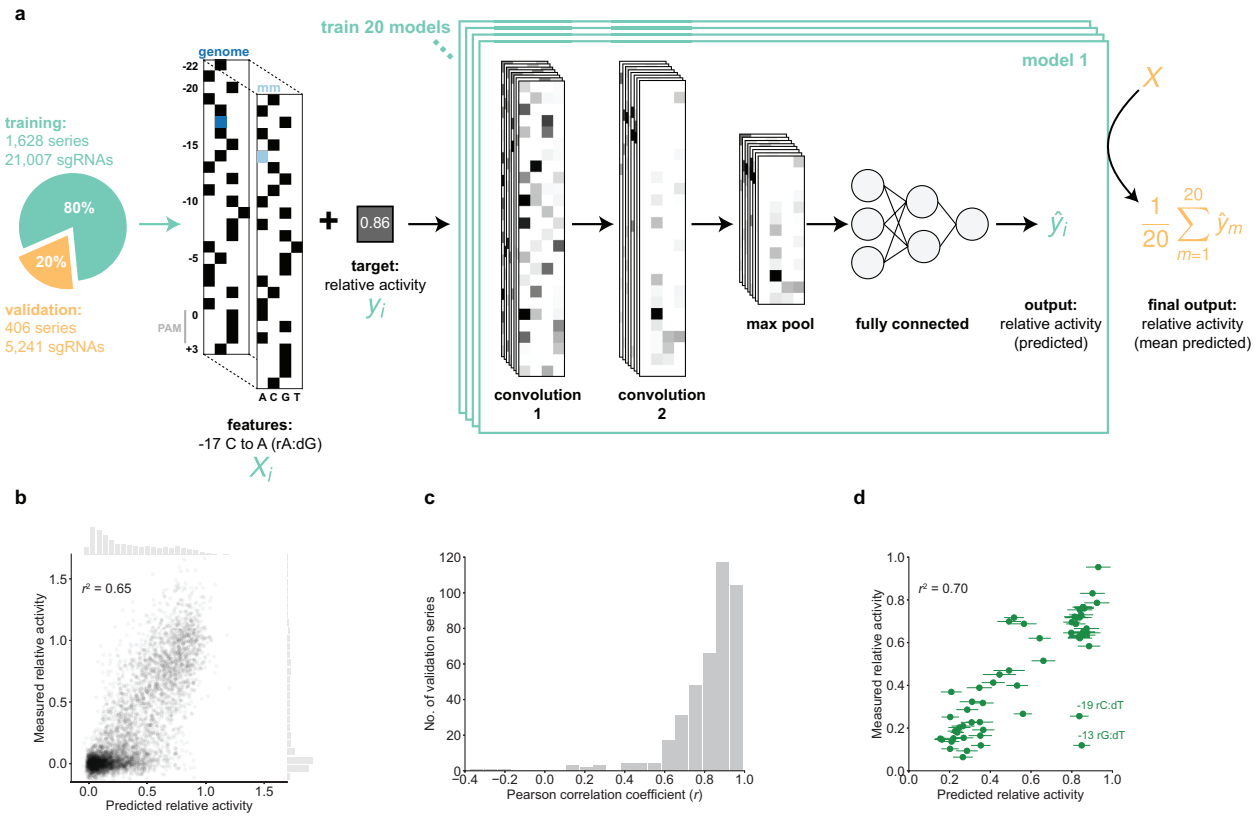


Figure 3.4. Neural network predictions of sgRNA activity. **(a)** Schematic of a singly-mismatched sgRNA feature array (X_i), and the convolutional neural network architecture trained on pairs of such arrays and their corresponding relative activities (y_i). Black squares in X_i represent the value 1 (the presence of a base at the indicated position); white represents 0. The mean prediction from 20 independently trained models was used to assign a final prediction ($\hat{y}$) to each sgRNA in the hold-out validation set (orange). **(b)** Measured relative growth phenotypes from the large-scale screen vs. predicted activities assigned by the neural network. Marginal histograms show distributions of relative activities along the corresponding axes. **(c)** Distribution of Pearson *r* values (predicted vs. measured relative activity) for each sgRNA series in the validation set. **(d)** Measured relative activity (i.e. relative knockdown) in the GFP experiment vs. predicted relative sgRNA activity. Two outliers with lower-than-predicted activity are annotated with their respective mismatch position and type. Predictions are shown as mean ± S.D. from the 20-model ensemble.

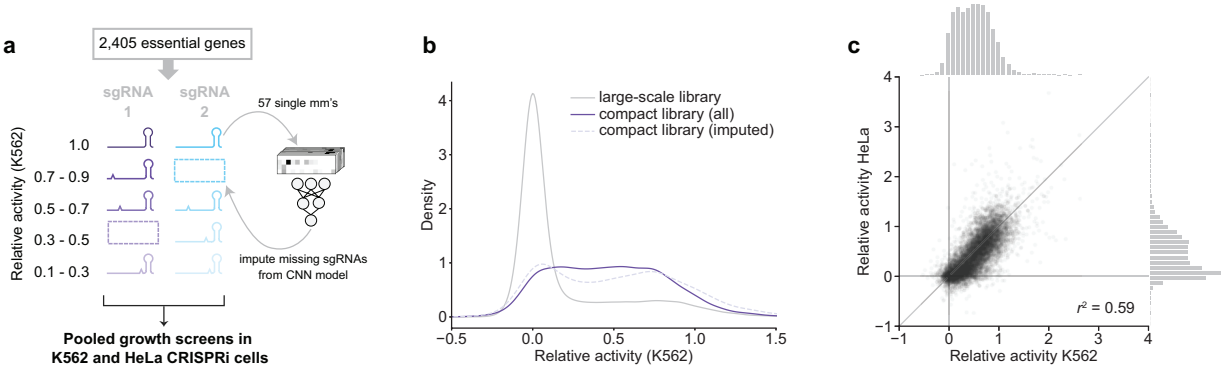


Figure 3.5. Compact mismatched sgRNA library targeting essential genes. **(a)** Design of library. For activity bins lacking a previously measured sgRNA, novel mismatched sgRNAs were included according to predicted activity. **(b)** Distribution of relative activities from the large-scale library (grey) and the compact library (purple) in K562 cells. **(c)** Comparison of relative activities of mismatched sgRNAs in HeLa and K562 cells. Marginal histograms show the distributions of relative activities along the corresponding axes.

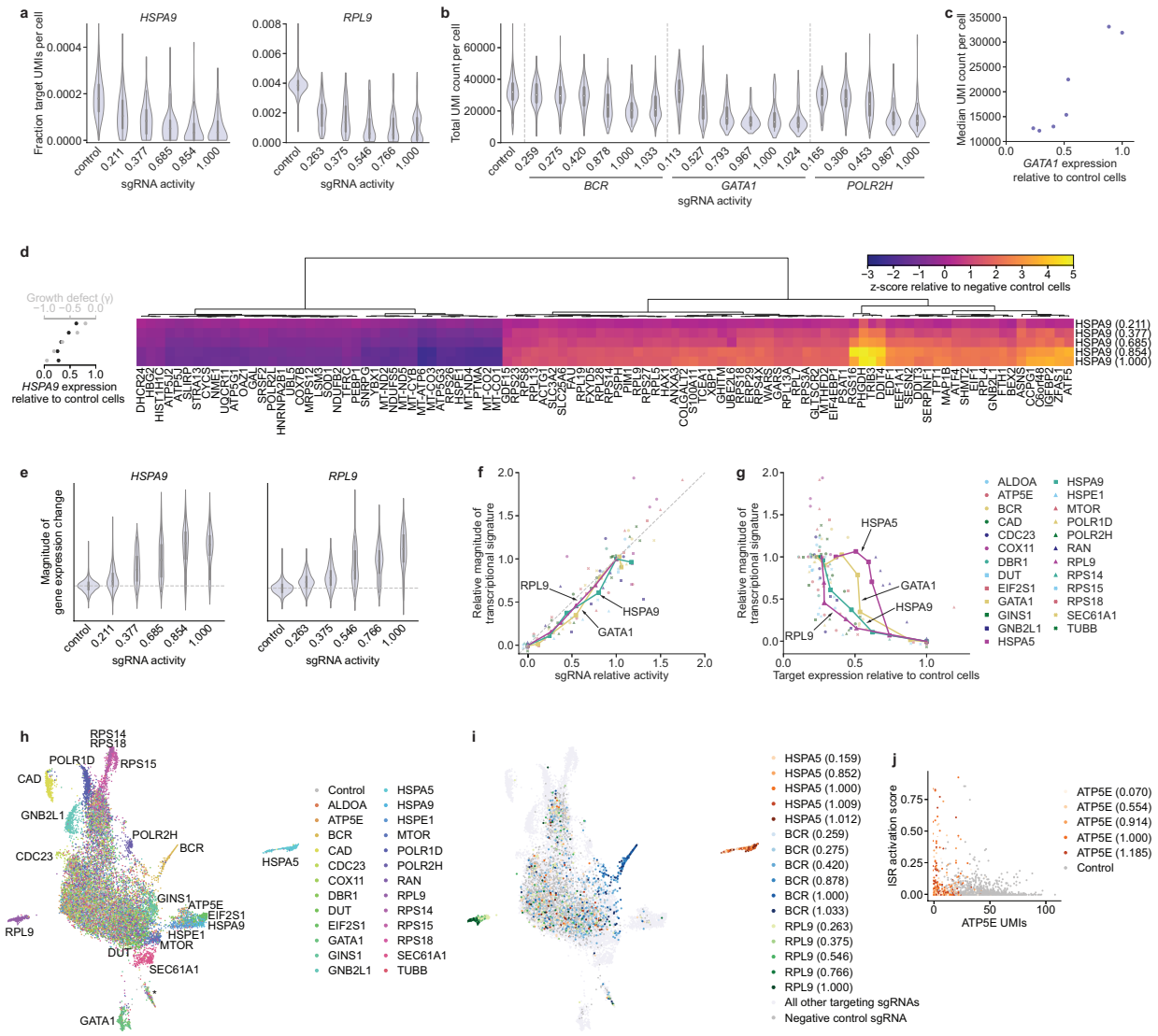
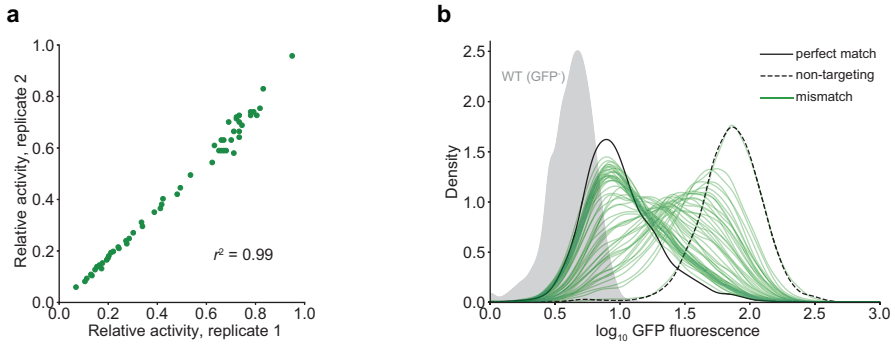
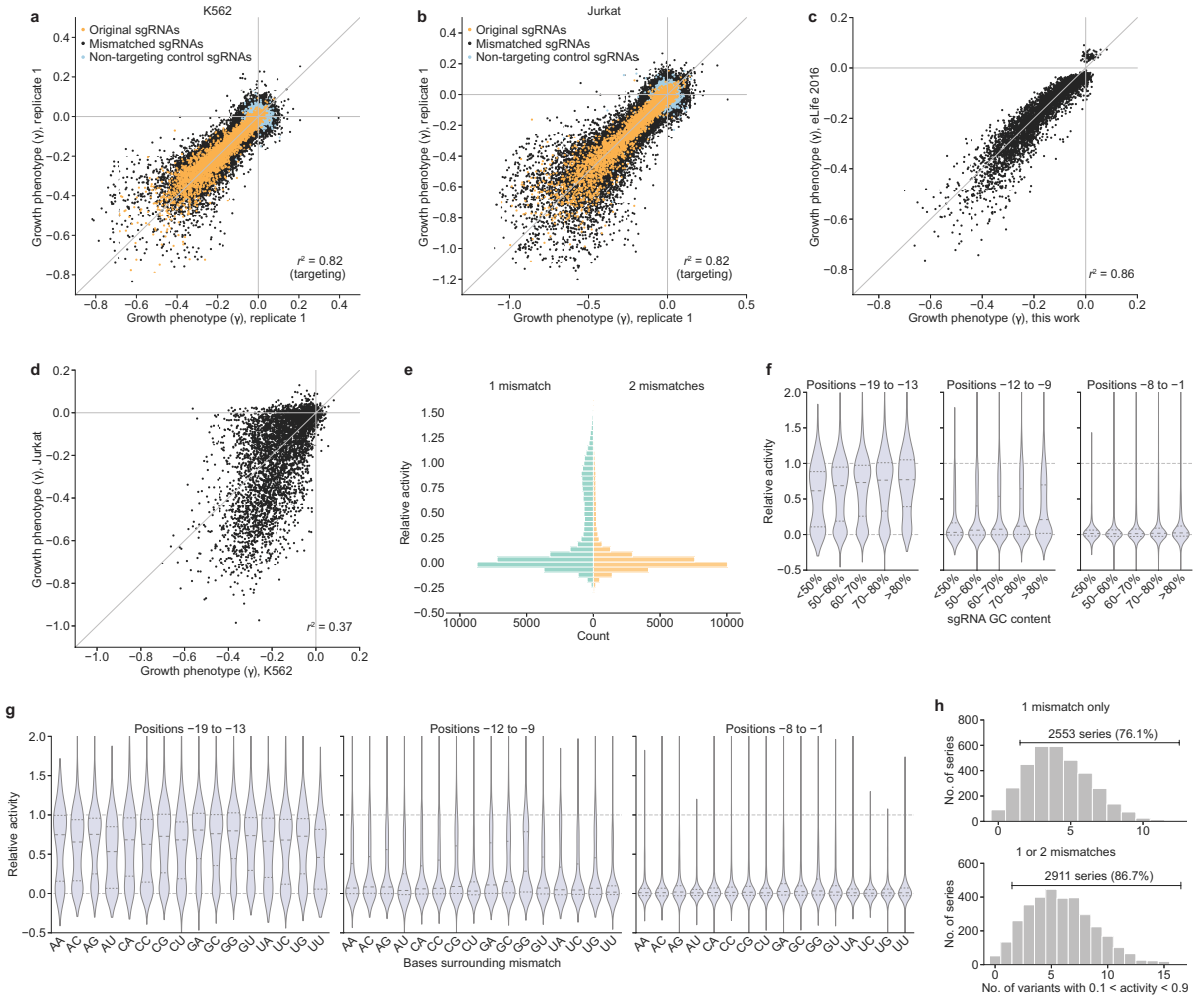


