Title
Mapping rich genotype-phenotype landscapes with single-cell CRISPR screens

Author
Replogle, Joseph Michael

Publication Date
2014

Peer reviewed|Thesis/dissertation

Mapping rich genotype-phenotype landscapes with single-cell CRISPR screens

by
Joseph Replogle

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

in

Genetics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Alexander Marson
7F25CDCE383C4A8...

Alexander Marson
Chair

DocuSigned by:
Jonathan S. Weissman
[illegible]4E2...

Jonathan S. Weissman

DocuSigned by:
Jimmie Ye
[illegible]429...

Jimmie Ye

DocuSigned by:
Jonathan K. Pritchard
06E0982F3C4F458...

Jonathan K. Pritchard

Committee Members

Dedication and Acknowledgments

I would like to thank my mentor Jonathan Weissman for his intellectual guidance and support over the past four years. I would also like to thank my thesis committee members, Alexander Marson, Jimmie Ye, and Jonathan Pritchard, for their insightful feedback, as well as the many faculty members who I interacted with during my coursework and qualifying exams. I am grateful to members of the Weissman lab for support, encouragement, and advice - both scientific and personal. I would especially like to thank: Britt Adamson, who I worked closely with on the development of direct capture Perturb-seq; Tom Norman, whose approach to data analysis and science I always try to emulate; Jeffrey Hussmann, who has been a great friend and roommate; Marco Jost, who has answered my questions about all things CRISPR; Alina Guna, who has been a constant companion thinking about mitochondria; Max Horlbeck and Luke Gilbert, who served as my mentors during my rotation and beyond. I'd also like to thank the collaborators that I have had the opportunity to work with during my PhD, both in academia and in industry. To my family and friends across the world – from Phoenix to Zurich – thank you for your continual love and support. To John Replogle and David Phizicky, I count myself incredibly lucky that you two have also been at MIT during these years. To my mom, your constant willingness to listen has allowed me to grow over the years. Finally, thanks to Sara – I cannot encapsulate how much I have enjoyed going through this challenge together.

Contributions

Chapter 1 *is reprinted largely as it appears in:*

Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Cogan JZ, Meer EJ, Terry JM, Riordan DP, Srinivas N, Fiddes IT, Arthur JG, Alvarado LJ, Pfeiffer KA, Mikkelsen TS, Weissman JS, Adamson B. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat Biotechnol. 2020 Aug;38(8):954-961. doi: 10.1038/s41587-020-0470-y. PMID: 32231336.

Chapter 2 *includes contributions from:*

Reuben Saunders, Angela Pogson, Eric Wagner, Karen Adelman, Thomas Norman, and Jonathan Weissman

Abstract

"Mapping rich genotype-phenotype landscapes with single-cell CRISPR screens"

Joseph M. Replogle

A central goal of genetics is to define the phenotypic consequences of genetic perturbations. Single-cell CRISPR screens such as Perturb-seq (pooled single-cell RNA-sequencing CRISPR screens) represent an emerging tool to systematically construct genotype-phenotype maps by pairing high-dimensional genetic perturbations with rich phenotypic readouts in single cells. However, to date, these screens have been deployed at a limited scale.

The first chapter of this thesis addresses a major technological hinderance to the scalable application of Perturb-seq: reliance on indirect indexing of single-guide RNAs (sgRNAs). I present direct-capture Perturb-seq, a versatile screening approach in which expressed sgRNAs are sequenced alongside single-cell transcriptomes. Direct-capture Perturb-seq enables detection of multiple distinct sgRNA sequences from individual cells and thus allows pooled single-cell CRISPR screens to be easily paired with combinatorial perturbation libraries that contain dual-guide expression vectors. I then demonstrate the utility of this approach for high-throughput investigations of genetic interactions and, leveraging this ability, dissect epistatic interactions between cholesterol biogenesis and DNA repair. Using direct capture Perturb-seq, I also show that targeting individual genes with multiple sgRNAs per cell improves the efficacy of CRISPR interference and activation, facilitating the use of compact, highly active CRISPR libraries for large-scale single-cell screens. Last, I show that hybridization-based target enrichment permits sensitive, specific sequencing of informative transcripts from single-cell RNA-seq experiments.

The second chapter builds on the first chapter by applying direct capture Perturb-seq and multiplexed CRISPR interference (CRISPRi) to perform the first genome-scale Perturb-seq screens. From these data, I yield a blueprint for the construction and analysis of rich genotype-phenotype maps. I show that genes can be clustered by transcriptional phenotypes across many

essential cellular processes and reveal new roles for poorly characterized genes in ribosome biogenesis, transcription, and respiration. Beyond clustering genes, these data enable in-depth dissection of the functional consequences of genetic perturbations on a remarkable array of complex, composite phenotypes—including RNA processing, differentiation, and chromosomal instability. Leveraging this ability, I comprehensively identify genetic drivers and consequences of aneuploidy, and I uncover unanticipated perturbation-specific regulation of the mitochondrial genome. This thesis establishes Perturb-seq as a scalable tool for the principled exploration of multidimensional cellular behaviors, gene function, and regulatory networks.

Table of Contents

List of Figures

INTRODUCTION

This thesis work takes a new technological approach to address a central goal of genetics and functional genomics: the comprehensive mapping between genotypes and phenotypes. The first chapter presents the development of a genetic screening platform that combines high-dimensional genotypes with rich phenotypic readouts at single-cell resolution. The second chapter applies this platform to perform large-scale genetic screens in human cell lines, which yields a blueprint for the construction and analysis of rich genotype-phenotype maps. Each chapter contains distinct introductory material, results, figures, and discussion sections.

In *Chapter 1*, I present the design and experimental validation of a new platform for rich phenotypic screens, termed “direct capture Perturb-seq”, which was conducted in collaboration with Britt Adamson at Princeton University and a team at 10x Genomics. Previous Perturb-seq platforms were inflexible in their application due to relying on proxy transcripts to read out the CRISPR single guide RNA (sgRNA) within each cell. Depending on the experimental platform, this inflexibility either reduced experimental scale or prevented the application to certain questions, such as the study of genetic interactions. I designed a platform that allowed for directly sequencing sgRNAs from each cell, rather than proxy transcripts, and I extensively validated this system to study the Unfolded Protein Response (UPR). This simple difference in technical approach opened the door to many new experimental applications. To demonstrate this, I used my platform to study the genetic interaction between DNA repair genes and genes involved in cholesterol biosynthesis. Importantly, I also demonstrated that this new approach enabled the use of multiplexed sgRNA libraries that targeted each gene with multiple distinct sgRNAs to improve CRISPR efficacy. This improvement in sgRNA library design ended up playing an important role in allowing for large-scale Perturb-seq experiments as described in *Chapter 2*. Finally, I showed that select transcripts could be enriched from single-cell RNA-

sequencing libraries using biotin-ligated oligonucleotides for hybridization-based pulldown. In sum, this work included both the development of a new screening platform and also highlighted a number of new concepts to aid the application of this platform to functional genomic studies.

In *Chapter 2,* I build on the work presented in *Chapter 1* using the direct capture Perturb-seq to perform the first genome-scale single-cell CRISPR screens in K562 and RPE1 cells. The data from these screens enabled the construction of a rich genotype-phenotype amp, where the transcriptional consequences of loss of every expressed gene was catalogued for the first time. The unique structure of this dataset enabled a range of relationships to be studied in a principled, systematic way, including gene-gene, gene-phenotype, and phenotype-phenotype relationships. Examining gene-gene relationships based on their transcriptional responses, I found that transcriptional phenotypes provided remarkably detailed insight into gene function and was able to group genes into pathways with numerous essential roles in the cell. From this analysis, I was able to predict the function for poorly-characterized genes with roles in translation and ribosome biogenesis, transcription, and respiration. Next, I examined the relationship between loss of particular genes with downstream phenotypes. I found that single-cell RNA-sequencing could be used to dissect many complex, composite cellular phenotypes, including genetically driven deviations in RNA splicing, expression of transposable elements, chromosomally instability, and cellular differentiation. Finally, I used this rich-phenotype map to discover a new cellular phenotype: perturbation-specific regulation of the expression of the mitochondrial genome. Personally, this was my favorite part of this work as it uncovered a framework with implications for understanding why the mitochondrial genome has been conserved through evolution.

The text, figures and tables in this thesis have been adapted with minor changes from work that is published (*Chapter 1; Replogle et al., Nature Biotechnology 2020*) and in preparation for publication (*Chapter 2*).

CHAPTER 1

BACKGROUND

CRISPR-based genetic tools have recently been paired with high-resolution phenotypic profiling to enable genetic screens with information rich readouts (Feldman et al., 2019; Packer and Trapnell, 2018; Rubin et al., 2018). These efforts have dramatically expanded our ability to investigate genetic control over complex cellular processes. One such approach, independently implemented as Perturb-seq (Adamson et al., 2016; Dixit et al., 2016), CRISP-seq (Jaitin et al., 2016), Mosaic-seq (Xie et al., 2017), and CROP-seq (Datlinger et al., 2017) and herein referred to as single-cell CRISPR screening, combines pooled CRISPR screens with single-cell RNA-sequencing (scRNA-seq) readouts to facilitate unbiased exploration of gene function and systematic delineation of genetic regulatory networks. However, current implementations face technical and practical limitations that unnecessarily restrict their use. Here, we present advances that address these limitations, specifically poor scalability, dependence on specialized vector systems, and high cost (Adamson et al., 2018; Feldman et al., 2018; Hill et al., 2018; Xie et al., 2018), and by doing so, we enable facile and scalable single-cell analysis of both single and combinatorial genetic perturbations. In particular, we establish a method for interrogating programmed pairs of CRISPR sgRNAs by scRNA-seq, thus enabling efforts to study redundant gene isoforms or paralogs, investigate cis-regulatory genome architecture (Gasperini et al., 2019), evade knockout rescue (Smits et al., 2019), generate precise genetic edits (Anzalone et al., 2019; Ran et al., 2013), or map genetic interactions (GIs) (Norman et al., 2019).

The technological crux of all single-cell CRISPR screens is the assignment of perturbation identities to single-cell phenotypes. To achieve this, scRNA-seq screening platforms typically rely on polyadenylated indexes. These indexes are co-expressed with non-polyadenylated sgRNAs,

but unlike the sgRNAs, they can be recorded on standard scRNA-seq platforms that capture only polyadenylated RNAs (**Supplementary Figure 1.1a,b**). However, recombination of indexed sgRNA libraries during lentiviral delivery can uncouple indexes from their assigned sgRNAs (Adamson et al., 2018; Feldman et al., 2018; Hill et al., 2018; Xie et al., 2018). This means that such platforms are limited to arrayed use and restricted scale (Hill et al., 2018; Xie et al., 2018). Notably, one method, CROP-seq, has minimized this problem (Datlinger et al., 2017). CROP-seq uses a clever vector system to deliver sgRNAs to cells. This vector duplicates the sequence of a single encoded sgRNA during lentiviral transduction to produce two expression cassettes on the same construct: one that expresses a functional sgRNA and another that expresses a polyadenylated transcript carrying the sgRNA sequence at the 3' end. In this way, CROP-seq ensures delivery of pooled guide libraries to cells with faithful pairing of sgRNAs and polyadenylated “indexes”. However, due to constraints on cassette size, CROP-seq is thought to be incompatible with delivery of multiple sgRNAs.

RESULTS

To establish tools for more versatile single-cell CRISPR screens, we sought to directly sequence sgRNAs alongside single-cell transcriptomes in a method we refer to as “direct capture Perturb-seq”. Breifly, droplet-based scRNA-seq uses molecular barcoding to identify transcripts from individual cells. This barcoding occurs during reverse transcription (RT), when both unique molecular identifiers (UMIs) and cell barcodes (CBCs) are added to the 3' or 5' ends of mRNA sequences (**Supplementary Figure 1.1a,b**) (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017). For direct capture Perturb-seq, we extended this barcoding to non-polyadenylated sgRNAs by addition of guide-specific primers during RT (**Figure 1.1a,b**). To maximize flexibility, we designed platforms for direct capture with both 5' and 3' scRNA-seq. For 5' scRNA-seq, this required the simple addition of an unbarcoded guide-specific RT primer to standard protocols (**Figure 1.1a** and **Supplementary Figure 1.1b**), an approach also reported by Mimitou *et al*. while this work was under review (Mimitou et al., 2019). For 3' scRNA-seq, the RT configuration necessitated that we implement an entirely new scRNA-seq platform (**Figure 1.1b**). This platform concurrently delivers target-specific, barcoded primers to single-cell reactions alongside barcoded oligo-dT (**Figure 1.1b** and **Supplementary Figure 1.1a**). These target-specific primers anneal to capture sequences (cs1 and cs2) in modified sgRNA constant regions and thus enable RT of sgRNAs and efficient recording of sgRNA sequences (**Supplementary Figure 1.1c,d,e,f**). We selected the capture sequences for our platform carefully, to ensure for example, that their incorporation into an optimized sgRNA constant region (CR1) would not compromise guide activity. However, these capture sequences are not guide-specific, and thus, in principle, will enable multiplexed capture of additional features, such as antibodies (Peterson et al., 2017; Stoeckius et al., 2017) and other oligo-tagged markers (Zhang et al., 2018). Herein, we refer to guides with cs1 incorporated in a stem loop of our standard *Streptococcus pyogenes* Cas9 sgRNAs as sgRNA-CR1^{cs1} and guides with cs2 incorporated at the 3' end as sgRNA-CR1^{cs2}. We

note that an alternate configuration with incorporation of cs1 at the 3' end compromises activity and therefore is not recommended (**Supplementary Figure 1.1f**).

To test the performance of guide capture, we next performed 5 parallel CRISPRi-based (Gilbert et al., 2014, 2013) screens in K562 cells designed to compare 3' direct capture and 5' direct capture to indexing by a polyadenylated barcode transcript (hereafter referred to as guide barcode or GBC Perturb-seq). On each platform, we screened one or more sgRNA libraries containing the same 32 targeting sequences (against 30 genes whose depletion leads to activation of the unfolded protein response, UPR, and including 2 non-targeting controls (Adamson et al., 2016)). To enable comparison, we prepared each of these libraries using arrayed cloning and lentiviral packaging, and after performing our screens, used custom protocols to amplify index molecules (GBCs or guides) for deep sequencing alongside mRNA sequences (**Supplementary Figure 1.1g**; see *Methods*). At a constant sequencing depth, screens using both direct capture platforms gave higher index capture than the GBC-based method (4.1-fold higher for 3' sgRNA-CR1^{cs1} capture; 15.5-fold higher for 5' sgRNA-CR1^{cs1} capture; 7.8-fold higher for 5' sgRNA-CR1 capture), with the exception of 3' capture of sgRNA-CR1^{cs2}, which had modestly lower capture (0.56-fold) (**Supplementary Figure 1.2a**). To assign guide identities to cells, we then fit a two-component Poisson-Gaussian mixture model to the $\log_2$-transformed guide UMIs per cell for each guide (**Supplementary Figure 1.2b**; see *Methods*). This approach aims to separate true guide-expressing cells from “background cells” which arise from spurious cell barcode-sgRNA pairing (potentially due to PCR chimeras or capture of ambient guides). Unlike capture of GBCs, we found that guide capture was sequence-dependent with capture rates varying across guides by targeting sequence (**Figure 1.1c** and **Supplementary Figure 1.2c,d**). Importantly, this variation was correlated across screens and was related to the nucleotides at the 5' ends of guide RNAs but not to overall GC content (**Supplementary Figure 1.2e-g**). Nevertheless, our assignment procedure robustly assigned guide identities to 84-94% of cells (compared to 89% for GBC Perturb-seq) with roughly expected guide distributions across all

platforms (**Figure 1.1d** and **Supplementary Figure 1.2h**). Moreover, indicative of robust assignment, we found strong (and comparable) target depletion across platforms (median knockdown: 90% for GBC capture, 94% for 3' sgRNA-CR1^{cs1} capture, 93% for 3' sgRNA-CR1^{cs2} capture, 95% for 5' sgRNA-CR1 capture, 93% for 5' sgRNA-CR1^{cs1} capture) (**Supplementary Figure 1.3a**).

We then sought to benchmark the performance of direct capture Perturb-seq for the study of genes and genetic networks. High-content Perturb-seq phenotypes should enable (1) functional clustering of target genes, (2) identification of transcriptional phenotypes caused by individual perturbations, (3) delineation of gene expression regulons, and (4) identification of cell-to-cell heterogeneities. We therefore asked how phenotypes from our direct capture screens with the highest guide assignment rates (3' capture of sgRNA-CR1^{cs1} and 5' capture of standard CR1 sgRNAs) performed on each of these tasks (compared to GBC Perturb-seq). First, we hierarchically clustered target genes based on their pseudo-bulk expression profiles (**Figure 1.1e**). This recapitulated known functional and physical interactions and, when compared to results generated with GBC Perturb-seq, produced highly similar relationships (cophenetic correlation with GBC Perturb-seq: r=0.95 for 3' sgRNA-CR1^{cs1}; r=0.95 for 5' sgRNA-CR1). Next, we evaluated transcriptional responses and found good agreement across screens / platforms (for the top 100 differentially expressed genes, r = 0.88 for 3' sgRNA-CR1^{cs1} capture compared to GBC Perturb-seq and r=0.87 for 5' sgRNA-CR1 capture compared to GBC Perturb-seq) with especially high correlations for perturbations that led to differential expression of >100 genes (**Supplementary Figure 1.3b,c**). This result confirms our ability to accurately assign guide identities and suggests that our guide-specific RT primers do not globally alter single-cell gene expression profiles. Next, we tested the utility of direct capture Perturb-seq for the discovery of genetic networks. For this, we relied on our prior empirical classification of genes regulated by the three separate signaling branches of the UPR (Adamson et al., 2016). Examining the covariance of these genes across single cells in our current data, we found gene expression

modules that were conserved across platforms (cophenetic correlation with GBC Perturb-seq: r=0.93 for 3' sgRNA-CR1^{cs1} capture; r=0.95 for 5' sgRNA-CR1 capture), with modules tending to cluster functionally based on their regulation by the three UPR branches (**Figure 1.1f**). Lastly, we quantitatively evaluated the single-cell performance of our platforms by training a random forest classifier to classify perturbed and unperturbed (control) cells for each targeting guide. Despite the intrinsic noise of scRNA-seq data, prediction accuracies were highly similar across platforms (correlation with GBC Perturb-seq: r=0.91 for 3' sgRNA-CR1^{cs1} capture, r=0.90 for 5' sgRNA-CR1 capture) (**Figure 1.1g** and **Supplementary Figure 1.3d**).

To demonstrate the versatility of direct capture Perturb-seq, we next performed a 3' direct capture Perturb-seq experiment in induced pluripotent stem cells (iPSCs) with Cas9 (Mandegar et al., 2016), now using pooled lentiviral packaging and transduction of 40 sgRNAs. In iPSCs, we again found high guide capture rates (mean capture of 999 UMIs/cell; **Supplementary Figure 1.4a,b**) and transcriptional phenotypes that were correlated for guides targeting the same gene (**Supplementary Figure 1.4c,d).**

Recently, we showed that GBC Perturb-seq can be coupled with epistasis analysis to provide mechanistic insights into how genes interact (Norman et al., 2019). However, GBC Perturb-seq is not scalable. Motivated by this limitation, we next explored the use of direct capture Perturb-seq to study genetic interactions, specifically a complex set of GIs we recently identified between genes that control cholesterol biosynthesis (e.g. *FDPS*, *MVD*, and *IDI1*) and genes that facilitate DNA repair (e.g. *ATR* and genes encoding components of the 9-1-1 complex) (Horlbeck et al., 2018). For this, we cloned a CRISPRi library of 92 programmed sgRNA pairs targeting 41 genes and 81 gene pairs using a strategy for pooled cloning of dual-guide vectors (**Figure 1.2a** and **Supplementary Figure 1.5a**). Notably, we made and tested this library in two configurations, using two combinations of guide constant region sequences (CR3^{cs1}/CR1^{cs1} and CR2^{cs2}/CR1^{cs1}). Screening this library in K562s by 3' direct capture Perturb-seq revealed adequate capture of guides from both vector positions (position A, median of 776 UMIs/cell; position B, median of 511

UMIs/cell; **Supplementary Figure 1.5b,c**), and after mapping these guides to cells, we observed >90% guide assignment with >67% of guide-bearing cells expressing two sgRNAs, as expected given multiple infections, doublets from cell loading, and imperfect guide calling (**Figure 1.2b**; see *Methods*). Importantly, consistent with similar dual-guide expression systems (Adamson et al., 2016; Horlbeck et al., 2018; Norman et al., 2019), we also observed comparable knockdown between the two positions in our dual-guide vector. Specifically, for three guides in our library (sgHUS1, sgFDPS, and sgTOPBP1) encoded in both positions paired with a non-targeting guide, we achieved target knockdown of 84% and 84% (position A and position B), 81% and 73% (position A and position B), 70% and 74% (position A and position B), respectively. Lastly, we note that the design of our dual-guide expression system minimizes intramolecular recombination between linked sgRNA sequences by using distinct U6 promoters and sgRNA constant regions, as previously demonstrated by Adamson *et al.* (Adamson et al., 2016); however, it does not prevent paired sgRNA pairs from shuffling due to intermolecular recombination events. Nevertheless, because direct capture allows us to assign sgRNAs to cells in an unbiased way, we were able to identify novel sgRNA pairs and excluded them from downstream analysis.

We previously proposed a model for GIs between cholesterol biosynthesis and DNA repair genes wherein repression of the former leads to the buildup of toxic metabolic intermediates in cells, which then cause replicative stress and genotoxin-activated cell cycle arrest (Horlbeck et al., 2018). This model emerged from a set of low-content and low-throughout biochemical and functional experiments that primarily investigated the relationship between two genes: *FDPS* and a key mediator of the replication checkpoint machinery, *HUS1*. By contrast, direct capture Perturb-seq allows simultaneous interrogation of many single and double genetic perturbations with information rich phenotypes. The method therefore enabled us to thoroughly examine how cells respond to depletion of several enzymes in the cholesterol biosynthesis pathway (**Figure 1.2c**). From this, we made three clear observations. First, as expected, perturbation of early pathway steps led to feedback marked by upregulation of cholesterol biosynthesis genes (**Figure 1.2d**).

Second, repression of intermediate pathway genes (*FDPS*, *MVD*, and *IDI1*), which are synthetic lethal with DNA repair genes, led to an accumulation of cells in S-phase of the cell cycle (**Figure 1.2d**). Notably, the difference in phenotypes among genes within this linear biosynthetic pathway supports the idea that buildup of toxic intermediates, rather than depletion of cholesterol itself, leads to S-phase arrest. Lastly, as predicted by our model, we observed a buffering relationship between genes that regulate the early and intermediate steps in cholesterol biosynthesis. Specifically, when we fit a regression model to decompose dual-gene perturbations into linear combinations of single-gene effects, we observed that *PMVK* repression suppressed the *FDPS*-specific transcriptional response, while maintaining the cholesterol feedback response shared by both perturbations (**Figure 1.2e**). This further suggests that S-phase arrest is caused by loss of the enzymatic activity of the intermediate genes, not from a loss of cholesterol itself.

While repression of *HUS1* causes modest cell-cycle aberration alone, in combination with *FDPS* knockdown, we observed a substantial bypass of the S-phase checkpoint and an accumulation of cells in G2/M (**Figure 1.2f, Supplementary Figure 1.5d**). Additionally, we found that these cells (with perturbation of both *HUS1* and *FDPS*) demonstrate a neomorphic phenotype characterized by a transcriptional response not induced by either perturbation alone (**Figure 1.2g**). At the single-cell level, this generated a population of G2/M-arrested cells with notably decreased total mRNA content, likely representing dying cells (**Figure 1.2h**). Based on this, we propose an updated model where synthetic lethality in these cells is caused specifically by failure to detect replication stress in HUS1-depleted cells, resulting in inappropriate cell cycle progression and mitotic catastrophe from unresolved damage. Broadly, this example highlights the power of high-resolution, single-cell phenotypes for the mechanistic dissection of GIs and demonstrates how direct capture Perturb-seq can be used to understand GIs in a comprehensive, unbiased fashion without the need for specific hypotheses.

To further enable single-cell CRISPR screening efforts, we next tested the idea that genetic perturbation libraries that co-deliver multiple sgRNAs per gene to the same cell could

increase screening efficiency by requiring fewer constructs per gene. To test this, we selected pairs of CRISPRi and CRISPR activation (CRISPRa) sgRNAs with high predicted activity (Horlbeck et al., 2016) against individual genes that span a range of biological functions and expression levels (87 for CRISPRi; 49 for CRISPRa). Then, using our pooled library cloning strategy and direct capture Perturb-seq in K562 cells, we compared the activity (knockdown or activation) of the guide pairs expressed from a dual-guide vector to single sgRNAs (expressed from the same dual-guide vector paired with non-targeting control sgRNAs). For both CRISPRi and CRISPRa, the multiplexed sgRNAs nearly doubled CRISPR activity over what was achieved with the best single guide (CRISPRi: sgRNAs 1+control, median relative target expression=0.20; sgRNAs 1+2, median relative target expression=0.11; Wilcoxon signed-rank two-sided test n=87 genes, W=378, *p*=8e-11; CRISPRa: sgRNAs 1+control, median fold-activation=2.9; sgRNAs 1+2, median fold-activation=4.7; Wilcoxon signed-rank two-sided test n=49 genes, W=162, p=7e-6; **Figure 1.3a,b** and **Supplementary Figure 1.6a,b**). Moreover, in both cases, the multiplexed sgRNAs appeared to perform better than expected based on a dominant model of guide activity, suggesting some degree of synergy between multiplexed sgRNAs (CRISPRi Wilcoxon signed-rank two-sided test: n=87 genes, W=698, *p*=3e-7; CRISPRa Wilcoxon signed-rank two-sided test: n=49 genes, W=233, *p*=0.0002; **Supplementary Figure 1.6c,d,e,f**) which is consistent with previous reports (Moreno et al., 2018; Savell et al., 2019). These results show that compact, highly active libraries (expressing multiple sgRNAs per gene) can be used to scale single-cell experiments with direct capture Perturb-seq, interrogating more genes while minimizing false negatives due to insufficient expression modulation.

Lastly, we addressed the fact that current droplet-based scRNA-seq implementations are constrained by sequencing the whole transcriptome for phenotyping, which can be prohibitively expensive. This requirement is compounded by the fact that the distribution of gene expression is skewed (i.e. 2% of expressed genes consume >50% of sequencing reads; **Supplementary Figure 1.7a**) and biased across gene functions. Indeed, genes with important biological functions

(e.g., transcription factors, cell-surface receptors, kinases) are often lowly-expressed and difficult to measure (**Supplementary Figure 1.7b**). However, given a suitable method for targeted enrichment of scRNA-seq libraries, many transcriptional states could be faithfully inferred from a subset of gene expression measurements (Cleary et al., 2017; Subramanian et al., 2017). Diverse approaches exist for enriching transcripts using multiplexed PCR (Salomon et al., 2019), custom RT beads (Saikia et al., 2019), or linear amplification (Vallejo et al., 2019); however, each of these is limited by number of target genes, quality, and/or *a priori* gene selection. Instead, we hypothesized that we could use hybridization-based target enrichment to specifically sequence thousands of select transcripts, thereby limiting sequencing while maintaining high-content phenotypes.

To test target enrichment, we therefore empirically-designed hybridization baits for 978 genes, the L1000 landmark genes (Subramanian et al., 2017). We chose these genes because they can serve as a reduced representation of the whole transcriptome and their expression levels span four orders of magnitude, providing ample range to examine potential biases introduced by hybridization capture. In our test, we performed a pulldown on 3' scRNA-seq library and deeply sequenced recovered molecules. Hybridization capture increased the percentage of mRNA molecules aligning to target genes by >14-fold, from 6% in an unenriched control to 87% after target enrichment (**Figure 1.3c**). Thus, at only ~0.1x sequencing depth of the original library, the enriched library contains more UMIs per cell for most targeted genes (**Supplementary Figure 1.7c**). Enriched gene expression profiles were highly correlated with unenriched profiles at the global (r=0.98), single-cell (median r=0.93), and single-gene (median r=0.75) levels (**Figure 1.3d** and **Supplementary Figure 1.7d,e,f**), and perturbation-dependent differential gene expression was highly similar before and after enrichment (median r=0.71; **Supplementary Figure 1.7g**). Given these results, we next tested the ability of our reduced transcriptome subset to functionally cluster genes. Hybridization capture on our multiplexed CRISPRi Perturb-seq libraries revealed that L1000-targeted gene expression profiles can recapitulate relationships between genetic

perturbations established by sequencing the entire transcriptome (cophenetic correlation r=0.95; **Figure 1.3e,f** and **Supplementary Figure 1.7h**). Altogether, these results demonstrate that hybridization capture is a simple and sensitive procedure for informative enrichment of tailored gene sets from scRNA-seq libraries. Moreover, because our target enrichment procedure is performed on final libraries, target genes do not need to be selected *a priori* and can be iteratively refined for a single experiment. This technology motivates the future optimization of gene sets that maximize biological information while minimizing sequencing requirements (analogous to the L1000 landmark genes for hybridization-based fluorescent assays (Subramanian et al., 2017)).

DISCUSSION

Since its inception, single-cell CRISPR screening has made it possible to simultaneously examine high-dimensional genotypic and phenotypic landscapes. Here, we described an improved Perturb-seq approach that substantially expands the scale and flexibility of this technology (**Figure 1.3g**). Importantly, our 5' and 3' direct capture Perturb-seq platforms (now commercially available from 10x Genomics) have crucial advantages under different circumstances. For example, 5' direct capture is compatible with standard sgRNAs, has higher guide capture rates, and allows for V(D)J clonotype analysis, whereas 3' direct capture is compatible with many molecular recording and lineage tracing approaches (Chan et al., 2019). We specifically demonstrated the value of direct capture for the mechanistic dissection of GIs, which is a laborious undertaking with other methods, and for generating compact, highly-active CRISPR libraries. Additionally, to decrease the cost of Perturb-seq experiments, we implemented hybridization-based target enrichment. With our target enrichment strategy, biologically meaningful gene panels (e.g., immune, developmental, metabolic, tumor suppressors/oncogenes, etc.) can be probed without unnecessary sequencing of housekeeping genes. Taken together, direct capture Perturb-seq and target enrichment greatly expand the accessibility, scalability, and flexibility of single-cell CRISPR screens.

MATERIALS AND METHODS

Cell culture and viral production. RPMI-1640 with 25mM HEPES, 2.0 g/L NaHCO3, 0.3 g/L L-Glutamine supplemented with 10% FBS, 2 mM glutamine, 100 units/mL penicillin and 100 µg/mL streptomycin was used to grow K562 cells. HEK293T cells, used for packaging lentivirus, were grown in Dulbecco's modified eagle medium (DMEM) in 10% FBS, 100 units/mL penicillin and 100 µg/mL streptomycin. Induced pluripotent stem cells (iPSCs) expressing Cas9 (WTC CRISPRn Gen1C(Mandegar et al., 2016)) were maintained under feeder-free conditions on growth factor-reduced Matrigel (Corning) in mTeSR medium (STEMCELL Technologies). Accutase (STEMCELL Technologies) was used to enzymatically dissociate iPSCs into single cells to passage with 10 µM p160-Rho-associated coiled-coil kinase (ROCK) inhibitor Y-27632 (Selleckchem) added to promote cell survival. Lentivirus was produced by co-transfecting HEK293T cells with transfer plasmids and standard packaging vectors using *Trans*IT®-LTI Transfection Reagent (Mirus, MIR 2306).

Plasmid construction and development of sgRNA capture sequences. Direct capture guide RNAs were designed by appending non-random capture sequences to the 3' end of standard guide sequences or by inserting these sequences into the loop region of the so-called "stem loop 2" (**Supplementary Figure 1.1c**). Expression vectors encoding these guides are available at Addgene. To test the activity of modified guides and guide expression vectors, expression constructs carrying GFP-targeting guides with variant constant regions were transduced into GFP+ K562 dCas9-KRAB cells(Adamson et al., 2016) with centrifugation (2 hours at 1000 x g at 33°C). Cells were analyzed by flow cytometry on an LSR II flow cytometer (BD Biosciences). Data in **Supplementary Figure 1.1d** were processed as follows: Measurements of median GFP were recorded from GFP+ K562 dCas9-KRAB cells transduced with the indicated guides. These measurements were adjusted by subtracting background fluorescence (collected from control

cells that do not express GFP) and then divided by measurements of median GFP (also background-subtracted) recorded from cells without a GFP-targeting guide (untransduced cells grown in the same wells). These “GFP remaining” ratios were then normalized to those derived from cells transduced with a positive control guide, our standard sgRNA-CR1 (on plate control) and are reported as averages of triplicates from separate infections. Data in **Supplementary Figure 1.5a** are shown as the Gaussian kernel density estimates of normalized flow-cytometry measurements representing GFP expression of all cells with the indicated guide RNAs. We chose final guide designs based on these GFP depletion results. The reverse complements of our final capture sequences (cs1 5'-GCTTTAAGGCCGGTCCTAGCAA-3' and cs2 5'-GCTCACCTATTAGCGGCTAAGG-3') were incorporated into gel beads in the Chromium Single Cell 3' Reagent Kits v3 with Feature Barcoding technology.

Pilot UPR direct capture Perturb-seq. For these experiments, we constructed three CRISPRi libraries (the UPR GBC, UPR sgRNA-CR1^{cs1}, and UPR sgRNA-CR1^{cs2} libraries) by arrayed cloning. Each of these libraries encodes guide RNAs programmed with 32 unique guide RNA targeting regions: 30 which target genes whose depletion was previously shown to activate the unfolded protein response (UPR) by GBC Perturb-seq (Adamson et al., 2016) and 2 non-targeting controls (sgNegCtrl2 and sgNegCtrl3). Our sequence-verified libraries were then packaged into lentiviruses by arrayed transfection of individual vectors, and lentiviral preparations from each library were pooled for co-transduction into K562 dCas9-KRAB cells (Gilbert et al., 2014) (spinfected 2 hours at 1000 x g at 33 C). To ensure representation of guides at the time of scRNA-seq (7 days after transduction), lentiviral pooling was performed in a manner that accounted for both packaging variability and guide effects on cell growth after transduction (as determined by individual test infections). Pooling ratios were designed to ensure even representation among targeting guides and delivery of sgNegCtrl2 and sgNegCtrl3 at 4-fold excess. Three days post infection, we measured BFP expression (a marker for guide transduction) on an LSR II flow

cytometer (BD Biosciences) and calculated the multiplicity of infection (MOI) for each library (0.04 for the GBC library, 0.05 for the sgRNA-CR1^{cs1} library, and 0.05 for the sgRNA-CR1^{cs2} library). Transduced cells were sorted to near purity (FACSAria2, BD Biosciences). Up to this point, cell viability for all three libraries remained >87%.

Seven days post infection, cells were separated into droplet emulsions using the Chromium Controller (10x Genomics) across 5 lanes (cell pools >94% BFP+). Cells transduced with the UPR GBC library were loaded on two lanes. On one lane, cell capture was performed with Chromium Single Cell 3' Gel Beads v2 (GBC Perturb-seq), while on the other, cell capture was performed with Chromium Single Cell 5' Gel Beads and a spike-in of 5 pmols of oJR160 (5'-AAGCAGTGGTATCAACGCAGAGTACCAAGTTGATAACGGACTAGCC-3') to the RT Master Mix (5' direct capture Perturb-seq using standard CR1 sgRNAs). Similarly, cells transduced with the UPR sgRNA-CR1^{cs1} library were loaded onto two lanes. On one lane, cell capture was performed with Chromium Single Cell 3' Gel Beads v3 (GBC Perturb-seq), while on the other, cell capture was performed with Chromium Single Cell 5' Gel Beads and a 5 pmol spike-in of oJR161 (5'- AAGCAGTGGTATCAACGCAGAGTACTTGCTAGGACCGGCCTTAAAGC-3') to the RT Master Mix (5' direct capture Perturb-seq using sgRNA-CR1^{cs1}). Finally, cells transduced with UPR sgRNA-CR1^{cs2} were loaded onto a single lane with Chromium Single Cell 3' Gel Beads v3 (3' direct capture Perturb-seq using sgRNA-CR1^{cs2}). For all lanes, cells were loaded to recover ~10,000 cells (~260 cells per guide). Approximately 100 pmols of 10x RT oligo (poly-dT RT Primer PN 2000007) are added to each 10x RT reaction based on quantification by NanoDrop spectrophotometer (Thermo Scientific). Therefore, for 5' direct capture Perturb-seq, we chose to add our guide capture oligos at ~5%. The recovered cells and subsamples thereof are analyzed in **Figure 1.1c-g** and **Supplementary Figures 1.2** and **1.3**.

iPSC 3' direct capture Perturb-seq. To test direct capture Perturb-seq with iPSC cells, we constructed a sequence-verified library of 40 guides using the sgRNA-CR1^{cs1} design by arrayed

cloning (**Supplementary Note 2** and **Supplementary Table 3**). This "iPSC sgRNA-CR1^{cs1}" library was then packaged into lentivirus (pooled format), and transduced into iPSCs carrying inducible Cas9[29] at an MOI of 10%. iPSCs were treated daily with 2 μM doxycycline (Sigma) to drive Cas9 expression, and after two days, BFP+ cells were enriched on a BD FACS Aria2 (BFP is a marker of guide vector transduction). Seven days post-infection, cells were separated into droplet emulsions using the Chromium Controller (10x Genomics) with Chromium Single Cell 3' Gel Beads v3. The recovered cells and subsamples thereof are analyzed in **Supplementary Figure 1.4**.