Figure 3.6. Rich phenotyping of cells with intermediate-activity sgRNAs by perturb-seq. **(a)** Distributions of *HSPA9* and *RPL9* expression in cells with indicated perturbations. Expression is quantified as target gene UMI count normalized to total UMI count per cell. sgRNA activity is calculated using relative γ measurements from the perturb-seq cell pool after 5 days of growth. **(b)** Distributions of total UMI counts in cells with indicated perturbations. **(c)** Comparison of median UMI count per cell and *GATA1* expression in cells with *GATA1*-targeting sgRNAs or control cells. **(d)** Right: Expression profiles of 100 genes in populations with *HSPA9*-targeting sgRNAs of various strength. Expression is quantified as z-score relative to population of cells with non-targeting sgRNAs. Left: Growth phenotype and knockdown for each sgRNA. **(e)** Distribution of gene expression changes in populations with indicated sgRNAs. Magnitude of gene expression change is

calculated as sum of z-scores of genes differentially expressed in the series, with z-scores of individual genes signed by the direction of change in cells with the perfectly matched sgRNA. Distribution for negative control sgRNAs is centered around 0 (dashed line). **(f)** Comparison of relative growth phenotype and magnitude of gene expression change for all individual sgRNAs. Growth phenotype and magnitude of gene expression change are normalized in each series to those of the sgRNA with the strongest knockdown. Individual series highlighted as indicated. **(g)** Comparison of magnitude of gene expression and target gene knockdown, as in **f**. **(h)** UMAP projection of all single cells with assigned sgRNA identity in the experiment, colored by targeted gene. Clusters clearly assignable to a genetic perturbation are labeled. Cluster labeled * contains a small number of cells with residual stress response activation and could represent apoptotic cells. **(i)** UMAP projection, as in **h**, with selected series colored by sgRNA activity. **(k)** Comparison of extent of ISR activation to *ATP5E* UMIs in cells with knockdown of *ATP5E* or control cells.

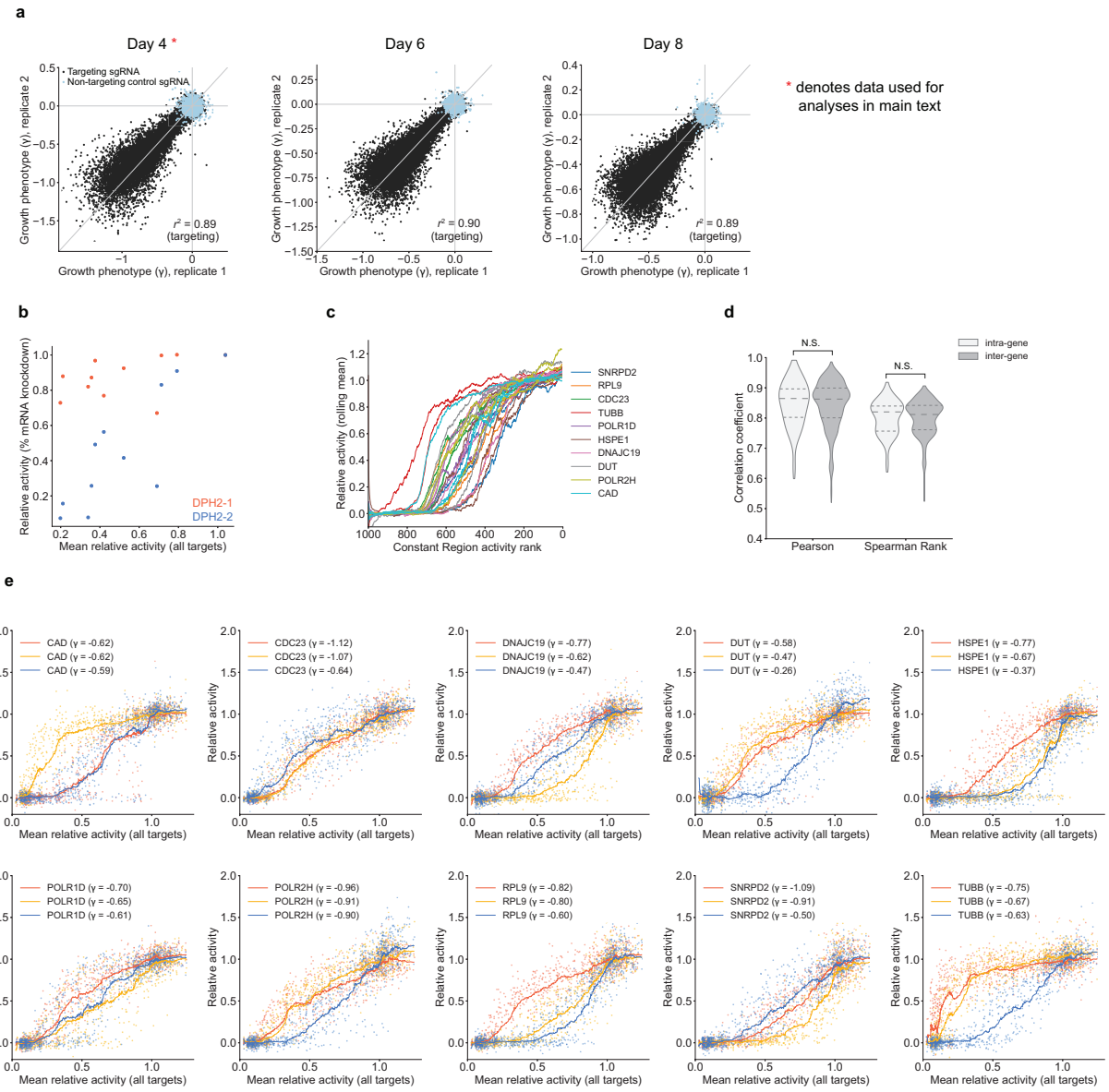


Supplementary Figure S3.1. Details of the GFP mismatch experiment. **(a)** Comparison of relative activities measured in two biological replicates. Relative activity was defined as the fold-knockdown of each mismatched variant ($GFP_{sgRNA[non\text{-}targeting]}$ / $GFP_{sgRNA[variant]}$) divided by the fold-knockdown of the perfectly-matched sgRNA. The background fluorescence of a GFP^- strain was subtracted from all GFP values prior to other calculations. **(b)** KDE plots of GFP distributions 10 days after transducing K562 GFP^+ cells with the perfectly-matched sgRNA, a non-targeting sgRNA, and each of the 57 singly-mismatched variants. Fluorescence of a GFP^- K562 strain is shown in grey.