Dual-guide 3' direct capture Perturb-seq to evaluate genetic interactions. To dissect the interaction between cholesterol biosynthesis and DNA repair, we constructed two sequence-verified dual-guide libraries of manually curated guide pairs by pooled cloning (CR3^{cs1}/CR1^{cs1} and CR2^{cs2}/CR1^{cs1}). For each library, lentivirus was prepared in a pooled format and transduced into K562 dCas9-KRAB cells (spinfection for 2 hours at 1000 x g). Three days post infection, we calculated MOIs (using BFP expression) and sorted transduced cells to near purity (LSR II and FACSAria2, BD Biosciences). To maximize the probability of observing interpretable transcriptional responses, we sampled cells at two time points (day 6 and day 9 post-transduction). At 6 days post infection, we separated cells transduced with CR3^{cs1}/ CR1^{cs1} (MOI=0.15, 89% BFP+) and CR2^{cs2}/ CR1^{cs1} (MOI=0.18, 93% BFP+) libraries into droplet emulsions using the Chromium Controller with Chromium Single Cell 3' Gel Beads v3. At 9 days post infection, we did the same for a second population of cells transduced with the CR3^{cs1}/ CR1^{cs1} library (MOI=0.1, 83% BFP+), both times aiming to recover 15,000 cells per lane.

Multiplexed CRISPRi and CRISPRa 3' direct capture Perturb-seq. To determine whether guide multiplexing can be used to construct compact CRISPRi and CRISPRa libraries, we built and analyzed dual-guide libraries wherein vectors contain either one targeting guide (paired with

a negative control guide) or two guides targeting a single gene. For these libraries, we manually chose gene targets representing a broad range of biological functions and expression levels (87 for CRISPRi, 49 for CRISPRa). We selected guide RNA targeting sequences predicted to be highly active (the top two by rank in hCRISPRi v2.1 and hCRISPRa v2[30]). In this manuscript, we refer to guides containing the top ranked targeting sequence as “sgRNA 1” and the next best as “sgRNA 2”. Of note, when the selected targeting sequence pairs targeted genomic sequence <80 bp apart, we also included the next best-ranked guide RNA spaced >80 bp away from the first. We refer to guides containing these targeting sequences as “sgRNA 3”. Additionally, for genes with two annotated transcription start sites (TSSs), we paired the top sgRNAs targeting each TSS. We cloned these libraries in pooled format. We then packaged each library into lentivirus (pooled format) and transduced K562 dCas9-KRAB cells(Gilbert et al., 2014) (CRISPRi) and K562 dCas9-SunTag/scFV-VP64 cells (Gilbert et al., 2014) (CRISPRa) with the appropriate library. Three days post infection, we calculated MOIs (by BFP expression) of 0.1 and 0.045, respectively, and sorted transduced cells to near purity (LSR II and FACSAria2, BD Biosciences). Then, 8 days post infection, we separated cells (CRISPRi at 90% BFP+ and CRISPRa at 88% BFP+) into droplet emulsions using the Chromium Controller with Chromium Single Cell 3' Gel Beads v3, aiming to recover 15,000 cells per lane.

Sequencing library preparation. GBC Perturb-seq sequencing libraries were prepared according to the Chromium Single Cell 3' Reagent Kits v2 User Guide (10x Genomics CG00052) with 11 cycles of PCR during cDNA amplification and 11 cycles of Sample Index PCR. Library molecules containing guide barcodes (GBCs) were specifically amplified using KAPA HiFi ReadyMix with 30 ng of the final library as template, 0.6 mM 052-P5 (5'-AATGATACGGCGACCACCGAGATCTACAC-3'), and 0.6 mM of i7 barcoded 055-N708 (5'-CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGGACCTCCCTAGCAAACTGGGGCACAAG-3'). PCR cycling was performed according

to the following protocol: (1) 95 C for 3 min, (2) 14 cycles of 98 C for 15 s, then 70 C for 10 s, (3) 72 C for 1 min. The resulting GBC sequencing library was purified via a 0.8X SPRI selection.

Our 3' direct capture Perturb-seq libraries were prepared using a protocol modified from the Chromium Single Cell 3' Reagent Kits v3 User Guide (10x Genomics, CG000184) (**Supplementary Figure 1.1b**). Briefly, following 11 cycles of cDNA amplification, library amplicons were size separated into two fractions (**Supplementary Figure 1.1g**): one enriched for amplicons containing guide sequences (by performing a 0.6X-1.2X double-sided SPRI), and the other (eluted from the 0.6X left-sided SPRI) containing larger cDNA amplicons. We processed the latter into gene expression sequencing libraries according to the Chromium Single Cell 3' Reagent Kits v3 User Guide (in this case, using 10 cycles of Sample Index PCR). In parallel, we used the guide-enriched cDNA amplicons to make perturbation index sequencing libraries. For this, guide-enriched cDNAs amplicons were purified by an additional 1X SPRI selection (30 uL elution). The eluted material (5 uL) was then used as template in the following nested PCR strategy: PCR1 with 50 uL Amp Mix (10x Genomics, PN#2000047), 45 uL Feature SI Primers 1 (10x Genomics, PN#2000098), and cycling by (1) 98 C for 45 s, (2) 12 cycles of 98 C for 20 s, then 60 C for 5 s, then 72 for 5 s, (3) 72 C for 1 min. PCR2 with the products of the first PCRs (5 uL after cleanup using a 1X SPRI selection and elution in 30 uL), 50 uL of Amp Mix (10x Genomics, PN#2000047), 35 uL of Feature SI Primers 2 (10x Genomics, PN#2000098), and cycling by (1) 98 C for 45 s, (2) 5 cycles of 98 C for 20 s, then 54 C for 30 s, then 72 for 20 s, (3) 72 C for 1 min. Finally, the resulting guide sequencing libraries were cleaned up via a double-sided 0.7X-1.0X SPRI selection.

Our 5' direct capture Perturb-seq sequencing libraries were prepared using a protocol modified from the Chromium Single Cell V(D)J Reagent Kits User Guide (10x Genomics CG000086) (**Supplementary Figure 1.1g**). For this, we used two direct capture spike-in oligos oJR160 and oJR161, each with an adapter identical to the adapter sequence contained on the Poly-dT RT Primer from 10x Genomics (PN-2000007). This adapter serves as a primer binding

site for the Non-Poly(dT) primer (10x Genomics, PN-220106) during cDNA amplification, and thus allows amplification of reverse transcribed guides to occur concurrently with standard cDNA amplification. Following 11 cycles of amplification, cDNAs amplicons were size separated into two fractions as described immediately above. Following this, the fractions were processed into gene expression libraries (according to the Chromium Single Cell V(D)J Reagent Kits User Guide with 14 cycles of Sample Index PCR) and index sequencing libraries. For our 5' sgRNA-CR1^{cs1} experiment, guide molecules were amplified using 0.6 mM oJR163 and oJR166 (5'-CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGTGTATAAG AGACAGGTACTTGCTAGGACCGGCCTTAAAGC-3'). Because the resulting index library had a contaminating low-molecular weight species (suspected primer dimers) an additional selection for 248-302 bp fragments was performed using a BluePippin (Sage Science) prior to sequencing.

Hybridization-based target enrichment. To enrich select transcripts from single-cell gene expression libraries for deep sequencing (so-called “target enrichment”), we developed a hybridization capture protocol. Briefly, using 120 nt biotinylated oligos, we performed streptavidin pulldowns of target sequencing amplicons from indexed 10x Genomics gene expression libraries. We tested our enrichment approach in two scenarios: (1) a single library from our dual-guide 3' direct capture Perturb-seq experiment to evaluate genetic interactions, and (2) all 16 libraries (in two pools of 8) from our multiplexed CRISPRi 3' direct capture Perturb-seq experiment. Using the former, we compared deep sequencing of both the unenriched and enriched libraries, and using the latter, we tested the functional utility of L1000 transcriptomes. For pooling libraries, we mixed 187.5 ng from each to obtain a total mass of 1500 ng and dried the library with a SpeedVac. We then followed steps 4 to 7 of the published Twist Biosciences protocol for hybrid capture. This protocol consists of a 16-hour probe hybridization at 70°C, pulldown of hybridized probes using streptavidin beads, and 5-cycles of post-capture PCR prior to sequencing.

Sequencing. For our UPR direct capture Perturb-seq experiments, both the gene expression and index sequencing libraries were sequenced using a NovaSeq 6000 S2 Reagent kit (Illumina) and a custom sequencing strategy (26 bp Read 1, 125 bp Read 2, and 8 bp Index Read 1) where the extended Read 2 was used to sequence guide RNA targeting regions in our 5' guide sequencing libraries. For all other libraries, we sequenced using the standard format for scRNA-seq from 10x Genomics (28 bp Read 1, 98 bp Read 2, and 8 bp Index Read 1) on a NovaSeq 6000 System (Illumina) with NovaSeq 6000 S4 Reagent kits (Illumina).

Data processing, statistics, and analysis. We used Cell Ranger 3.0 software (10x Genomics) for alignment of scRNA-seq reads, collapsing reads to unique molecular identifier (UMI) counts, cell calling, and depth normalization of transcriptome libraries. Index reads were aligned to expected sequences using bowtie for GBC Perturb-seq and bowtie2 for direct capture Perturb-seq. We observed index alignment rates of 0.82 for GBCs, 0.35 for 3' sgRNA-CR1^{cs2}, 0.62 for 3' sgRNA-CR1^{cs1}, 0.71 for 5' sgRNA-CR1, and 0.62 for 5' sgRNA-CR1^{cs1} in our UPR direct capture Perturb-seq experiments. Downstream analyses were performed in Python, using a combination of Numpy, Scipy, Pandas, scikit-learn, pomegranate, polo, and seaborn libraries.

Tests for differences in distributions (for example, of capture rates or correlations of guides within and between gene targets) were conducted with a two-sided Mann-Whitney U test (scipy.stats mannwhitneyu with use_continuity=True, alternative='two-sided'). Tests for differences in distributions for paired samples (for example, knockdown by single versus multiplexed guides) were carried out with a two-sided Wilcoxon signed rank test (scipy.stats wilcoxon with zero_method='wilcox', correction=False, alternative='two-sided'). Tests for differential gene expression were performed with a two-sample, two-sided Kolmogorov–Smirnov test and corrected for multiple-hypothesis testing at an FDR of 0.01 using the Benjamini–Yekutieli procedure. As indicated in the text, differentially expressed genes were also identified by random forest classifiers (scikit-learn extremely randomized trees with 1000 trees in the forest to predict

perturbation status). The advantage of this approach is that we assess the similarity of average expression profiles across platforms regardless of the strength of the perturbation because we do not employ a strict cutoff. Correlations reported are Pearson correlation coefficients unless otherwise indicated. Sample sizes used to calculate statistics are provided in the figures and legends.

Perturbation identity mapping. Within our sequencing data, we found evidence of perturbation index reads containing spurious cell barcode (CBC) / index pairs. We attribute these to the droplet encapsulation of ambient indexes (GBC transcripts or guides) and PCR chimeras. Therefore, to accurately assign guide identities to cells, true CBC/index pairs had to be determined. For GBC Perturb-seq, we did this using a threshold that separates the bimodal distribution of GBC coverage (reads per UMI) as previously described (Adamson et al., 2016). However, for direct capture Perturb-seq, we found that coverage distributions were not bimodal, at least not at the downsampled sequencing depth we used to compare libraries in our UPR experiments (25 million aligned indexing reads per GBC or guide sequencing library). At this sequencing depth, saturation of the index libraries is 0.75 for GBC, 0.96 for 3' sgRNA-CR1^{cs2}, 0.71 for 3' sgRNA-CR1^{cs1}, 0.28 for 5' sgRNA-CR1, and 0.60 for 5' sgRNA-CR1^{cs1}. Instead, we found that each guide had a bimodal distribution of the number of UMIs per CBC (capture rates) and that these rates vary across guide RNA targeting regions (perhaps influenced by targeting region-dependent variability in guide stability, Cas9 binding, and RT efficiency) (**Supplementary Figure 1.2e,f**).

Given targeting region-variable capture rates, to assign guide identities to cells, we fit a two-component mixture model, consisting of a Poisson (lower) and Gaussian (upper) distribution, to the $\log_2$ transformed capture rates (UMIs per CBC) for each guide RNA targeting region, as exemplified in **Supplementary Figure 1.2b**. These mixture models enabled us to separate the upper modes (representing transduced cells) from the lower modes (representing background) and thus assign guides to cells. Each cell with a posterior probability >0.5 of belonging to an upper

mode component was assigned a given guide identity. This procedure produced a coherent proportion of cells assigned to each guide identity (**Supplementary Figure 1.2h**) and a coherent multiplet rate across platforms–within 1-1.6 fold of expectations based on library transduction (assuming Poisson infection distribution) and published multiple encapsulation rates (**Figure 1.1d**).

In our UPR experiments, only cells with a single assigned guide were considered for downstream analysis; however, for dual-guide experiments, cells with two assigned guides were used. Across all dual-guide direct capture Perturb-seq experiments, we observed that >67% of cells contained exactly two sgRNAs. We attribute many of the cells with less than two assigned guides to stringent guide assignment cutoffs. For example, given ~90% assignment rate in single-guide experiments (on par with GBC Perturb-seq and CROP-seq), we expect only ~81% of dual-guide cells to be assigned exactly 2 guides. Yet our mapping strategy is clearly overly conservative as it assumes that guides are independently paired. To increase assignments rates in future applications, our mapping framework could be extended to fit a multivariate mixture model that jointly calls guides by leveraging shared information. Cells with more than two guides, on the other hand, may arise from either multiple infection events or double loading into droplets. Our loading scheme (designed to recover ~15,000 cells per lane) increased these doublets but notably also minimized reagent cost per recovered cell. Lastly, with direct capture Perturb-seq, we can identify cells bearing undesired sgRNA pairs (generated from intermolecular lentiviral recombination between programmed pairs (Adamson et al., 2018; Feldman et al., 2018; Hill et al., 2018; Xie et al., 2018)) and computationally exclude them from downstream analysis. In our data, we observed rates of novel pairs varying from 0.09-0.15 across experiments, which is roughly consistent with a previous report (Horlbeck et al., 2018).

Expression normalization, average expression profiles, and target knockdown or activation. We normalized for differences in capture and sequencing coverage across cells by

rescaling each cell to have the same total gene expression UMIs (i.e., each row of the raw expression matrix is rescaled to have the same sum). We then *z*-normalized expression of each gene with respect to the mean and standard deviation of that gene in the control cell population. We generated pseudo-bulk RNA-seq phenotypes for individual guides or guide pairs by averaging the normalized expression profiles of well-expressed genes–excluding genes with a mean expression <1 UMI per cell (**Figure 1.1e** and **Figure 1.3e**) and <0.5 UMI per cell (**Supplementary Figure 1.3b**) across all cells assigned that guide or guide pair (and excluding multiplets). We computed on-target gene knockdown as the ratio of the mean number of target UMIs in perturbed cells versus the mean number of target UMIs in control cells (bearing non-targeting sgRNAs), and we computed on-target gene activation as the ratio of the mean number of target UMIs in perturbed cells vs the mean number of target UMIs in controls.

FIGURES

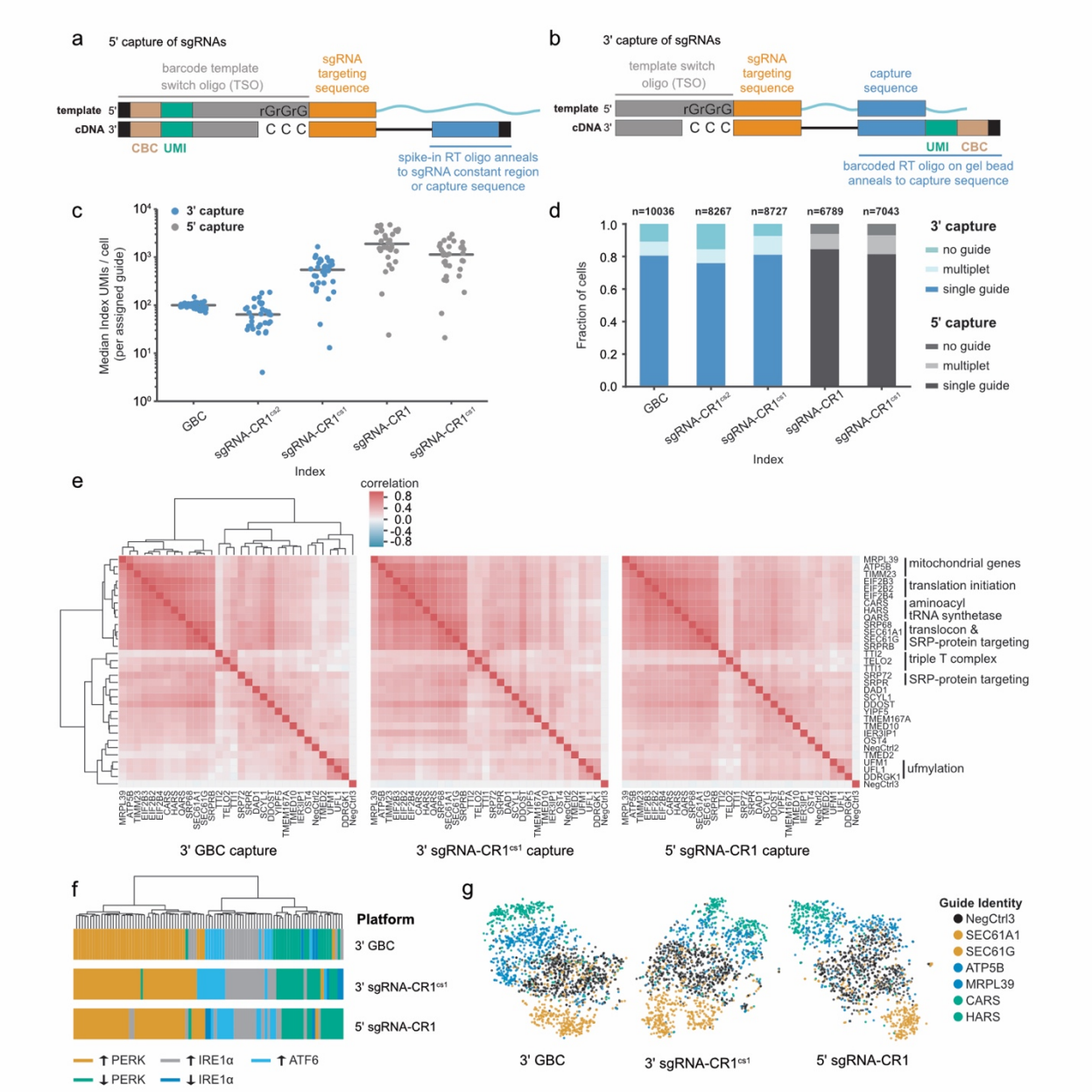


Figure 1.1: Design and validation of direct capture Perturb-seq for 3' and 5' single-cell RNA-sequencing. a) Schematic of sgRNA capture during 5' scRNA-seq. An sgRNA containing a standard constant region (top) anneals to a guide-specific RT oligo. Indexing of reverse transcribed cDNA (bottom) occurs after template switch. This strategy is compatible with unmodified sgRNAs (shown) or with sgRNAs with an integrated capture sequence. **b)** Schematic of sgRNA capture via an integrated capture sequence by 3' scRNA-seq. A capture sequence within the constant region of the sgRNA (top) anneals to a barcoded, target-specific RT primer. Indexed cDNA (bottom) is produced by reverse transcription. **c)** Index (GBC or guide) capture rates per cell across experiments conducted with GBC Perturb-seq and direct capture Perturb-seq. Data represent median index UMI counts per cell for cells bearing each of n=32 sgRNAs across platforms. Grey lines indicate median values. “sgRNA-CR1” indicates 5' capture of

standard sgRNAs without a capture sequence. **d)** Index (GBC or guide) assignment rates across experiments conducted with GBC Perturb-seq and direct capture Perturb-seq. The total number of cells per experiment as well as the fractions of cells assigned no guide, a single guide, or more than one guide are indicated. “sgRNA-CR1” indicates 5' capture of standard sgRNAs without a capture sequence. **e)** Clustering of perturbations from UPR Perturb-seq experiments conducted with GBC Perturb-seq and direct capture Perturb-seq. Heatmaps represent Spearman’s rank correlations between pseudo-bulk expression profiles for each of n=32 perturbations. For visual comparison, the rows and columns of all three heatmaps are ordered identically based on the hierarchical clustering of GBC Perturb-seq data. Functional annotations are indicated. **f)** Hierarchical clustering of UPR-regulated genes based on co-expression in each of the indicated Perturb-seq experiments. Colors indicate membership in different UPR-regulated groups as determined by Adamson *et al*.(Adamson et al., 2016) **g)** Single-cell projections are based on t-sne visualization of 10 independent components (n=1795 cells for 3' GBC Perturb-seq, n=1595 cells for 3' sgRNA-CR1^{cs1} Perturb-seq, and n=1424 cells for 5' sgRNA-CR1 Perturb-seq). Colors indicate functional similarities among targeted genes.

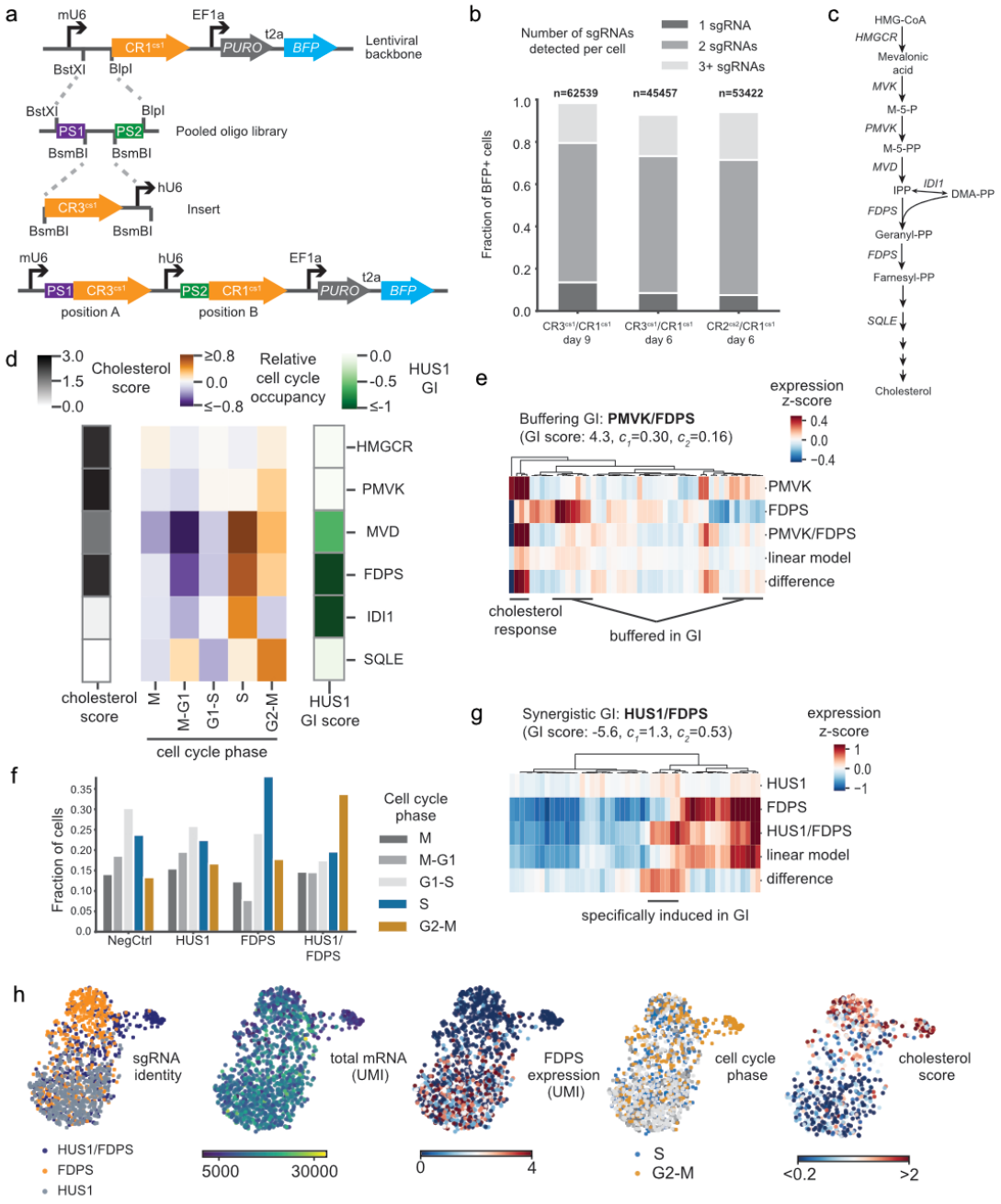


Figure 1.2: Direct capture Perturb-seq and pooled dual-guide cloning allows systematic dissection of genetic interactions between cholesterol biosynthesis and DNA repair genes. **a)** Schematic of programmed dual-guide library cloning strategy. Paired sgRNA targeting regions are synthesized on a single oligo and cloned into a direct capture Perturb-seq vector by ligation. Then, an sgRNA constant region and hU6 promoter are inserted between the sgRNA targeting regions to generate a dual-guide array in a lentiviral backbone. This example shows a $CR3^{cs1}/CR1^{cs1}$ library design. **b)** Guide assignment rates for dual-guide direct capture Perturb-seq experiments. The fraction of cells carrying sgRNAs (marked by BFP) varied due to strong CRISPRi growth defects; the total number of cells were therefore first scaled by BFP positivity. The total number of cells and fraction of cells assigned a single guide, two guides, or more than

two guides are indicated. **c)** Schematic of the cholesterol biosynthesis pathway. **d)** Heatmap of cell cycle and cholesterol phenotypes for cells with depletion of enzymes in the cholesterol biosynthesis pathway. Cell cycle occupancy for each perturbation depicted indicates the relative enrichment or depletion of cells in each phase relative to unperturbed cells. The cholesterol score is the mean z-scored expression of enzymes in the cholesterol biosynthesis pathway. The "HUS1 GI" is a metric of the growth defect caused by an genetic perturbations paired with *HUS1* knockdown relative to the genetic perturbation alone as determined by Horlbeck *et al.* (Horlbeck et al., 2018) All genes were significantly depleted by CRISPRi (percent knockdown: *HMGCR* 94%; *PMVK* 92%; *MVD* 83%; *FDPS* 78%; *IDI1* 82%; *SQLE* 84%). Number of cells per perturbation: non-targeting control n=527, HMGCR n=608, PMVK n=389, MVD n=184, FDPS n=439, IDI1 n=131, SQLE n=255. **e)** Heatmap of gene expression for the 50 most differentially expressed genes between cells carrying each indicated perturbation. Expression values are the z-scored expression relative to unperturbed cells (n=389 PMVK cells, n=1921 FDPS cells, and n=517 PMVK/FDPS cells). Cells were combined to generate the expression signatures. Knockdown was consistent between single-gene and dual-gene targeting (*FDPS* knockdown 73% alone vs. 82% paired; *PMVK* knockdown 92% alone vs. 86% paired). The indicated GI score was previously determined by Horlbeck *et al*. (Horlbeck et al., 2018), where GI scores >3 are considered strongly buffering interactions. **f)** Fraction of cells in each cell cycle phase across cells with the indicated perturbations. Number of cells per perturbation: non-targeting control n=780, HUS1 n=905, FDPS n=439, HUS1/FDPS n=831. **g)** Heatmap of gene expression for the 50 most differentially expressed genes between cells carrying each indicated perturbation. Expression values are the z-scored expression relative to unperturbed cells (n=905 HUS1 cells, n=439 FDPS cells, and n=831 HUS1/FDPS cells). Cells were combined to generate the expression signatures. Knockdown was consistent between single-gene and dual-gene targeting (*FDPS* knockdown 78% alone vs. 72% paired; *HUS1* knockdown 95% alone vs. 85% paired) The indicated GI score was previously determined by Horlbeck *et al.* (Horlbeck et al., 2018), where GI scores <-3 are considered strongly synergistic interactions. **h)** Single-cell UMAP projections with informative cell features highlighted (n=2175 cells).

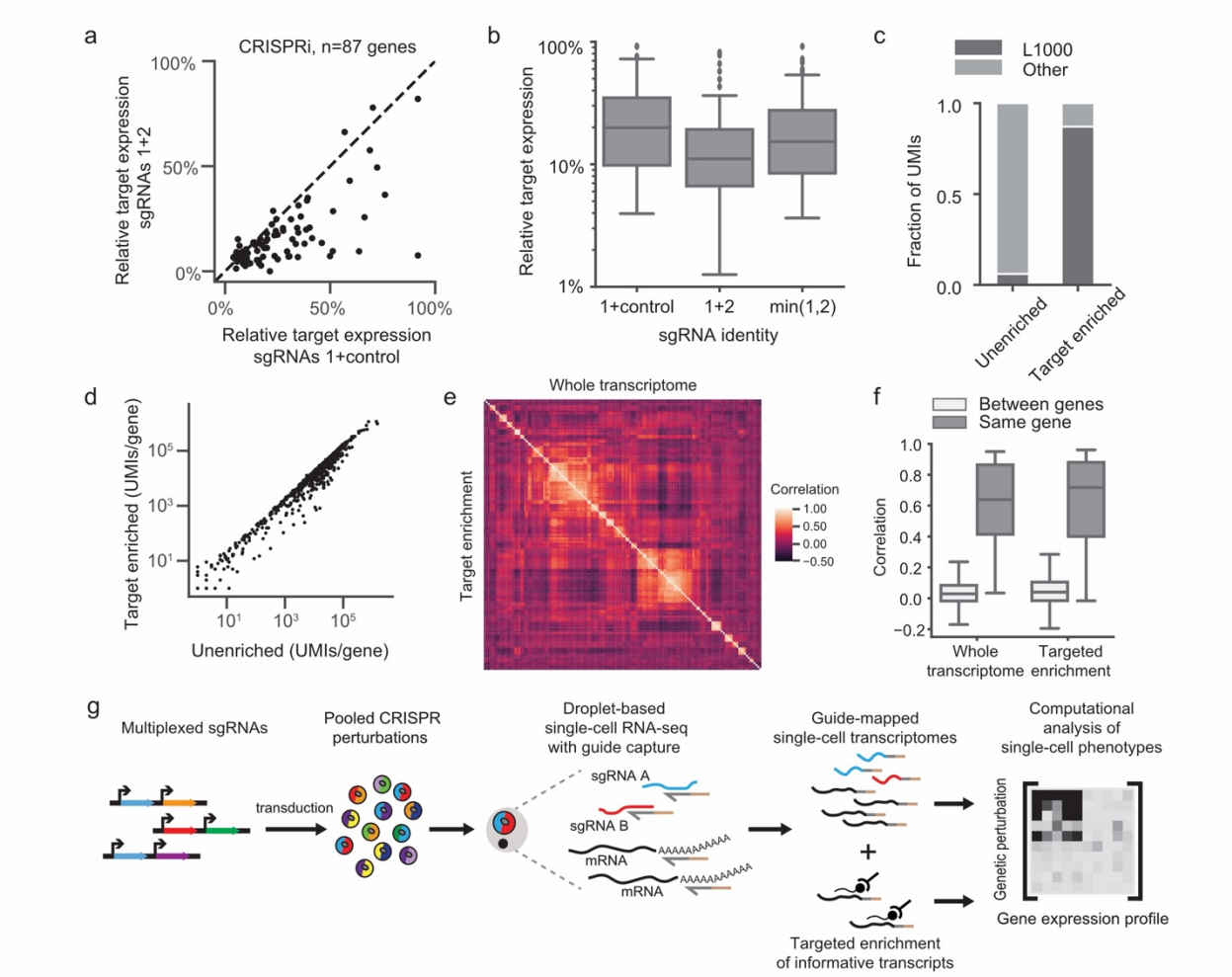
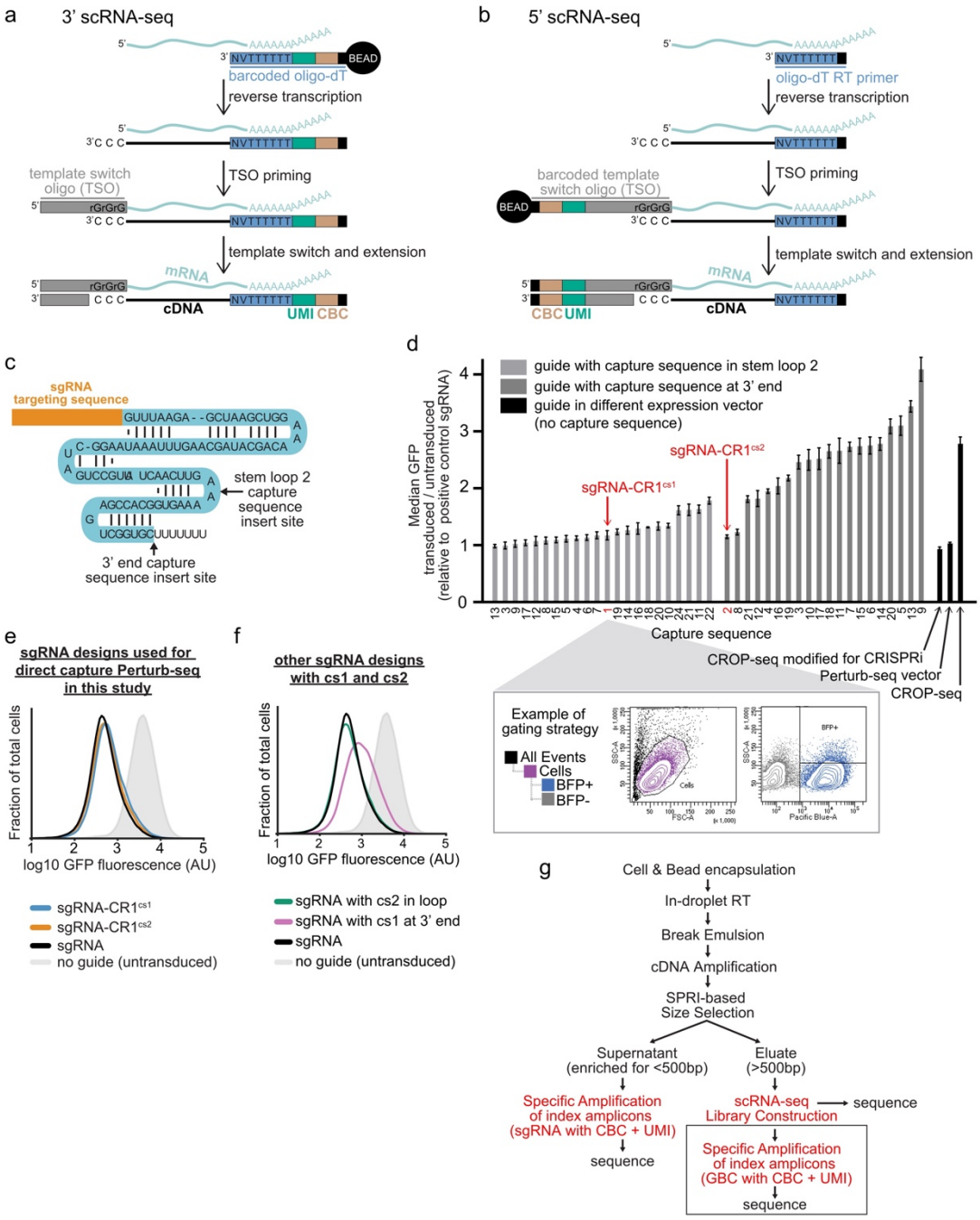


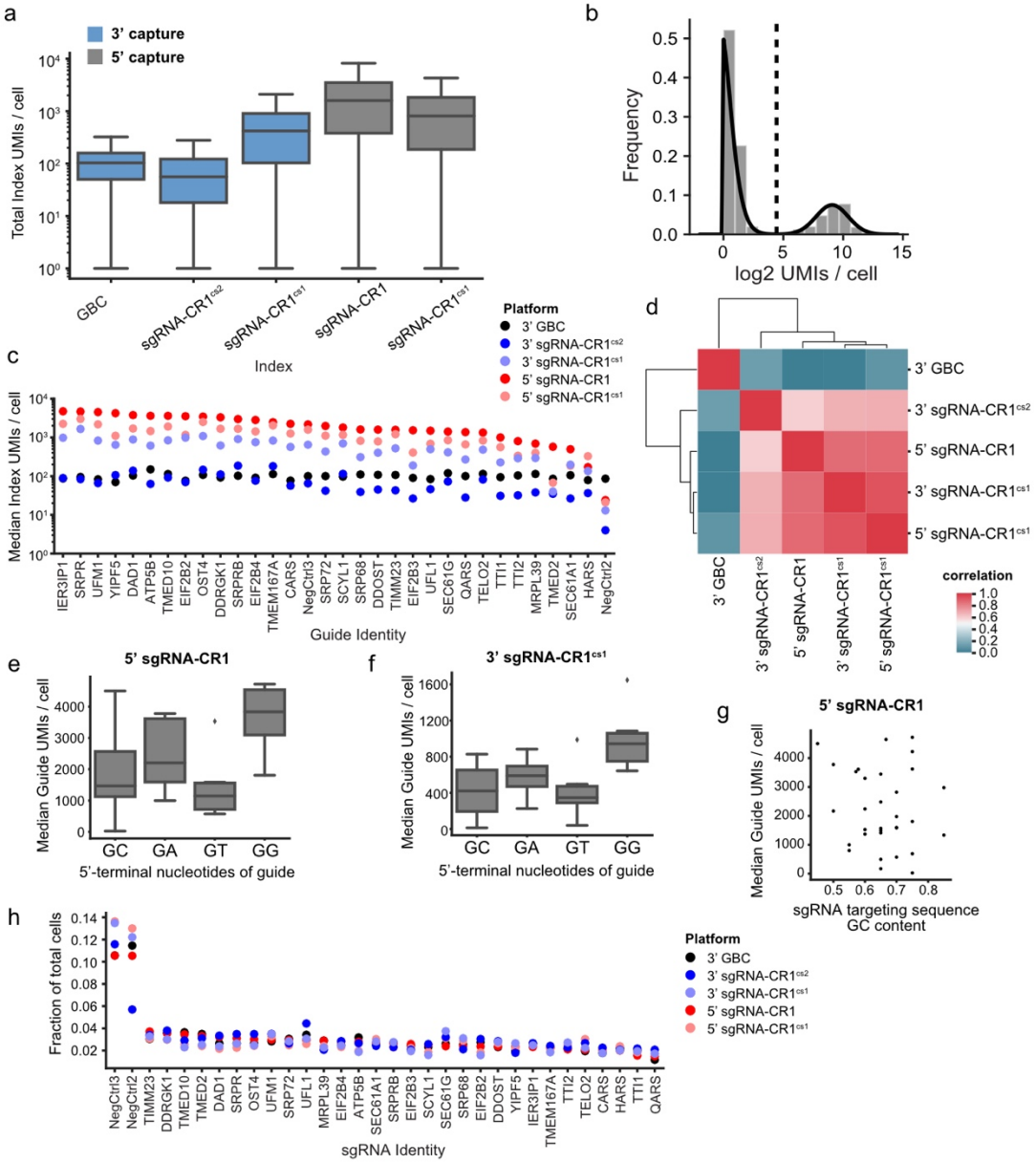
Figure 1.3: Multiplexed CRISPRi/CRISPRa and hybridization-based target enrichment enable scalable and versatile single-cell CRISPR screens. a) Scatterplot of the relative target expression per gene comparing CRISPRi knockdown with a single sgRNA (expressed from a dual-guide vector paired with a non-targeting control) versus multiplexed sgRNAs. Multiplexed sgRNAs significantly improve knockdown (sgRNAs 1+control median relative target expression=0.20; sgRNAs 1+2 median relative target expression=0.11; Wilcoxon signed-rank two-sided test n=87 genes, W=378, *p*=8e-11). sgRNA 1, best predicted sgRNA for each gene. sgRNA 2, second best predicted sgRNA for each gene. **b)** Box plots of the relative target expression per gene in the multiplexed CRISPRi experiment denoting quartile ranges (box), median (center mark), and 1.5 × interquartile range (whiskers). "min(1,2)" indicates the minimum remaining target expression between sgRNA 1 (paired with negative control) and sgRNA 2 (paired with negative control), ie. the predicted multiplexed sgRNA knockdown based on a dominant model of knockdown. The multiplexed sgRNAs performed better than the dominant model (Wilcoxon signed-rank two-sided test n=87 genes, W=698, *p*=3e-7). **c)** The fraction of total UMIs for L1000 genes (n=978) versus other genes, before and after target enrichment. **d)** Scatterplot of the total number of UMIs for each gene, before and after target enrichment (n=978 genes). The Pearson correlation of $\log_{10}$ normalized UMIs is r=0.98. **e)** Heatmap depicts clustering of guides in our multiplexed CRISPRi experiment. Heatmap represents Spearman's rank correlations between pseudo-bulk expression profiles of well-expressed genes (>1 UMI/cell). Data from all perturbations with >10 differentially expressed genes compared to controls are included (n=145 genes). The upper triangle (correlation matrix) was calculated on the whole transcriptome while

the lower triangle (correlation matrix) was calculated on the target-enriched transcriptome. Both triangles were identically ordered based on hierarchical clustering of the whole transcriptome correlation matrix. **f)** Pearson correlations of pseudo-bulk differential expression profiles of well-expressed genes (>1 UMI/cell) caused by sgRNAs targeting the same gene (for n=39 genes whose knockdown led to differential gene expression) versus sgRNAs targeting different genes (n=111592 pairs). sgRNAs targeting the same gene had significantly more similar profiles than sgRNAs targeting different genes, both before and after target enrichment (unenriched median r=0.64, Mann-Whitney U two-sided test U=117224, p=1.4e-24; enriched median r=0.72, Mann-Whitney U two-sided test U=259898, p=1.7e-21). Box plots denote quartile ranges (box), median (center mark), and 1.5 × interquartile range (whiskers). **g)** Schematic overview of direct capture Perturb-seq workflow.



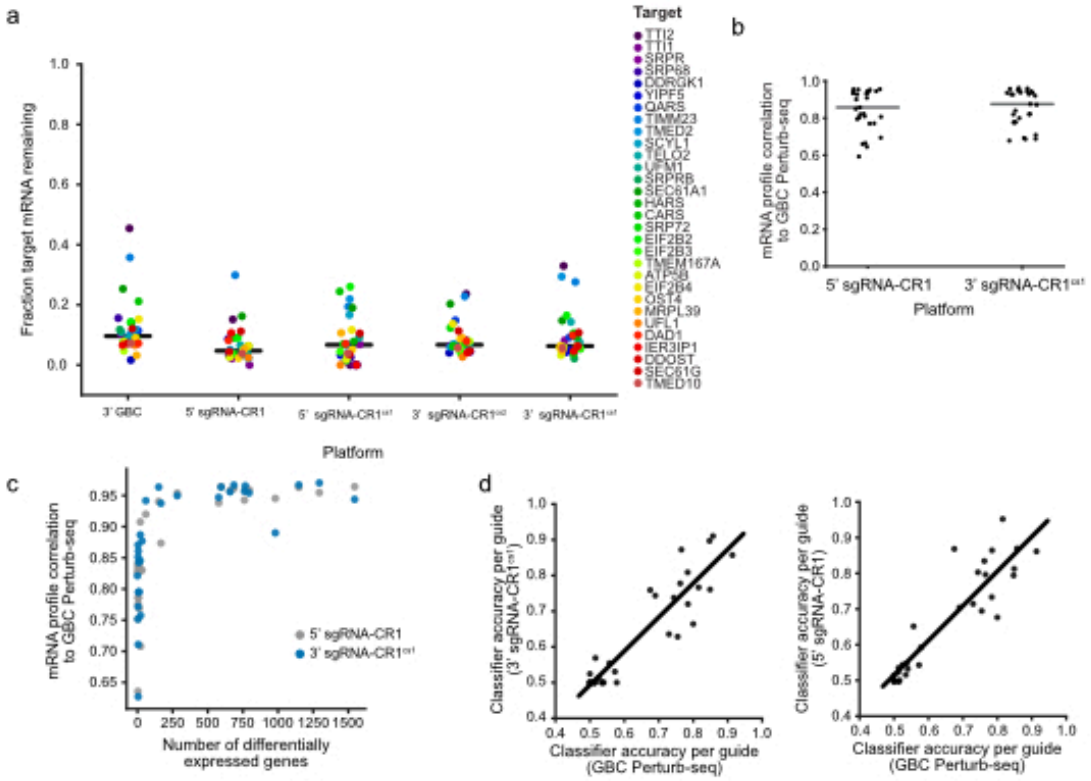
Supplementary Figure 1.1: Optimization of modified of guide constant regions to enable 3' direct capture Perturb-seq. a) Schematic of 3' single-cell RNA-sequencing (3' scRNA-seq). Polyadenylated mRNAs from individual cells (top, light blue) anneal to barcoded oligo-dT primers in emulsion droplets (delivered to droplets on gel beads) and are reverse transcribed into indexed cDNA (bottom). TSO, template switch oligo. UMI, unique molecular identifier. CBC, cell barcode. **b)** Schematic of 5' single-cell RNA-sequencing (5' scRNA-seq). Polyadenylated mRNAs from individual cells (top, light blue) anneal to unbarcoded oligo-dT primers in emulsion droplets (delivered to droplets as free oligos) and are reverse transcribed. Indexing of cDNA (bottom) occurs when template switching allows for extension of barcoded TSOs (delivered to droplets on

gel beads). **c)** Schematic of constant region 1 (CR1) guide RNAs. Arrows indicate the positions of capture sequence insertions. **d)** CRISPRi activity of guides carrying the indicated capture sequences (all programmed with an identical GFP targeting region) in GFP+ K562 dCas9-KRAB cells 10 days post-transduction. Data from guides selected for direct capture experiments (sgRNA-CR1^{cs1} and sgRNA-CR1^{cs2}) are indicated. For comparison, data from standard guides targeting GFP (programmed with the same targeting region but without capture sequences and expressed from 3 other vectors) were also included. One of these, indicated as "CROP-seq", has a different previously published (Datlinger *et al.*, *Nature Methods*, 14-3, 2017) constant region and is expressed from a different promoter. Data represents the average of independently infected triplicates normalized to controls ± standard deviation. The data was collected in two separate batches (independently controlled). Representative flow cytometry gating for one sample is also shown. **e)** Gaussian kernel density estimates of normalized flow-cytometry measurements representing GFP expression demonstrate CRISPRi activity of the indicated sgRNAs (programmed with an identical GFP targeting region). Data was collected in three independent biological replicates and a representative replicate is shown. AU, arbitrary units. **f)** Gaussian kernel density estimates of normalized flow-cytometry measurements representing GFP expression demonstrate CRISPRi activity of the indicated sgRNAs (programmed an identical GFP targeting region). Data was collected in three independent biological replicates and a representative replicate is shown. AU, arbitrary units. **g)** Schematic of experimental workflow for direct capture Perturb-seq (3' or 5') based on protocols from 10x Genomics. Red indicates generation of sequencing libraries. Box details construction of index sequencing library for GBC Perturb-seq, which is based on a previously published protocol (Adamson *et al*., *Cell*, 167-7, 2016).

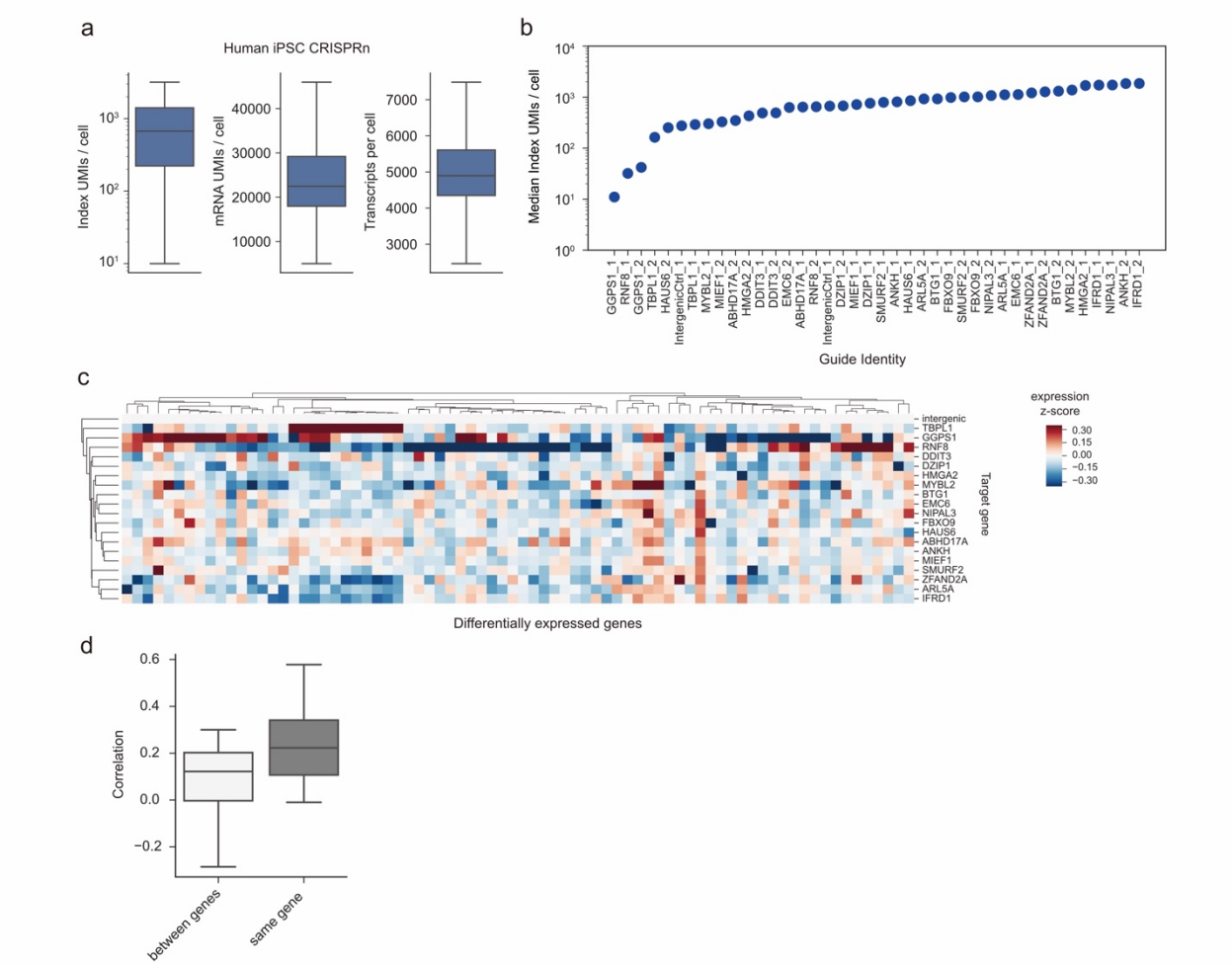


Supplementary Figure 1.2: Cell indexing by direct guide capture is robust and comparable to indexing by GBC capture. a) Box plot of total index (GBC or guide) UMI counts per cell for all cells (prior to guide identity mapping). Box plots denote quartile ranges (box), median (center mark), and 1.5 × interquartile range (whiskers). Data represent n=10036 cells for 3' GBC, n=8267 cells for 3' sgRNA-CR1^{cs2}, n=8727 cells for 3' sgRNA-CR1^{cs1}, n=6789 cells for 5' sgRNA-CR1, and n=7043 cells for 5' sgRNA-CR1^{cs1}. Several direct capture methods gave higher index capture than the GBC-based method (Mann-Whitney two-sided U test: U=65, *p*=2e-9 for 3' sgRNA-CR1^{cs1} capture; U=64, *p*=2e-9 for 5' sgRNA-CR1^{cs1} capture; U=32, *p*=1e-10 for 5' sgRNA-CR1 capture), while 3' capture of sgRNA-CR1^{cs2} had modestly lower capture (Mann-Whitney two-sided U test: U=788, *p*=0.0002) **b)** Representative guide identity mapping. Data correspond to NegCtrl3 in 3' sgRNA-CR1^{cs1} Perturb-seq (n=8727 cells). Guide identity mapping relies on fitting a 2-component Poisson and Gaussian mixture model (black line), where cells with a posterior probability >0.5 (dotted line) of belonging to the upper mode

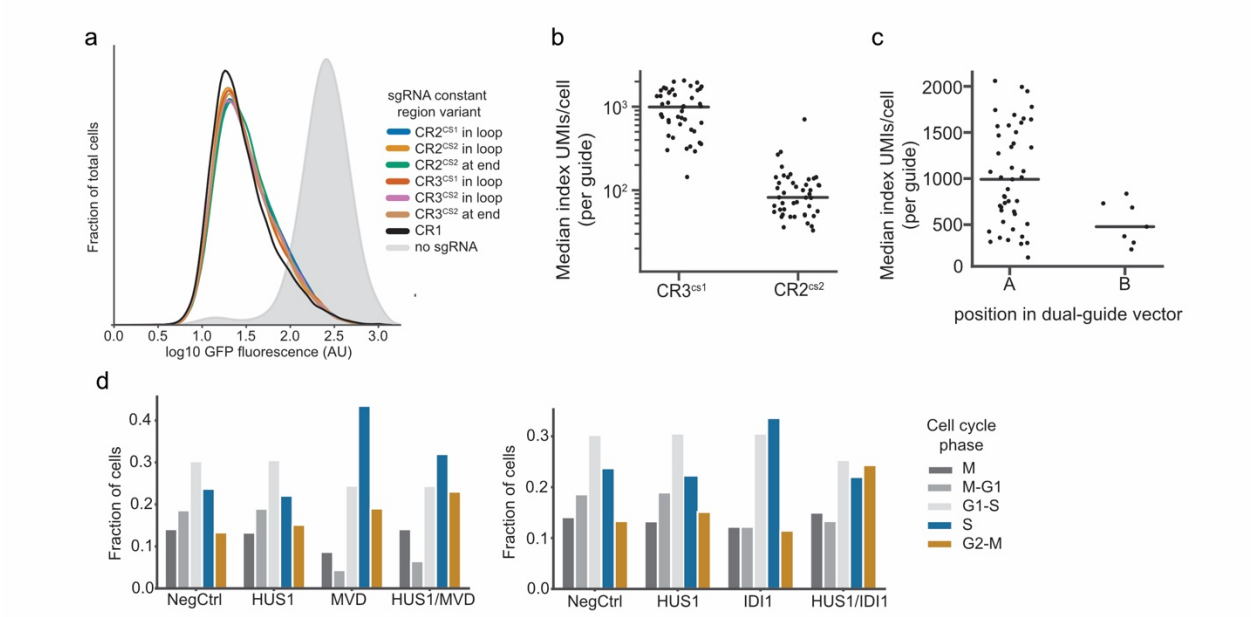
component are assigned to NegCtrl3. **c)** Median index UMI counts per cell (capture rate) for cells assigned to each guide identity in Perturb-seq experiments (n=32 guides per experiment). Across platforms, NegCtrl2 has the worst capture rate, which may be explained by the fact that this negative control guide has a targeting region containing an extended run of guanine nucleotides (5'-GCGATGGGGGGGTGGGTAGC-3'). Data plotted here are also plotted in **Figure 1.1c,d)** For each pairwise comparison of Perturb-seq experiments, we calculated a Pearson correlation of guide capture rates (n=32 guides). Across experiments performed with direct capture Perturb-seq, guide capture rates are correlated (r>0.6), suggesting that targeting region-dependent features influence guide capture. **e)** Box plots of median guide UMIs per cell stratified by the 5' terminal nucleotides of the targeting region. Box plots denote quartile ranges (box), median (center mark), and 1.5 × interquartile range (whiskers). The displayed data is from Perturb-seq by 5' sgRNA-CR1 capture. There is a significant relationship between capture rate and the 5' terminal nucleotides (Kruskal-Wallis H-test: n=32 guides, H=10.2, p=0.017 for 5' sgRNA-CR1) **f)** Box plots of median guide UMIs per cell stratified by the 5' terminal nucleotides of the targeting region. Box plots denote quartile ranges (box), median (center mark), and 1.5 × interquartile range (whiskers). The displayed data is from Perturb-seq by 3' sgRNA-CR1^{cs1} capture. There is a significant relationship between capture rate and the 5' terminal nucleotides (Kruskal-Wallis H-test: n=32 guides, H=10.9, p=0.012 for 3' sgRNA-CR1^{cs1}) **g)** Scatterplot of the median number of UMIs per cell and targeting region GC content for each of n=32 guides. The displayed data is from Perturb-seq by 5' sgRNA-CR1 capture. For all platforms, we observed no significant Pearson correlation between capture rate and targeting region GC content (p>0.4 for all platforms). **h)** Identity assignment rates per guide for Perturb-seq experiments. Balanced representation among cells assigned to each of n=32 guides (with intentionally 4-fold overrepresented negative controls) was achieved by titering lentiviruses prior to pooling.



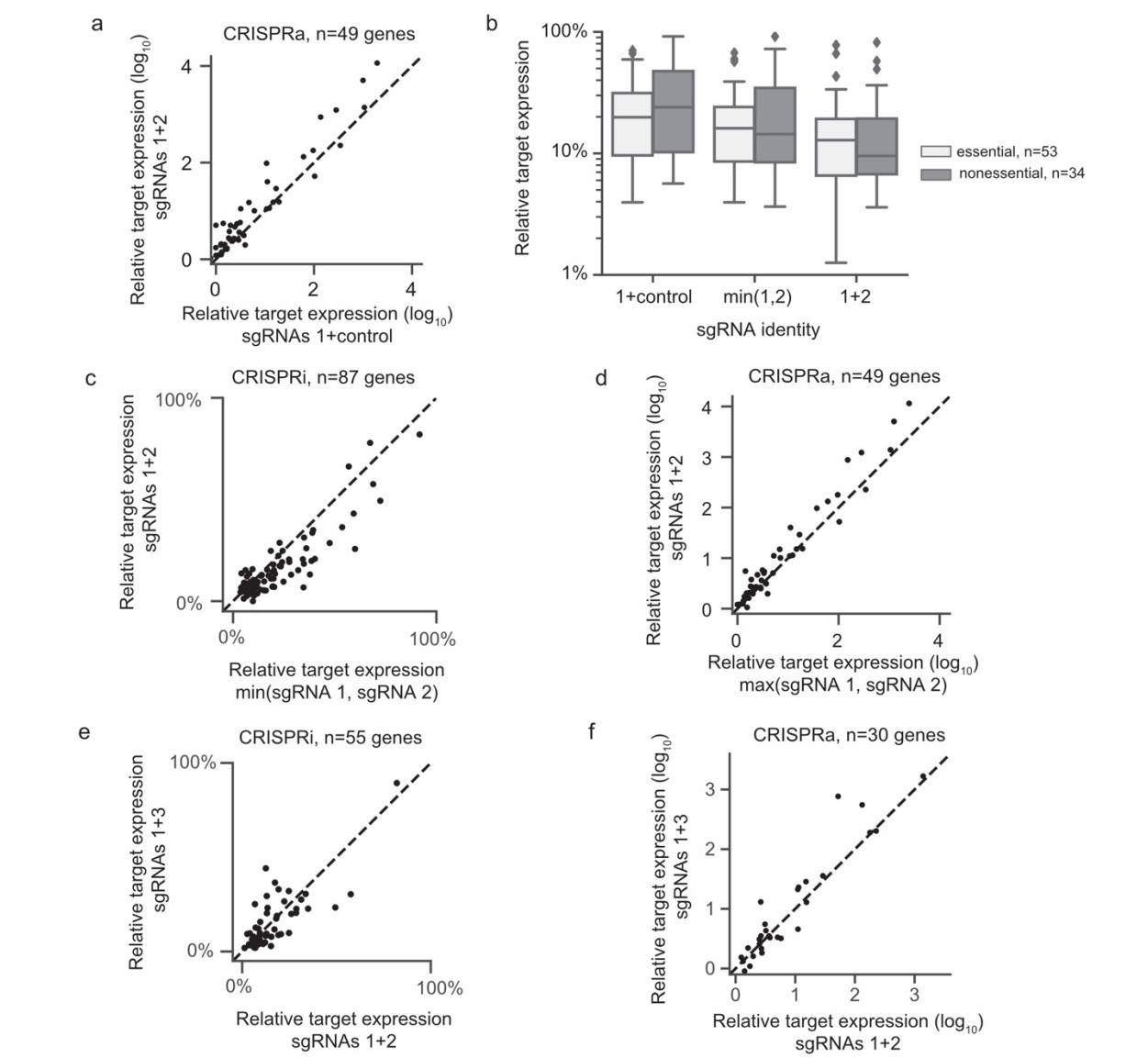
Supplementary Figure 1.3: Direct capture Perturb-seq performs comparably to GBC Perturb-seq for phenotypic analysis. a) Mean target knockdown (fraction mRNA remaining) for each targeting guide (n=30) in the indicated experiments. For each guide, the data point represents the mean normalized expression level of the target gene across cells bearing the corresponding guide divided by the mean normalized expression level of the target gene in control cells (NegCtrl3). **b)** The Pearson correlation of pseudo-bulk expression profiles from direct capture Perturb-seq and GBC Perturb-seq experiments for each perturbation (n=30 targeting guides). Profiles were generated from the top 100 most differentially expressed genes in GBC Perturb-seq. Grey lines indicate medians. **c)** Scatterplot indicates the relationship between the number of differentially expressed genes for each guide (determined by a two-sided, two-sample Kolmogorov-Smirnov test using GBC Perturb-seq data) and the Pearson correlation of pseudo-bulk expression profiles between GBC Perturb-seq and direct capture Perturb-seq on the indicated platform (n=30 targeting guides per platform). **d)** Scatterplots of the balanced accuracy of random forest classifiers trained to distinguish perturbed and unperturbed (NegCtrl3) cells for each of n=30 targeting guides on the indicated platforms. Direct capture Perturb-seq accuracies were highly correlated with GBC Perturb-seq (Pearson correlation: r=0.91 for 3' sgRNA-CR1^{cs1} capture; r=0.90 for 5' sgRNA-CR1 capture). We failed to detect significant differences in performance between direct capture Perturb-seq and GBC Perturb-seq (Wilcoxon signed-rank two-sided test: p=0.2 for 3' sgRNA-CR1^{cs1}; p=0.6 for 5' sgRNA-CR1 capture).



Supplementary Figure 1.4: Direct capture Perturb-seq allows for robust guide assignment and phenotypic analysis in iPSCs with CRISPR cutting. a) Box plots displaying the index (guide) UMIs per cell, mRNA UMIs per cell, and transcripts per cell in iPSCs expressing Cas9 (n=5300 cells). Box plots denote quartile ranges (box), median (center mark), and 1.5 × interquartile range (whiskers). **b)** Median index (guide) UMI counts per cell (capture rate) for cells assigned to each guide identity. **c)** Heatmap represents gene expression of most differentially expressed genes across all cells with the indicated genetic perturbation, as determined by a random forest classifier. Expression values are the z-scored expression relative to unperturbed cells. Profiles for each gene are calculated by averaging the pseudo-bulk expression profiles of the two independent sgRNAs targeting the gene (n=19 genes targeted by two sgRNAs each). **d)** Box plot of Pearson correlations of pseudo-bulk expression profiles caused by sgRNAs targeting the same gene (n=19) versus sgRNAs targeting different genes (n=684 pairs). Box plots denote quartile ranges (box), median (center mark), and 1.5 × interquartile range (whiskers). sgRNAs targeting the same gene cause significantly more similar profiles than sgRNAs targeting different genes (Mann-Whitney two-sided U test U=1636.0, p=0.0002). Differences in expression profiles caused by sgRNAs targeting the same genes are likely due to variation of CRISPR cutting efficacy, indel profiles, and/or genetic compensation.

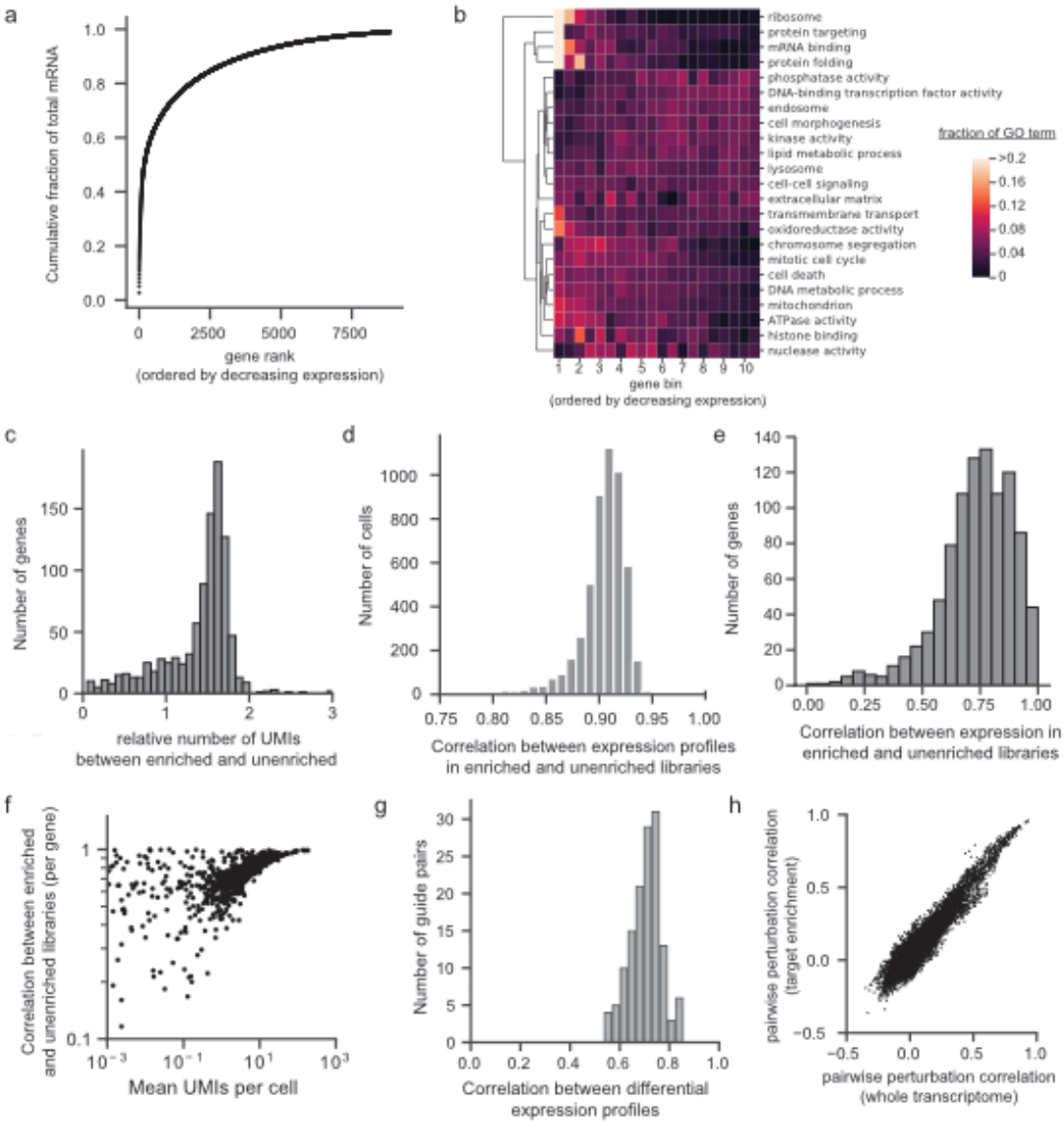


Supplementary Figure 1.5: Optimization of additional guide constant regions to enable dual-guide 3' direct capture Perturb-seq. **a)** Gaussian kernel density estimates of normalized flow-cytometry measurements representing GFP expression demonstrate CRISPRi activity of sgRNAs with indicated constant regions (programmed with an identical GFP targeting region). Data was collected in three independent biological replicates and a representative replicate is shown. In this experiment, sgRNAs were expressed from a single-guide vector. As all CR2 and CR3 sgRNA variants were highly active, we used CR3 with cs1 in the stem loop ($CR3^{cs1}$) and CR2 with cs2 at the 3' end ($CR2^{cs2}$) for downstream experiments. AU, arbitrary units. **b)** Median index (guide) UMIs per cell for each of n=45 sgRNAs where grey lines indicate medians. Data are from our dual-guide genetic interaction 3' direct capture Perturb-seq experiments. **c)** Median index (guide) UMIs per cell for each of n=45 sgRNAs where grey lines indicate medians. Data are from our dual-guide genetic interaction 3' direct capture Perturb-seq experiment with an mU6-$CR3^{cs1}$-hU6-$CR1^{cs1}$ design. **d)** Fraction of cells in each cell cycle phase across cells with the indicated perturbation.



Supplementary Figure 1.6: Multiplexed sgRNAs improve CRISPRi and CRISPRa activity. a) Scatterplot depicts the relative target expression per gene to compare the CRISPRa activity of a single sgRNA (expressed from a dual-guide vector paired with a non-targeting control) with CRISPRa activity from multiplexed sgRNAs. Multiplexing sgRNAs significantly improves activation (sgRNAs 1+control, median fold-activation=2.9; sgRNAs 1+2, median fold-activation=4.7; Wilcoxon signed-rank two-sided test n=49 genes, W=162, p=7e-6). **b)** Box plots depict the relative target expression per gene for genes stratified by essentiality in K562 cells (Horlbeck *et al*., *Elife*, 2016). "min(1,2)" indicates the minimum remaining target expression between sgRNA 1 (paired with negative control) and sgRNA 2 (paired with negative control), ie. the predicted multiplexed sgRNA knockdown based on a dominant model. For essential genes, sgRNA multiplexing improves median knockdown from 80% to 87%. For nonessential genes, sgRNA multiplexing improves median knockdown from 76% to 90%. Box plots denote quartile ranges (box), median (center mark), and 1.5 × interquartile range (whiskers). Data from multiplexed CRISPRi experiment. **c)** Scatterplot depicts relative target expression per gene to compare observed CRISPRi-based gene knockdown with predicted knockdown using multiplexed sgRNAs (assuming a dominant model). Multiplexing sgRNAs performs better than predicted

based on the dominant model (Wilcoxon signed-rank two-sided test n=87 genes, W=698, p=3e-7). **d)** Scatterplot depicts relative target expression per gene to compare predicted CRISPRa activity from multiplexed sgRNAs (assuming a dominant model) to observed activity. Multiplexing sgRNAs performs better than predicted based on the dominant model (Wilcoxon signed-rank two-sided test n=49 genes, W=233, *p*=0.0002). **e)** Scatterplot depicts relative target expression per gene to compare CRISPRi-based gene knockdown using sgRNAs 1+2 (<80bp apart) with the activity of sgRNAs 1+3 (>80bp apart). We failed to detect a significant increase in CRISPRi activity with increased distance between sgRNAs (Wilcoxon signed-rank two-sided test n=55 genes, W=643, *p*=0.3). **f)** Scatterplot depicts relative target expression per gene to compare the CRISPRa activity of sgRNAs 1+2 (<80bp apart) with sgRNAs 1+3 (>80bp apart). We failed to detect a significant increase in CRISPRa activity with increased distance between sgRNAs (Wilcoxon signed-rank two-sided test n=30 genes, W=190, *p*=0.4).



Supplementary Figure 1.7: Target enriched gene expression libraries are well correlated with deeply sequenced, unenriched libraries. a) Cumulative density function of gene expression from K562 cells. In K562 cells, 2% of expressed genes consume >50% of sequencing reads. **b)** Heatmap compares gene ontology (GO) terms by their expression level in K562 cells. While some GO terms are enriched for highly expressed genes (e.g. ribosome, protein targeting, mRNA binding, protein folding) others are enriched for lowly expressed genes (e.g. DNA-binding transcription factor activity, phosphatase activity). **c)** Histogram depicts ratio of total UMIs in the target enriched library compared to total UMIs in the unenriched library for each of n=978 targeted genes. **d)** Histogram depicts Pearson correlations of the L1000 gene expression profiles per cell, before and after target enrichment (n=6349 cells with identified guide identities). **e)** Histogram depicts Pearson correlations of gene expression across cells per each gene, before and after target enrichment (n=978 genes). **f)** Scatterplot shows gene expression level (mean UMIs per cell) compared to the per gene Pearson correlation before and after target enrichment (n=978 genes). **g)** Histogram depicts Pearson correlations between differential expression profiles for the same guide pairs in enriched and unenriched libraries. Perturbations leading to >10 differentially

expressed genes by two-sided, two-sample Kolmogorov-Smirnov test were included (n=137). Genes expressed at >1 UMI/cell were considered (n=263 genes), and gene expression profiles were z-scored with respect to the population of cells expressing non-targeting control guides in order to determine differential gene expression profiles. The median Pearson correlation of r=0.71 shows that genetic perturbation-dependent differential expression patterns are overall conserved before and after target enrichment. **h)** Scatterplot of the guide-guide Spearman's rank correlations, calculated before and after target enrichment (n=10440 pairwise comparisons, cophenetic correlation r=0.95).