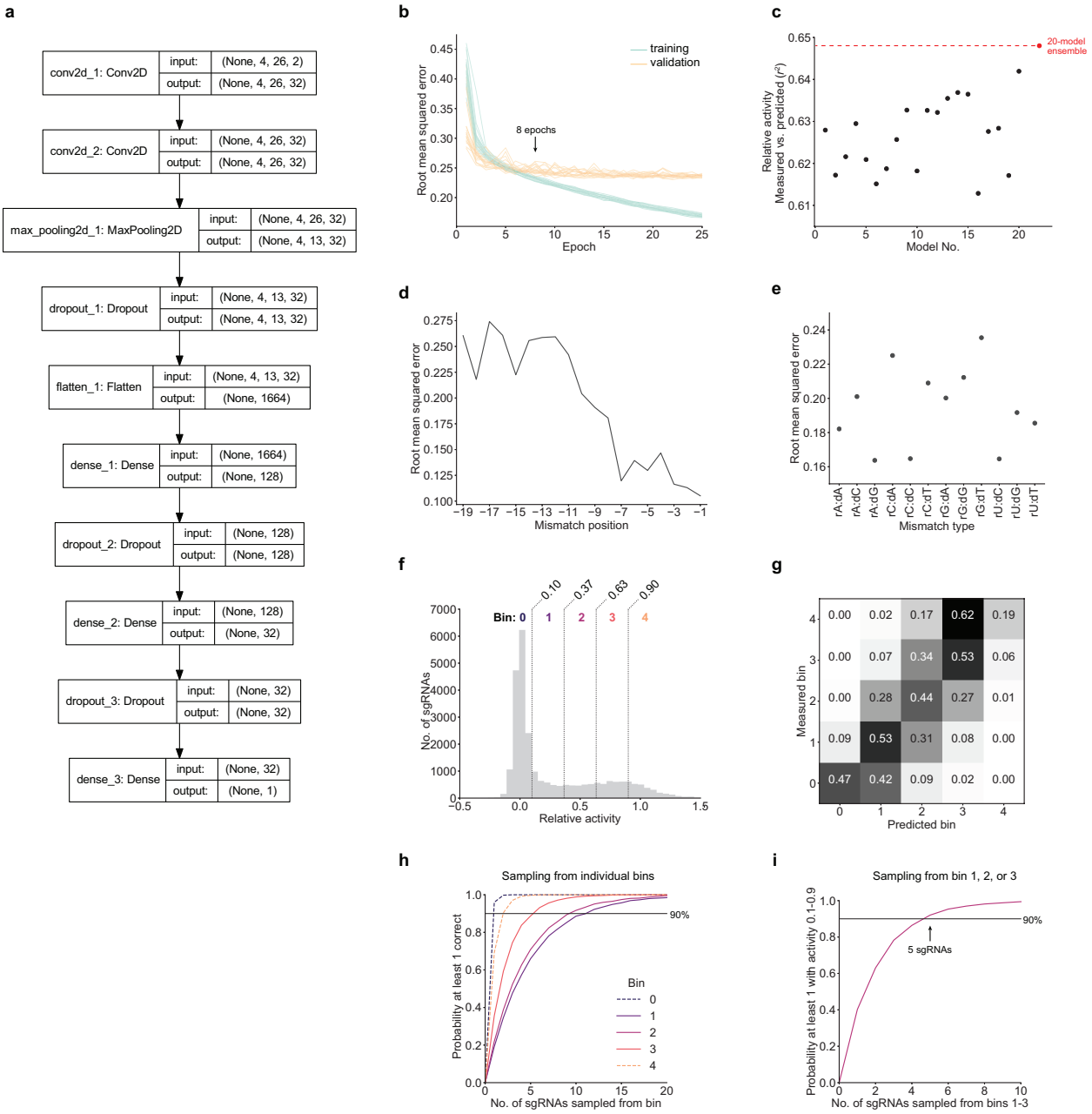


Supplementary Figure S3.2. Additional analysis of large-scale mismatched sgRNA screen. **(a,b)** Comparison of growth phenotypes (γ) of all sgRNAs derived from biological replicates of the **(a)** K562 and **(b)** Jurkat screens. **(c)** Comparison of growth phenotypes (γ) of perfectly matched sgRNAs from the K562 screen in this work and a previously published K562 screen (7). **(d)** Comparison of growth phenotypes (γ) of perfectly matched sgRNAs in K562 and Jurkat cells. **(e)** Distribution of mismatched sgRNA relative activities for sgRNAs with 1 mismatch (left) or 2 mismatches (right). **(f)** Distribution of mismatched sgRNA relative activities stratified by sgRNA GC content, grouped by mismatches located in positions –19 to –13 (PAM-distal region), positions –12 to –9 (intermediate region), and positions –8 to –1 (PAM-proximal/seed region). **(g)** Distribution of mismatched sgRNA relative activities stratified by the identity of the 2 bases flanking the mismatch, grouped by mismatches located in positions –19 to –13 (PAM-distal region), positions –12 to –9 (intermediate region), and positions –8 to –1 (PAM-proximal/seed region). **(h)**

Distribution of sgRNA series by number of sgRNAs with intermediate activity (0.1 < relative activity < 0.9), using only sgRNAs with a single mismatch (top) or all mismatched sgRNAs (bottom).

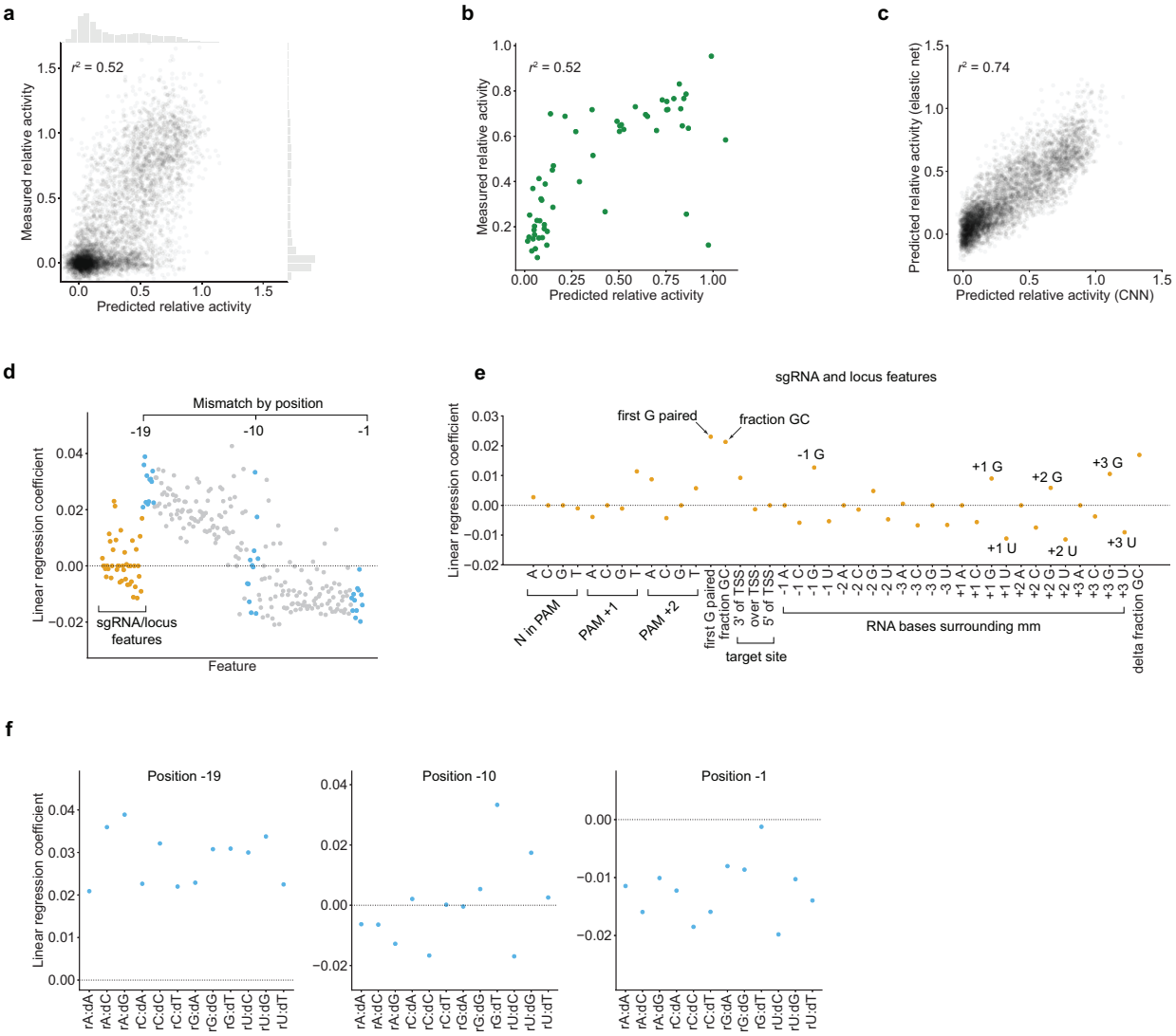


Supplementary Figure S3.3. Additional analysis of modified constant regions. **(a)** Comparison of growth phenotypes measured in each biological replicate after 4, 6, or 8 days of growth from t_0. Data from Day 4 was used for all subsequent analyses. **(b)** Comparison of relative % knockdown (quantified via RT-qPCR) and mean relative growth phenotype for 10 intermediate-activity constant region variants paired with two targeting sequences against *DPH2*. **(c)** Relative activities of constant regions paired with all 30 targeting sequences, ranked by the average strength of each constant region and displayed as rolling means with a window size of 50. **(d)** Distribution of all pairwise correlations of constant region relative activities within and between gene targets. N.S.; no significant difference according to two-tailed Student's t-test ($p >> 0.05$). **(e)** Relative activity of each indicated target sequence:constant region pair vs. the mean relative activity of the respective constant region for all targets. Growth phenotypes (γ) with the unmodified constant region are indicated in the figure legends. Lines represent rolling means of individual data points.

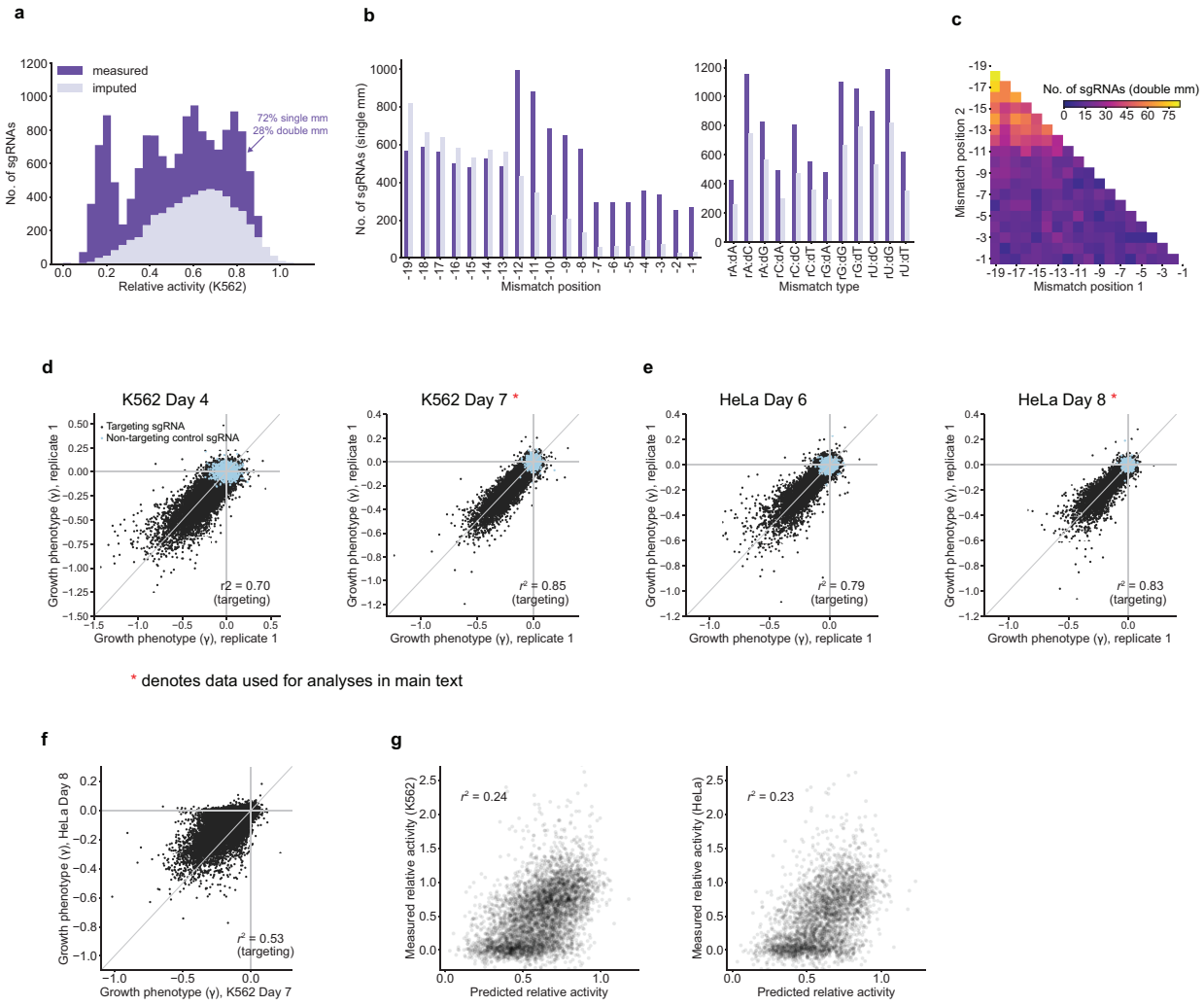


Supplementary Figure S3.4. Additional details for the neural network. **(a)** Graph of the CNN model architecture. **(b)** Model loss, measured as root mean squared error, for training and validation data over 25 training epochs. Each line represents one of 20 models trained. The final models used for our predictions were only trained for 8 epochs, as additional cycles only reduced training loss without significant improvement in validation loss (i.e., the model becomes over-fit). **(c)** Explained variance (r^2) of validation sgRNA relative activities for each individual model (black), and for the mean prediction of all 20 models (red). **(d)** Validation error stratified by mismatch position. **(e)** Validation error stratified by mismatch type. **(f)** Partitioning of sgRNAs into bins based on relative activity in the large-scale K562 screen. **(g)** Confusion matrix showing the fraction of sgRNAs in each

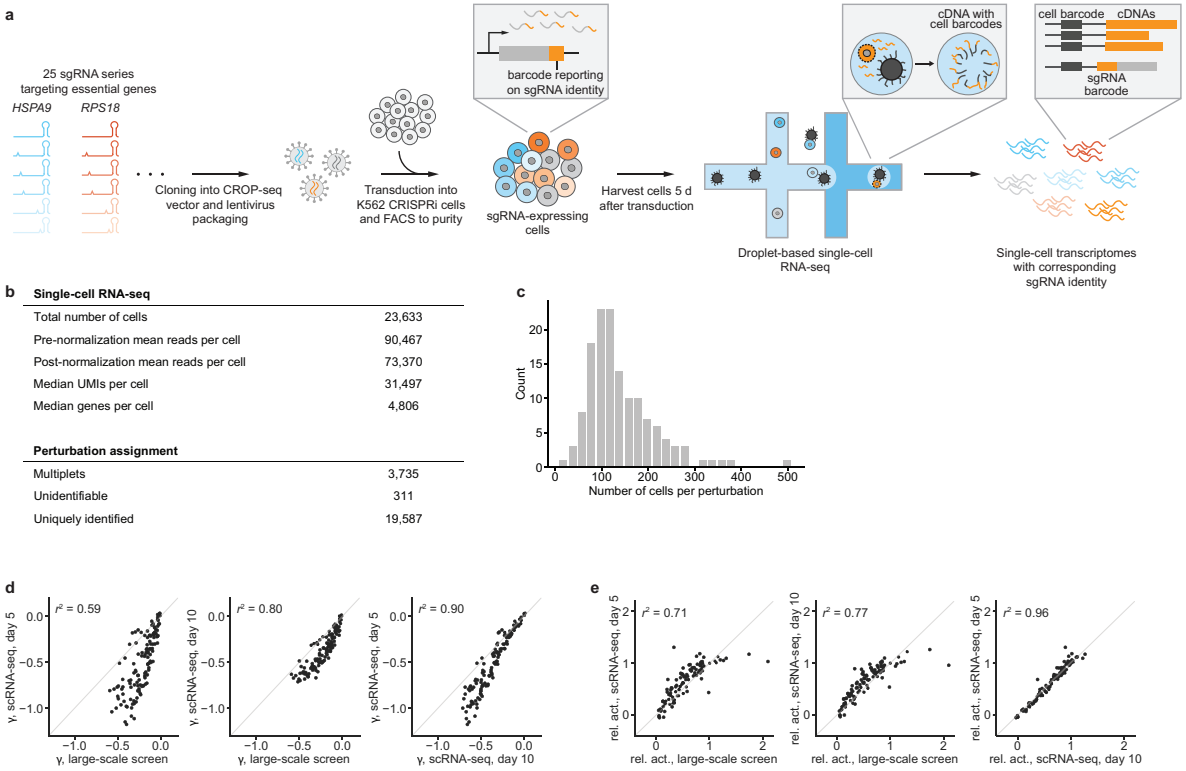
actual (measured) activity bin that were assigned to each predicted bin by the CNN model. Each row sums to 1. **(h)** Statistics indicating the requisite number of randomly sampled sgRNAs from each activity bin to have a given probability of selecting at least one sgRNA with true activity in that bin. Simulations are based on the probabilities outlined in the confusion matrix (panel **e**). **(g)** Similar to panel **f**, with random sampling from any of the intermediate activity bins (1-3) to yield at least one sgRNA with intermediate activity (0.1-0.9).



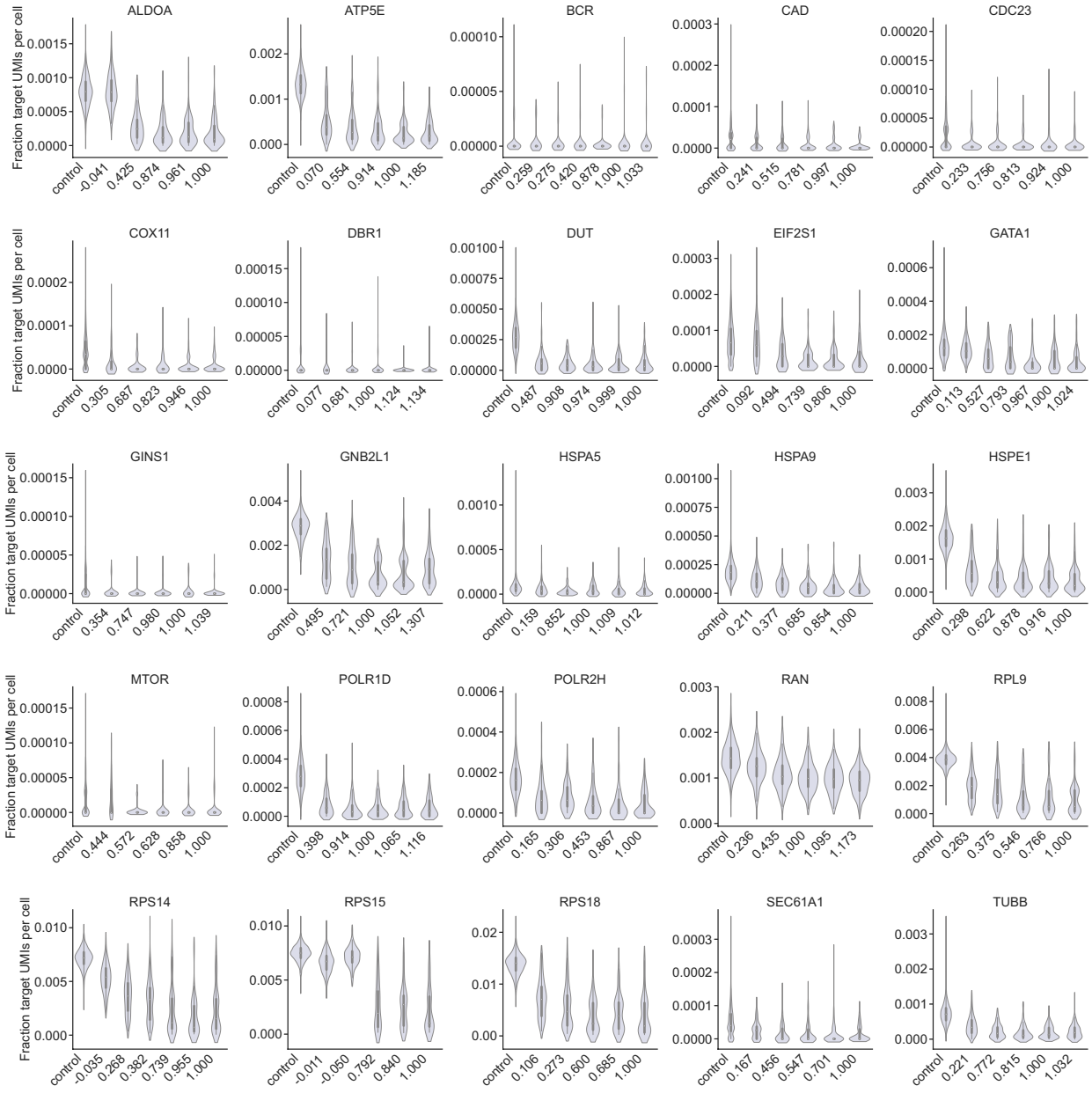
Supplementary Figure S3.5. Additional details for the linear model. **(a)** Measured relative growth phenotypes from the large-scale screen vs. predicted activities assigned by the elastic net linear model. Marginal histograms show distributions of relative activities along the corresponding axes. **(b)** Measured relative activity (relative knockdown) in the GFP experiment vs. predicted relative sgRNA activity. **(c)** Comparison of predicted relative activities from the linear model vs. the neural network, based on the validation set of singly-mismatched sgRNAs. **(d)** Regression coefficients assigned to each feature in the linear model. 228 features (grey, blue) describe the position and type of mismatch; 42 features (gold) carry other information about the sgRNA and genomic context surrounding the protospacer. These features are detailed in subsequent panels. **(e)** Linear coefficients for features of the sgRNA and targeted locus. TSS; transcription start site. **(f)** Linear coefficients for features covering positions in the distal, intermediate, and seed regions of the targeting sequence (highlighted blue in panel **d**).



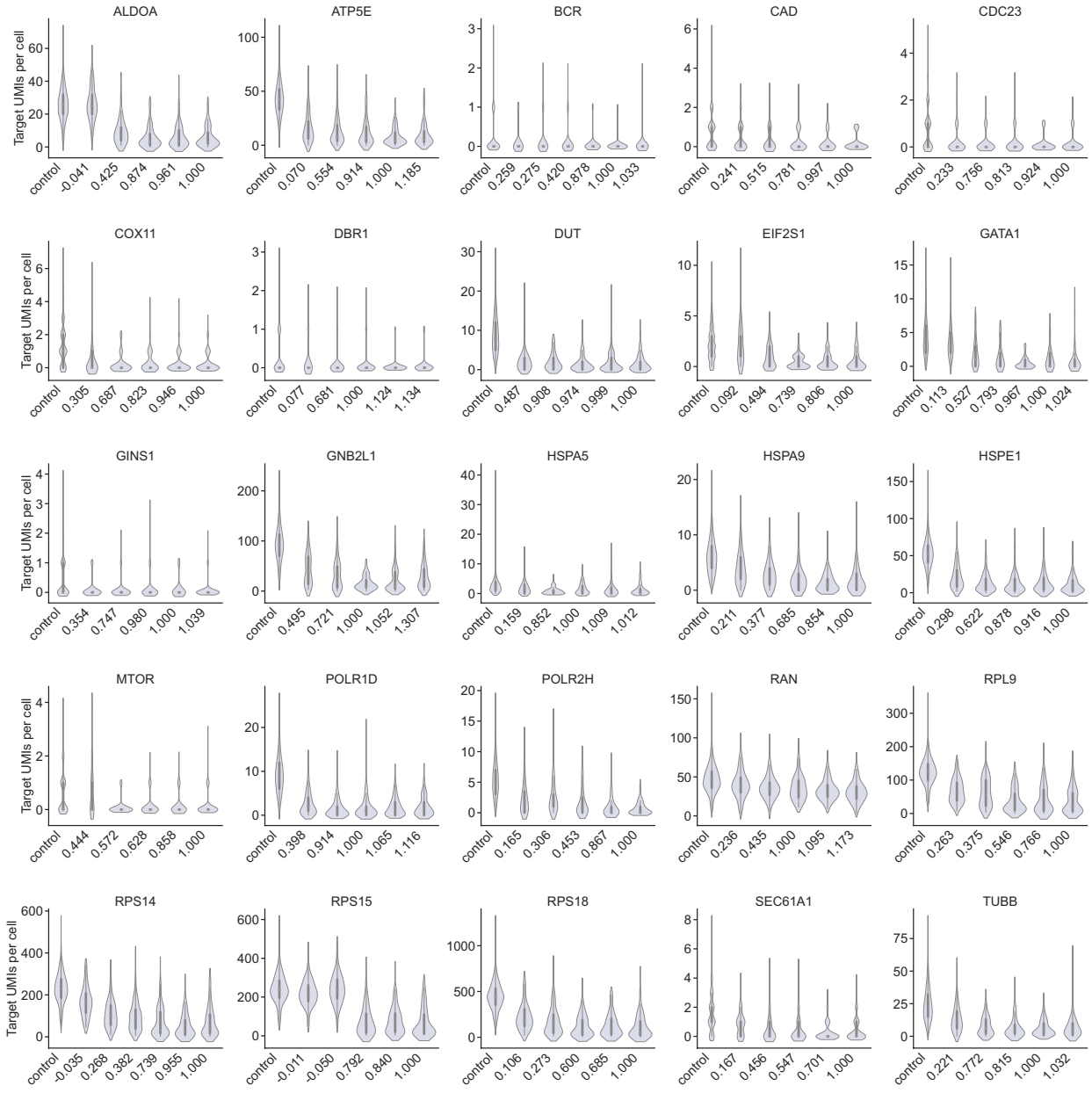
Supplementary Figure S3.6. Additional analysis of the compact allelic series screen. **(a)** Composition of the compact library, in terms of previously measured relative activities in the large-scale screen (dark purple), or predicted relative activities assigned by the CNN model ensemble (light purple). Perfectly matched guides, which by definition have relative activities of 1.0, comprise 20% of the library but were not included in the histogram. **(b)** Distribution of mismatch positions and types for singly-mismatched sgRNAs in the compact library, for previously measured (dark purple) and CNN-imputed (light purple) sgRNAs. **(c)** Heatmap showing the distribution of mutated positions for doubly-mismatched sgRNAs in the compact library. **(d)** Comparison of growth phenotypes measured in each K562 biological replicate 4- and 7-days post-transduction. Data from Day 7 was used for all subsequent analyses. **(e)** Comparison of growth phenotypes measured in each HeLa biological replicate 6- and 8-days post-transduction. Data from Day 8 was used for all subsequent analyses. **(f)** Comparison of growth phenotypes in HeLa and K562 cells (γ expressed as the average of biological replicate measurements). **(g)** Measured vs. predicted relative activities of CNN-imputed sgRNAs in K562 cells (left) and HeLa cells (right). A small number of points lying beyond the y-axis limits were excluded to more clearly display the bulk of the distribution.