REFERENCES

Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, Villalta JE, Gilbert LA, Horlbeck MA, Hein MY, Pak RA, Gray AN, Gross CA, Dixit A, Parnas O, Regev A, Weissman JS. 2016. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* 167:1867-1882.e21. doi:10.1016/j.cell.2016.11.048

Adamson B, Norman TM, Jost M, Weissman JS. 2018. Approaches to maximize sgRNA-barcode coupling in Perturb-seq screens. *Biorxiv* 298349. doi:10.1101/298349

Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, Liu DR. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*. doi:10.1038/s41586-019-1711-4

Chan MM, Smith ZD, Grosswendt S, Kretzmer H, Norman TM, Adamson B, Jost M, Quinn JJ, Yang D, Jones MG, Khodaverdian A, Yosef N, Meissner A, Weissman JS. 2019. Molecular recording of mammalian embryogenesis. *Nature* 570:77–82. doi:10.1038/s41586-019-1184-5

Cleary B, Cong L, Cheung A, Lander ES, Regev A. 2017. Efficient Generation of Transcriptomic Profiles by Random Composite Measurements. *Cell* 171:1424-1436.e18. doi:10.1016/j.cell.2017.10.023

Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. 2017. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* 14:297–301. doi:10.1038/nmeth.4177

Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, Adamson B, Norman TM, Lander ES, Weissman JS, Friedman N, Regev A. 2016. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167:1853-1866.e17. doi:10.1016/j.cell.2016.11.038

Feldman D, Singh A, Garrity AJ, Blainey PC. 2018. Lentiviral co-packaging mitigates the effects of intermolecular recombination and multiple integrations in pooled genetic screens. *Biorxiv* 262121. doi:10.1101/262121

Feldman D, Singh A, Schmid-Burgk JL, Carlson RJ, Mezger A, Garrity AJ, Zhang F, Blainey PC. 2019. Optical Pooled Screens in Human Cells. *Cell* 179:787-799.e17. doi:10.1016/j.cell.2019.09.016

Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS, Trapnell C, Ahituv N, Shendure J. 2019. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* 176:377-390.e19. doi:10.1016/j.cell.2018.11.029

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, Qi LS, Kampmann M, Weissman JS. 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* 159:647–61. doi:10.1016/j.cell.2014.09.029

Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, Lim WA, Weissman JS, Qi LS. 2013. CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* 154:442–51. doi:10.1016/j.cell.2013.06.044

Hill AJ, McFaline-Figueroa JL, Starita LM, Gasperini MJ, Matreyek KA, Packer J, Jackson D, Shendure J, Trapnell C. 2018. On the design of CRISPR-based single-cell molecular screens. *Nat Methods* 15:271. doi:10.1038/nmeth.4604

Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M, Weissman JS. 2016. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* 5:e19760. doi:10.7554/elife.19760

Horlbeck MA, Xu A, Wang M, Bennett NK, Park CY, Bogdanoff D, Adamson B, Chow ED, Kampmann M, Peterson TR, Nakamura K, Fischbach MA, Weissman JS, Gilbert LA. 2018. Mapping the Genetic Landscape of Human Cells. *Cell* 174:953-967.e22. doi:10.1016/j.cell.2018.06.010

Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, Salame TM, Tanay A, Oudenaarden A van, Amit I. 2016. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* 167:1883-1896.e15. doi:10.1016/j.cell.2016.11.039

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. 2015. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* 161:1187–1201. doi:10.1016/j.cell.2015.04.044

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161:1202–14. doi:10.1016/j.cell.2015.05.002

Mandegar MA, Huebsch N, Frolov EB, Shin E, Truong A, Olvera MP, Chan AH, Miyaoka Y, Holmes K, Spencer CI, Judge LM, Gordon DE, Eskildsen TV, Villalta JE, Horlbeck MA, Gilbert LA, Krogan NJ, Sheikh SP, Weissman JS, Qi LS, So P-L, Conklin BR. 2016. CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. *Cell Stem Cell* 18:541–553. doi:10.1016/j.stem.2016.01.022

Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, Legut M, Roush T, Herrera A, Papalexi E, Ouyang Z, Satija R, Sanjana NE, Koralov SB, Smibert P. 2019. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat Methods* 16:409–412. doi:10.1038/s41592-019-0392-0

Moreno AM, Fu X, Zhu J, Katrekar D, Shih Y-RV, Marlett J, Cabotaje J, Tat J, Naughton J, Lisowski L, Varghese S, Zhang K, Mali P. 2018. In Situ Gene Therapy via AAV-CRISPR-Cas9-Mediated Targeted Gene Regulation. *Mol Ther* 26:1818–1827. doi:10.1016/j.ymthe.2018.04.017

Norman TM, Horlbeck MA, Replogle JM, Ge AY, Xu A, Jost M, Gilbert LA, Weissman JS. 2019. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* 365:786–793. doi:10.1126/science.aax4438

Packer J, Trapnell C. 2018. Single-Cell Multi-omics: An Engine for New Quantitative Models of Gene Regulation. *Trends Genet* 34:653–665. doi:10.1016/j.tig.2018.06.001

Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, Moore R, McClanahan TK, Sadekova S, Klappenbach JA. 2017. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* 35:936–939. doi:10.1038/nbt.3973

Ran FA, Hsu PD, Lin C-Y, Gootenberg JS, Konermann S, Trevino AE, Scott DA, Inoue A, Matoba S, Zhang Y, Zhang F. 2013. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* 154:1380–9. doi:10.1016/j.cell.2013.08.021

Rubin AJ, Parker KR, Satpathy AT, Qi Y, Wu B, Ong AJ, Mumbach MR, Ji AL, Kim DS, Cho SW, Zarnegar BJ, Greenleaf WJ, Chang HY, Khavari PA. 2018. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* 176:361-376.e17. doi:10.1016/j.cell.2018.11.022

Saikia M, Burnham P, Keshavjee SH, Wang MFZ, Heyang M, Moral-Lopez P, Hinchman MM, Danko CG, Parker JSL, Vlaminck ID. 2019. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. *Nat Methods* 16:59–62. doi:10.1038/s41592-018-0259-9

Salomon R, Kaczorowski D, Valdes-Mora F, Nordon RE, Neild A, Farbehi N, Bartonicek N, Gallego-Ortega D. 2019. Droplet-based single cell RNAseq tools: a practical guide. *Lab Chip* 19:1706–1727. doi:10.1039/c8lc01239c

Savell KE, Bach SV, Zipperly ME, Revanna JS, Goska NA, Tuscher JJ, Duke CG, Sultan FA, Burke JN, Williams D, Ianov L, Day JJ. 2019. A Neuron-Optimized CRISPR/dCas9 Activation System for Robust and Specific Gene Regulation. *Eneuro* 6:ENEURO.0495-18.2019. doi:10.1523/eneuro.0495-18.2019

Smits AH, Ziebell F, Joberty G, Zinn N, Mueller WF, Clauder-Münster S, Eberhard D, Savitski MF, Grandi P, Jakob P, Michon A-M, Sun H, Tessmer K, Bürckstümmer T, Bantscheff M,

Steinmetz LM, Drewes G, Huber W. 2019. Biological plasticity rescues target activity in CRISPR knock outs. *Nat Methods* 16:1087–1093. doi:10.1038/s41592-019-0614-5

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 14:865–868. doi:10.1038/nmeth.4380

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, Lahr DL, Hirschman JE, Liu Z, Donahue M, Julian B, Khan M, Wadden D, Smith IC, Lam D, Liberzon A, Toder C, Bagul M, Orzechowski M, Enache OM, Piccioni F, Johnson SA, Lyons NJ, Berger AH, Shamji AF, Brooks AN, Vrcic A, Flynn C, Rosains J, Takeda DY, Hu R, Davison D, Lamb J, Ardlie K, Hogstrom L, Greenside P, Gray NS, Clemons PA, Silver S, Wu Xiaoyun, Zhao W-N, Read-Button W, Wu Xiaohua, Haggarty SJ, Ronco LV, Boehm JS, Schreiber SL, Doench JG, Bittker JA, Root DE, Wong B, Golub TR. 2017. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171:1437-1452.e17. doi:10.1016/j.cell.2017.10.049

Vallejo AF, Davies J, Grover A, Tsai C-H, Jepras R, Polak ME, West J. 2019. Resolving cellular systems by ultra-sensitive and economical single-cell transcriptome filtering. *Biorxiv* 800631. doi:10.1101/800631

Xie S, Cooley A, Armendariz D, Zhou P, Hon GC. 2018. Frequent sgRNA-barcode recombination in single-cell perturbation assays. *Plos One* 13:e0198635. doi:10.1371/journal.pone.0198635

Xie S, Duan J, Li B, Zhou P, Hon GC. 2017. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol Cell* 66:285-299.e5. doi:10.1016/j.molcel.2017.03.007

Zhang S-Q, Ma K-Y, Schonnesen AA, Zhang M, He C, Sun E, Williams CM, Jia W, Jiang N. 2018. High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nat Biotechnol* 36:1156–1159. doi:10.1038/nbt.4282

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8:ncomms14049. doi:10.1038/ncomms14049

CHAPTER 2

BACKGROUND

A central goal of genetics is to map the relationship between genotypes and phenotypes. This mapping is traditionally carried out in either of two ways. A phenotype-centric, “forward genetic” approach reveals the genetic changes that drive a given phenotype of interest. Conversely, a gene-centric, “reverse genetic” approach catalogs the diverse phenotypes caused by a defined genetic change.

Recent technological developments have advanced both forward and reverse genetic efforts, especially as applied to cellular functional genomics (Przybyla and Gilbert, 2021). CRISPR-Cas tools now enable the deletion, mutation, repression, or activation of genes at will (Doench, 2018). In forward genetic screens, CRISPR-Cas systems can be used to generate cells bearing diverse genetic perturbations. These genetically heterogeneous cells can then be subjected to a selective pressure, with phenotypes assigned to genetic perturbations by sequencing. Forward genetic screens provide powerful tools for the identification of cancer dependencies, essential cellular machinery, differentiation factors, and suppressors of genetic diseases (Kramer et al., 2018; Tsherniak et al., 2017; Wang et al., 2021, 2015). In parallel, dramatic improvements in molecular phenotyping now allow for single-cell readouts of epigenetic, transcriptomic, proteomic, and imaging information (Stuart and Satija, 2019). Applied to reverse genetics, single-cell profiling can refine the understanding of how select genetic perturbations affect cell types and cell states (Montoro et al., 2018; Plasschaert et al., 2018).

However, both phenotype-centric and gene-centric approaches suffer conceptual and technical limitations. Pooled forward genetic screens typically use low-dimensional phenotypes (e.g., growth, marker gene expression, drug resistance) for selection. The use of simple phenotypes can conflate genes acting via different mechanisms, requiring extensive follow-up studies to disentangle genetic pathways (Przybyla and Gilbert, 2021). Additionally, in forward

genetics, serendipitous discovery is prevented by the prerequisite of selecting phenotypes prior to screening. On the other hand, while reverse genetic approaches enable the study of multidimensional and complex phenotypes, they have typically been restricted in scale to rationally chosen targets, limiting the ability to make systematic comparisons.

Single-cell CRISPR screens present a solution to these problems. These screens simultaneously read out the genetic perturbation and high-dimensional phenotype of individual cells in a pooled screening format, thus combining the throughput of forward genetic screens with the rich phenotypes of reverse genetics. While these approaches initially focused on transcriptomic phenotypes (e.g., Perturb-seq, CROP-seq) (Adamson et al., 2016; Datlinger et al., 2017; Dixit et al., 2016; Jaitin et al., 2016; Replogle et al., 2020), technical advances have enabled their application to epigenetic (Rubin et al., 2019), imaging (Feldman et al., 2019), or multimodal phenotypes as well (Frangieh et al., 2021; Mimitou et al., 2019; Papalexi et al., 2021). From these rich data, it is possible to identify genetic perturbations that cause a specific behavior as well as to catalog the spectrum of phenotypes associated with each genetic perturbation. Despite the promise of single-cell CRISPR screens, to date their use has been limited to studying at most a few hundred genetic perturbations, typically chosen with a bias towards predefined biological questions.

We reasoned that there would be unique value to genome-scale single-cell CRISPR screens. For example, while the number of perturbations scales linearly with experimental cost, the number of pairwise comparisons in a screen—and thus its utility for unsupervised classification of gene function—scales quadratically. Similarly, in large-scale screens, the diversity of perturbations allows one to more fully explore the remarkable range of cell states that can be revealed by rich phenotypes, and as many human genes are well-characterized, these genes serve as natural controls to anchor the interpretation of observations. Finally, genome-scale experiments could help address fundamental biological questions, such as what fraction of genetic changes elicit global transcriptional phenotypes and how transcriptional programs are

rewired between cell types, with implications for understanding the organizing principles of cellular systems (Tanay and Regev, 2017).

Motivated by this potential, we perform genome-scale single-cell CRISPR screens via Perturb-seq. We use a compact, multiplexed CRISPR interference (CRISPRi) library to assay thousands of loss-of-function genetic perturbations with single-cell RNA-sequencing (scRNA-seq) in chronic myeloid leukemia (K562) and retinal pigment epithelial (RPE1) cell lines. Leveraging the scale and diversity of these perturbations across millions of cells, we show that Perturb-seq can be used to study numerous complex cellular phenotypes—from RNA splicing to differentiation to chromosomal instability—in a single screen. We demonstrate how the interpretability of scRNA-seq phenotypes enables the discovery of gene function, extensively validating our findings with orthogonal experiments. Finally, we invert our analysis to focus on regulatory networks rather than genetic perturbations and uncover unanticipated stress-specific regulation of the mitochondrial genome. In sum, we use Perturb-seq to reveal a multidimensional portrait of cellular behavior, gene function, and regulatory networks that advances the goal of creating comprehensive genotype-phenotype maps.

RESULTS

A multiplexed CRISPRi strategy for genome-scale Perturb-seq

Perturb-seq uses scRNA-seq to concurrently read out the CRISPR single-guide RNAs (sgRNA) (ie., genetic perturbation) and transcriptome (ie., high-dimensional phenotype) of single-cells in a pooled format (**Figure 2.1A**). To enable genome-scale Perturb-seq, we considered key parameters that would increase scalability and data quality, such as the genetic perturbation modality and sgRNA library.

Although Perturb-seq is compatible with a range of CRISPR-based perturbations including knockout (Datlinger et al., 2017; Dixit et al., 2016; Jaitin et al., 2016), knockdown (CRISPRi) (Adamson et al., 2016), or activation (CRISPRa) (Norman et al., 2019), we elected to use CRISPRi for several reasons. First, CRISPRi allows the efficacy of the genetic perturbation, knockdown, to be directly measured from scRNA-seq. Exploiting this feature allowed us to target each gene in our library with a single element and empirically exclude unperturbed genes from downstream analysis. Second, CRISPRi tends to yield more homogeneous genetic perturbation than nuclease-based CRISPR knockout, which can generate a subset of cells bearing active in-frame indels (Smits et al., 2019). The relative homogeneity of CRISPRi limits selection for unperturbed cells, especially when studying essential genes. Third, unlike nuclease-based gene knockout, CRISPRi does not lead to activation of the DNA damage response which can alter cell state and transcriptional signatures (Haapaniemi et al., 2018).

To improve scalability, we optimized our CRISPRi sgRNA libraries. In order to maximize CRISPRi efficacy, we used multiplexed CRISPRi libraries in which each construct contains two distinct sgRNAs targeting the same gene (Replogle et al., 2020; see *Methods*). To avoid low representation of essential genes, we performed preliminary growth screens and purposefully overrepresented constructs that caused strong growth defects during library synthesis (**Supplementary Figure 2.1 A-D**).

Next, we devised a three-pronged Perturb-seq screening approach encompassing multiple timepoints and cell types (**Figure 2.1A**). As a primary cell line, we studied chronic myeloid leukemia (CML) K562 cells engineered to express the CRISPRi effector protein dCas9-KRAB (Gilbert et al., 2014). In this cell line, we performed two Perturb-seq screens: one targeting all expressed genes at day 8 after transduction (n=9,866 genes; n=10,673 total perturbations; some genes have multiple independent transcripts) and another targeting common essential genes at day 6 after transduction (n=2,057 genes; n=2,176 total perturbations). As a secondary cell line, we used RPE1 cells engineered to express dCas9 fused to a KRAB domain derived from the gene ZIM3, which was recently shown to yield improved transcriptional repression compared to the KOX1 KRAB domain used in previous CRISPRi experiments (Alerasool et al., 2020). In contrast to K562 cells, RPE1 cells are a non-cancerous retinal pigment epithelial cell line that are hTERT-immortalized, near-euploid, adherent, and p53-positive. In RPE1 cells, we performed a screen targeting common essential genes plus a subset of nonessential genes that produced phenotypes in K562 cells (n=2,393 genes; n=2,549 total perturbations).

Across these three screens, we ran 377 lanes of 10x Genomics droplet-based 3' scRNA-seq with direct sgRNA capture. After sequencing and read alignment, we performed sgRNA identification and removed any cells bearing sgRNAs targeting different genes, which are an expected byproduct of lentiviral recombination between sgRNA cassette or doublet encapsulation during scRNA-seq. In total, we obtained >2.5 million high-quality cells with a median coverage of >100 cells per perturbation (**Supplementary Figure 2.1E-G**). We observed a median target knockdown of 85.5% in K562 cells and 91.6% in RPE1 cells (**Figure 2.1B**), confirming both the efficacy of our CRISPRi libraries and the fidelity of sgRNA assignment. The difference in performance between these cell lines was likely due to the use of an optimized KRAB domain in the RPE1 cells, suggesting that future efforts would benefit from improved CRISPRi efficacy.

A robust computational framework to detect transcriptional phenotypes

The scale of our experiment provided a unique opportunity to ask what fraction of genetic perturbations cause a transcriptional phenotype, a preliminary requirement for inferring gene function. Significant transcriptional phenotypes can take many forms, ranging from altered occupancy of a given cell state to focused changes in the expression level of a small number of target genes. To contend with this diversity, we created a robust framework capable of detecting transcriptional changes between groups of cells in our data. Our experimental design included many control cells bearing diverse non-targeting sgRNAs. These allow for internal *z*-normalization of expression measurements, and we found that this procedure corrected for batch effects that resulted from parallelized scRNA-seq and sequencing (**Supplementary Figure 2.2**). As Perturb-seq captures single-cell genetic perturbation identities in a pooled format, we can use powerful statistical approaches that treat each cell as an independent experimental sample. In general, we chose to use conservative, non-parametric statistical tests to detect transcriptional changes rather than making specific assumptions about the underlying distribution of gene expression levels.

First, we examined global transcriptional changes using a permuted energy distance test (see *Methods*). We compared cells bearing each genetic perturbation to non-targeting control cells at the level of principal components (approximating global transcriptional features like cell state and gene expression programs). Relative to a permuted null distribution, this test asks whether cells carrying a given genetic perturbation could have been drawn from the control population. By this metric, we found that 2,987 of 9,608 genetic perturbations targeting a primary transcript (31.1%) compared to 11 of 585 non-targeting controls (1.9%) caused a significant transcriptional phenotype in K562 cells.

While powerful, the energy distance test assays global shifts in expression without providing insight into what specific transcripts are altered. To detect individual differentially expressed genes, we applied the Anderson-Darling (AD) test to compare the distribution of expression levels for each gene in cells bearing each genetic perturbation against control cells.

Importantly, the AD test is sensitive to transcriptional changes in a subset of cells, enabling us to find differences even when phenotypes are incompletely penetrant. With the AD test, we found 2,935 of 9,608 genetic perturbations targeting a primary transcript (30.5%) compared to 12 of 585 non-targeting controls (2.1%) caused >10 differentially expressed genes in K562 cells. These results were well-correlated between time points and cell types (**Supplementary Figure 2.3A,B**) and concordant with the energy distance test (78.7% concordance by Jaccard index).

Next we wanted to understand features of genetic perturbations that predict a transcriptional phenotype. We found that the strength of the transcriptional response was correlated with the strength of the growth defect (Spearman's rho = –0.51) with 86.6% of essential genetic perturbations (gamma < –0.1) leading to a significant transcriptional response in K562 cells (**Figure 2.1C; Supplementary Figure 2.3C,D**). Critically, a substantial number of genetic perturbations that cause a transcriptional phenotype nonetheless have a negligible growth phenotype (n=771; **Supplementary Figure 2.3E**), showing that many genetic perturbations influence cell state but not growth or survival. We also found that highly expressed genes are more likely to produce transcriptional phenotypes (Spearman's rho = 0.42) (**Supplementary Figure 2.3C**).

Considering that some of our genetic perturbations did not produce strong on-target knockdown, our estimate of the fraction of genetic perturbations that cause a transcriptional phenotype is likely to be a lower bound. While off-target effects may explain some fraction of phenotypes, an advantage of Perturb-seq is the ability to directly detect a major source of CRISPRi off-target effects: knockdown of neighboring genes. Consistent with earlier studies showing that CRISPRi can silence bidirectional promoters, we found that ~7.5% of perturbations caused significant knockdown of a neighboring gene in K562 cells, but neighbor gene knockdown was not enriched in genetic perturbations with a negligible growth defect that produced a transcriptional phenotype (**Supplementary Figure 2.4**). Taken together, these results present a

coherent picture where knockdown of a significant fraction of expressed genes causes a transcriptional response.

Because Perturb-seq is a single-cell assay, we could also assess the penetrance of perturbation-induced phenotypes. As a single-cell metric of phenotypic magnitude, we used SVD-based leverage scores (see *Methods*). In this formulation, leverage scores quantify how outlying each perturbed cell's transcriptome is relative to non-targeting control cells without assuming that each perturbation drives a single axis of variation in all cells. We found that mean leverage scores averaged over all cells for each genetic perturbation were correlated with the number of differentially expressed genes (Spearman's rho = 0.71; **Supplementary Figure 2.5A**), and were reproducible across the day 6 and day 8 K562 experiments (Spearman's rho = 0.79; **Supplementary Figure 2.5B,C**). We then examined the response to perturbations targeting two large essential complexes associated with strong transcriptional phenotypes, Mediator and the cytosolic small ribosomal subunit. In both cases, equivalent knockdown of different subunits could yield variable effects in terms of the number of differentially expressed genes (**Supplementary Figure 2.5D,F**) and the fraction of outlying cells as quantified by leverage scores (**Supplementary Figure 2.5E,G**). This result could be explained by biological variation in subcomplex function (see below) or dosage imbalance. In some cases, apparently unperturbed subpopulations of cells emerged over time, suggesting one technical source of incomplete penetrance is selection to escape toxic perturbations (**Supplementary Figure 2.5E,G**). In total, these results show how Perturb-seq can quantify both the average and cell-to-cell heterogeneity of effects induced by genetic perturbations.

Annotating gene function from transcriptional phenotypes

Previous Perturb-seq screens have focused on targeted sets of genetic perturbations that are often related biologically, such as hits from a forward genetic screen. Our large-scale screen

therefore presented a unique opportunity to assess how well transcriptional phenotypes can resolve gene function when deployed in an unbiased manner.

We focused on a subset of 1,973 perturbations that had strong transcriptional phenotypes (>50 differentially expressed genes by AD test) (**Figure 2.2A**). Because related perturbations could have different magnitudes of effect, we used the correlation between mean expression profiles as a scale-invariant metric of similarity.

To assess the extent to which correlated mean expression profiles between genetic perturbations indicated common function, we compared our results to two curated sources of biological relationships. First, among the 1,973 targeted genes, there were 327 protein complexes from the CORUM 3.0 database with at least two thirds of the complex members present, representing 14,165 confirmed protein-protein interactions (Giurgiu et al., 2019). The corresponding expression profile correlations were markedly stronger (median correlation 0.61) than the background distribution of all possible gene pairs (median correlation 0.10) (**Figure 2.2B**). Second, we compared the correlation between genetic perturbations to the STRING database of known and predicted protein-protein interactions, which had scores for 243,558 of the possible gene-gene relationships within our dataset (Szklarczyk et al., 2019). High STRING scores, reflecting high-confidence interacting proteins, were also associated with high expression correlations (**Figure 2.2C**).

We next performed an unbiased search for global structure to group similar perturbations within the dataset. We identified 64 discrete clusters based on strong intra-cluster correlations and annotated their function using CORUM, STRING, and manual searches. To visualize the dataset, we constructed a minimum distortion embedding that places genes with correlated expression profiles close to each other in the plane and labeled the location of gene clusters (**Figure 2.2D**).

Both the clusters and the embedding showed clear organization by biological function spanning a diverse array of different processes including: chromatin modification; transcription;

mRNA splicing, capping, polyadenylation, and turnover; nonsense-mediated decay; translation; post-translational modification, trafficking, and degradation of proteins; central metabolism; mitochondrial transcription and translation; DNA replication; cell division; microRNA biogenesis; and major signaling pathways active in K562 cells such as BCR-ABL and mTOR. We further annotated the embedding visualization by labeling CORUM complexes and STRING clusters whose members were placed in nearby positions, revealing structure at finer resolution such as identifying the SMN complex, exon junction complex, U6 snRNP, and methylosome within the spliceosome and the association of ribosome biogenesis factors with the 40S ribosomal subunit.

In our dataset, we identified many poorly annotated genes whose perturbation appeared highly similar to genes of known function, naturally predicting a role for these uncharacterized genes. To orthogonally test a subset of these predictions, we selected ten poorly annotated genes whose perturbation correlated with subunits and biogenesis factors of either the large or small subunit of the cytosolic ribosome, which formed distinct clusters in our data. This included genes that had no previous association with ribosome biogenesis (CCDC86, CINP, SPATA5L1, ZNF236, C1orf131) as well as genes that had not been associated with functional defects in a particular subunit (SPOUT1, TMA16, NOPCHAP1, ABCF1, and NEPRO). We used CRISPRi to target these genes in K562 cells and looked for evidence of ribosome biogenesis defects by assessing the ratio of 28S to 18S rRNA by Bioanalyzer electrophoresis. With the exception of ABCF1, nine of the ten candidate factors showed evidence of substantial defects in ribosome biogenesis (**Figure 2.2E**). In every case, the affected ribosomal subunit corresponded to the Perturb-seq clustering across two independent sgRNAs. While this study was in progress, another group used cryo-EM to identify C1orf131 as a core structural component of the pre-A1 small subunit processome, complementing our functional evidence (Singh et al., 2021). This validation suggests that many poorly characterized genes can be assigned functional roles through Perturb-seq, although a subset of these relationships might be explained by off-target effects (**Supplementary Figure 2.6**).

In total, these results show that transcriptional phenotypes have utility beyond studying gene regulation or transcriptional programs, and can serve as valuable tools for resolving and interrogating many of the most central processes within cell biology.

Delineating functional modules of the Integrator complex

In general, perturbations to members of known protein complexes produced similar transcriptional phenotypes in our dataset. Therefore, we were surprised by the wide spectrum of transcriptional responses to knockdown of subunits of Integrator, a metazoan-specific essential nuclear complex with roles in small nuclear RNA (snRNA) biogenesis and transcription termination at paused RNA polymerase II (**Figure 2.3A**) (Kirstein et al., 2021). Each of the fourteen core subunits of Integrator was targeted in our experiment, allowing us to systematically compare their transcriptional phenotypes in K562 and RPE1 cells (**Figure 2.3B**; **Supplementary Figure 2.7A**). INTS1, INTS2, INTS5, INTS7, and INTS8 formed a tight cluster which weakly correlated with INTS6 and INTS12. Separately, INTS3, INTS4, INTS9, and INTS11 clustered together alongside splicing regulators involved in snRNP assembly and the tri-snRNP. Finally, INTS10, INTS13, and INTS14 formed another discrete cluster together with C7orf26, an uncharacterized gene.

These distinct functional modules bore similarities to recent structures of the Integrator complex (Fianu et al., 2021; Zheng et al., 2020). The INTS1-2-5-7-8 functional module contained the subunits identified as the structural shoulder and backbone of Integrator. The INTS3-4-9-11 functional module contained the subunits identified as the endonuclease domain. While INTS10, INTS13, and INTS14 were not resolved in the recent Integrator structures, these subunits have been identified as a stable biochemical subcomplex (Pfleiderer and Galej, 2021; Sabath et al., 2020).

Integrator is an essential and well-studied complex, so we were intrigued by the robust clustering of the uncharacterized gene C7orf26 with Integrator subunits 10, 13, and 14. To explore

this, we tested whether loss of C7orf26 impacted the abundance of Integrator subunits. CRISPRi-based depletion of C7orf26 destabilized INTS10 in K562 cells, confirming either a direct or indirect protein-level relationship (**Figure 2.3C**). Next, we checked for a direct biochemical interaction between these proteins. Pulldown of His-INTS10 from cell lysates recovered endogenous C7orf26 alongside INTS13 and INTS14 (**Figure 2.3D**). Additionally, overexpression of C7orf26 with INTS10, INTS13, and INTS14 enabled the purification of a stable INTS10-13-14-C7orf26 complex by size-exclusion chromatography (**Figure 2.3E**; **Supplementary Table 2.7B and 2.8**). We also detected a physical interaction between the C7orf26 homologue and Integrator in Drosophila and co-essentiality between C7orf26 and INTS10, INTS13, INTS14 in the Cancer Dependency Map, suggesting that this relationship is conserved across species and cell types (**Supplementary Figure 2.9**). Together, these results suggest that C7orf26 is a core subunit of a novel INTS10-13-14-C7orf26 Integrator module.

We sought to better understand the distinct transcriptional phenotype of INTS10-13-14-C7orf26 compared to the shoulder/backbone and endonuclease modules. We began by examining genes that were differentially expressed between the modules (**Supplementary Figure 2.7C,D**). While robust, the transcriptional differences between modules did not reveal function in an obvious way. Next, we explored the canonical role of Integrator in snRNA biogenesis. As mature snRNAs are not captured in 3' scRNAs-seq, we monitored changes in global splicing as a proxy for snRNA biogenesis defects. In our Perturb-seq data, we quantified changes in splicing by comparing the ratio of intronic (unspliced) to exonic (spliced) reads for each gene. Validating our approach, depletion of known splicing factors as well as subunits of the endonuclease and shoulder/backbone modules led to gross splicing defects (**Figure 2.3F**). By contrast, depletion of subunits of the INTS10-13-14-C7orf26 module did not cause a substantial splicing defect. To definitively test the effect of the INTS10-13-14-C7orf26 module on snRNA biogenesis, we used PRO-seq to probe active RNA-polymerase positioning. This data confirmed that the endonuclease and backbone/shoulder modules, but not INTS10, INTS13, or C7orf26,

caused a dramatic increase in transcriptional readthrough past the 3' cleavage site of snRNAs (**Figure 2.3G**). In addition, the PRO-seq data confirmed that the INTS10-13-14-C7orf26 module has a distinct transcriptional phenotype from other modules, showing that the module influences not only mRNA levels but also active RNA polymerase activity (**Supplementary Figure 2.7E**).

In sum, our results show that INTS10-13-14-C7orf26 represents a functionally and biochemically distinct module of the Integrator complex, and we propose that C7orf26 be renamed INTS15 for future studies (**Figure 2.3H**). Although Integrator has been subjected to extensive structural analyses, it has been difficult to resolve the INTS10-13-14 components in relation to the rest of the complex. Inclusion of C7orf26 may facilitate future structural efforts. Broadly, this example highlights the utility of high-dimensional functional phenotypes for the unsupervised classification of protein complex subunits into functional modules.

Systematic analysis of transcriptional programs and composite phenotypes

While clustering can organize genetic perturbations into pathways or complexes, it does not reveal the functional consequences of perturbations. A key strength of Perturb-seq is the ability to systematically classify transcriptional phenotypes to understand the cellular effects of perturbations.

To globally summarize the genotype-phenotype relationships in our dataset, we: (i) clustered genes into expression programs based on their co-regulation across perturbations; (ii) clustered perturbations with strong phenotypes based on their transcriptional profiles (as described above); and (iii) computed the average activity of each gene expression program within each perturbation cluster (**Figure 2.4A,B**; see *Methods*). This map uncovered many known gene expression programs associated with genetic perturbations, including upregulation of proteasomal subunits due to proteosome dysfunction (Radhakrishnan et al., 2010), activation of NFkB signaling upon loss of ESCRT proteins (Mamińska et al., 2016), downregulation of growth-related genes in response to essential genetic perturbations, and upregulation of the cholesterol

biosynthesis pathway in response to defects in vesicular trafficking (Luo et al., 2020). Beyond these large-scale relationships, we could also score the effects of individual genetic perturbations on different expression programs. For example, our analysis delineated the canonical branches of the cellular stress response into the independently regulated Unfolded Protein Response (UPR) and Integrated Stress Response (ISR) (**Figure 2.4C**) (Adamson et al., 2016). The ISR was highly activated by loss of mitochondrial proteins, aminoacyl-tRNA synthetases, and translation initiation factors, whereas the UPR was activated by loss of ER-resident chaperones and translocation machinery. Collectively, this analysis establishes the ability of Perturb-seq to learn regulatory circuits by leveraging the variability of responses across perturbations.

Interestingly, our unbiased clustering uncovered many perturbations that drove the expression of markers of erythroid or myeloid differentiation, consistent with the known multilineage potential of K562 cells (**Figure 2.4D**) (Leary et al., 1987). The scale of our experiment allowed us to comprehensively search for genes whose modulation promotes cellular differentiation, an application of major interest in both developmental and cancer biology. As expected, loss of central regulators of erythropoiesis (GATA1, LDB1, LMO2, and the differentiation therapy target KDM1A) caused myeloid differentiation while knockdown of BCR-ABL and its downstream adaptor GAB2 induced erythroid differentiation (Orkin and Zon, 2008). Surprisingly, loss of a number of common essential genes (i.e. essential across cell lines in the Cancer Dependency Map) also caused expression of either myeloid (e.g., Integrator subcomplex) or erythroid (e.g., NuRD complex, DNA replication machinery) markers. Next, we investigated selectively essential genes that drove differentiation, as in principle these could be promising targets for differentiation therapy, analogous to KDM1A (Maes et al., 2018; Yu et al., 2021). We observed that loss of PTPN1, a tyrosine phosphatase selectively essential in K562 cells, drove myeloid differentiation. While inhibitors of PTPN1 have been developed for use in diabetes and certain cancers (Sharma et al., 2020), to our knowledge they have not been tested as a differentiation therapy. Remarkably, in targeted experiments, we found that knockdown of PTPN1

and KDM1A in combination caused a substantial increase in differentiation and growth defect compared to either single genetic perturbation, suggesting that these targets act via different cellular mechanisms (**Figure 2.4E**; **Supplementary Figure 2.10A**). These results highlight the utility of rich phenotypes for understanding differentiation as well as nominating promising therapeutic targets or combinations.

Finally, we recognized that our scRNA-seq readout could be used to study phenotypes that integrate data from across the transcriptome and, therefore, would be difficult to study in traditional forward genetic screens . Examples of these “composite phenotypes” include total cellular RNA content and the fraction of RNA derived from transposable elements (TE). We found that these phenotypes were under strong genetic control, with highly reproducible effects across screen replicates and cell types (**Figure 2.4F**). In the specific case of TE regulation, two major classes of perturbations increased the fraction of TE RNA in single cells, affecting broad classes of elements including Alu, L1, and ERV1 (**Figure 2.4G**). First, loss of the exosome led to a substantial increase in the fraction of TE RNA, suggesting that transcripts deriving from TEs might be constitutively degraded. Second, loss of the CPSF cleavage and polyadenylation complex and parts of the Integrator complex produced a similar phenotype, suggesting that many of the TE RNAs observed in K562 cells may be derived from failure of normal RNA transcription termination, rather than representing independent reactivation of TE transcription.

Turning to total RNA content (**Figure 2.4H**), we found that loss of many essential regulators of S-phase and mitosis increased the RNA content of cells. This is consistent with the observation that cells tend to increase their size, and thus their RNA content, as they progress through the cell cycle (**Supplementary Figure 2.10B**), so perturbations that arrest cells in later cell cycle stages yield increased total RNA content on average. By contrast, loss of essential transcriptional machinery, including general transcription factors, the Mediator complex, and transcription elongation factors, decreased total RNA content. In sum, these analyses show that

genome-scale Perturb-seq enables robust hypothesis-driven and data-driven exploration of complex cellular features which would be difficult to study through other means.

Exploring genetic drivers and consequences of aneuploidy in single-cells

In our data, we noted cases of heterogenous responses to genetic perturbations at the level of individual cells. For example, knockdown of CCNF (cyclin F) primarily induced transcriptional changes in G2/M but not in other cell cycle phases, consistent with its role as an inhibitor of centrosome reduplication in G2 (**Supplementary Figure 2.11A**) (D'Angiolella et al., 2010). While heterogeneous responses can have a biological (e.g., incomplete penetrance) or technical (e.g., variable knockdown) basis, we reasoned that systematically exploring sources of heterogeneity could uncover mechanistic insights into cellular phenotypes.