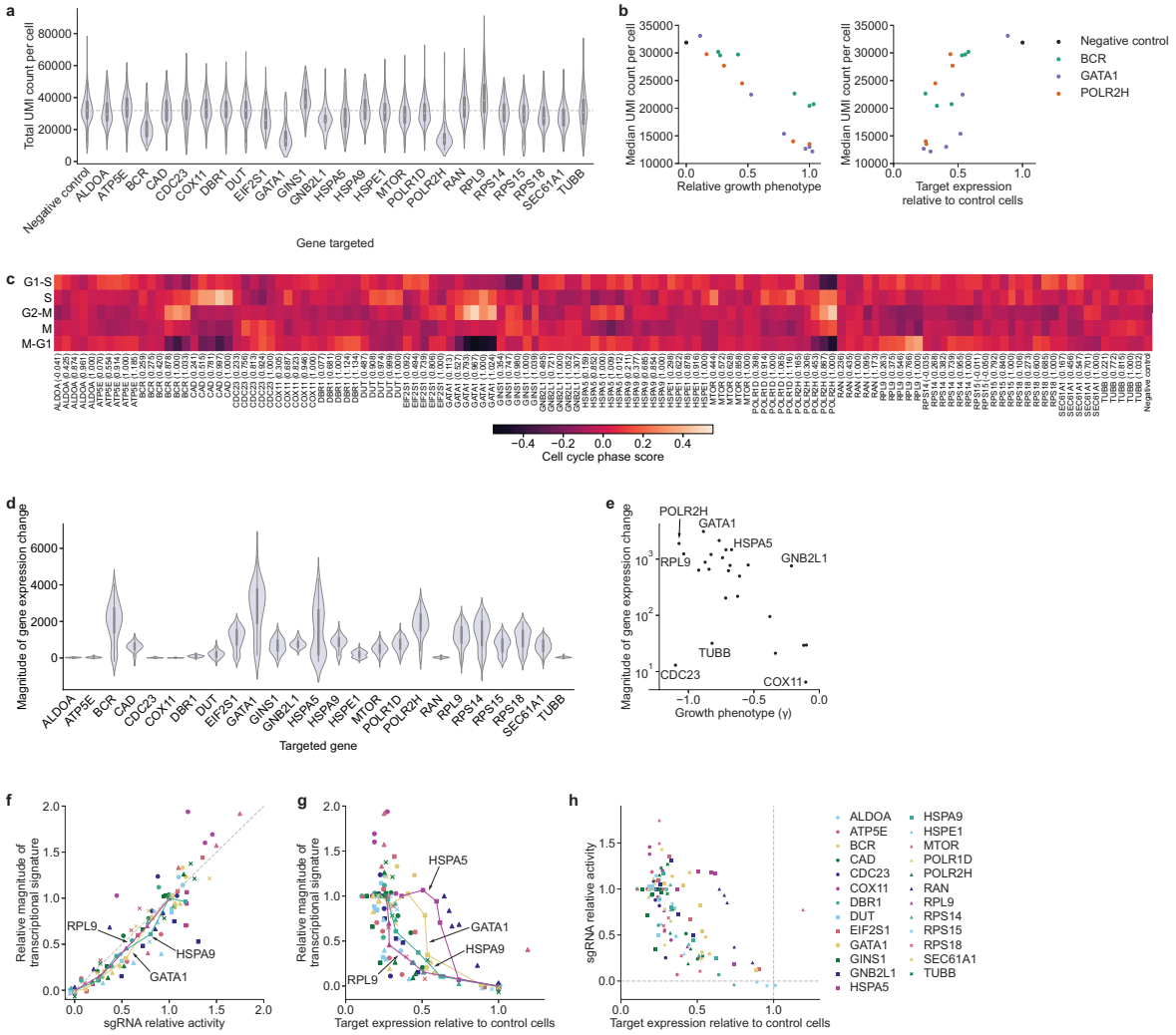
Supplementary Figure S3.7. Summary of perturb-seq experiment. **(a)** Schematic of perturb-seq strategy to capture single-cell transcriptomes with matched sgRNA identities. **(b)** Summary of sequencing and perturbation assignment statistics. **(c)** Distribution of number of cells captured per perturbation. Median: 122 cells per perturbation; 5th to 95th percentile: 66 – 277 cells per perturbation. **(d,e)** Comparison of **(d)** growth phenotypes (γ) and **(e)** relative activities measured in the large-scale mismatched sgRNA screen and in the perturb-seq experiment. Differences are likely due to the different timescales and the different vectors used.



Supplementary Figure S3.8. Distributions of target gene expression in cells with indicated perturbations. Expression is quantified as target gene UMI count normalized to total UMI count per cell.

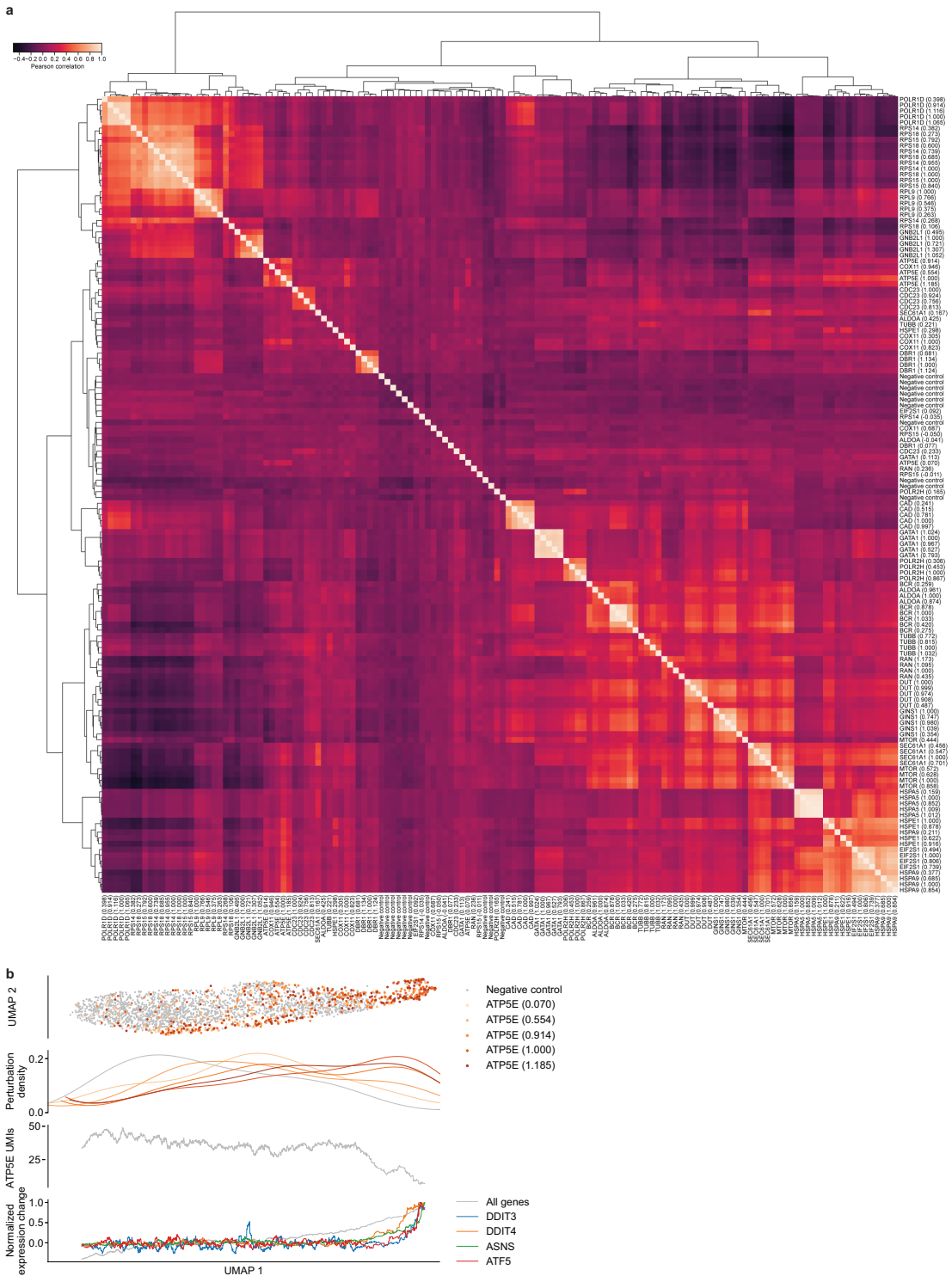


Supplementary Figure S3.9. Distributions of target gene expression in cells with indicated perturbations. Expression is quantified as raw target gene UMI count.



Supplementary Figure S3.10. Phenotypes resulting from target gene titration. **(a)** Distributions of total UMI counts in cells with the perfectly matched sgRNA against the indicated genes. **(b)** Left: Comparison of median UMI count per cell and relative growth phenotype in cells with sgRNAs targeting *BCR*, *GATA1*, or *POLR2H* or control cells. Right: Comparison of median UMI count per cell and target gene expression. **(c)** Cell cycle scores for populations of cells with individual sgRNAs. **(d)** Magnitudes of gene expression change of populations with perfectly matched sgRNAs targeting indicated genes. Magnitude of gene expression change is calculated as sum of z-scores of genes differentially expressed in the series, with z-scores of individual genes signed by the average direction of change in cells with the perfectly matched sgRNA. **(e)** Comparison of magnitude of gene expression change to growth phenotype (γ) for all perfectly matched sgRNAs in the experiment. **(f)** Comparison of relative growth phenotype and magnitude of gene expression change for all individual sgRNAs, as in Fig. 3.6f but without increased transparency for individual series. **(g)** Comparison of magnitude of gene expression and target gene knockdown, as in Fig. 3.6g but

without increased transparency for individual series. **(h)** Comparison of relative growth phenotype and target gene expression, as in Fig. 3.6f.



Supplementary Figure S3.11. Diverse phenotypes resulting from essential gene depletion. **(a)** Clustered correlation heatmap of perturbations. Gene expression profiles for genes with mean UMI

count > 0.25 in the entire population were z-normalized to expression values in cells with negative control sgRNAs and then averaged for populations with the same sgRNA. Crosswise Pearson correlations of all averaged transcriptomes were clustered by the Ward variance minimization algorithm implemented in scipy. **(b)** UMAP projection, distribution of cells with indicated sgRNAs, target gene expression (rolling mean over 50 cells), and magnitudes of transcriptional changes for all differentially expressed genes and selected ISR regulon genes (rolling mean over 50 cells) for cells with knockdown of *ATP5E* or control cells.

Supplementary Tables S3.1-S3.10 are too large to be published here but will be available online pending publication in a peer-reviewed journal.

Supplementary Table S3.11. Oligonucleotide sequences used in this study.

NAME	SEQUENCE
constant_region_1_fw	TAAGCTGGAAACAGCATAGCAAGCTCAAATAAGACTAGTTC GTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT TTC
constant_region_1_rv	TCGAGAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGT TGATAACGAACTAGTCTTATTTGAGCTTGCTATGCTGTTT CCAGC
constant_region_2_fw	TAAGCTGGAAACAGCATAGCAAGTTCAAATAAGGCTAGTCC GTTATGTACTTCAAAAAGTGGCACCGAGTCGGTGCTTTTT TTC
constant_region_2_rv	TCGAGAAAAAAAGCACCGACTCGGTGCCACTTTTTGAAGT ACATAACGGACTAGCCTTATTTGAACTTGCTATGCTGTTTC CAGC
constant_region_3_fw	TAAGCTGGAAACAGCATAGCGAGTTCAAATAAGGCTCGTC CGTTATCCACTTGAAAAAGTGGCACCGAGTCGGTGCTTTT TTTC
constant_region_3_rv	TCGAGAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGT GGATAACGGACGAGCCTTATTTGAACTCGCTATGCTGTTT CCAGC
constant_region_4_fw	TAAGCTGGAAACAGCATAGCAAGTTCAAATAAAGTTAATCT GTTATCAACTCGAAAGAGTGGCACCGAGTCGGTGCTTTTT TTC
constant_region_4_rv	TCGAGAAAAAAAGCACCGACTCGGTGCCACTCTTTCGAGT TGATAACAGATTAACTTTATTTGAACTTGCTATGCTGTTTC CAGC
constant_region_5_fw	TAAGCTGGAAACAGCATAGCAAGTTCAAATAAGGCTAGCCC GTTATGAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT TTC
constant_region_5_rv	TCGAGAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGT TCATAACGGGCTAGCCTTATTTGAACTTGCTATGCTGTTT CCAGC
constant_region_6_fw	TAAGCTGGAAACAGCATAGCAAGTTCAAATAAGGCTAGTCC GTTATCAACTTGAAAAAGTGGCACCGGGGCGGTGCTTTTT TTC
constant_region_6_rv	TCGAGAAAAAAAGCACCGCCCCGGTGCCACTTTTTCAAGT TGATAACGGACTAGCCTTATTTGAACTTGCTATGCTGTTT CCAGC
constant_region_7_fw	TAAGCTGGAAACAGCATAGCAAGTTCAAATATGGCTAGTCC GTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT TTC
constant_region_7_rv	TCGAGAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGT TGATAACGGACTAGCCATATTTGAACTTGCTATGCTGTTT CCAGC
constant_region_8_fw	TAAGCTGGAAACAGCATAGCAAGTTCAAATAAGGATATTCC GTTATCAAGTTGAAAAACTGGCACCGAGTCGGTGCTTTTT TTC

NAME	SEQUENCE
constant_region_8_rv	TCGAGAAAAAAAGCACCGACTCGGTGCCAGTTTTTCAACT TGATAACGGAATATCCTTATTTGAACTTGCTATGCTGTTTC CAGC
constant_region_9_fw	TAAGCTGGAAACAGCATAGCAAGTTCAAATAAGGCTAGTCC GTTATCAACTTGAGAAAGTGGCACCGGGTCGGTGCTTTTT TTC
constant_region_9_rv	TCGAGAAAAAAAGCACCGACCCGGTGCCACTTTCTCAAGT TGATAACGGACTAGCCTTATTTGAACTTGCTATGCTGTTT CCAGC
constant_region_10_fw	TAAGCTGGAAACAGCATAGCAAGTTCAAATAAGGCTAGTCC GTTATCAACTTGAAAAAGTGGCACCGCGTCGGTGCTTTTT TTC
constant_region_10_rv	TCGAGAAAAAAAGCACCGACGCGGTGCCACTTTTTCAAGT TGATAACGGACTAGCCTTATTTGAACTTGCTATGCTGTTT CCAGC
DPH2_qPCR_fw	ACCTGGACGGAGTGTACGAG
DPH2_qPCR_rv	TCTCCCAATAGCTGGTCAGG
ACTB_qPCR_fw	GCTACGAGCTGCCTGACG
ACTB_qPCR_rv	GGCTGGAAGAGTGCCTCA
oCRISPRi_seq_V5	GTGTGTTTTGAGACTATAAGTATCCCTTGGAGAACCACCT TGTTG
oCRISPRi_seq_V4_3'	CCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTAAACT TGCTATGCTGT
oCRISPRi_PE_constant_region _common_primer	AATGATACGGCGACCACCGAGATCTACACGCACAAAAGGAA ACTCACCCT
oCRISPRi_PE_constant_region _indexing_primer	CAAGCAGAAGACGGCATACGAGATNNNNNNGTCTCGTGG GCTCGGAGATGTGTATAAGAGACAGGCCGCCTAATGGATC CTAG
oBA503	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTG GGCTCGGAGATGTGTATAAGAGACAGGTGTTTTGAGACTA TAAGTATCCCTTGGAGAACCACCTTGTTG
PCR_perturb-seq_P5	AATGATACGGCGACCACCGAGATCTACAC

Supplementary Table S3.12. Genes targeted in perturb-seq experiment.

GENE	DESCRIPTION
ALDOA	Aldolase A; glycolytic enzyme
ATP5E	ATP synthase subunit
BCR-ABL	Fusion gene; drives CML-derived K562 cells
CAD	Pyrimidine nucleotide biosynthesis enzyme; catalyzes multiple pathway steps
CDC23	Anaphase promoting complex/cyclosome component
COX11	Mitochondrial respiratory chain; cytochrome c oxidase assembly factor
DBR1	Lariat debranching enzyme; required for lariat intron degradation after splicing
DUT	dUTP pyrophosphatase; involved in thymidine biosynthesis
EIF2S1	eIF2α; Translation initiation factor; translational control factor
GATA1	Erythroid-lineage transcription factor
GINS1	DNA replication initiation factor
GNB2L1	RACK1; 40s ribosomal protein; associated with numerous signalling processes
HSPA5	BiP; ER chaperone involved in protein import and folding
HSPA9	Mortalin; Mitochondrial chaperone and import factor
HSPE1	Mitochondrial chaperone
MTOR	Kinase; regulates growth, metabolism, and autophagy
POLR1D	RNA polymerase I and III subunit
POLR2H	RNA polymerase I, II, and III subunit
RAN	G-protein that controls protein and RNA transport through the nuclear pore
RPL9	Ribosomal protein L9
RPS14	Ribosomal protein S14
RPS15	Ribosomal protein S15
RPS18	Ribosomal protein S18
SEC61A1	ER translocon component
TUBB	beta-tubulin; structural component of microtubules

Supplementary Table S3.13. sgRNA sequences used in this study.

EXPERIMENT	NAME	SEQUENCE	TARGET
GFP single mismatches	EGFP-NT2	GACCAGGATGGGCACCACCC	EGFP
constant region RT-qPCR	DPH2_+_44435896.24-all	GAGTAAGCAGTCCTGGCACCC	DPH2
constant region RT-qPCR	DPH2_-_44435877.23-all	GATGTTTAGCAGCCCTGCCG	DPH2
constant region RT-qPCR	non-targeting_00564	GCCGATGGTCTTGTACTACA	N/A
constant region screen	RPL9_+_39460483.23-P1P2	GGATGTTTCTGTGCTCGTGG	RPL9
constant region screen	RPL9_+_39460504.23-P1P2	GCTGCGTCTACTGCGAGGTA	RPL9
constant region screen	RPL9_+_39460476.23-P1P2	GCTGTGCTCGTGGGGGTACT	RPL9
constant region screen	HSPE1_-_198365117.23-P1P2	GCGGACTGCGAGTCTCTTTG	HSPE1
constant region screen	HSPE1_+_198365089.23-P1P2	GGAGACTCGCAGTCCGGCCC	HSPE1
constant region screen	HSPE1_-_198365304.23-P1P2	GGCCCGATGGCACCTTGGAG	HSPE1
constant region screen	POLR1D_+_28196016.23-P1	GGGAAGCAAGGACCGACCGA	POLR1D
constant region screen	POLR1D_+_28196036.23-P1	GCGAGGCGCGGAGGCGAAGC	POLR1D
constant region screen	POLR1D_+_28196012.23-P1	GGCAAGGACCGACCGACGGA	POLR1D
constant region screen	SNRPD2_+_46195119.23-P1P2	GAGGCCGGGCTAGGCTTAGG	SNRPD2
constant region screen	SNRPD2_+_46195138.23-P1P2	GGCGTAGTGACCATCATGTG	SNRPD2
constant region screen	SNRPD2_-_46195150.23-P1P2	GCTAGCCCGGCCTCACATGA	SNRPD2
constant region screen	CDC23_+_137548970.23-P1P2	GAGTACCTCCATGGTCCCGG	CDC23
constant region screen	CDC23_-_137548987.23-P1P2	GACAGCCACCGGGACCATGG	CDC23
constant region screen	CDC23_-_137548622.23-P1P2	GCCAGTGACAGGGCACTCAG	CDC23
constant region screen	CAD_+_27440280.23-P1P2	GGCTGGAGAGAAGCCGGGCG	CAD
constant region screen	CAD_+_27440373.23-P1P2	GCGAGTACGGAGAAGCGGGA	CAD
constant region screen	CAD_+_27440253.23-P1P2	GTAGGAGCCTCGGGCGCGCT	CAD

EXPERIMENT	NAME	SEQUENCE	TARGET
constant region screen	TUBB_+_30688126.23-P1	GCGGCAGGAAGGTTCTGAGA	TUBB
constant region screen	TUBB_+_30688173.23-P1	GAGGTTGGAATGCGCCCCAG	TUBB
constant region screen	TUBB_+_30688145.23-P1	GCAGCGAGGTGCAAACGCGA	TUBB
constant region screen	POLR2H_-_184081237.23-P1P2	GGTGCACGTACTCCCAACTG	POLR2H
constant region screen	POLR2H_+_184081227.23-P1P2	GTGAGAGCGCGACCACAGTT	POLR2H
constant region screen	POLR2H_+_184081251.23-P1P2	GGGGCCACGAGAGCAGCAGA	POLR2H
constant region screen	DUT_+_48624414.23-P1P2	GAGGCGAGCGAGGAGACCAC	DUT
constant region screen	DUT_-_48624041.23-P1P2	GCGTCTGGAAGGAATCCACG	DUT
constant region screen	DUT_-_48623651.23-P1P2	GCAGGACGGGCGCGTCTTCA	DUT
constant region screen	DNAJC19_+_180707414.23-P1P2	GGGATGAGCCGTGCTCCCGG	DNAJC19
constant region screen	DNAJC19_+_180707118.23-P1P2	GCTTGCCTGGAACTCCTGTA	DNAJC19
constant region screen	DNAJC19_+_180707491.23-P1P2	GGGCGCCTGTGCTTGAGGTT	DNAJC19
constant region screen	non-targeting_03786	GTGGCCGTTCATGGGACCGG	N/A
constant region screen	non-targeting_03636	GACAATATCTGGATCGCCAA	N/A
constant region screen	non-targeting_03478	GGATGGGCTCGCCTGGCCAG	N/A
constant region screen	non-targeting_03229	GGTCCCACGGCGAAGCGACT	N/A
constant region screen	non-targeting_00564	GCCGATGGTCTTGTACTACA	N/A
constant region screen	non-targeting_00763	GGCGCGGGCCCCATAAAAAC	N/A
perturb-seq	RPS18_+_33239917.23-P1P2_00	GCTGCGATGCCGCTGGATCA	RPS18
perturb-seq	RPS18_+_33239917.23-P1P2_01	GCTGCAATGCCGCTGGATCA	RPS18
perturb-seq	RPS18_+_33239917.23-P1P2_02	GCTGGGATGCCGCTGGATCA	RPS18
perturb-seq	RPS18_+_33239917.23-P1P2_08	GCTGCGATTCCGCTGGATCA	RPS18
perturb-seq	RPS18_+_33239917.23-P1P2_04	GCTGCGATCCCGCTGGATCA	RPS18