To compare the heterogeneity in response to genetic perturbations, we scored each perturbation by its variation in single-cell leverage scores (**Figure 2.5A**). Intriguingly, many genes implicated in chromosome segregation were among the top drivers of heterogeneity, including TTK, SPC25, and DSN1 (**Figure 2.5B**) (Musacchio and Salmon, 2007). We hypothesized that the extreme transcriptional variability caused by these genetic perturbations might in fact result from acute changes in the copy number of individual chromosomes due to mitotic mis-segregation. To explore this, we used inferCNV (Patel et al., 2014) to estimate single-cell DNA copy number along the genome by quantifying the change in moving average gene expression compared to control cells. Consistent with our hypothesis, knockdown of TTK, a core component of the spindle assembly checkpoint (Jelluma et al., 2008), led to dramatic changes in estimated DNA copy number in both K562 and RPE1 cells (**Figure 2.5C**; **Supplementary Figure 2.11B**). Specifically, in RPE1 cells, we found that 61/80 (76%) of TTK knockdown cells had evidence of karyotypic changes compared to 274/13140 (2%) of unperturbed cells. Notably, TTK knockdown cells bore highly variable karyotypes due to the stochastic gain or loss of chromosomes, accounting for the phenotypic heterogeneity observed in these cells.

An important advantage of the rich data provided by Perturb-seq is the ability to dissect not only perturbation-phenotype associations but also relationships between cellular phenotypes. We were curious how chromosomal instability (CIN) would affect cell cycle progression in euploid, p53-positive RPE1 cells versus constitutively aneuploid, p53-deficient K562 cells. Expanding our analysis to all cells in our experiment independent of genetic perturbation, we found that RPE1 cells with likely karyotypic changes tended to arrest in G1/G0 of the cell cycle, while K562 cells with likely karyotypic changes had less significant shifts in cell cycle occupancy (**Figure 2.5D,E**; see *Methods*). Interestingly, within the population of RPE1 cells bearing a chromosomal loss, the likelihood of cell cycle arrest depended on the magnitude of karyotypic abnormality (**Supplementary Figure 2.11C**). Additionally, we observed that cells with the most severe karyotypic changes—those bearing both chromosomal gains and losses—had marked upregulation of the ISR (**Figure 2.5F**; **Supplementary Figure 2.11D**). These results are consistent with models in which cell cycle checkpoints are activated by the secondary consequences of aneuploidy (e.g., DNA damage or proteostatic stress) rather than changes in chromosome number *per se* (Santaguida et al., 2017; Santaguida and Amon, 2015).

Finally, we looked across all perturbations to systematically identify genetic drivers of CIN. We assigned a score to each perturbation based on the average magnitude of induced karyotypic abnormalities. Validating our approach, we found that CIN scores were strongly correlated across K562 and RPE1 cell lines (r=0.69) and nominated many known regulators of chromosomal segregation, including components of the spindle assembly checkpoint, centromere, and NDC80 complex (**Figure 2.5G**). Remarkably, we uncovered CIN regulators with diverse cellular roles, from cytoskeletal components to DNA repair machinery (**Figure 2.5H**). While many of these genes have previously been associated with chromosomal instability through targeted studies, the scale and single-cell resolution of Perturb-seq allowed us to identify numerous genetic drivers of CIN in a single, unbiased experiment. This analysis also shows the potential of single-cell CRISPR screens to dissect phenotypes that were not predefined endpoints of the experiment.

Discovery of stress-specific regulation of the mitochondrial genome

Mitochondria arose from the engulfment and endosymbiotic evolution of an ancestral alphaproteobacterium by the precursor to eukaryotic cells (Friedman and Nunnari, 2014). While the vast majority (~99%) of mitochondrially-localized proteins are encoded in the nuclear genome, mitochondria contain a small (~16.6 kilobase) remnant of their ancestral genome encoding 2 rRNAs, 22 tRNAs, and 13 protein-coding genes. An open question is how the nuclear and mitochondrial genomes coordinate their expression to cope with mitochondrial stress (Quirós et al., 2016). The scale of our experiment provided a unique opportunity to investigate this question.

We began by comparing the nuclear transcriptional responses to CRISPRi-based depletion of nuclear-encoded mitochondrial genes (i.e., mitochondrial perturbations). We found that mitochondrial perturbations elicited relatively homogeneous nuclear transcriptional responses, illustrated by well-correlated transcriptional phenotypes across mitochondrial perturbations (**Figure 2.6A**; **Supplementary Figure 2.12A**). While there was some variation in the magnitude of transcriptional responses (e.g., proteostatic injury drove an especially strong ISR activation), nuclear transcriptional responses generally failed to discriminate genetic perturbations by function. Although this result was broadly consistent with recent literature which has highlighted the role of the ISR as response to mitochondrial stress (Fessler et al., 2020; Guo et al., 2020; Mick et al., 2020; Münch and Harper, 2016; Quirós et al., 2017), the lack of functional specificity of the transcriptional response was puzzling in light of: (i) the multifaceted roles of mitochondria in diverse processes such as respiration, intermediary metabolism, iron-sulfur cluster biogenesis, and apoptosis and (ii) the high-resolution separation of cytosolic perturbations by transcriptional response.

In contrast to the nuclear transcriptional response, we observed that expression of the 13 mitochondrial-encoded genes was highly variable between mitochondrial perturbations (**Figure 2.6B**; **Supplementary Figure 2.12B**). When we clustered mitochondrial perturbations based solely on expression levels of the 13 mitochondrial-encoded genes, a remarkably intricate and

coherent pattern emerged: the clustering separated perturbations to Complex I, Complex IV, Complex III, Complex V (ATP synthase), the mitochondrial large ribosomal subunit, the mitochondrial small ribosomal subunit, chaperones/import machinery, and RNA processing factors (**Figure 2.6C**; **Supplementary Figure 2.12C**). To quantitatively support this observation, we trained a random forest classifier to distinguish cells with perturbations to different mitochondrial complexes and found that the mitochondrial transcriptome was far more predictive than the nuclear transcriptome (mitochondrial accuracy 0.64; nuclear accuracy: 0.25) (**Supplementary Figure 2.12D**). We then visualized the gene expression signatures of a subset of representative perturbations (**Figure 2.6D**). The coregulation of mitochondrial genes tended to reflect function, with the exception of the bicistronic mRNAs ND4L/ND4 and ATP8/ATP6 (Mercer et al., 2011). However, we did not identify a clear logic to map genetic perturbations to their observed transcriptional consequences. While previous studies have described distinct regulation of mitochondrial translation in response to specific perturbations (notably, loss of complex III and complex IV assembly factors (Richter-Dennerlein et al., 2016; Salvatori et al., 2020)), our data generalizes this phenomenon to a comprehensive set of stressors.

Next, we wanted to shed light on the mechanistic basis for this unappreciated complexity of mitochondrial genome responses. Given its singular origin, the mitochondrial genome is expressed by unique processes (**Figure 2.7A**) (Kummer and Ban, 2021). Mitochondrial-encoded genes are transcribed as part of three polycistronic transcripts punctuated by tRNAs. These transcripts are then processed into rRNAs and mRNAs by tRNA excision, and individual mRNAs can be polyadenylated, expressed, or degraded. This complex system limits the potential for distinct transcriptional control but presents multiple opportunities for post-transcriptional regulation. To identify the modes of regulation used to carry out the newly described perturbation-specific responses to mitochondrial stressors, we examined the distribution of Perturb-seq reads along the mitochondrial genome (**Figure 2.7B**). As our scRNA-seq used poly-A selection, most reads aligned to the 3’ ends of mRNAs. To validate the utility of this position-based analysis, we

confirmed that knockdown of known regulators of mitochondrial transcription (TEFM) and RNA degradation (PNPT1) led to major shifts in the position of reads along the mitochondrial genome. By contrast, many of the perturbation-specific responses discovered in the present study appeared to cause shifts in the relative abundance of mRNAs rather than gross shifts in positional alignments. As scRNA-seq primarily sequences polyadenylated transcripts, the observed mitochondrial genome responses could have been regulated at the level of RNA abundance or polyadenylation. To distinguish these, we performed a bulk RNA-sequencing experiment without poly-A selection, confirming that the majority of regulation influenced the level of RNA abundance (cophenetic correlation r=0.79; **Figure 2.7C**). Given the complexity of the observed responses, we propose that there are likely to be multiple mechanisms that regulate the levels of mitochondrial-encoded transcripts in response to diverse stressors.

Finally, we asked whether we could use the detailed clustering produced by the mitochondrial genome to predict gene function. Knockdown of an unannotated gene, TMEM242, produced a transcriptional signature resembling loss of ATP synthase in both K562 and RPE1 cells (**Figure 2.7D**; **Supplementary Figure 2.12E**). Supporting this relationship, the top five co-essential genes with TMEM242 were components of ATP synthase in the Cancer Dependency Map. To examine the effect of TMEM242 on respiratory chain function, we performed a Seahorse assay and confirmed that basal respiration was decreased in TMEM242 knockdown cells (**Figure 2.7E**). While this work was in progress, another group used a biochemical approach to show that TMEM242 physically interacts with ATP synthase subunits and regulates ATP synthase complex assembly in cells (Carroll et al., 2021). Together, these experiments discover a novel factor required for ATP synthase activity and point to the precision of mitochondrial genome regulation.

DISCUSSION

Single-cell CRISPR screens represent an emerging tool to generate rich genotype-phenotype maps. However, to date, their use has been limited to the study of preselected genes focused on discrete, predefined biological questions. Here, we perform genome-scale single-cell CRISPR screens using Perturb-seq and demonstrate how these screens enable data-driven dissection of a remarkable breadth of complex biological phenomena. Reflecting on this study, we highlight key biological insights and derive principles to guide future discoveries from rich genotype-phenotype maps.

A primary aim of large-scale functional screens is to organize genes into pathways or complexes. To this end, we used Perturb-seq to perform high-resolution clustering of genetic perturbations. From a single assay, we recapitulated thousands of known relationships while also assigning new, experimentally validated roles to genes involved in ribosome biogenesis or translation (CCDC86, CINP, SPATA5L1, ZNF236, C1orf131, SPOUT1, TMA16, NOPCHAP1, NEPRO), transcription (C7orf26), and respiration (TMEM242). However, other large-scale experimental techniques, such as protein-protein interaction mapping, genetic interaction mapping, and co-essentiality analysis, similarly group genes or proteins by function. How then are single-cell CRISPR screens unique?

We argue that these screens are particularly powerful because of the intrinsic interpretability of comprehensive genotype-phenotype maps, enabling in-depth dissection of the functional consequences of genetic perturbations that impinge on many distinct aspects of cell biology. Of particular note is the ability to use the information-rich readouts to study complex, composite phenotypes which are difficult to measure by other modalities. These composite phenotypes can be created in a hypothesis-driven manner (e.g., measuring intron/exon ratios to studying splicing) or data-driven manner (e.g., deriving stress-specific transcriptional changes of mitochondrial encoded genes), resulting in an enormous breath of measured phenotypes. In the case of scRNA-seq, we show that it measures not only features such as differential gene

expression and the activity of critical transcriptional programs, but also RNA splicing and processing, expression of transposable elements, differentiation, transcriptional heterogeneity, cell cycle progression, and chromosomal instability. Once a phenotype is defined, the genotype-phenotype map can be used to explore its genetic underpinnings, in a manner analogous to a forward genetic screen, as well as its relationship to other cellular phenotypes.

An illustrative example of this process is our study of chromosomal instability. Based on an initial observation of heterogeneous responses to specific perturbations, we suspected that some cells carried genetically-induced chromosomal gains or losses. In a hypothesis-driven manner, we then used our rich phenotypic data to discover a large collection of perturbations—which were only loosely connected by clustering on average transcriptional phenotypes—that promote chromosomal instability. Importantly, the single-cell nature of our Perturb-seq data also allowed us to explore the relationship between karyotypic changes and other phenotypes, including cell cycle progression and stress induction. While aneuploidy is an important hallmark of most cancers, it has not been easy to study with traditional genetic screens as it requires both a single-cell and multimodal readout. In future work, this platform could be used to investigate interactions between genetic perturbations and specific karyotypes, karyotype-dependent stress responses, or the temporal evolution of karyotypes (Ben-David and Amon, 2020).

Because composite phenotypes can be generated and explored without being preregistered at the time of data collection, rich genotype-phenotype maps provide a powerful resource for the discovery of new cellular behaviors. Genetic perturbations can push cells into extreme states that are not observed in unperturbed cells. With data-driven exploration, we can uncover new phenotypes in a principled manner and interpret these emergent phenotypes through the lens of their genetic dependencies.

Using this ability, we discovered a remarkable degree of stress-specific regulation of the mitochondrial genome. It was only possible to appreciate the functional specificity of this regulation by pairing a defined set of mitochondrial perturbations with a high-dimensional readout.

This discovery suggests a framework to help explain how cells cope with diverse insults to mitochondria: a general nuclear response is layered over perturbation-specific regulation of the mitochondrial genome (**Figure 2.7F**). Building on this observation, we can ask new questions about the mitochondrial stress response. The transcriptional changes we observed may reflect adaptive responses or, alternatively, complex patterns of dysfunction owing to disruption of the intricate system of mitochondrial gene expression. Understanding how and in what contexts this regulation is adaptive may have important implications for diseases associated with mitochondrial stress. An intriguing additional question is whether individual mitochondria are able to regulate their expression autonomously. Combined with the nuanced responses observed here, this would support and substantially extend the "co-location for redox regulation" (CoRR) hypothesis which holds that the endosymbiotically-derived mitochondrial genome has been retained through evolution to enable localized regulation of mitochondrial gene expression (Allen, 2017).

A final theme emerging from our work is the technical advantage of single-cell CRISPR screens compared to other functional genomic approaches. Because these screens extract rich information from each cell in a pooled format, they require only a fraction of the number of cells used by other approaches and thus are well suited to the study of iPSC-derived cells and *in vivo* samples. As technologies for single-cell, multimodal phenotyping advance, single-cell screens will continue to become more powerful. At present, the major limitation of single-cell CRISPR screens is cost. Careful experimental designs, such as multiplexed libraries or compressed sensing (Cleary et al., 2017), together with technical advances in single-cell phenotyping (Datlinger et al., 2021; Martin et al., 2021) and DNA sequencing promise to greatly increase the scale of these experiments. To this point, we concluded our work by sequencing our genome-scale K562 libraries on a lower-cost, ultra-high throughput sequencing platform developed by Ultima Genomics, generating results equivalent to those sequenced on Illumina instruments.

In sum, our study presents a blueprint for the construction and analysis of rich genotype-phenotype maps which serve as a driving force for the systematic and principled exploration of

genetic and cellular function. Our data will be available after publication in interactive and raw formats at https://weissman.wi.mit.edu/perturbseq/.

MATERIALS AND METHODS

Experimental Methods

Cell culture and lentiviral production K562 cells were grown in RPMI-1640 with 25 mM HEPES, 2.0 g/l NaHCO3, and 0.3 g/l L-glutamine supplemented with 10% FBS, 2 mM glutamine, 100 units/ml penicillin, and 100 µg/ml streptomycin. hTERT-immortalized RPE1 cells (ATCC, CRL-4000) were grown in DMEM:F12 supplemented with 10% FBS, 0.01 mg/ml hygromycin B, 100 units/ml penicillin, and 100 µg/ml streptomycin. HEK293T cells were used for generation of lentivirus, and grown in DMEM supplemented with 10% FBS, 100 units/ml penicillin and 100 µg/ml streptomycin. Lentivirus was produced by co-transfecting HEK293T cells with transfer plasmids and standard packaging vectors using TransIT-LTI Transfection Reagent (Mirus, MIR 2306).

Cell line generation CRISPRi K562 cells expressing dCas9-BFP-KRAB (KOX1-derived) were obtained from Gilbert et al., 2014. CRISPRi RPE1 cells expressing dCas9-BFP-KRAB (KOX1-derived) were obtained from Jost et al., 2017 and only used for growth screens. CRISPRi RPE1 cells were generated by stably transducing RPE1 cells (ATCC, CRL-4000) with lentivirus expressing ZIM3 KRAB-dCas9-P2A-BFP from a UCOE-SFFV promoter (pJB108) and sorting for BFP+ cells stably expressing the construct using fluorescence activated cell sorting. Cell lines were verified by monitoring BFP fluorescence over several generations to confirm stable integration and confirming knockdown of select surface markers by flow cytometry.

Library design and cloning A distinct set of genes was targeted for each of the three large-scale Perturb-seq experiments. For the K562 day 8 genome-scale experiment, we targeted (i) genes expressed in K562 cells (ii) transcription factors as determined by Lambert et al., 2018 (iii) Cancer Dependency Map common essential genes as defined in 20Q1 (iv) non-targeting control sgRNAs

accounting for 5% of the total library. To define expressed genes in K562 cells, we used a combination of bulk RNA-seq data from ENCODE (https://www.encodeproject.org/files/ENCFF717EVE/) and 10x Genomics 3' single-cell RNA-seq data (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE146194), selecting a set of genes accounting for ~99% of aligned reads in both datasets. For the K562 day 6 essential-scale experiment, we targeted (i) Cancer Dependency Map common essential genes as defined in 20Q1 (ii) non-targeting control sgRNAs accounting for 5% of the total library. For the RPE1 day 7 essential-scale experiment, we targeted (i) 20Q1 Cancer Dependency Map common essential genes (https://depmap.org/portal/download/) (ii) a number of hand-selected genes with interesting phenotypes in the K562 genome-wide Perturb-seq dataset (iii) non-targeting control sgRNAs accounting for 5% of the total library. To define control perturbations, we randomly sampled non-targeting control perturbations from Horlbeck et al., 2016. A small number of genes were lost in this pipeline due to changes in gene annotation between datasets.

To minimize library size while maximizing knockdown, multiplexed CRISPRi libraries were constructed which targeted each gene with two unique sgRNAs expressed from tandem U6 expression cassettes in a single lentiviral vector, as previously described in Replogle *et al.*, 2020. The Horlbeck *et al.*, 2016 CRISPRi sgRNA libraries were used as a source of sgRNAs targeting each gene, with the optimal sgRNA pair targeting each gene selected based on a balance of empirical data with computational predictions. For strong essential genes (defined by a p-value<0.001 and gamma<-0.2 in the Horlbeck *et al.*, 2016 CRISPRi growth screen), sgRNAs were ranked by growth. Then, for genes that produced a significant phenotype in previous CRISPRi screens, sgRNAs were ranked by a discriminant score multiplying the negative $\log_{10}$ p-value by the effect size. Finally, for genes without any empirical evidence, sgRNAs were ranked according to the Horlbeck *et al.,* 2016 hCRISPRi v2.1 algorithm.

We adapted the protocol previously described in Replogle *et al.,* 2020 to clone libraries with capture sequences for 3' direct capture Perturb-seq. Briefly, an sgRNA lentiviral expression

vector (pJR101) was derived from the parental pJR85 (Addgene #140095), modified to incorporate a GFP fluorescent marker to avoid spectral overlap with BFP+ CRISPRi constructs and a UCOE element upstream of the EF1alpha promoter to prevent silencing. A two-step restriction enzyme digestion and ligation cloning of oligos into pJR101 was performed to maintain coupling of sgRNAs targeting the same gene. Oligos encoding the targeting regions of dual-sgRNA pairs were synthesized as an oligonucleotide pool (Twist Bioscences) with the structure: 5'- PCR adapter - CCACCTTGTTG – targeting region A - gtttcagagcgagacgtgcctgcaggatacgtctcagaaacatg – targeting region B - GTTTAAGAGCTAAGCTG - PCR adapter-3'. When ordering oligos, the representation of essential genes was increased to compensate for growth phenotypes (see below). Oligo pools were amplified, digested with BstXI/BlpI, and ligated into pJR101. To add an sgRNA constant region and U6 promoter to the vector, pJR89 (Addgene #140096) was BsmBI-digested and ligated into the intermediate library.

K562 and RPE1 growth screens Pooled sgRNA growth screens in K562 cells were used to quantify growth phenotypes of sgRNA pairs targeting expressed genes. CRISPRi K562 cells expressing dCas9-BFP-KRAB were transduced with lentiviral particles encoding the dual-sgRNA library by spinfection (1000g) with polybrene (8 ug/ml) to obtain an infection rate of ~25%-35%. Screens were performed in biological replicate with the aim of maintaining 1000 cells per library element for the duration of the screen. Between day 2 and day 6 post-transduction, cells were selected for lentiviral infection using 1 ug/mL puromycin, replenished every 24 hours. On day 7 post-transduction, an aliquot of cells was harvested as an initial time point) The rest of the cell population was passaged for 10 more days and collected at final time point.

Pooled sgRNA growth screens in RPE1 cells were used to quantify growth phenotypes of sgRNA pairs targeting common essential genes. The CRISPRi RPE1 cell line expressing dCas9-BFP-KRAB was used for growth screens which took place before the publication of the next-

generation ZIM3 KRAB domain. Cells were transduced in biological replicate with lentiviral particles encoding the dual-sgRNA library by replating cells into virus-laden media with polybrene (8 ug/ml) to obtain an infection rate of ~45%. Because RPE1 cells are puromycin resistant, we performed the screen without selection for sgRNA-infected cells, nonetheless maintaining an infection rate-corrected 1000 cells per library element for the duration of the screen. On day 6 post-transduction, an aliquot of cells was harvested as an final time point for direct comparison to the abundances in the plasmid library.

For both K562 and RPE1 growth screens, DNA libraries of the initial and final samples were prepared for deep sequencing by genomic DNA isolation and PCR amplification of dual-sgRNA amplicons. First, a NucleoSpin Blood XL kit (Macherey–Nagel) was used to extract genomic DNA (gDNA) from cells. Then, isolated gDNA or plasmid DNA was amplified by 22 cycles (gDNA) or 13 cycles (plasmid DNA) of PCR using NEBNext Ultra II Q5 PCR MasterMix (NEB), appending Illumina adaptors and sample indices (oJR234 forward primer: 5'-AATGATACGGCGACCACCGAGATCTACACCGCGGTCTGTATCCCTTGGAGAACCACCT-3'; index primers 5'-CAAGCAGAAGACGGCATACGAGATnnnnnGCGGCCGGCTGTTTCCA GCTTAGCTCTTAAA-3'). Amplicons were isolated by a 0.5-0.65X SPRI bead selection (SPRIselect Beckman Coulter #B23318). Sequencing was performed on a NovaSeq 6000 (Illumina) using a 19 bp read 1, 19 bp read 2, and 5 bp index read 1 with custom sequencing primers oJR326 (custom read 1, 5'- CGCGGTCTGTATCCCTTGGAGAACCACCTTGTTGG-3'), oJR328 (custom read 2, 5'- GCGGCCGGCTGTTTCCAGCTTAGCTCTTAAAC-3'), and oJR327 (custom index read 1, 5'- GTTTAAGAGCTAAGCTGGAAACAGCCGGCCGC-3').

Perturb-seq experiments To perform our K562 day 8 genome-scale Perturb-seq experiment, library lentivirus was packaged into lentivirus in 293T cells and empirically measured in K562 cells to obtain viral titers. CRISPRi K562 cells were transduced via spinfection (1000g) with polybrene (8 ug/ml) with the target of obtaining an infection rate of ~10%. Cells were maintained at a viability

of >90%, a coverage of 1000 cells per library element, and a density of 250,000 to 1,000,000 cells/ml for the course of the experiment. Three days post-transduction, an infection rate of 14% was measured, and cells were sorted to near purity by FACS (FACSAria2, BD Biosciences), using GFP as a marker for sgRNA vector transduction. Eight days post infection, the cells were measured to be 97% GFP+ (LSR2, BD Biosciences), >90% viable, and at a concentration of ~800,000 cells/ml (Countess II, ThermoFisher). Cells were prepared for single-cell RNA-sequencing by resuspension in 1X PBS with 0.04% BSA as detailed in the 10x Genomics Single Cell Protocols Cell Preparation Guide (10x Genomics, CG00053 Rev C). Cells were then separated into droplet emulsions using the Chromium Controller (10x Genomics) with Chromium Single-Cell 3′ Gel Beads v3 (10x Genomics, PN-1000075 and PN-1000153) across 273 “lanes”/”GEM groups” following the 10x Genomics Chromium Single Cell 3' Reagent Kits v3 User Guide with Feature Barcode technology for CRISPR Screening (CG000184 Rev C) with the goal of recovering ~15,000 cells per GEM group before filtering. Because the formation of droplet emulsions occurred in batches of 8 GEM groups over several hours, fresh populations of cells were obtained every hour to prevent alterations in single-cell transcriptomes.

To perform our K562 day 6 essential-scale Perturb-seq experiment, library lentivirus was packaged into lentivirus in 293T cells and empirically measured in K562 cells to obtain viral titers. CRISPRi K562 cells were transduced via spinfection (1000g) with polybrene (8 ug/ml) with the target of obtaining an infection rate of ~10% with maintenance of cells as described above. Three days post-transduction, an infection rate of 15% was measured, and cells were sorted to near purity by FACS (FACSAria2, BD Biosciences), using GFP as a marker for sgRNA vector transduction. Six days post infection, the cells were measured to be 93% GFP+ (LSR2, BD Biosciences), >90% viable, and at a concentration of ~600,000 cells/ml (Countess II, ThermoFisher). Cells were prepared for single-cell RNA-sequencing by resuspension in 1X PBS with 0.04% BSA as detailed in the 10x Genomics Single Cell Protocols Cell Preparation Guide (10x Genomics, CG00053 Rev C). Cells were then separated into droplet emulsions using the

Chromium Controller (10x Genomics) with Chromium Single-Cell 3′ Gel Beads v3 (10x Genomics, PN-1000075 and PN-1000153) across 48 “lanes”/”GEM groups” following the 10x Genomics Chromium Single Cell 3' Reagent Kits v3 User Guide with Feature Barcode technology for CRISPR Screening (CG000184 Rev C) with the goal of recovering ~15,000 cells per GEM group before filtering.

To perform our RPE1 day 7 essential-scale Perturb-seq experiment, library lentivirus was packaged into lentivirus in 293T cells and empirically measured in RPE1 cells to obtain viral titers. CRISPRi RPE1 cells expressing ZIM3 KRAB-dCas9-P2A-BFP were transduced via replating into virus-laden media with polybrene (8 ug/ml) with the target of obtaining an infection rate of ~10%. Three days post-transduction, an infection rate of 7% was measured, and cells were sorted to near purity by FACS (FACSAria2, BD Biosciences), using GFP as a marker for sgRNA vector transduction. Seven days post infection, the cells were measured to be 86% GFP+ (LSR2, BD Biosciences) and >95% viable (Countess II, ThermoFisher). After trypsin dissociation, cells were prepared for single-cell RNA-sequencing by resuspension in 1X PBS with 0.04% BSA as detailed in the 10x Genomics Single Cell Protocols Cell Preparation Guide (10x Genomics, CG00053 Rev C). Cells were then separated into droplet emulsions using the Chromium Controller (10x Genomics) with Chromium Single-Cell 3′ Gel Beads v3 (10x Genomics, PN-1000075 and PN-1000153) across 56 “lanes”/”GEM groups” following the 10x Genomics Chromium Single Cell 3' Reagent Kits v3 User Guide with Feature Barcode technology for CRISPR Screening (CG000184 Rev C) with the goal of recovering ~15,000 cells per GEM group before filtering.

Perturb-seq library preparation and sequencing For preparation of gene expression and sgRNA libraries, samples were processed according to 10x Genomics Chromium Single Cell 3' Reagent Kits v3 User Guide with Feature Barcode technology for CRISPR Screening (CG000184 Rev C). To allow for parallel library preparation, samples were arranged in 96-well plates with magnetic selections conducted on an Alpaqua Catalyst 96 plate (#A000550). For sequencing,

mRNA and sgRNA libraries were pooled to avoid index collisions at a 10:1 ratio. Libraries were sequenced on a NovaSeq 6000 (Illumina) according to the 10x Genomics User Guide.

rRNA analyses K562s expressing Zim3-dCas9-2A-BFP were spinfected in biological duplicate (targeting sgRNAs) or quadruplicate (non-targeting sgRNAs) with lentivirus expressing GFP and an sgRNA. Two days after spinfection, the cells were sorted for GFP+ on a BD ARIA II. Sort purity was generally >95%. After the sort, cells were maintained in media supplemented with 4 ug/ml puromycin for four days and then recovered for two days. Cells were counted, collected by centrifugation, and harvested by vigorous vortexing in Tri Reagent (ThermoFisher AM9738).

RNA was extracted with chloroform according to the manufacturer's instructions, quantified by nanodrop, and snap frozen. Small samples were diluted to 200 ng/ul and run on Bioanalyzer RNA nano chips (Agilent 5067-1511) according to the manufacturer's instructions. Runs were aligned to the 18s peak and signal intensity was normalized to total RNA area.

Integrator co-depletion K562s expressing Zim3-dCas9-2A-BFP were spinfected with lentivirus expressing GFP and an sgRNA. Two days after spinfection, the cells were sorted for GFP+ on a BD ARIA II. Sort purity was generally >95%. After the sort, cells were maintained in media supplemented with 4 ug/ml puromycin for four days and then recovered for two days. Cells were counted, washed twice with DPBS, and collected as pellets. The pellets were resuspended in SDS lysis buffer (100 mM Tris pH 8.0, 1% SDS), thermomixed at 95˚/1500 RPM for thirty minutes, aliquoted, and snap-frozen.

Quantification for western blots. An equal amount of material was loaded, as assessed by lysate A280.

Integrator co-immunoprecipitation Human expression plasmids encoding codon-optimized INTS10 or His8-INTS10 were synthesized (Twist Bioscience) and transfected into HEK 293T/17

cells (ATCC CRL-11268) with FuGene HD (Promega E2311) according to the manufacturer's protocol. Two days later, the cells were washed twice with DPBS and harvested with IP lysis buffer (25 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% NP-40, 5% glycerol; ThermoFisher 87787) supplemented with protease inhibitors (ThermoFisher A32965). Lysates were nutated at 4˚ for 30 mins, clarified by centrifugation at 12,000xg for 10 minutes, and snap-frozen. Concentrations were measured with the BCA assay (ThermoFisher 23225).

Lysates were thawed on ice, supplemented with imidazole to 10 mM, and nutated at 4˚ for 30 minutes with cobalt magnetic beads (ThermoFisher 10103D) pre-equilibrated in IP lysis buffer + 10 mM imidazole. The beads were separated on a magnet, washed twice with lysis buffer + 10 mM imidazole, and eluted with lysis buffer + 300 mM imidazole.

Quantification for western blots. For input samples, an equal amount of material was loaded, as assessed by BCA. For IP samples, an equal volume of eluate was loaded.

Integrator purification Human expression plasmids encoding codon-optimized HIS-INTS10, INTS13, INTS14, and C7orf26 were synthesized (Twist Bioscience) and co-transfected with ExpiFectamine 293 (ThermoFisher A14524) into Expi293 cells (ThermoFisher A14527) maintained in Expi293 medium (ThermoFisher A1435101) according to the manufacturer's instructions. The cells were harvested after four days and snap frozen.

The pellets were resuspended in CHAPS Lysis Buffer (50 mM HEPES pH 8.0, 300 mM NaCl, 0.2% CHAPS, 10% glycerol, 1 mM TCEP, 1 mM EDTA, 0.5 mM PMSF, 1x protease inhibitors, 0.002% benzonase) and stirred at 4˚ for 30 minutes. The lysates were clarified at 120,000xg for 30 minutes, supplemented with 15 mM imidazole, and nutated for an hour with Ni-NTA agarose beads (ThermoFisher 25215) pre-equilibrated in CHAPS lysis buffer + 15 mM imidazole. The beads were loaded into a gravity column, washed with >10 volumes of wash buffer (50 mM HEPES pH 8.0, 300 mM NaCl, 10% glycerol, 1 mM TCEP, 1 mM EDTA, 15 mM imidazole), and eluted with wash buffer supplemented with 250 mM imidazole. The eluate was

concentrated and buffer exchanged into SEC buffer (50 mM HEPES pH 8.0, 150 mM KCl, 10% glycerol, 1 mM EDTA) by ultrafiltration, and snap frozen.

The eluate was thawed on ice, passed through a 0.2 μM PES filter, and loaded onto an Superdex 200 Increase 10/300 GL column pre-equilibrated with SEC buffer. Fractions were collected and flash frozen.

Quantification for gels and western blots. An equal volume of sample from each SEC fraction was loaded. Less Ni-NTA eluate was loaded to account for the dilution over SEC.

Integrator PRO-seq Pro-seq was conducted largely according to published protocols with slight modifications. K562s expressing dCas9-BFP-KRAB were spinfected with lentivirus expressing GFP and an sgRNA. Two days after spinfection, the cells were sorted for GFP+ on a BD ARIA II. Sort purity was generally >95%. After the sort, cells were maintained in media supplemented with 4 ug/ml puromycin for three days and then recovered for two days.

Cells were counted, harvested by centrifugation, and washed with cold DPBS. All subsequent steps took place at 4˚ or on ice. All solutions were made with RNase-free reagents and were 0.2 μm filtered and chilled before use. 12 million cells were pelleted by centrifugation, resuspended in 250ul of buffer W (10 mM Tris pH 8.0, 10 mM KCl, 250 mM sucrose, 5 mM MgCl2, 1 mM EGTA, 0.5 mM DTT, 10% glycerol, 1x protease inhibitor [ThermoFisher A32965], and 0.02% v/v SUPERase-In RNase inhibitor [AM2694], strained, and transferred to conical tubes that had been coated with 1% BSA in PBS overnight. The cells were permeabilized by dilution in 10 ml of buffer P (buffer W + 0.1% v/v Igepal CA-630 + 0.05% v/v Tween-20) and incubated for 5 minutes. The permeabilized cells were harvested by centrifugation at 400xg for 5 minutes, resuspended in 10 mL of buffer W, harvested by centrifugation at 400xg for 5 minutes, and resuspended in 250 ul buffer F (50 mM Tris pH 8.0, 40% v/v glycerol, 5 mM MgCl2, 1.1 mM EDTA, 0.5 mM DTT, 0.02% v/v SUPERase-In RNase inhibitor). ≥97% permeabilization efficiency was confirmed on a NucleoCounter NC-202 and permeabilized cells were snap frozen.

PRO-seq libraries were generated and sequenced by the Nascent Transcriptomics Core at Harvard Medical School according to their standard protocol. PRO-seq data were aligned and quantified using STAR (version 2.7.9a) with parameters alignEndsType=Local, outFilterMultimapNmax=20, outFilterScoreMinOverLread=0.3, and outFilterMatchNminOverLread=0.3. For comparison of gene-level expression profiles, gene counts were normalized for sequencing depth (reads per million), log-transformed, and subset to well-expressed genes (n=758 genes with >3000 rpm). Then, Spearman's correlation was used to compare the similarity of expression profiles.

SDS-PAGE and western blotting Samples were mixed with sample loading buffer (Licor 928-40004) supplemented with DTT and incubated at 95˚ for 5 minutes. SDS-PAGE was performed with pre-cast 4-12% gradient gels (ThermoFisher NW04127BOX) in MOPS (ThermoFisher B000102) according to the manufacturer's instructions.

For Coomassie staining, gels were washed thoroughly in water, incubated with ReadyBlue Protein Gel Stain (Sigma RSB-1L) overnight, and destained in water. For western blots, proteins were transferred to nitrocellulose membranes by semi-dry transfer (Biorad 1704158) according to the manufacturer's instructions. Membranes were rinsed in water and stained with Revert 700 Total Protein Stain according to the manufacturer's instructions. The membranes were then rinsed in TBS, rocked with Everyblot Blocking Buffer (Biorad 12010020) at room temperature for > 30 minutes, and rocked with primary antibody overnight at 4˚. The membranes were washed with TBST, and rocked with IR800CW-labeled secondary antibodies for 30-60 minutes, washed with TBST, and imaged on a Licor Odyssey CLx.

CD11b cell surface staining K562s expressing dCas9-BFP-KRAB were co-spinfected with lentiviruses expressing GFP-sgKDM1A and mCherry-sgPTPN1. Eight days after spinfection, the cells were counted and harvested by centrifugation. Cells were washed with PBE buffer (DPBS +

0.5% BSA + 2 mM EDTA) and resuspended with α-CD11b-AF647 antibody diluted 1:50 in PBE. Cells were incubated at 4˚ in the dark for 30 minutes, washed twice with PBE, and analyzed on a BD LSRFortessa. The populations were gated from a single sample as sgKDM1A (GFP+,mCherry-), sgPTPN1 (GFP-,mCherry+), and sgKDM1A/sgPTPN1 (GFP+,mCherry+). Unstained K562s expressing either GFP or mCherry were used as single color compensation controls. AF647 was compensated with UltraComp eBeads Plus (Thermo 01-3333-42) labeled with α-CD11b-AF647.

Internally controlled growth assays K562s expressing Zim3-dCas9-2A-BFP were co-spinfected in triplicate with lentiviruses expressing GFP-sgKDM1A and mCherry-sgPTPN1, or with lentivirus expressing GFP and a non-targeting sgRNA. Every two days, cells were analyzed for BFP, GFP, and mCherry on an Attune flow cytometer. Enrichment was calculated as sgKDM1A (BFP+,GFP+,mCherry-), sgPTPN1 (BFP+,GFP-,mCherry+), and sgKDM1A/sgPTPN1 (BFP+,GFP+,mCherry+) vs uninfected (BFP+,GFP-,mCherry-).