EXPERIMENT	NAME	SEQUENCE	TARGET
perturb-seq	RPS14_+_149829238.23-P1P2_00	GAGGCCCGGGCGCG ACAATC	RPS14
perturb-seq	RPS14_+_149829238.23-P1P2_01	GAGACCCGGGCGCG ACAATC	RPS14
perturb-seq	RPS14_+_149829238.23-P1P2_02	GAGGCCCTGGCGCG ACAATC	RPS14
perturb-seq	RPS14_+_149829238.23-P1P2_04	GAGGCCCGCGCGCG ACAATC	RPS14
perturb-seq	RPS14_+_149829238.23-P1P2_13	GAGGCCCGGGCGCG ACAGTC	RPS14
perturb-seq	RPS14_+_149829238.23-P1P2_08	GAGGCCCGGGCTCG ACAATC	RPS14
perturb-seq	RPL9_+_39460483.23-P1P2_00	GGATGTTTCTGTGC TCGTGG	RPL9
perturb-seq	RPL9_+_39460483.23-P1P2_01	GGATGATTCTGTGC TCGTGG	RPL9
perturb-seq	RPL9_+_39460483.23-P1P2_05	GGATGTTTCGGTGC TCGTGG	RPL9
perturb-seq	RPL9_+_39460483.23-P1P2_04	GGATGTTTCAGTGC TCGTGG	RPL9
perturb-seq	RPL9_+_39460483.23-P1P2_07	GGATGTTTCTGCGC TCGTGG	RPL9
perturb-seq	GNB2L1_+_180670873.23-P1P2_00	GTGCAAGGCGGCGG CAGGAG	GNB2L1
perturb-seq	GNB2L1_+_180670873.23-P1P2_08	GTGCAAGGTGGCGG CAGGAG	GNB2L1
perturb-seq	GNB2L1_+_180670873.23-P1P2_13	GTGCAAGGCGGCGG CGGGAG	GNB2L1
perturb-seq	GNB2L1_+_180670873.23-P1P2_07	GTGCAAGGCGGGGG CAGGAG	GNB2L1
perturb-seq	GNB2L1_+_180670873.23-P1P2_02	GTGCAAGACGGCGG CAGGAG	GNB2L1
perturb-seq	RPS15_-_1438413.23-P1P2_00	GACCAAAGCGATCTC TTCTG	RPS15
perturb-seq	RPS15_-_1438413.23-P1P2_07	GACCAAAGCGGTCTC TTCTG	RPS15
perturb-seq	RPS15_-_1438413.23-P1P2_02	GACCAAGGCGATCTC TTCTG	RPS15
perturb-seq	RPS15_-_1438413.23-P1P2_12	GACCAAAGCGATCTC TTGTG	RPS15
perturb-seq	RPS15_-_1438413.23-P1P2_01	GACCAAACCGATCTC TTCTG	RPS15
perturb-seq	HSPE1_+_198365089.23-P1P2_00	GGAGACTCGCAGTC CGGCCC	HSPE1
perturb-seq	HSPE1_+_198365089.23-P1P2_01	GGAGACACGCAGTC CGGCCC	HSPE1

EXPERIMENT	NAME	SEQUENCE	TARGET
perturb-seq	HSPE1_+_198365089.23-P1P2_03	GGTGACTCGCAGTC CGGCCC	HSPE1
perturb-seq	HSPE1_+_198365089.23-P1P2_02	GGAGACTGGCAGTC CGGCCC	HSPE1
perturb-seq	HSPE1_+_198365089.23-P1P2_14	GGAGACTCGCAGTC CTGCCC	HSPE1
perturb-seq	RAN_+_131356438.23-P1P2_00	GGCGGTCGCTGCGC TTAGGG	RAN
perturb-seq	RAN_+_131356438.23-P1P2_02	GGCGGCCGCTGCGC TTAGGG	RAN
perturb-seq	RAN_+_131356438.23-P1P2_03	GGGGGTCGCTGCGC TTAGGG	RAN
perturb-seq	RAN_+_131356438.23-P1P2_04	GGCGGTCGCGGCGC TTAGGG	RAN
perturb-seq	RAN_+_131356438.23-P1P2_12	GGCGGTCGCTGCGC TTAGGT	RAN
perturb-seq	POLR1D_+_28196016.23-P1_00	GGGAAGCAAGGACC GACCGA	POLR1D
perturb-seq	POLR1D_+_28196016.23-P1_08	GGGAAGCAGGGACC GACCGA	POLR1D
perturb-seq	POLR1D_+_28196016.23-P1_03	GGTAAGCAAGGACC GACCGA	POLR1D
perturb-seq	POLR1D_+_28196016.23-P1_01	GGGAAGCCAGGACC GACCGA	POLR1D
perturb-seq	POLR1D_+_28196016.23-P1_07	GGGAAGCAAGGAGC GACCGA	POLR1D
perturb-seq	DBR1_+_137893744.23-P1P2_00	GTTTGCAGGAGTCTA CACCC	DBR1
perturb-seq	DBR1_+_137893744.23-P1P2_01	GATTGCAGGAGTCTA CACCC	DBR1
perturb-seq	DBR1_+_137893744.23-P1P2_07	GTTTGCAGGGGTCT ACACCC	DBR1
perturb-seq	DBR1_+_137893744.23-P1P2_05	GTTTGCAGGAGTGT ACACCC	DBR1
perturb-seq	DBR1_+_137893744.23-P1P2_08	GTTTGCAGTAGTCTA CACCC	DBR1
perturb-seq	SEC61A1_-_127771295.23-P1_00	GGCACTGACGTGTC TCTCGG	SEC61A1
perturb-seq	SEC61A1_-_127771295.23-P1_02	GGCGCTGACGTGTC TCTCGG	SEC61A1
perturb-seq	SEC61A1_-_127771295.23-P1_01	GGCACTGTCGTGTC TCTCGG	SEC61A1
perturb-seq	SEC61A1_-_127771295.23-P1_03	GGTACTGACGTGTC TCTCGG	SEC61A1
perturb-seq	SEC61A1_-_127771295.23-P1_04	GGCACTGAAGTGTC TCTCGG	SEC61A1

EXPERIMENT	NAME	SEQUENCE	TARGET
perturb-seq	HSPA5_+_128003624.23-P1P2_00	GAGCCGAGTAGGCG ACGGTG	HSPA5
perturb-seq	HSPA5_+_128003624.23-P1P2_04	GAGCCGAGAAGGCG ACGGTG	HSPA5
perturb-seq	HSPA5_+_128003624.23-P1P2_08	GAGCCGAGTGGGCG ACGGTG	HSPA5
perturb-seq	HSPA5_+_128003624.23-P1P2_01	GAACCGAGTAGGCG ACGGTG	HSPA5
perturb-seq	HSPA5_+_128003624.23-P1P2_06	GAGCCGAGTAGACG ACGGTG	HSPA5
perturb-seq	GINS1_-_25388381.23-P1P2_00	GGACTAGAACGAAAG GAGTG	GINS1
perturb-seq	GINS1_-_25388381.23-P1P2_08	GGACTAGAGCGAAAG GAGTG	GINS1
perturb-seq	GINS1_-_25388381.23-P1P2_06	GGACTAGAACGGAAG GAGTG	GINS1
perturb-seq	GINS1_-_25388381.23-P1P2_03	GGACTATAACGAAAG GAGTG	GINS1
perturb-seq	GINS1_-_25388381.23-P1P2_14	GGACTAGAACGAAAG GAGCG	GINS1
perturb-seq	CDC23_-_137548987.23-P1P2_00	GACAGCCACCGGGA CCATGG	CDC23
perturb-seq	CDC23_-_137548987.23-P1P2_02	GACAGCTACCGGGA CCATGG	CDC23
perturb-seq	CDC23_-_137548987.23-P1P2_08	GACAGCCATCGGGA CCATGG	CDC23
perturb-seq	CDC23_-_137548987.23-P1P2_04	GACAGCCAACGGGA CCATGG	CDC23
perturb-seq	CDC23_-_137548987.23-P1P2_11	GACAGCCACCGGGA CCACGG	CDC23
perturb-seq	CAD_+_27440280.23-P1P2_00	GGCTGGAGAGAAGC CGGGCG	CAD
perturb-seq	CAD_+_27440280.23-P1P2_03	GGCTGGTGAGAAGC CGGGCG	CAD
perturb-seq	CAD_+_27440280.23-P1P2_07	GGCTGGAGCGAAGC CGGGCG	CAD
perturb-seq	CAD_+_27440280.23-P1P2_06	GGCTGGAGAGTAGC CGGGCG	CAD
perturb-seq	CAD_+_27440280.23-P1P2_13	GGCTGGAGAGAAGC CTGGCG	CAD
perturb-seq	TUBB_+_30688126.23-P1_00	GCGGCAGGAAGGTT CTGAGA	TUBB
perturb-seq	TUBB_+_30688126.23-P1_01	GCAGCAGGAAGGTT CTGAGA	TUBB
perturb-seq	TUBB_+_30688126.23-P1_06	GCGGCAGGACGGTT CTGAGA	TUBB

EXPERIMENT	NAME	SEQUENCE	TARGET
perturb-seq	TUBB_+_30688126.23-P1_03	GCGGCAGCAAGGTTCTGAGA	TUBB
perturb-seq	TUBB_+_30688126.23-P1_10	GCGGCAGGAAGGTTCAGAGA	TUBB
perturb-seq	DUT_+_48624411.23-P1P2_00	GCGAGCGAGGAGACCACCGG	DUT
perturb-seq	DUT_+_48624411.23-P1P2_01	GCCAGCGAGGAGACCACCGG	DUT
perturb-seq	DUT_+_48624411.23-P1P2_08	GCGAGCGAGGAGGCCACCGG	DUT
perturb-seq	DUT_+_48624411.23-P1P2_07	GCGAGCGAGGAGCCCACCGG	DUT
perturb-seq	DUT_+_48624411.23-P1P2_10	GCGAGCGAGGAGACCAACGG	DUT
perturb-seq	POLR2H_+_184081251.23-P1P2_00	GGGGCCACGAGAGCAGCAGA	POLR2H
perturb-seq	POLR2H_+_184081251.23-P1P2_11	GGGGCCACGAGAGCAGCGGA	POLR2H
perturb-seq	POLR2H_+_184081251.23-P1P2_08	GGGGCCACGCGAGCAGCAGA	POLR2H
perturb-seq	POLR2H_+_184081251.23-P1P2_12	GGGGCCACGAGAGCAGGAGA	POLR2H
perturb-seq	POLR2H_+_184081251.23-P1P2_07	GGGGCCACGAGTGCAGCAGA	POLR2H
perturb-seq	GATA1_-_48645022.23-P1P2_00	GTGAGCTTGCCACATCCCCA	GATA1
perturb-seq	GATA1_-_48645022.23-P1P2_03	GTGCGCTTGCCACATCCCCA	GATA1
perturb-seq	GATA1_-_48645022.23-P1P2_04	GTGAGCTTACCACATCCCCA	GATA1
perturb-seq	GATA1_-_48645022.23-P1P2_08	GTGAGCTTTCCACATCCCCA	GATA1
perturb-seq	GATA1_-_48645022.23-P1P2_06	GTGAGCTTGCGACATCCCCA	GATA1
perturb-seq	GATA1_-_48645022.23-P1P2_12	GTGAGCTTGCCACATCCGCA	GATA1
perturb-seq	BCR_+_23523092.23-P1P2_00	GCGCGCGGGGCCCGTCTCAG	BCR
perturb-seq	BCR_+_23523092.23-P1P2_07	GCGCGCGGGGCTCGTCTCAG	BCR
perturb-seq	BCR_+_23523092.23-P1P2_04	GCGCGCGGAGCCCGTCTCAG	BCR
perturb-seq	BCR_+_23523092.23-P1P2_05	GCGCGCGGCGCCCGTCTCAG	BCR
perturb-seq	BCR_+_23523092.23-P1P2_15	GCGCGCGGGGCCCGTCGCAG	BCR

EXPERIMENT	NAME	SEQUENCE	TARGET
perturb-seq	BCR_+_23523092.23-P1P2_13	GCGCGCGGGGCCCATCTCAG	BCR
perturb-seq	HSPA9_-_137911079.23-P1P2_00	GGAGCTGCGCGATGCGGTGG	HSPA9
perturb-seq	HSPA9_-_137911079.23-P1P2_07	GGAGCTGCGGGATGCGGTGG	HSPA9
perturb-seq	HSPA9_-_137911079.23-P1P2_02	GGAGTTGCGCGATGCGGTGG	HSPA9
perturb-seq	HSPA9_-_137911079.23-P1P2_08	GGAGCTGCTCGATGCGGTGG	HSPA9
perturb-seq	HSPA9_-_137911079.23-P1P2_04	GGAGCTGCGCAATGCGGTGG	HSPA9
perturb-seq	EIF2S1_-_67827080.23-P1P2_00	GAGCGAAGCGCACGCTGAGG	EIF2S1
perturb-seq	EIF2S1_-_67827080.23-P1P2_06	GAGCGAAGCGCGCGCTGAGG	EIF2S1
perturb-seq	EIF2S1_-_67827080.23-P1P2_02	GAGCGCAGCGCACGCTGAGG	EIF2S1
perturb-seq	EIF2S1_-_67827080.23-P1P2_01	GAGCGAAACGCACGCTGAGG	EIF2S1
perturb-seq	EIF2S1_-_67827080.23-P1P2_07	GAGCGAAGCGCTCGCTGAGG	EIF2S1
perturb-seq	COX11_+_53045977.23-P1P2_00	GGCTCTGGCGTCCTGGATGG	COX11
perturb-seq	COX11_+_53045977.23-P1P2_03	GGCTCTGTCGTCCTGGATGG	COX11
perturb-seq	COX11_+_53045977.23-P1P2_04	GGCTCTGGCGCCCTGGATGG	COX11
perturb-seq	COX11_+_53045977.23-P1P2_05	GGCTCTGGCGTCTTGGATGG	COX11
perturb-seq	COX11_+_53045977.23-P1P2_10	GGCTCTGGCGTCCCGGATGG	COX11
perturb-seq	MTOR_+_11322547.23-P1P2_00	GGGCAGGGGGCCTGAAGCGG	MTOR
perturb-seq	MTOR_+_11322547.23-P1P2_07	GGGCAGGGGGTCTGAAGCGG	MTOR
perturb-seq	MTOR_+_11322547.23-P1P2_05	GGGCAGGGGGCTTGAAGCGG	MTOR
perturb-seq	MTOR_+_11322547.23-P1P2_06	GGGCAGGGGGGCTGAAGCGG	MTOR
perturb-seq	MTOR_+_11322547.23-P1P2_10	GGGCAGGGGGCCTGAAGCAG	MTOR
perturb-seq	ATP5E_-_57607036.23-P1P2_00	GGTGTCCAGGGGCACTCTGT	ATP5E
perturb-seq	ATP5E_-_57607036.23-P1P2_01	GGTGTCCTGGGGCACTCTGT	ATP5E

EXPERIMENT	NAME	SEQUENCE	TARGET
perturb-seq	ATP5E_-_57607036.23-P1P2_16	GGTGTCCAGGGGCG CTCTGT	ATP5E
perturb-seq	ATP5E_-_57607036.23-P1P2_04	GGTGTCCAGGAGCA CTCTGT	ATP5E
perturb-seq	ATP5E_-_57607036.23-P1P2_14	GGTGTCCAGGGGCA CTGTGT	ATP5E
perturb-seq	ALDOA_+_30077139.23-P1P2_00	GGTCACCAGGACCC CTTCTG	ALDOA
perturb-seq	ALDOA_+_30077139.23-P1P2_06	GGTCACCAGGATCC CTTCTG	ALDOA
perturb-seq	ALDOA_+_30077139.23-P1P2_07	GGTCACCAGGCCCC CTTCTG	ALDOA
perturb-seq	ALDOA_+_30077139.23-P1P2_14	GGTCACCAGGACCG CTTCTG	ALDOA
perturb-seq	ALDOA_+_30077139.23-P1P2_13	GGTCACCAGGACCC CTTTTG	ALDOA
perturb-seq	non-targeting_00001	GTGCACCCGGCTAG GACCGG	N/A
perturb-seq	non-targeting_00028	GGTGGCCTTTGCAA TTGGCG	N/A
perturb-seq	non-targeting_00054	GGGCCTGGACGAGC CTAAAA	N/A
perturb-seq	non-targeting_00089	GGGGTGAGGGTCCA ATTCGG	N/A
perturb-seq	non-targeting_00217	GTGAACTCAAAAATC CCGAC	N/A
perturb-seq	non-targeting_00283	GGGCCGACGGATAG GAGGGA	N/A
perturb-seq	non-targeting_00406	GGCGCCGGACTGGA CCTCGA	N/A
perturb-seq	non-targeting_00527	GTGGGAGCAGATCAA GACTC	N/A
perturb-seq	non-targeting_00802	GCACGACGCTCCGG CACGCG	N/A
perturb-seq	non-targeting_01040	GTACGGCATGGCGC ACTGCG	N/A

References

1. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, **337**, 816–821.

2. Sternberg,S.H., Redding,S., Jinek,M., Greene,E.C. and Doudna,J.A. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, **507**, 62–67.

3. Szczelkun,M.D., Tikhomirova,M.S., Sinkunas,T., Gasiunas,G., Karvelis,T., Pschera,P., Siksnys,V. and Seidel,R. (2014) Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *PNAS*, **111**, 9798–9803.

4. 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. *Cell*, **159**, 647–661.

5. Nishimasu,H., Ran,F.A., Hsu,P.D., Konermann,S., Shehata,S.I., Dohmae,N., Ishitani,R., Zhang,F. and Nureki,O. (2014) Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell*, **156**, 935–949.

6. 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.

7. Horlbeck,M.A., Gilbert,L.A., Villalta,J.E., Adamson,B., Pak,R.A., Chen,Y., Fields,A.P., Park,C.Y., Corn,J.E., Kampmann,M., *et al.* (2016) Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife*, **5**, e19760.

8. 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. *Nature Methods*, **6**, 443–445.

9. 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. *PNAS*, **110**, E2317–E2326.

10. 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. *Cell*, **155**, 1479–1491.

11. Grevet,J.D., Lan,X., Hamagami,N., Edwards,C.R., Sankaranarayanan,L., Ji,X., Bhardwaj,S.K., Face,C.J., Posocco,D.F., Abdulmalik,O., *et al.* (2018) Domain-focused CRISPR screen identifies HRI as a fetal hemoglobin regulator in human erythroid cells. *Science*, **361**, 285–290.

12. Datlinger,P., Rendeiro,A.F., Schmidl,C., Krausgruber,T., Traxler,P., Klughammer,J., Schuster,L.C., Kuchler,A., Alpar,D. and Bock,C. (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, **14**, 297–301.

13. Replogle,J.M., Xu,A., Norman,T.M., Meer,E.J., Terry,J.M., Riordan,D., Srinivas,N., Mikkelsen,T.S., Weissman,J.S. and Adamson,B. (2018) Direct capture of CRISPR guides enables scalable, multiplexed, and multi-omic Perturb-seq. *bioRxiv*, 10.1101/503367.

14. Adamson,B., Norman,T.M., Jost,M., Cho,M.Y., Nuñez,J.K., Chen,Y., Villalta,J.E., Gilbert,L.A., Horlbeck,M.A., Hein,M.Y., *et al.* (2016) A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, **167**, 1867-1882.e21.

15. Norman,T.M., Horlbeck,M.A., Replogle,J.M., Ge,A.Y., Xu,A., Jost,M., Gilbert,L.A. and Weissman,J.S. (2019) Exploring genetic interaction manifolds constructed from rich phenotypes. *bioRxiv*, 10.1101/601096.

16. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M., *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202–1214.

17. Boyle,E.A., Andreasson,J.O.L., Chircus,L.M., Sternberg,S.H., Wu,M.J., Guegler,C.K., Doudna,J.A. and Greenleaf,W.J. (2017) High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *PNAS*, **114**, 5461–5466.

18. 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**, 280.

19. Briner,A.E., Donohoue,P.D., Gomaa,A.A., Selle,K., Slorach,E.M., Nye,C.H., Haurwitz,R.E., Beisel,C.L., May,A.P. and Barrangou,R. (2014) Guide RNA Functional Modules Direct Cas9 Activity and Orthogonality. *Molecular Cell*, **56**, 333–339.

20. Eraslan,G., Avsec,Ž., Gagneur,J. and Theis,F.J. (2019) Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 10.1038/s41576-019-0122-6.

21. Kim,H.K., Min,S., Song,M., Jung,S., Choi,J.W., Kim,Y., Lee,S., Yoon,S. and Kim,H. (Henry) (2018) Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. *Nature Biotechnology*, **36**, 239–241.

22. Dixit,A., Parnas,O., Li,B., Chen,J., Fulco,C.P., Jerby-Arnon,L., Marjanovic,N.D., Dionne,D., Burks,T., Raychowdhury,R., *et al.* (2016) Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, **167**, 1853-1866.e17.

23. Jaitin,D.A., Weiner,A., Yofe,I., Lara-Astiaso,D., Keren-Shaul,H., David,E., Salame,T.M., Tanay,A., Oudenaarden,A. van and Amit,I. (2016) Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*, **167**, 1883-1896.e15.

24. Harding,H.P., Zhang,Y., Zeng,H., Novoa,I., Lu,P.D., Calfon,M., Sadri,N., Yun,C., Popko,B., Paules,R., *et al.* (2003) An Integrated Stress Response Regulates Amino Acid Metabolism and Resistance to Oxidative Stress. *Molecular Cell*, **11**, 619–633.

25. McInnes,L., Healy,J. and Melville,J. (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*.

26. Semenova,E., Jore,M.M., Datsenko,K.A., Semenova,A., Westra,E.R., Wanner,B., Oost,J. van der, Brouns,S.J.J. and Severinov,K. (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *PNAS*, **108**, 10098–10103.

27. Wiedenheft,B., Duijn,E. van, Bultema,J.B., Waghmare,S.P., Zhou,K., Barendregt,A., Westphal,W., Heck,A.J.R., Boekema,E.J., Dickman,M.J., *et al.* (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *PNAS*, **108**, 10092–10097.

28. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O., *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, **31**, 827–832.

29. Keren,L., Hausser,J., Lotan-Pompan,M., Vainberg Slutskin,I., Alisar,H., Kaminski,S., Weinberger,A., Alon,U., Milo,R. and Segal,E. (2016) Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell*, **166**, 1282-1294.e18.

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

6/13/2019

Date