Bulk RNA-seq K562s expressing dCas9-BFP-KRAB were spinfected in biological duplicate with lentivirus expressing GFP and an sgRNA. Two days after spinfection, the cells were sorted for GFP+ on a BD ARIA II. Sort purity was generally >95%. After the sort, cells were maintained in media supplemented with 4 ug/ml puromycin for four days and then recovered for two days. Cells were counted, collected by centrifugation, and harvested by vigorous vortexing in Qiazol (Qiagen 79306).

Total RNA was extracted with miRNeasy Mini columns (Qiagen 217004) according to the manufacturer's instructions and sequencing libraries were prepared with TruSeq Stranded Total RNA Library Prep Human/Mouse/Rat kits (Illumina 20020596) according to the manufacturer’s instructions. Libraries were sequenced 2x150 on a NovaSeq (Illumina).

Bulk RNA-seq data were aligned and quantified using STAR (version 2.7.9a) with parameters alignEndsType=Local and outFilterMultimapNmax=20. For comparison of gene-level expression profiles, gene counts were corrected for sequencing depth (reads per million), and the $\log_2$ fold-change for each gene was calculated relative to within-replicate non-targeting control expression. The two replicates for each genetic perturbation were averaged in order to produce the final data.

Seahorse experiment K562s expressing dCas9-BFP-KRAB were spinfected with lentivirus expressing GFP and an sgRNA. Two days after spinfection, the cells were sorted for GFP+ on a BD ARIA II. Sort purity was generally >95%. After the sort, cells were maintained in media supplemented with 4 ug/ml puromycin for four days and then recovered for three days. On the 9th day post spinfection, seahorse assay were plates were treated with Cell-Tak (Corning 354240) according to the manufacturer's instructions. Cells were counted, collected by centrifugation, and resuspended in supplemented Seahorse XF RPMI (Agilent 103576-100). 150,000 cells were added to the Seahorse assay plate and attached via centrifugation at 200xg for 1 minute with no brake. After 30 minutes of recovery at 37˚, the cells were subjected to a Mito Stress Test on a Seahorse XFe96 analyzer according to the manufacturer's instructions.

Computational Methods

Alignment, cell calling, and guide assignment Cell Ranger 4.0.0 software (10x Genomics) was used for alignment of scRNA-seq reads to the transcriptome, alignment of sgRNA reads to the library, collapsing reads to UMI counts, and cell calling. The 10x Genomics GRCh38 version 2020-A genome build was used as a reference transcriptome. For specific applications discussed below, STARsolo (STAR version 2.7.9a) was used extract transcript features, including intronic and exonic alignments and alignment of reads to transposable elements.

Reads from the sgRNA libraries were mapped with Cell Ranger. To account for differences in sequencing depths across GEM groups from the same experiment, reads were downsampled to produce a more even distribution of the number of reads per cell across gemgroups, with a threshold of 1000 reads per cell in the K562 day 8 experiment, 800 reads per cell in the K562 day 6 experiment, and 3000 reads per cell in the RPE1 experiment. Guide calling was performed with a Poisson-Gaussian mixture model as previously described. For each guide, the mixture model was fit 100 times, selecting the maximum likelihood model from among the fits. After guide calling, each cell was categorized according to its guide identities as representing a a single genetic perturbation or a multiplet (which may arise from lentiviral recombination or multiple cell encapsulation during droplet generation). Only cells bearing two guides targeting the same gene or a single guide were used for downstream analysis.

Downstream analyses were performed in Python, using a combination of numpy, scipy, Pandas, scikit-learn, pomegranate, infercnvpy, pygenometracks, scanpy and seaborn libraries.

Filtering and internal normalization of gene expression measurements Our internal normalization approach is similar to the one described in (Adamson et al., 2016)

First, we identified “core” control sgRNAs. That is, within each experiment there are tens to hundreds of possible negative control sgRNAs that were synthesized to have similar base compositions to targeting sgRNAs (Horlbeck et al., 2016). Some of these by chance induce detectable phenotypes. We constructed a minimal set of control sgRNAs that are largely indistinguishable from each other using the following procedure: (1) We take all cells bearing all possible non-targeting sgRNAs and represent them by the vector of genes with mean >1 UMI count per cell. (2) We *z*-normalize the expression of these genes: i.e. we subtract the mean and divide by the standard deviation. (3) For each gene, we test for equality of distribution using the Anderson-Darling test (scipy.stats.anderson_ksamp) between all possible pairs of non-targeting control sgRNAs. (E.g. In the genome-scale dataset, there are 585 possible control sgRNAs and

therefore $\binom{585}{2} = 170280$ pairwise comparisons.) (4) We adjust the resulting *p*-values for multiple hypothesis testing using the Benjamini-Hochberg procedure. (5) For each potential control sgRNA, we compute the average number of differentially expressed genes relative to all other potential control sgRNAs. (6) We set a dataset-dependent threshold on the number of differentially expressed genes (8 in the genome-scale dataset and 30 in the "K562 essentials" and "RPE1 essentials" datasets, which were more deeply sequenced and so had more genes passing the expression threshold) and kept all potential control sgRNAs that fell below the threshold. E.g. in the genome-scale dataset this resulted in 514 control sgRNAs.

Next, we filtered cells based on quality metrics. We first computed scale factors to adjust for variable sequencing depths across gemgroups: we examined all core control cells (which make up ~4% of all cells), computed factors that equalized the mean UMI counts within these cells across gemgroups, and then applied these factors to all cells in the gemgroup to produce adjusted UMI counts. We then applied two quality filters, ensuring that cells passed a minimum adjusted UMI content filter (genome-scale dataset: 2000 UMIs, K562 essentials/RPE1 essentials: 3000 UMIs) and a maximum mitochondrial RNA filter (genome-scale dataset: <25%, K562 essentials: <20%, RPE1 essentials: <11%). (Mitochondrial RNA content is the fraction of total UMIs derived from mitochondrially-encoded genes.) These filter parameters were chosen by plotting adjusted UMI content vs. mitochondrial RNA content and setting thresholds that captured the (obvious) mode containing most cells.

Finally, we computed a normalized gene expression matrix for cells passing the quality filters via two normalization steps: (1) *UMI count normalization*: We scale expression within all cells so that their total UMI counts equal the median UMI count of core control cells within the experiment. (2) *Relative z-normalization*: Within each gemgroup, for each gene, we compute the mean and standard deviation of expression within control cells and use these to *z*-normalize

expression. (I.e. if x is the expression of a given gene, it is represented by the score $z = (x - \mu_{\text{control}})/\sigma_{\text{control}}$, where the mean and standard deviation are separately computed within each gemgroup.) The resulting scores should therefore be interpreted as "fraction of transcriptional effort" due to the UMI count normalization, with a scale set relative to control cells. (E.g. An expression score of +2 thus represents a gene expressed at a level 2 standard deviations above the mean in control cells.)

Examining effects of normalization on batch effects As described in the main text, we observed batch effects in the data (**Supplementary Figure 2.2**). This variation appeared to track mostly with sets of 8 samples that went through the 10x Chromium instrument and library prep together, though the precise origin is unclear. To construct this figure, we normalized the data in two ways. *Raw data normalization:* (1) To adjust for variable sequencing depth, scale cellular UMI counts by factors chosen so that so that core control cells have the same total UMI counts across all gemgroups. (2) Construct a gemgroup mean expression profile of all genes with mean >2 UMI counts per cell by averaging counts following normalization in previous step across all cells in the gemgroup. (3) Scale the gemgroup mean expression profiles by dividing by their mean across all gemgroups. An expression value of 1 is then the mean across all cells across all gemgroups. *Internal z-normalization*. Normalize expression as described in previous section.

Supplementary Figure 2.2 compares the two normalization schemes. In both cases the ranges of the plot are chosen using seaborn's robust option (which sets the min and max to the 2nd and 98th percentile of the data). Genes are clustered based on the raw data normalization and are in the same order in both panels. The gemgroups are presented in order based on how samples were multiplexed while performing the experiment as indicated by the color groupings at the top.

Energy distance test for identifying perturbations that induce altered transcriptional states

To compare distributions of expression states, we used tests derived from energy statistics, which allow for testing of equality of distributions when data are high-dimensional. In short, each cell is represented by a vector composed of its top 20 principal component scores, and we compare whether the distribution of these 20-dimensional vectors is equal or not between unperturbed control cells and cells bearing each perturbation. When these distributions differ, we can infer that the perturbation is causing some change either in the structure or distribution of transcriptional states within the perturbed cells.

To construct the distributions to compare, we first applied a series of filtering steps: (1) we removed cells that did not pass the UMI or mitochondrial RNA filters described in *Internal normalization of gene expression measurements*; (2) as features, we took the *z*-normalized expression of all genes with mean expression >0.5 UMIs per cell; (3) to dampen the effects of a handful of strongly induced outlier genes, we clipped any measure with a *z*-score greater than 10 to 10 (this only affects a handful of genes); (4) finally, we applied principal components analysis (using sklearn's PCA implementation, which will use randomized algorithms for datasets of this scale) and kept only the top 20 principal components. The test should therefore be interpreted as assessing gross changes in cellular transcriptional state.

To construct a null distribution, we randomly subsampled 5,000 control cells bearing non-targeting sgRNAs. (Subsampling was necessary for performance reasons.) For each perturbation, we then compute an estimator of the energy distance:

$$\mathcal{E}(x,y) = \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|\boldsymbol{x_i}-\boldsymbol{y_j}\| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|\boldsymbol{x_i}-\boldsymbol{x_j}\| - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|\boldsymbol{y_i}-\boldsymbol{y_j}\|$$

where each $\boldsymbol{x_i}$ is one of the control cells and each $\boldsymbol{y_j}$ is one of the perturbed cells.

In the limit of infinite data, the energy distance will be 0 between identical distributions and positive between non-identical distributions. We assess statistical significance in practice using a

permutation test by permuting the labels of control and perturbed cells 10,000 times and estimating how frequently a larger energy distance would be observed by chance. The specific implementation is based on the python package torch-two-sample, modified to use numba for improved performance.

Gene-level differential expression testing using the Anderson-Darling and Mann-Whitney tests Because of biological differences in expression characteristics across different genes, the batch effects described above, incomplete penetrance of some perturbations, and heterogeneity of some gene expression programs, we opted to use non-parametric statistical tests rather than tests based on specific distributional assumptions about gene expression. Specifically, we *z*-normalize gene expression relative to control cells as described (*Internal normalization of gene expression measurements*) and for each gene test whether the distribution of normalized expression is identical between control cells bearing non-targeting sgRNAs and cells bearing each perturbation. We used two tests implemented in scipy: the Anderson-Darling test (scipy.stats.anderson_ksamp), which is broadly sensitive to changes in distribution, and the Mann-Whitney U test (scipy.stats.mannwhitneyu), which tests whether one distribution is stochastically greater than another. For the Anderson-Darling test we extended the range of *p* values beyond those available in scipy's implementation by computing the *p*-value for many values of the test statistic using R's kSamples package and interpolating any intermediate values using scipy.interpolate.interp1d. For the Mann-Whitney test we used the asymptotic *p* values and excluded any perturbation with fewer than 10 cells. *p*-values in both cases were adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure to produce the final results.

Leverage scores for quantifying perturbation penetrance and variability Our use of non-parametric tests in differential expression testing is in part to accommodate perturbations that may be incompletely penetrant or heterogeneous in effect. To attempt to quantify these features,

we developed a scalar single-cell score to summarize how outlying each cell's transcriptional state was relative to control cells. We used an approach based on leverage scores, which measure how outlying the rows or columns of a matrix are (Ma et al., 2013). Specifically, we: (1) Construct an expression matrix consisting of all cells that pass the quality filters described in *Internal normalization of gene expression measurements*, and all genes with mean expression >0.25 UMI counts per cell. (2) To dampen the effects of a handful of strongly induced outlier genes, we clipped any measure with a *z*-score greater than 10 to 10 (this only affects a handful of genes) (3) To avoid the influence of gemgroup-level batch effects and variable sequencing depth, we then compute leverage scores separately within each gemgroup. Row leverage scores, corresponding to the cell axis of the expression matrix, are calculated as the squared norm of the top 20 left singular vectors within each gemgroup (computed via the truncated SVD routine scipy.sparse.svds with the arpack solver with k=20). We then normalize these scores so that the sum over all cells in the gemgroup is 1 (i.e. compute the leverage sampling probability distribution). (4) Finally, to integrate leverage scores across gemgroups, we then take logs, and rescale by *z*-normalizing relative to the scores of control cells (subtracting their mean and dividing by their standard deviation). All the leverage scores presented in the figures are the leverage scores after this normalization procedure.

In Supplementary Figure 2.5 we conducted various analyses to validate leverage scores as measures of phenotype and to use them to study penetrance of perturbations. We considered all perturbations that passed the following criteria: (1) >5 differentially expressed genes by Anderson-Darling test; (2) detected in at least 25 cells that passed our quality filters; and (3) the gene targeted by the perturbation was either undetectable in the expression data or knocked down by at least 30% if detected (i.e. we removed perturbations that appear non-functional). For these perturbations we compared the mean leverage scores to the number of differentially expressed genes found using the Anderson-Darling test. To assess reproducibility we then subset

to perturbations that were present in both the genome-scale dataset and the K562 essentials dataset, including non-targeting controls.

In the analyses of perturbations targeting Mediator and the small subunit of the ribosome we only included perturbations that were (1) present in both the genome-scale dataset and the "K562 essentials" dataset and (2) targeted the principal "P1" transcript identified by the FANTOM consortium. (A handful of genes also had perturbations targeting the P2 transcript that did not generally have effects.) Knockdown was computed as the ratio of mean (unnormalized) expression of the target gene within perturbed cells vs. that in cells with non-targeting sgRNAs. The plots show kernel density estimates of the distributions of the leverage scores of all cells with these perturbations constructed using seaborn's violinplot (with cut set to 0 so that estimated distributions do not extend beyond the range of the data). The gray bars represent the 10%-90% quantiles of cells with non-targeting control sgRNAs for comparison.

Finally, in Figure 2.5B we used leverage scores to search for perturbations that had highly variable phenotypes. We considered all perturbations using the same criteria as in Supplementary Figure 2.5. We used the standard deviation of the leverage scores as a metric for variability, as diagrammed in Figure 5A. The two examples in this figure are derived from actual data. The 20 labeled genes are the most outlying from the lowess local regression between the standard deviation of leverage scores and the log of the number of differentially expressed genes detected by the Anderson-Darling test.

Global analysis and clustering of strong perturbations The analysis presented covers 1973 perturbations that met three criteria: (1) at least 50 differentially expressed genes at a significance of $p < 0.05$ by Anderson-Darling test following Benjamini-Hochberg correction; (2) at least 25 perturbed cells that passed our quality filters; and (3) an on-target knockdown, if measured, of at least 30% (i.e. the target of perturbation was either knocked down by at least 30% or was not detected, a broad attempt to remove non-functional perturbations). As features, we used a union

of two sets of genes: (1) the top 10 differentially expressed genes for all perturbations (ordered by the value of the Anderson-Darling test statistic) and (2) all genes of mean >0.25 UMIs per cell with variance in the top 30% of the dataset. We represented perturbations by their mean normalized expression profile across these 2319 highly variable genes. To prevent the direct targets of knockdown influencing results, the target gene value was replaced by 0 for the corresponding perturbation. (E.g. RPS5 gene normalized expression was set to 0 in the expression profile of the RPS5 perturbation, which is equal to the mean in control cells by construction.)

Because clearly related perturbations sometimes showed variable absolute phenotypic strengths, we used correlation as a metric to compare profiles, since it is scale-invariant. We conducted two global assessments of the ability of these expression profile correlations to recall known biological relationships. First, curated complexes were obtained from the 03.09.2018 CORUM3.0 database (Giurgiu et al., 2019). We identified all complexes that had at least 66% of genes represented within the 1973 perturbations (based on matching gene symbols between the datasets), leading to 327 complexes. Each represented complex was then split into a series of links (e.g. if a complex contained genes A, B, and C, then it would be split into links A-B, B-C, and A-C). The figure plots the distribution of expression profile correlations of these annotated links versus the distribution of all possible links among the 1973 targeted genes. A similar analysis was then conducted using predicted protein links from the v11.5 of STRING (Szklarczyk et al., 2019)(9606.protein.links.v11.5.txt.gz) after mapping STRING protein IDs to gene names (using the “preferred_name” field in 9606.protein.info.v11.5.txt.gz). Among the 1973 genes in the figure there are 1,945,378 possible pairwise links between genes, 243,558 of which have scores within STRING. We binned these represented links into 6 equally spaced bins based on observed expression profile correlation. The figure shows kernel density estimates of the STRING score distribution within each bin made using seaborn’s violinplot (with cut set to 0 so that density estimates do not extend past observed data).

To identify clusters of related perturbations, we manually computed correlation distances between all pairs of expression profiles, and used HDBSCAN (metric='precomputed', min_cluster_size=4, min_samples=1, cluster_selection_method='eom') to identify 63 clusters. This procedure is intrinsically conservative due to the choice of metric and clustering algorithm, so many perturbations are not assigned to any cluster—our emphasis was on identifying the strongest signals rather than the most comprehensive. We then annotated the possible function of these clusters using a combination of manual lookup of related genes and automated annotation using CORUM complexes and STRING clusters. (STRING clusters are derived from 9606.clusters.info.v11.5.txt.gz and are labeled according to the "best_described_by" field.) We only assigned automated annotations when a cluster contained 75% or more of the members of a CORUM complex or STRING cluster.

Minimum distortion embedding of strong perturbations The visualization in Figue 2.2D is a minimum distortion embedding (MDE) of the 1973 strong perturbations created using pymde v0.1.13. pyMDE solves MDE problems based on minimizing Euclidean distances. To adapt it to correlation distances, we first *z*-normalized each expression profile. (I.e. If a perturbation is represented by the vector $\boldsymbol{x}$ of 2319 highly variable genes, we computed $\hat{\boldsymbol{x}} = (\boldsymbol{x} - \langle \boldsymbol{x} \rangle)/\sigma_x$.) Because of the polarization identity, minimizing the Euclidean distance between these normalized profiles is equivalent to minimizing the (square root of) the correlation distance between the unnormalized profiles:

$$\|\hat{\boldsymbol{x}} - \hat{\boldsymbol{y}}\|^2 = \|\hat{\boldsymbol{x}}\|^2 + \|\hat{\boldsymbol{y}}\|^2 - 2(\hat{\boldsymbol{x}} \cdot \hat{\boldsymbol{y}}) = 2(1 - \mathrm{corr}(\boldsymbol{x}, \boldsymbol{y}))$$

where the second equality follows from the *z*-normalization and the scale-invariance of correlation.

We created two embeddings. First, we used pymde to embed the dataset into 20 dimensions. This "high-dimensional embedding" serves as an imputation step, as it distorts the geometry of the dataset so that clusters of related genes that may be driven by weaker overall

correlations are allowed to form. Proximity within this embedding was used to identify genes, CORUM complexes, and STRING clusters that were near to the HDBSCAN clusters called on the raw data (which were well-preserved in the embedding). To construct the embedding, we initialized pymde using the spectral embedding of the dataset (using sklearn's SpectralEmbedding with n_components=20, affinity='nearest_neighbors', n_neighbors=7, eigen_solver='arpack')), and then ran pymde's "preserve_neighbors" function function with embedding_dim=20, n_neighbors=7, and repulsive_fraction=5. pymde was run until convergence with a final average distortion of 0.0979 and final residual norm of 9.4e-06.

To produce the embedding in Figure 2.2D, we ran pymde with the same parameters but with the embedding dimension set to 2 (final average distortion 0.105, final residual norm 3.2e-06). The bold cluster labels in the figure correspond to the manual annotations mentioned in the previous section. A handful of changes were manually incorporated: (1) the cytochrome c-ubiquinol cluster was not detected by HDBSCAN, and was manually annotated (2) 4 clusters involving protein post-translational modifications (ubiquitination, sumoylation, acetylation, neddylation) were annotated with a single label of "post-translational modifications" (3) components of eIF3 split across two clusters that are next to each other in the embedding and are labeled as a single cluster (4) all clusters of unknown function were not labeled but are included in the supplemental tables. The complex labels come from CORUM or STRING. A complex/cluster label was placed if and only if 75% of the members of a complex or cluster were close to each other in the 20-dimensional pyMDE embedding ("close" meaning at or below the 5th percentile of all pairwise distances). Redundant/duplicated clusters were manually deduplicated. The locations of the labels on the figure were then adjusted for readability. Label location is a decent proxy for, but not an entirely accurate representation of, cluster and complex locations. The true embedding locations are in the supplemental tables.

Clustering of gene expression programs We next turned to identifying conserved gene expression programs using a similar pipeline applied to the transpose of the expression matrix from the previous sections. Initial HDBSCAN clustering on the raw data did not yield very many clusters, which we attributed to the broad range in gene expression program size and dynamic range. To attempt to equalize for these factors, we performed the clustering on a high-dimensional embedding of the data. Each gene was represented by its expression across the 1973 perturbations in Figure 2.2 and we masked the targets of knockdown as described there to avoid target gene knockdown influencing clustering. We used pyMDE with the same normalization as above to encourage genes with correlated expression to be placed nearby to each other (20 dimensions, n_neighbors=7 and repulsive_fraction=5, final average distortion 0.145, final residual norm 5.1e-06). We then identified clusters using HDBSCAN applied to the embedding (metric='euclidean', min_cluster_size=10, min_samples=10, cluster_selection_method='leaf'), producing 38 clusters. We performed similar analyses as in Figure 2.2 to annotate known CORUM complexes and STRING clusters. Cluster identities were then manually annotated using a combination of these automated annotations, manual gene searches, and gene set enrichment analyses conducted using Enrichr (Xie et al., 2021).

To produce Figure 2.3B, we averaged expression within the 64 perturbation clusters from Figure 2.2 across the genes within the 38 gene expression clusters. Each element in the heat map therefore represents an average over both multiple (related) perturbations and multiple (related) genes. The labels were manually selected to highlight interesting features.

Screens of gene expression programs To demonstrate the ability of Perturb-seq to conduct screens on aggregate phenotypes, we conducted two analyses to identify perturbations that strongly induced interesting expression programs. In the first comparison, we compared expression of genes associated with erythroid differentiation (gene expression cluster 15) to those associated with myeloid differentiation (gene expression cluster 21). We scored expression

programs by taking the mean normalized gene expression of all genes in the associated clusters. We computed scores for all perturbations in the genome-scale dataset that were detected in at least 25 cells, and then *z*-normalized these scores to make scales comparable. The figure has labels on the 15 most outlying genes across the two programs. We then conducted an identical analysis comparing expression of an unfolded protein response cluster (cluster 2) to an integrated stress response cluster (cluster 12).

Analysis of composite phenotypes: total RNA, fraction mtRNA, fraction TE, RNA splicing, chromosomal instability, and cell cycle Composite phenotypes integrate data from across the transcriptome to describe global cellular features. While some derive from simple metrics, others rely on extracting information from the transcriptome beyond gene expression levels. The number of UMIs aligned to the transcriptome (GRCh38 version 2020-A) was used to represent the total cellular RNA content. To calculate the fraction of mitochondrial RNA per cell (fraction mtRNA), the sum of the expression levels of the 13 mitochondrial genome protein-coding genes (MT-ND6, MT-ND1, MT-ND2, MT-ATP8, MT-ND4L, MT-ND5, MT-ND3, MT-CO1, MT-CO2, MT-ND4, MT-ATP6, MT-CO3, MT-CYB) was divided by the total cellular RNA content for each individual cell.

The scTE processing pipeline was used to quantify the expression of transposable elements in single cells. As transposable elements tend to be present in many degenerate copies throughout the genome, scTE allocates TE reads to TE metagenes rather than specific genomic positions. Reads were aligned to the genome using STARsolo (STAR version 2.7.9a) with the flags ‘ --outSAMattributes NH HI AS nM CR CY UR UY --soloFeatures Gene GeneFull SJ Velocyto --readFilesCommand zcat --outFilterMultimapNmax 100 --winAnchorMultimapNmax 100 --outMultimapperOrder Random --runRNGseed 777 --outSAMmultNmax 1’ to allow multimapping. To avoid incompatibilities in cell calling between STARsolo and Cell Ranger, the output of Cell Ranger cell calling was used to define the STARsolo cell barcode whitelist using ‘--soloCBwhitelist’. Next, aligned reads were allocated to genes and TEs using scTE. The flag ‘-o

nointron' was used to prevent the quantification of TEs in gene introns. From single-cell transcriptomes, quantification of all TEs in the classes LINE, SINE, LTR, DNA, and Retroposon based on RepeatMasker were extracted. To calculate the fraction of repetitive and transposable element RNA per cell (fraction TE), the sum of the expression level of these TEs was divided by the total cellular RNA content for each individual cell.

The alignments from STARsolo described above were also be used to quantify RNA splicing. Due to (i) the sparsity of single-cell data (ii) the relationship between the fraction of spliced reads and gene expression levels, gene-wise RNA splicing was quantified at the pseudobulk level. From the STARsolo Velocyto output, the levels of spliced and unspliced reads for each gene were averaged across all cells bearing each perturbation, ignoring ambiguous reads. Then, for each gene, the fraction of unspliced reads was calculated and divided by the mean fraction of unspliced reads for that gene across all non-targeting control perturbations.

The framework described in inferCNV and implemented as infercnvpy was used to detect evidence of chromosomal copy number changes. The raw single-cell gene expression matrix was first filtered to remove lowly expressed genes (<0.05 UMIs per cell across the population), normalized for total UMI content (using scanpy.pp.normalize_total with target_sum=1e6, exclude_highly_expressed=False, and max_fraction=0.05), and log-scaled (using scanpy.pp.log1p). Then we used infercnvpy to compute rolling average gene expression changes for windows of 100 genes with a dynamic threshold of 1.5 standard deviations for noise filtering (using infercnvpy.tl.infercnv). For each single-cell, this generated a vector of CIN values across the genome. To label cells with likely karyotypic abnormalities, unstable karyotypic cells were heuristically defined as having ≥1 chromosome with evidence of changes in chromosomal copy number (nonzero CIN values) for >80% of the chromosomal length. For genetic perturbations, the CIN score was calculated as the mean single-cell sum of squared CIN values, *z*-normalized relative to non-targeting control perturbations.

Previous approaches to cell cycle analyses in Perturb-seq have largely focused on supervised classification of cells into canonical cell cycle states. However, these approaches do not allow for aberrant cell cycle states sometimes generated by genetic perturbations. As a summary of single-cell cell cycle states, we performed a UMAP dimension reduction based on the expression n=199 known cell cycle genes (obtained from Seurat and Adamson Norman et al. 2016). From total UMI content normalized, log-scaled expression data, a neighborhood graph was computed (using scanpy.pp.neighbors with n_neighbors=30, method='umap', metric='correlation', and n_pcs=20) followed by UMAP embedding (using scanpy.tl.umap with default parameters). This UMAP revealed cells in canonical cell cycle stages when compared with other methods, but also naturally separated dying cells and putatively quiescent cells. Gates were drawn by manual inspection to approximately separate cells likely to be in S, G2/M, and G1/G0 cell cycle phases.

Perturbation-induced changes in transcriptional state (*CCNF* example) To create the figure, we took all cells (309) with sgRNAs targeting the *CCNF* gene that passed our UMI content and mitochondrial RNA content filters along with a spike-in of 618 randomly chosen control cells containing non-targeting sgRNAs. We represented each cell by the vector of its expression across the 178 differentially expressed genes detected by the Anderson-Darling test at $p < 0.05$. The plot shows a UMAP projection (metric='euclidean', random_state=100) of these cells, together with cell cycle phases computationally inferred from expression of cell-cycle regulated genes (Adamson et al., 2016).

FIGURES

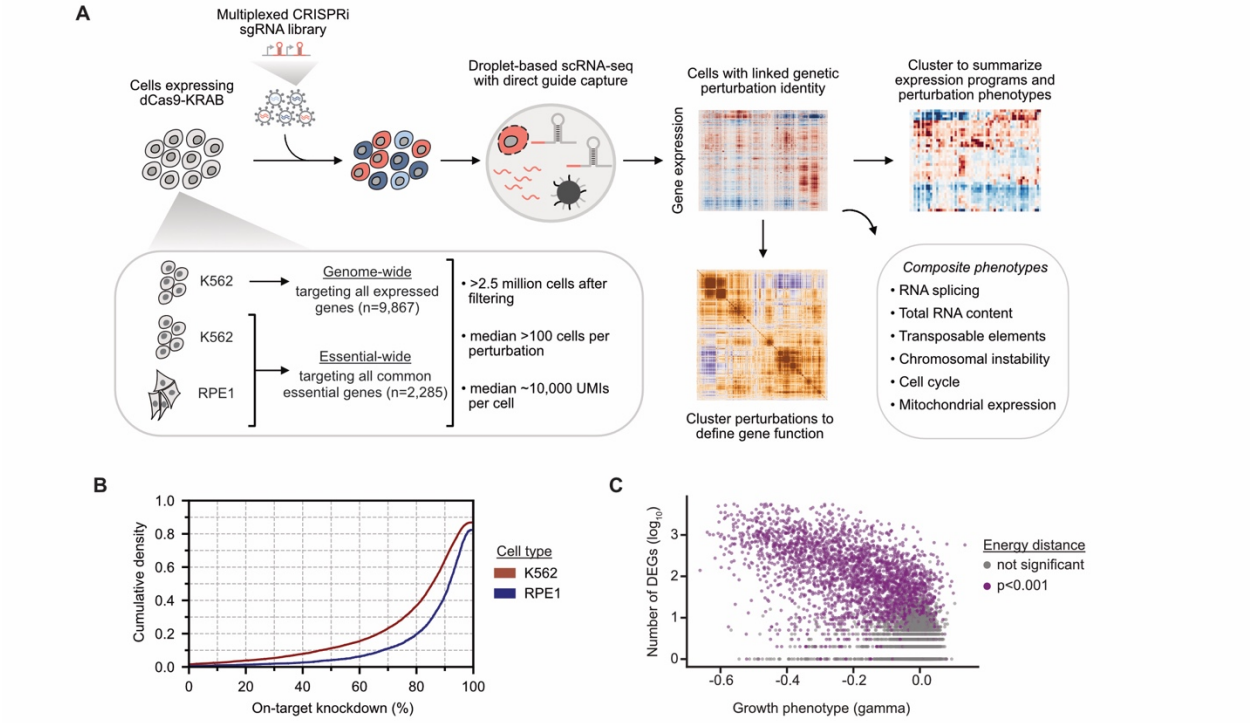


Figure 2.1: Genome-scale Perturb-seq via multiplexed CRISPRi.

A) Schematic experimental strategy. A multiplexed CRISPRi sgRNA library was used to knockdown all expressed genes (in K562 cells) or all common essential genes (in RPE1 and K562 cells). Cells were transcriptionally profiled using droplet-based single-cell RNA-sequencing, with genetic perturbations assigned to cells by direct capture and sequencing of sgRNAs.

B) On-target knockdown statistics. Cumulative density plot of on-target knockdown, for n=9,464 target genes in K562 cells (red) and n=2,333 target genes in RPE1 cells (blue).

C) Comparing growth phenotype versus the number of differentially expressed genes (DEGs) for each multiplexed guide pair in K562 cells. Growth phenotypes are reported as the $\log_2$ guide enrichment per cell doubling (gamma). DEGs were determined using a two-sample Anderson-Darling test compared against non-targeting guides. Dots are colored by Energy distance as either permutation significant (purple) or not significant (grey). The growth phenotype and number of DEGs are negatively correlated (Spearman's rho=-0.51, $p<10^{-16}$).

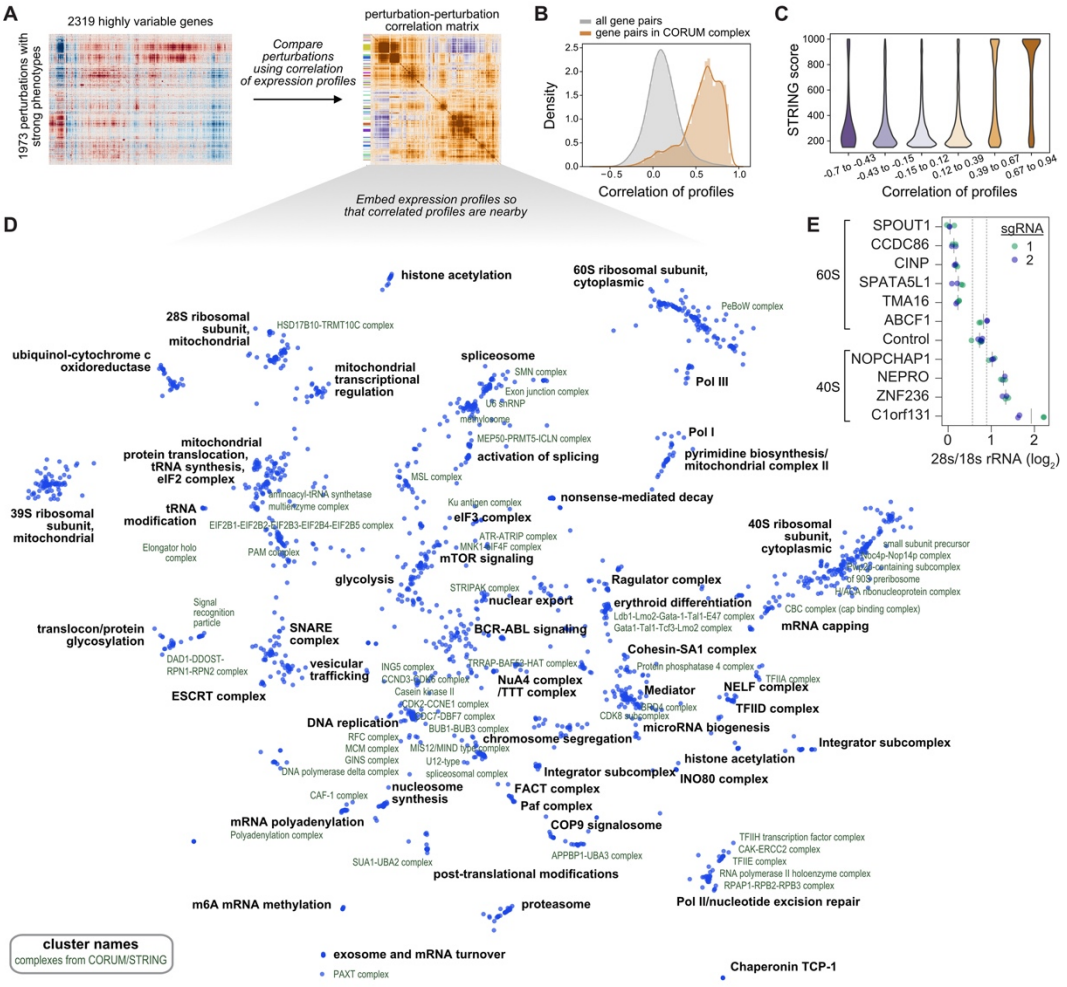


Figure 2.2: Data-driven inference of gene function from transcriptional phenotypes.

A) Schematic of analysis. To examine the ability of transcriptional phenotypes to assign gene function, we analyzed 1973 genetic perturbations that elicited strong responses. Perturbations were compared and clustered using the correlation of gene expression across 2319 highly variable genes.

B) Expression profile correlations among genes in curated complexes. 327 protein complexes from the CORUM3.0 database have at least two thirds of complex subunits within the dataset. Plot compares the distribution of pairwise expression profile correlations among genes in complexes vs. all possible gene-gene pairs.

C) Comparing expression profile correlations to predicted protein-protein interactions from STRING. 243,558 gene-gene relationships within the dataset are scored within STRING. The relationships were sorted into 6 equally spaced bins based on expression profile correlation. Plot shows kernel density estimates of STRING scores within each bin.

D) Minimum distortion embedding of dataset. Each dot represents a genetic perturbation, arranged so that perturbations with correlated expression profiles are nearby in the two dimensional embedding. Manual annotations (black labels) of cluster function are placed near the median location of genes within the cluster. CORUM complexes or STRING clusters (green labels) are annotated when involved genes are nearby within the embedding.

E) Quantification of 28s to 18s rRNA ratio. Poorly characterized genes with Perturb-seq predicted roles in ribosome biogenesis were targeted by CRISPRi. The 28s/18s rRNA ratio was measured by Bioanalyzer electrophoresis in biological duplicate with two distinct sgRNAs per gene (green and blue; solid grey lines represent mean). Dotted grey lines represent two standard deviations above and below the mean of non-targeting controls.

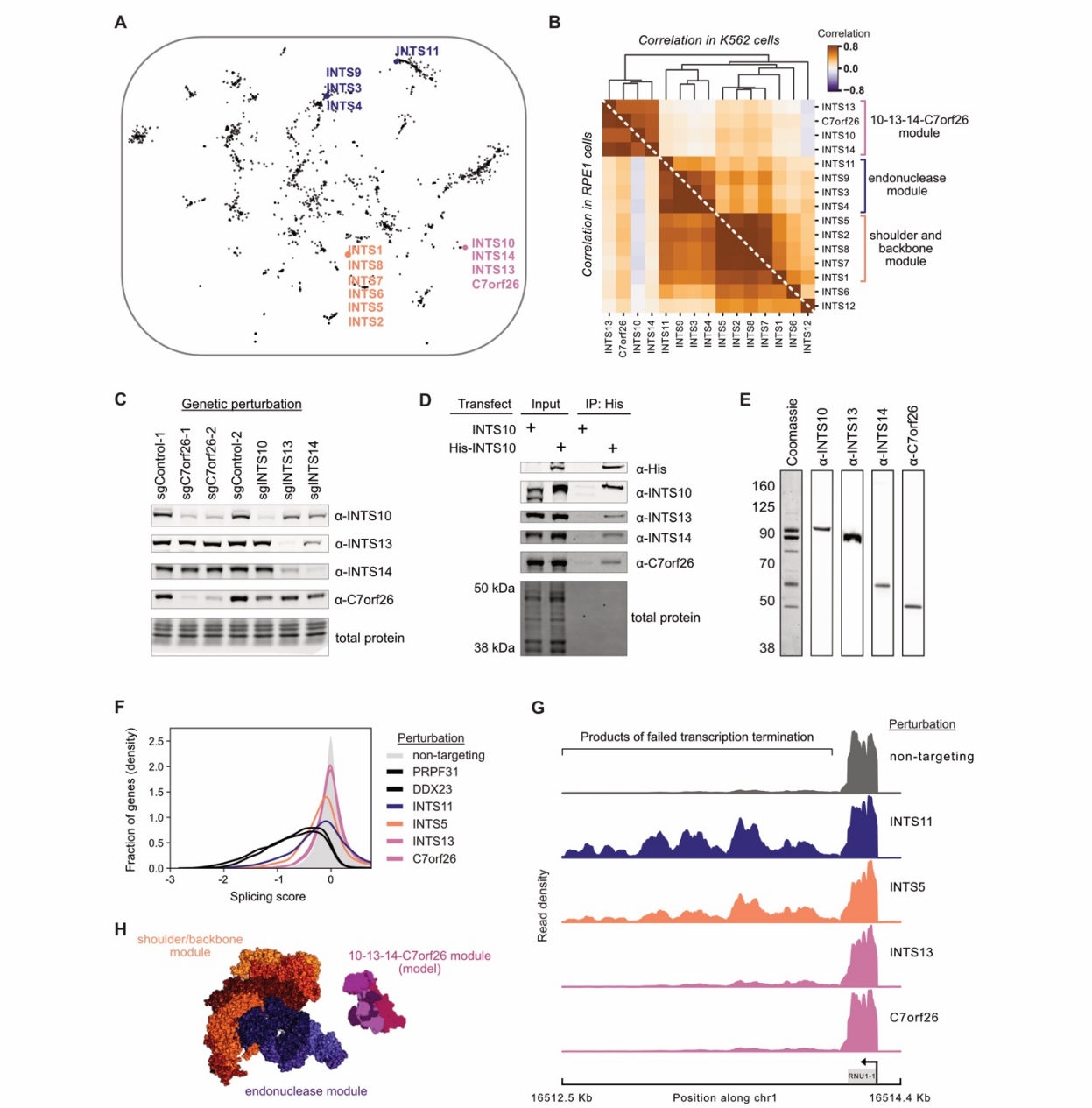


Figure 2.3: Perturb-seq discovers a novel gene member and functional submodules of the Integrator complex.

A) Location of known Integrator complex members in the minimum distortion embedding.

B) Relationship between Integrator complex members and C7orf26 in K562 cells (top) and RPE1 cells (bottom). The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of Integrator complex members. Genetic perturbations are ordered by average linkage hierarchical clustering based on correlation in K562 cells. Functional modules suggested by the clustering are highlighted.

C) Co-depletion of Integrator complex members. Individual Integrator complex members were depleted in CRISPRi K562 cells. Lysates were then probed for other module members by western blot.

D) Co-immunoprecipitation of endogenous C7orf26 with His-INTS10. HEK293T were transfected with His-INTS10 or INTS10. Cell lysates were affinity purified and select Integrator proteins were probed by western blot.
E) Purification of a INTS10-INTS13-INTS14-C7orf26 complex. His-INTS10, INTS13, INTS14, and C7orf26 were overexpressed in Expi293 cells, affinity purified, and separated via SEC. The INTS10-INTS13-INTS14-C7orf26 proteins co-fractionated as visualized by Western blotting.
F) Effects of Integrator modules on splicing from Perturb-seq data. Histogram (kernel density estimate) compares gene-level splicing scores. Splicing scores represent the change in the $\log_2$ ratio of total to unspliced reads for each gene relative to non-targeting control guides. Representative genetic perturbations from Integrator modules as well as the spliceosome are shown colored by module.
G) Density of PRO-seq reads at the snRNA RNU1-1 locus mapping actively engaged RNA polymerase II. For each perturbation, densities are shown relative to the maximum read count in the locus.
H) Structure of the Integrator complex colored by functional modules revealed by Perturb-seq. The endonuclease (blue) and shoulder/backbone (orange) modules were obtained from the cryo-EM structure (Zheng et al., 2020). The model of the newly discovered 10-13-14-C7orf26 module was built by docking the crystal structure of INTS13-INTS14 (Sabath et al., 2020) with an AlphaFold multimeric model of INTS10 and C7orf26.

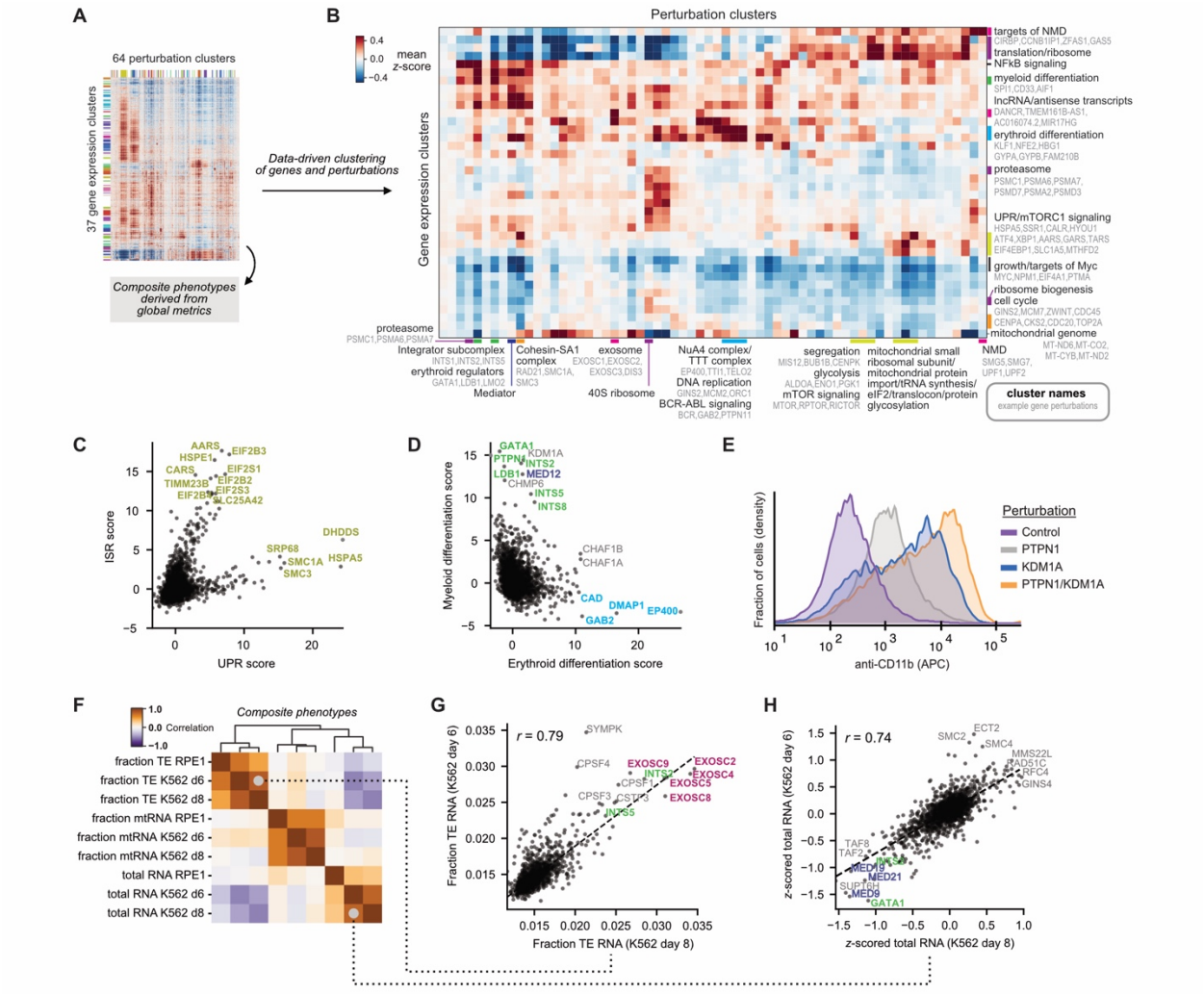


Figure 2.4: Summarizing genotype-phenotype relationships with Perturb-seq.

A) Schematic of analysis. To produce a high-level summary of genotype-phenotype relationships in K562 cells, 1973 genetic perturbations that elicited strong responses and 2319 highly variable genes were clustered using HDBSCAN after nonlinear embedding. Alternatively, composite phenotypes were derived from global metrics in a hypothesis-driven manner.

B) Heatmap of the high-level genotype-phenotype map. The heatmap represents the mean z-scored expression for gene expression and perturbation clusters. For a subset of clusters, clustered are labelled with manual annotations (black labels) of cluster function along with example genes within the cluster (light gray labels).

C) Comparison of ISR and UPR scores for genetic perturbations. Scores were recovered from unbiased clustering of genes by genetic dependency, and manually annotated.

D) Comparison of erythroid and myeloid differentiation scores for genetic perturbations. Scores were recovered from unbiased clustering of genes by genetic dependency, and manually annotated. Genetic perturbations are colored to reflect cluster identity.

E) Expression of CD11b/ITGAM in K562 cells upon knockdown of PTPN1 or KDM1A. CD11b was labelled by cell surface staining with anti-CD11b antibody and measured by flow cytometry.

F) Correlation of composite phenotypes across time points and cell types. Composite phenotypes were defined in a hypothesis-driven manner. Fraction TE (repetitive and transposable element) represents the number of non-intronic reads mapped to TEs over total, averaged over all cells bearing each perturbation (both collapsed on UMIs). Fraction mtRNA represents the mean number of reads mapped to mitochondrial genome protein-coding genes over total. Total RNA represents the mean total RNA content (number of UMIs).

G) Comparison of TE expression across time points. The mean fraction TE reads per perturbation is highly correlated across time points in K562 cells (r=0.79). Genetic perturbations are colored to reflect cluster identity.

H) Comparison of total RNA content across time points. Total RNA content per perturbation is highly correlated across time points in K562 cells (r=0.74). Genetic perturbations are colored to reflect cluster identity.

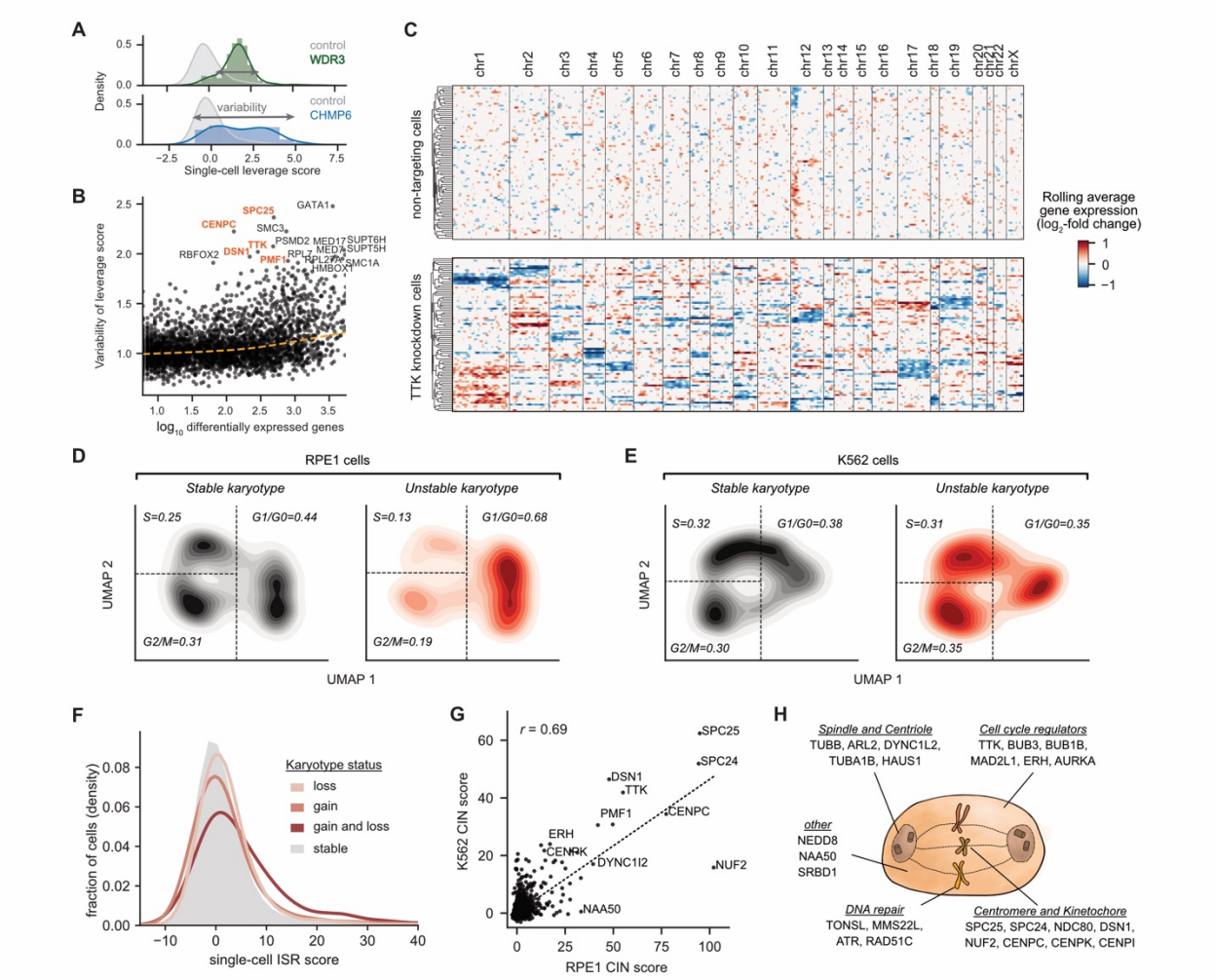


Figure 2.5: Exploring acute consequences and genetic drivers of aneuploidy in single-cells.

A) Schematic of heterogeneity statistic. Single-cell leverage scores quantify how outlying each cell is relative to non-targeting control cells by PCA. For each perturbation, heterogeneity of single-cell phenotypes is quantified as the standard deviation of leverage scores.

B) Identifying heterogeneous perturbations. Known regulators of chromosome segregation were among the perturbations with the highest single-cell heterogeneity (high variability of leverage scores), especially compared to their number of differentially expressed genes (based on Anderson-Darling test).

C) Heatmap of chromosomal copy number inference from Perturb-seq data. For all genes (expressed >0.05 UMI per cell), the log-fold change in expression is calculated with respect to the average of non-targeting control cells, and genes are ordered along the genome. A weighted moving average of 100 genes is used infer copy number changes (columns) in single-cells (rows) with noise and median filtering. 80 TTK knockdown RPE1 cells and 80 randomly sampled non-targeting control RPE1 cells are shown. Cells are ordered by average linkage hierarchical clustering based on correlation of chromosomal copy number profiles.

D) and E) Comparison of cell cycle occupancy upon acute karyotypic changes. Unstable karyotypic cells were defined as having ≥1 chromosome with evidence of changes in chromosomal copy number for >80% of the chromosomal length. For single-cells, cell-

cycle positioning was inferred by UMAP dimension reduction on differential expression profiles of 199 selected cell-cycle regulated genes. Cell cycle occupancy is shown as a 2D kernel density estimate of a random subset of 1000 cells per karyotypic status. Approximate gates between cell cycle phases (G1 or G0; S; G2 or M) are shown as dotted lines, and the fraction of cells in each cell cycle phase are indicated.

E) Effect of chromosomal instability (CIN) on activation of the Integrated Stress Response (ISR). Histogram (kernel density estimate) compares the ISR score versus CIN status in RPE1 cells. CIN status is defined as evidence of gain or loss of chromosomal copy number for >80% of the chromosomal length, with 240,768 stable cells, 5,522 cells bearing chromosomal loss, 1987 cells bearing chromosomal gain, and 904 cells bearing gain and loss of chromosomes. ISR score is defined as the sum of *z*-normalized expression of ISR marker genes where increased values indicate stronger ISR activation.

E) Comparison of the effect of genetic perturbations on the CIN score across cell types. For each genetic perturbation, the CIN score is calculated as the mean single-cell sum of squared CIN values, *z*-normalized relative to non-targeting control perturbations. The CIN score is correlated across cell types (r=0.69, $p<10^{-16}$).

F) Schematic of a subset of genetic perturbations that drive CIN. CIN drivers play diverse roles in mitosis, cell cycle regulation, and DNA repair.

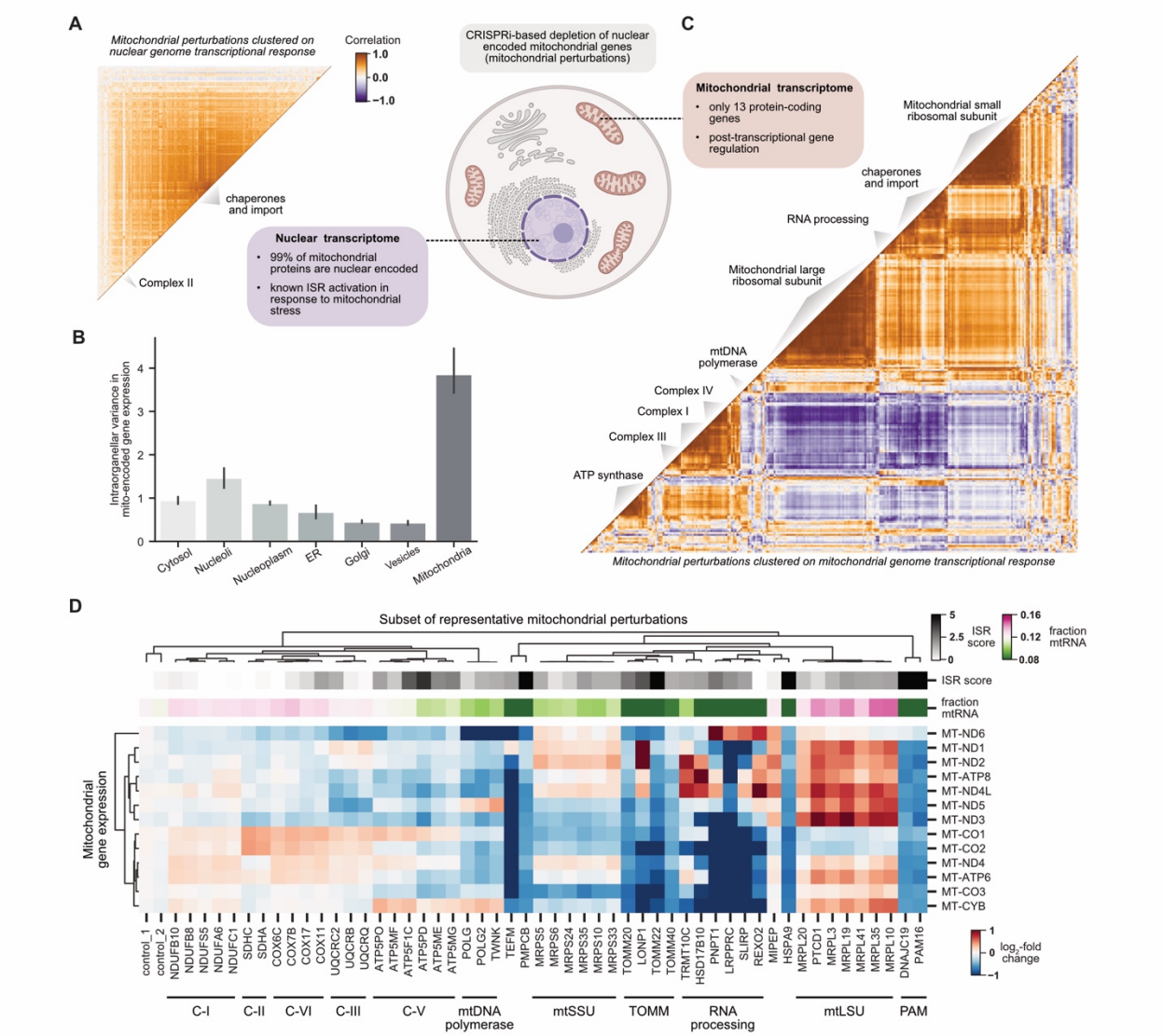


Figure 2.6: Global organization of the transcriptional response to mitochondrial stress.

A) Clustering mitochondrial perturbations by nuclear transcriptional response. CRISPRi enables knockdown of nuclear-encoded genes whose protein products are targeted to mitochondria (mitochondrial perturbations). Mitochondrial perturbations were annotated by MitoCarta3.0 and subset to those with a strong transcriptional phenotype (n=268 mitochondrial perturbations). Gene expression profiles were restricted to nuclear encoded genes which includes 99% of mitochondrial proteins. The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of mitochondrial perturbations in K562 cells. Genetic perturbations were clustered by HDBSCAN with a correlation metric.

B) Comparing variability in the mitochondrial transcriptome by perturbation localization. The mitochondrial genome encodes 13 protein-coding genes. Genetic perturbations were grouped based on localization of their protein products as determined by the Human Protein Atlas. For each of these 13 mitochondrial-encoded genes, the variance in pseudobulk z-normalized expression profiles was calculated between all perturbations with the same localization. Barplots represent the average across genes with 95% confidence interval obtained by bootstrapping.

C) Clustering mitochondrial perturbations by mitochondrial transcriptional response. Mitochondrial perturbations were annotated by MitoCarta3.0 and subset to those with a strong transcriptional phenotype as above (n=268 mitochondrial perturbations). Gene expression profiles were restricted to the 13 mitochondrial-encoded genes. The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of mitochondrial perturbations in K562 cells. Genetic perturbations are clustered by HDBSCAN with a correlation metric. Clusters were manually annotated.

D) Heatmap visualizing the mitochondrial genome transcriptional response to diverse mitochondrial stressors. The expression (log_2 fold-change relative to non-targeting controls) of the 13 mitochondrial-encoded genes is shown for a subset of perturbations representative of different mitochondrial complexes or function. Neither the ISR score nor mean fraction of mitochondrial RNA (mtRNA) would allow for high-resolution clustering by function as provided by the mitochondrial genome response. Genetic perturbations and genes are ordered by average linkage hierarchical clustering with a correlation metric.

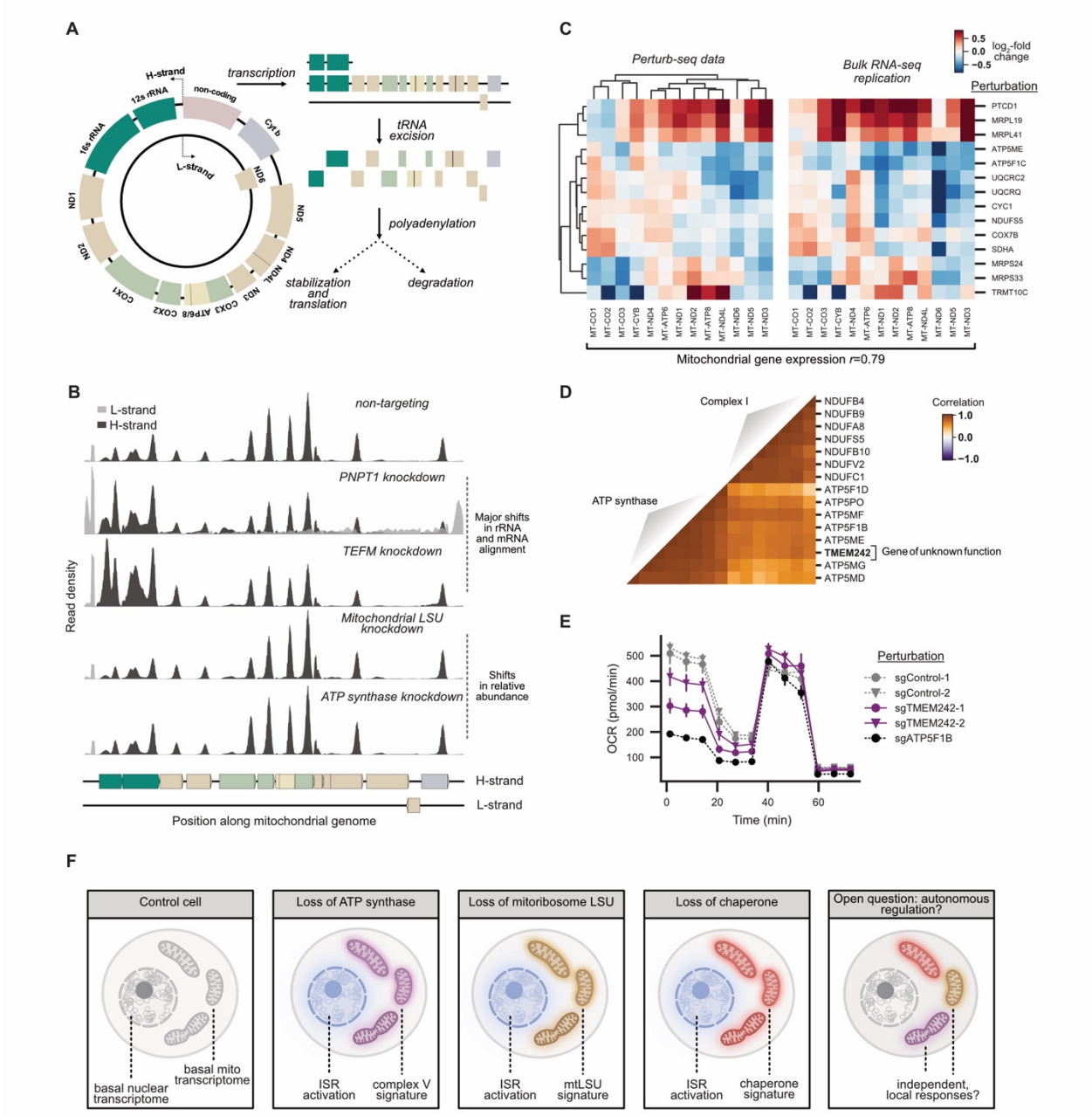


Figure 2.7: Investigating regulation of the mitochondrial genome in stress.

A) Schematic of the mitochondrial transcriptome. Each human cell contains many copies of the circular 16.6 kb mitochondrial genome distributed throughout the mitochondrial network. The human mitochondrial genome encodes 2 rRNAs, 22 tRNAs, and 13 protein-coding genes. Both the heavy (H) and light (L) strand of the genome are transcribed as polycistronic transcripts punctuated by tRNAs. Excision of tRNAs from transcripts generates nascent mRNA precursors (colored by complex membership). mRNA precursors can then be polyadenylated, stabilized, or degraded.

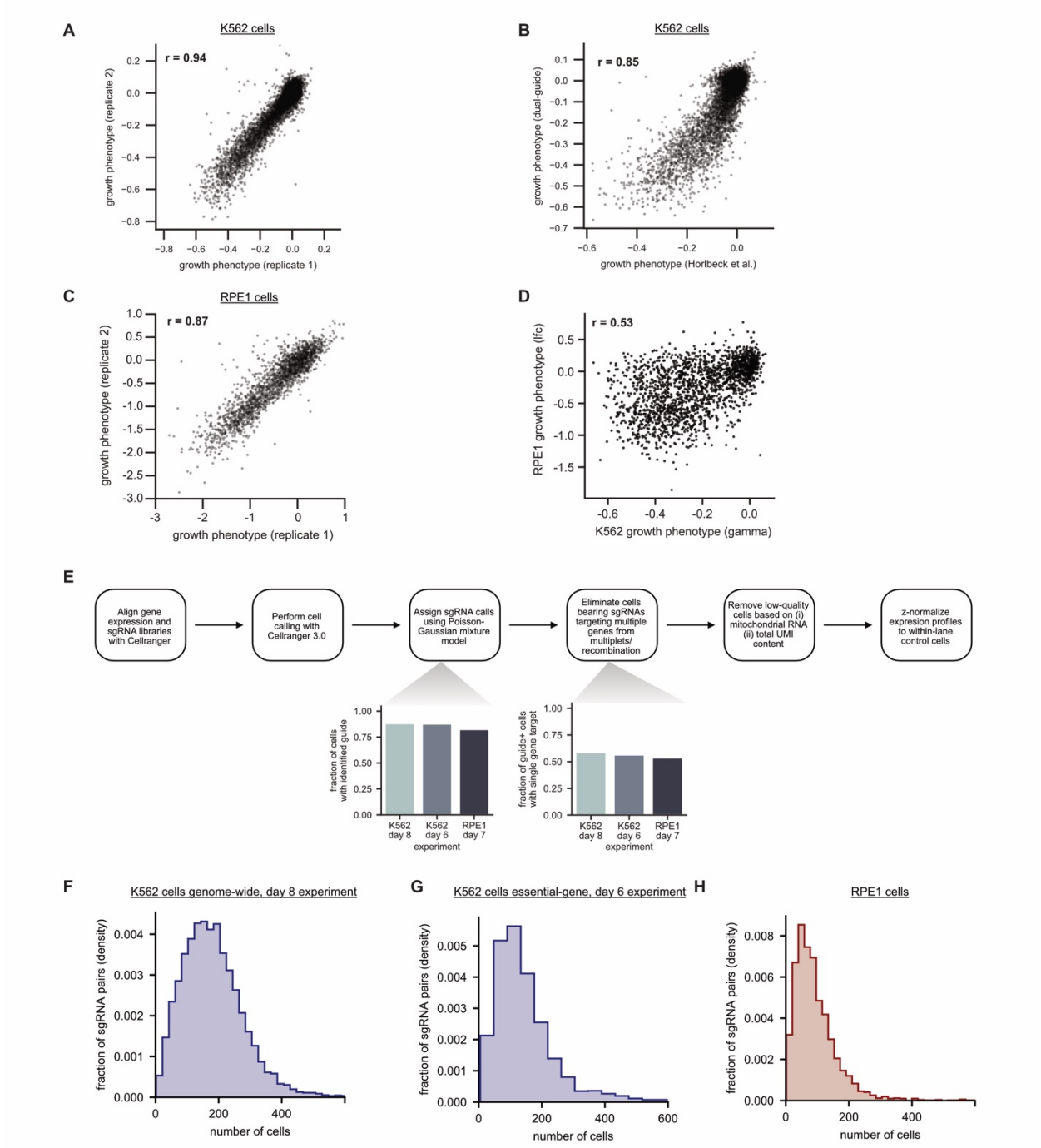
B) Density of Perturb-seq reads along the mitochondrial genome from select genetic perturbations. Reads are aligned to both the H-strand (dark grey) and L-strand (light grey). For each perturbation, densities are shown relative to the maximum read count in the locus.

C) Comparison of mitochondrial gene expression profiles between Perturb-seq and bulk RNA-seq. Heatmap displays $\log_2$-fold changes in expression of the 13 mitochondrial encoded genes (columns) for genetic perturbations (rows) in Perturb-seq and bulk RNA-seq data collected from K562 cells. Bulk RNA-seq was conducted to analyze total RNA (including non-polyadenylated RNA), with data representing the average of biological replicates. Genetic perturbations and genes are ordered by average linkage hierarchical clustering with a Euclidean distance metric. The profiles are strongly correlated (r=0.79, $p<10^{-39}$).

D) Clustering of TMEM242 genetic perturbation based on the mitochondrial transcriptome. Genetic perturbations to members of ATP synthase and Complex I of the respiratory chain were compared to knockdown of TMEM242, a mitochondrial gene of unknown function. Gene expression profiles were restricted to the 13 mitochondrial-encoded genes. The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of mitochondrial perturbations in K562 cells. Genetic perturbations are ordered by HDBSCAN with a correlation metric.

E) Effect of TMEM242 knockdown on mitochondrial respiration. A Seahorse analyzer was used to monitor oxygen consumption rate (OCR). The Mito Stress Test consists of sequential addition of oligomycin (an ATP synthase inhibitor that enables measurement of ATP-productive respiration), FCCP (an uncoupling agent that enables measurement of maximal respiratory capacity), and a mixture of rotenone and antimycin A (inhibitors of Complex I and Complex III, respectively, that enable measurement of non-mitochondrial respiration). Data is presented as average ± SEM, n=6.

F) Schematic diagram of mitochondrial stress response.



Supplementary Figure 2.1: Growth screens, filtering, and coverage.

A) Comparing the growth phenotypes of dual-sgRNA constructs between growth screen replicates in K562 cells. Growth phenotypes are reported as the $\log_2$ guide enrichment per cell doubling (gamma) between day 6 and day 16 post library transduction. Replicates are strongly correlated (n=11,056 dual sgRNA constructs; r = 0.94). For 50 outlier genes (where the residual from a regression comparing replicates was >0.2), the growth phenotype was set to missing.

B) Benchmarking the growth phenotypes of dual-sgRNA constructs to single-sgRNA screens. Growth phenotypes (gammas) are compared between the dual-sgRNA library

compared to the mean of the best three sgRNAs from Horlbeck et al. The screens are strongly correlated (n=9386 genes after excluding constructs mapping to secondary TSSs; r = 0.85) but with stronger growth phenotypes observed for the dual-sgRNA library.

C) Comparing the growth phenotypes of dual-sgRNA constructs between growth screen replicates in RPE1 cells. Growth phenotypes are reported as the $\log_2$ guide enrichment between the plasmid library and day 7 post library transduction. Replicates are strongly correlated (n=2203 constructs targeting common essential genes; r = 0.87). For 19 outlier genes (where the residual from a regression comparing replicates was >1), the growth phenotype was set to missing.

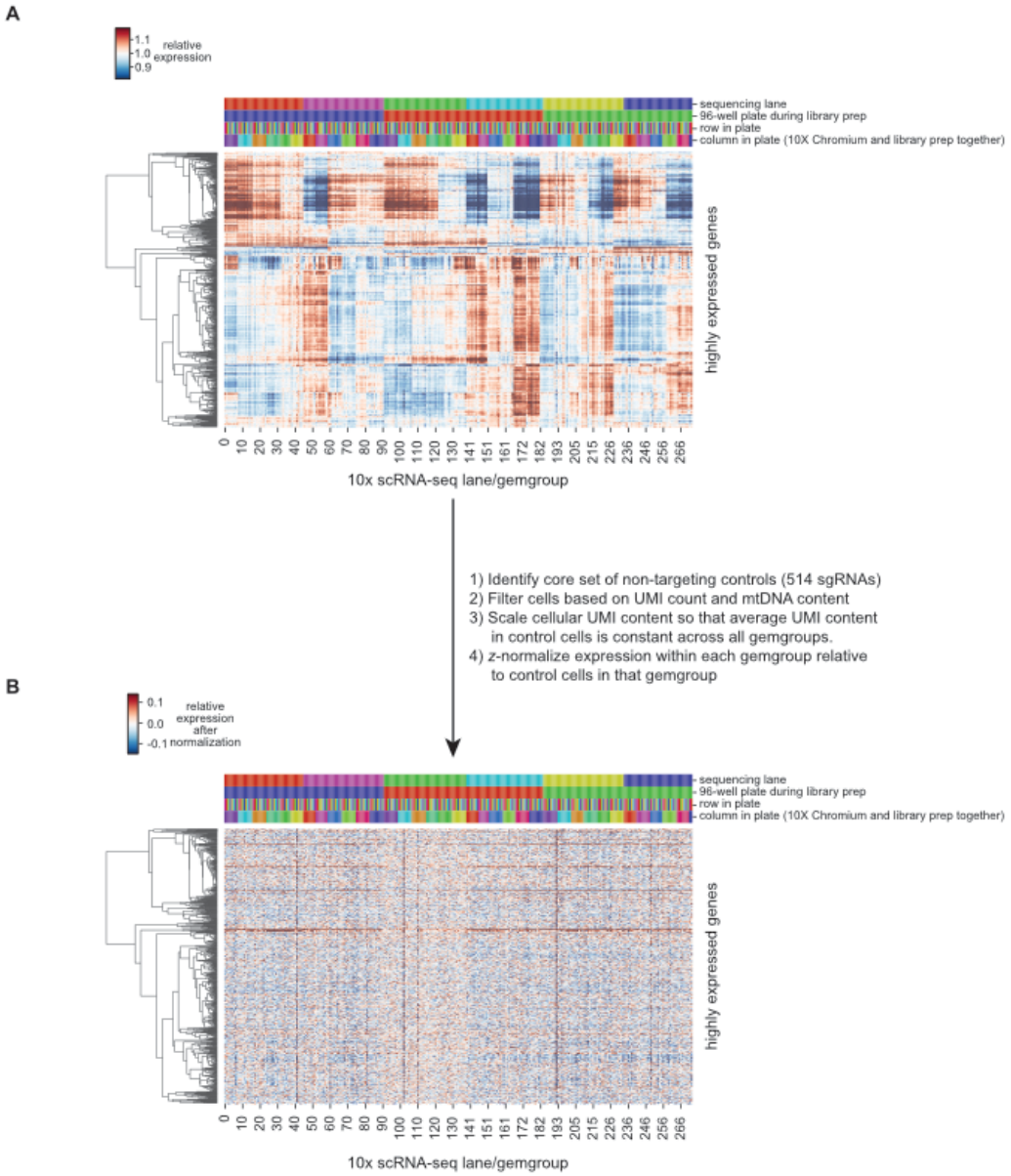
D) Comparing growth phenotypes between K562 and RPE1 cells. Growth phenotypes are correlated (n=1951 constructs; r=0.53) despite substantial differences in screen timepoint (day 6 to day 16 for K562 cells versus day 0 to day 7 in RPE1 cells).

E) Schematic overview of data alignment, cell calling, sgRNA assignment, and filtering.

F) Histogram of the number of cells per genetic perturbation in the K562 day 8 genome-wide Perturb-seq experiment. The number of detected genetic perturbations (expected sgRNA pairs) was n=11,258, with a mean coverage 183 cells per perturbation and a median coverage of 171 cells per perturbation after filtering.

G) Histogram of the number of cells per genetic perturbation in the K562 day 6 essential-wide Perturb-seq experiment. The number of detected genetic perturbations (expected sgRNA pairs) was n=2,285, with a mean coverage 148 cells per perturbation and a median coverage of 124 cells per perturbation after filtering.

H) Histogram of the number of cells per genetic perturbation in the RPE1 cell day 7 essential-wide Perturb-seq experiment. The number of detected genetic perturbations (expected sgRNA pairs) was n=2,679, with a mean coverage 101 cells per perturbation and a median coverage of 79 cells per perturbation after filtering.

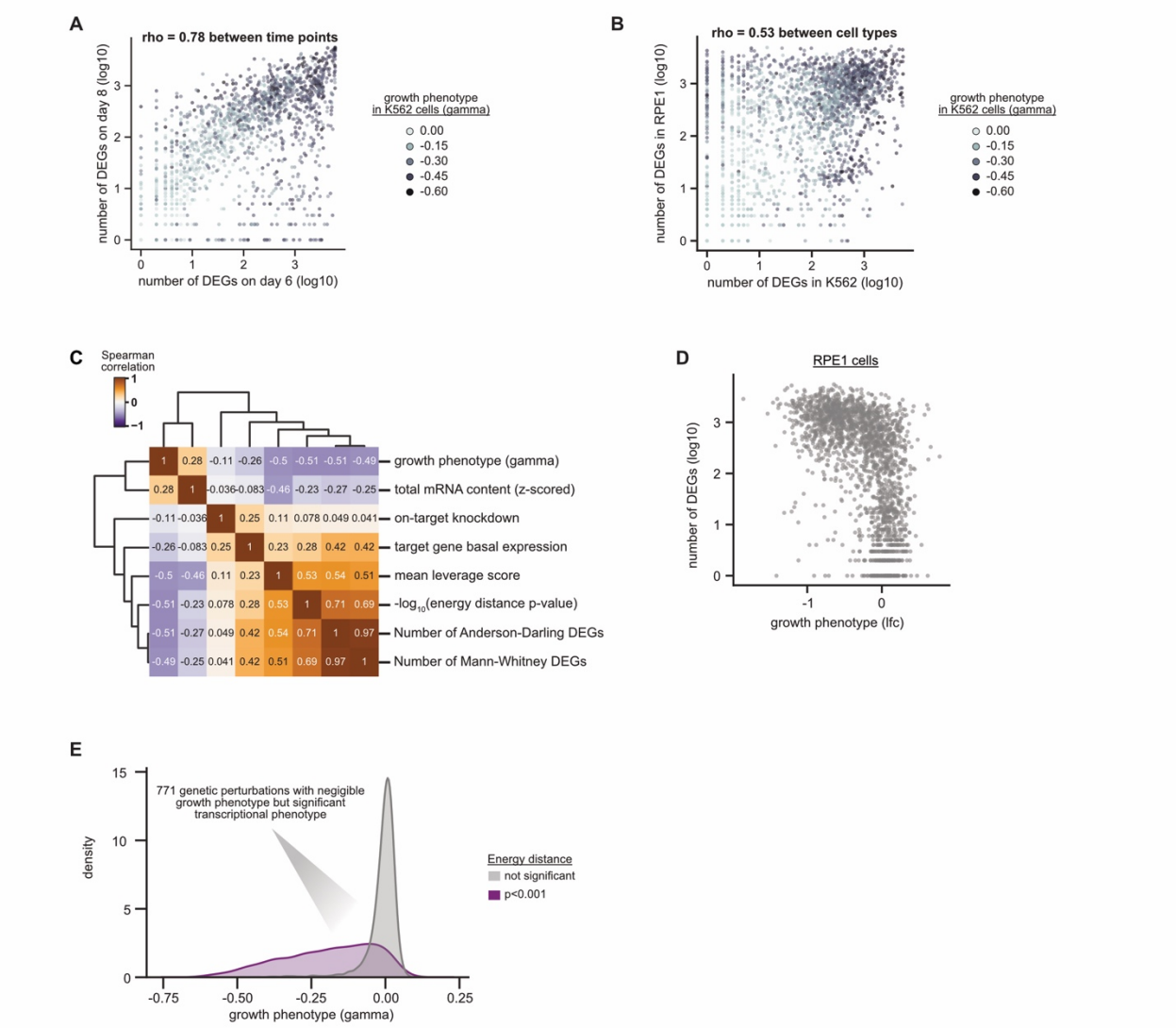


Supplementary Figure 2.2: Schematic and performance of internal normalization of gene expression measurements.

A) Batch effects in raw data. The K562 day 8 genome-wide experiment was conducted across 273 separate lanes of 10x Genomics droplet single-cell RNA sequencing ("gemgroups"). The plot shows mean expression profiles of highly expressed genes (>2 UMI per cell) within all cells in each gemgroup. The data is normalized (i) for sequencing depth of each gemgroup and (ii) so that the mean expression of each gene is 1 across all gemgroups. Colors at the top indicate different levels of multiplexing that were present within the experiment: including groups of samples (generally in sets of 8) that went through scRNA-seq together, 96 well plates used for library preparation, and separate lanes used during Illumina sequencing. The range of the heatmap is set according to the 2%-98% quantiles of the data.

B) Expression following internal normalization. We rescale gene expression by *z*-normalizing relative to the control cells (containing non-targeting sgRNAs, ~4% of all cells) within the

same gemgroup. The plot shows the average normalized expression of all cells within each gemgroup following this procedure, with the range of the heatmap set according to the 2%-98% quantiles of the data. By construction control cells have mean expression 0 and standard deviation 1 in this scale. Genes are presented in the same order as in panel (A).



Supplementary Figure 2.3: Growth screens, filtering, and coverage.

A) Relationship between the number of differentially expressed genes (DEGs) for a genetic perturbation in K562 cells at day 8 versus day 6 post-transduction. DEGs were determined using a two-sample Anderson-Darling test comparing against non-targeting guides (n=2276 common genetic perturbations, Spearman's rho=0.78).

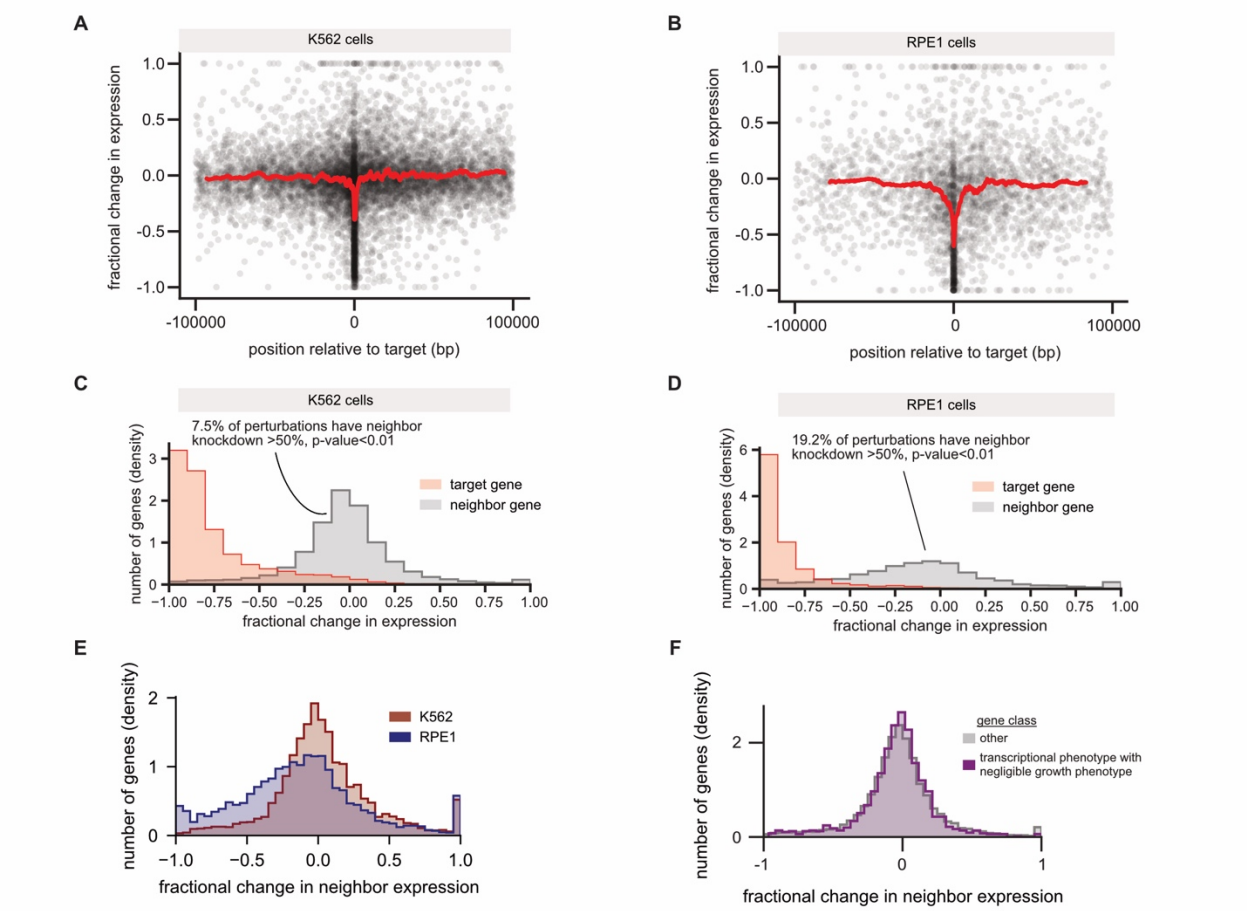
B) Relationship between the number of DEGs for a genetic perturbation in K562 cells (day 8 genome-wide dataset) versus RPE1 cells. DEGs were determined using a two-sample Anderson-Darling test comparing against non-targeting guides (n=2636 common genetic perturbations, Spearman's rho=0.53).

C) Relationship between features of genetic perturbations in K562 cells genome-wide day 8 Perturb-seq. The features were calculated as detailed in Methods. The heatmap displays Spearman correlations between features.

D) Comparing the growth phenotype versus the number of DEGs for each multiplexed guide pairs in RPE1 cells. Growth phenotypes are reported as the $\log_2$ guide enrichment between day 0 and day 7 post-lentiviral transduction. DEGs were determined using a two-sample Anderson-Darling test comparing against non-targeting guides.

E) The distribution of growth phenotypes in genetic perturbations with a transcriptional phenotypes in K562 cells genome-wide day 8 Perturb-seq. Histogram (kernel density

estimate) comparing the growth phenotype in K562 cells (gamma) of genetic perturbations to the permuted energy distance test. 771 genetic perturbations had a gamma > -0.1 (considered a negligible effect on cellular growth) but a significant transcriptional phenotype.



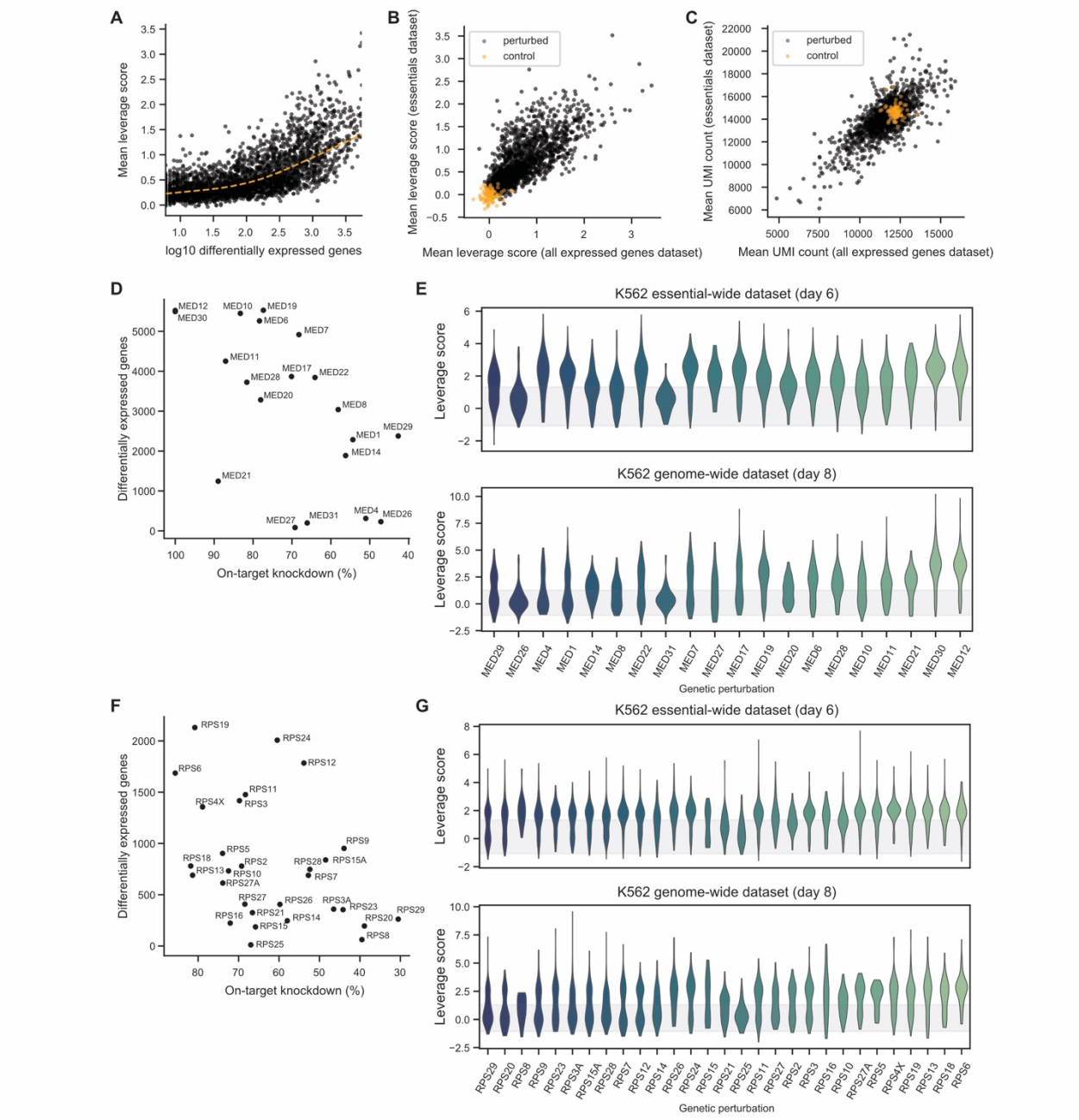
Supplementary Figure 2.4: Assessing neighbor gene off-target knockdown in Perturb-seq data.

A) and B) Relationship between neighbor gene off-target knockdown and position relative to the target gene in K562 cells (day 8) (A) and RPE1 cells (B). For each target genes, the two neighbor genes are defined as the gene immediately upstream and downstream (at an expression >0.1 UMI per cell). The position relative to the target is the distance of either the start or end of the neighbor gene (whichever is closer) to the start of the target gene. The fractional change in expression is defined as the expression in the targeted cells minus the expression in non-targeting cells, relative to the expression in the non-targeting cell population (-1 implies 100% knockdown).

C) and D) Comparison between target gene and neighbor gene knockdown in K562 cells (day 8) (C) and RPE1 cells (D). P-values are assigned by comparing a bootstrap test.

E) Comparison of neighbor gene knockdown in K562 cells (day 8) versus RPE1 cells.

F) Comparison of neighbor gene knockdown based on transcriptional phenotype in K562 cells (day 8). Perturbations with “transcriptional phenotype with negligible growth phenotype” are those perturbations where gamma > -0.1 that had a significant transcriptional phenotype by the permuted energy distance test.



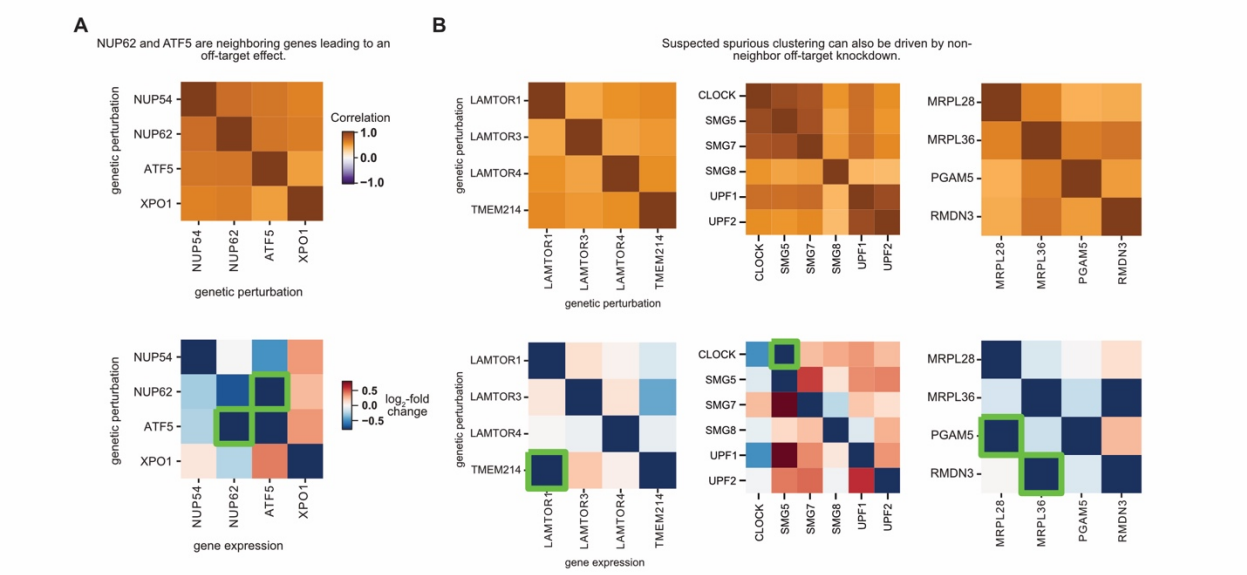
Supplementary Figure 2.5: Assessing the penetrance and heterogeneity of response to genetic perturbations.

A) We scored how outlying each perturbed cell was relative to non-targeting control cells using leverage scores. The plot compares the mean leverage score for each genetic perturbation to the number of differentially expressed genes detected by the Anderson-Darling test (Spearman's rho = 0.71).

B) To assess reproducibility of leverage scores, plot compares mean leverage scores of perturbations (black dots) in K562 cells between the essentials dataset (taken at day 6 post-infection) and the dataset targeting all expressed genes (taken at day 8 post-infection). Non-targeting control sgRNAs are in orange (Spearman's rho = 0.79).

C) For comparison to (B), the plot compares the average UMI counts between the two K562 cell Perturb-seq experiments (Spearman's rho = 0.61).

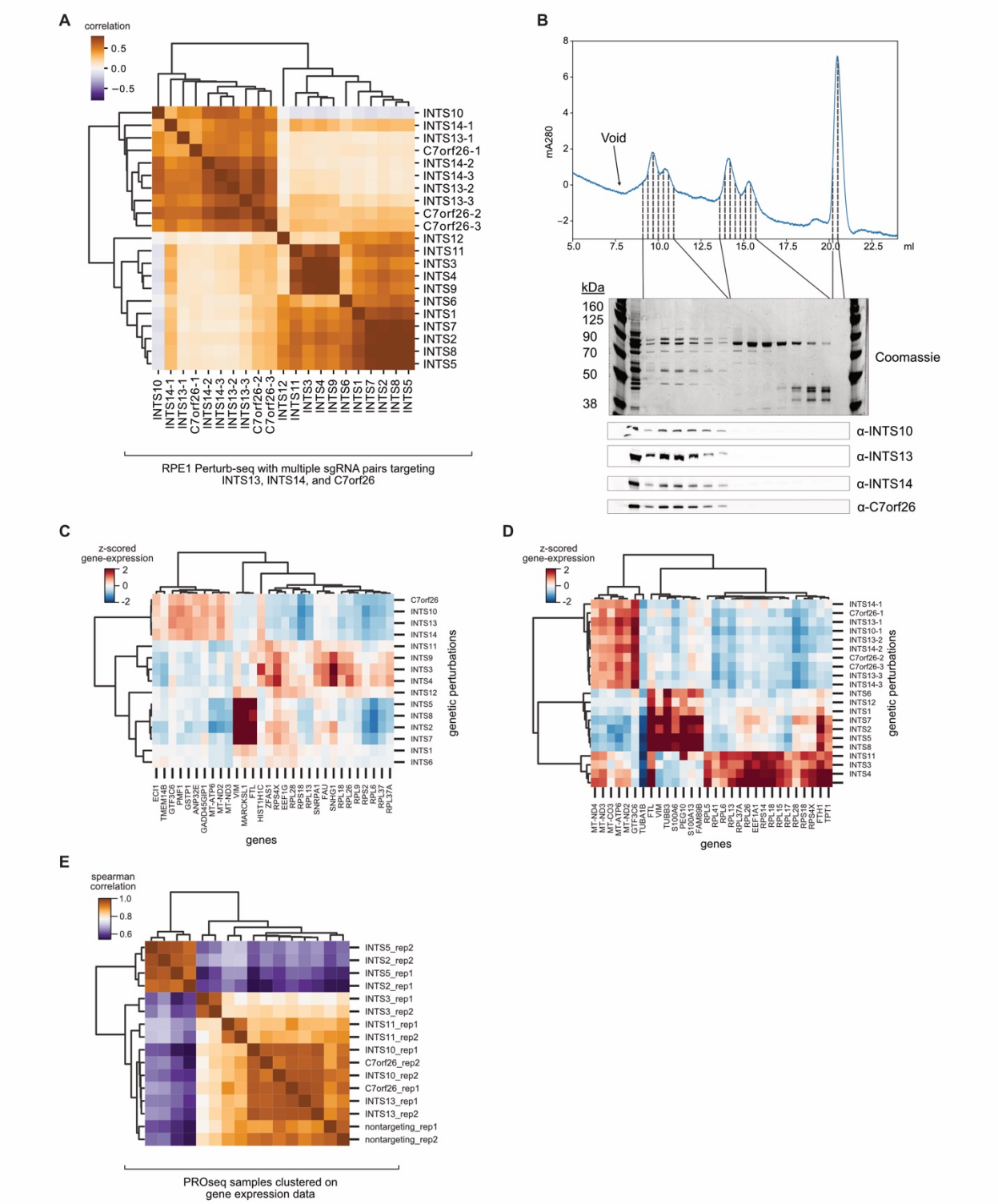
D) Relationship between knockdown of target gene (relative to expression in control cells bearing non-targeting sgRNAs) and number of differentially expressed genes detected by Anderson-Darling test for perturbations targeting subunits of the Mediator complex.
E) Leverage scores distributions of perturbations targeting subunits of the Mediator complex. Plot shows kernel density estimates for each perturbation ordered from least knocked down (left) to most (right). Top panel is within essentials dataset and bottom panel is within the all expressed genes dataset. The gray bar shows the 10%-90% range of leverage scores within control cells bearing non-targeting sgRNAs.
F-G) As in (D)-(E), but for perturbations targeting the small subunit of the ribosome.



Supplementary Figure 2.6: Examples of neighbor or distant off-target knockdown leading to phenotypic similarity in Perturb-seq.

A) Investigation of ATF5 phenotype in K562 cells (day 8). The upper heatmap shows the correlation of the expression profile of ATF5 knockdown with similar genetic perturbations. The sgRNA pair targeting ATF5 leads to a phenotype strongly correlated with knockdown of subunits of the nuclear pore complex (NUP54, NUP62) and nuclear export proteins (XPO1). The sgRNA pair targeting ATF5 leads to strong downregulation of NUP62 (bottom heatmap, shown as the $\log_2$-fold change compared to control cells). ATF5 is bidirectionally expressed with NUP62 on chromosome 19, explaining this neighbor gene off-target knockdown and similar phenotypes.

B) Investigation of other surprising phenotypes in K562 cells (day 8). We investigated three different surprising relationships between genetic perturbations: clustering of TMEM215 with LAMTOR subunits, clustering of CLOCK with NMD machinery, and clustering of PGAM5 and RMDN3 with the mitochondrial large ribosomal subunit (upper heatmaps). In all three cases, the relationship could be explained by suspected off-target knockdown of a component of the complex, directly detected in Perturb-seq (bottom heatmaps).



Supplementary Figure 2.7: Supplementary data related to the functional modules of the Integrator complex.

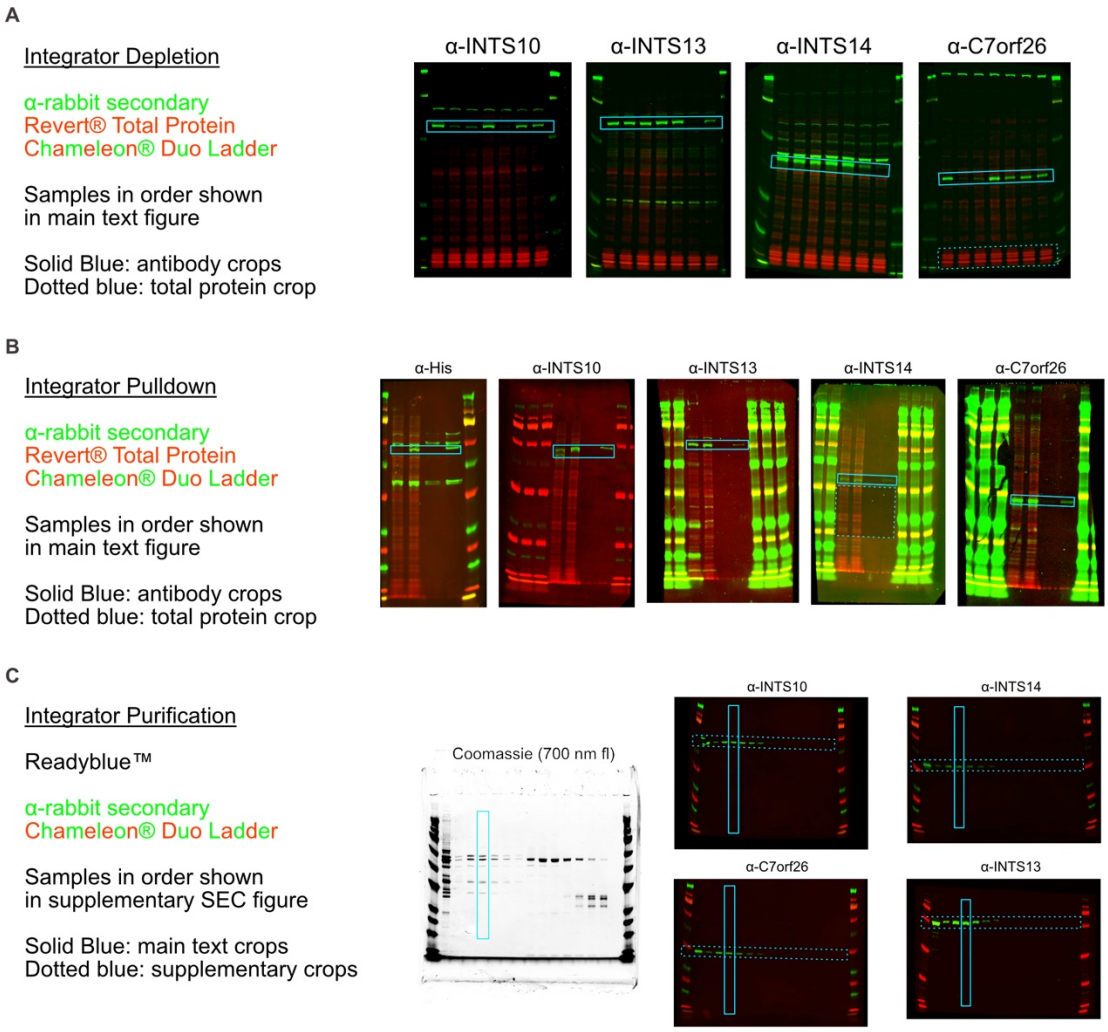
A) Relationship between Integrator complex members and C7orf26 in RPE1 cells. Multiple independent sgRNA pairs were used to target INTS13, INTS14, and C7orf26, with independent guides indicated by numbers next to gene names (e.g. C7orf26-1, C7orf26-

2, etc.). The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of Integrator complex members. Genetic perturbations are ordered by average linkage hierarchical clustering based on correlation.

B) SEC trace and full Western blots for urification of a INTS10-INTS13-INTS14-C7orf26 complex. His-INTS10, INTS13, INTS14, and C7orf26 were overexpressed in Expi293 cells, affinity purified, and separated via SEC. The INTS10-INTS13-INTS14-C7orf26 proteins co-fractionated as a higher molecular weight species as visualized by Western blotting.

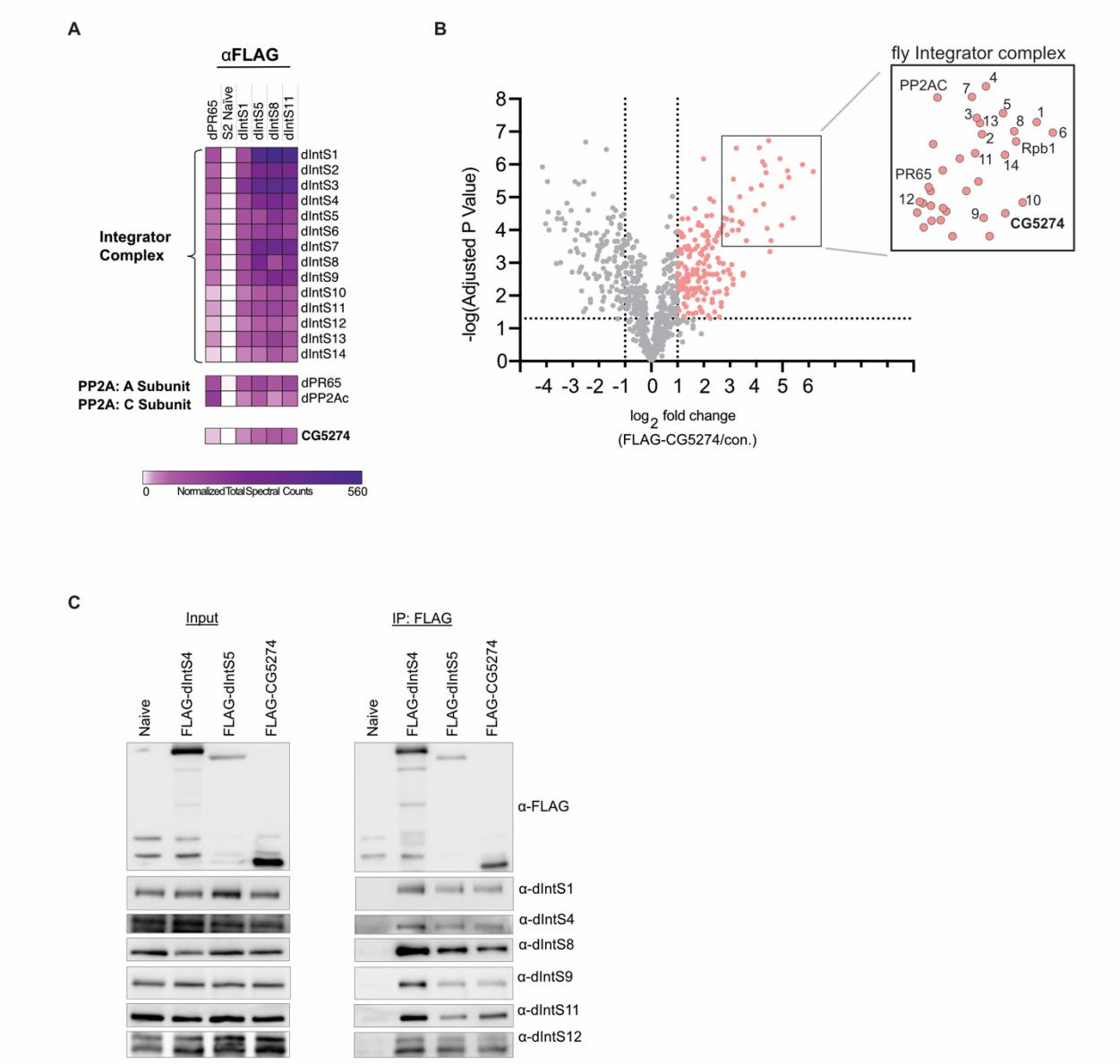
C) and D) Identification of differentially expressed genes between Integrator complex modules in K562 cells (C) and RPE1 cells (D). Random forest classifiers were trained on gene expression profiles to classify cells as having perturbation to one of three Integrator complex modules ("endonuclease", "shoulder and backbone" or "10-13-14-C7orf26"). The top 30 gene features led to an accuracy of 97% in K562 cells and 89% in RPE1 cells. Heatmap displays the z-scored gene expression of the top 30 features in each cell type.

E) Comparison of PRO-seq active RNA polymerase II gene expression profiles across genetic perturbations. PRO-seq reads were aligned to the transcriptome and depth normalized. Heatmap shows the Spearman correlation of expression profiles with biological replicates indicated (e.g. INTS5_rep1 and INTS5_rep2).



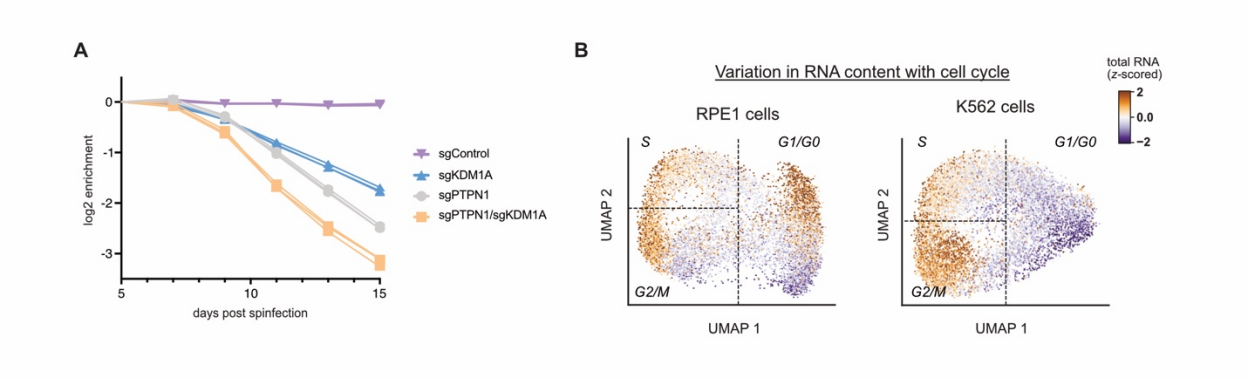
Supplementary Figure 2.8: Supplementary data related to Integrator biochemistry.

A) Full blots visualizing the effects of CRISPRi-based depletion of Integrator subunits with different probes.
B) Full blots visualizing INTS10 pulldown with different probes.
C) Full blots visualizing Integrator SEC purification with different probes.



Supplementary Figure 2.9: Supplementary data related to Integrator biochemistry in Drosophila.

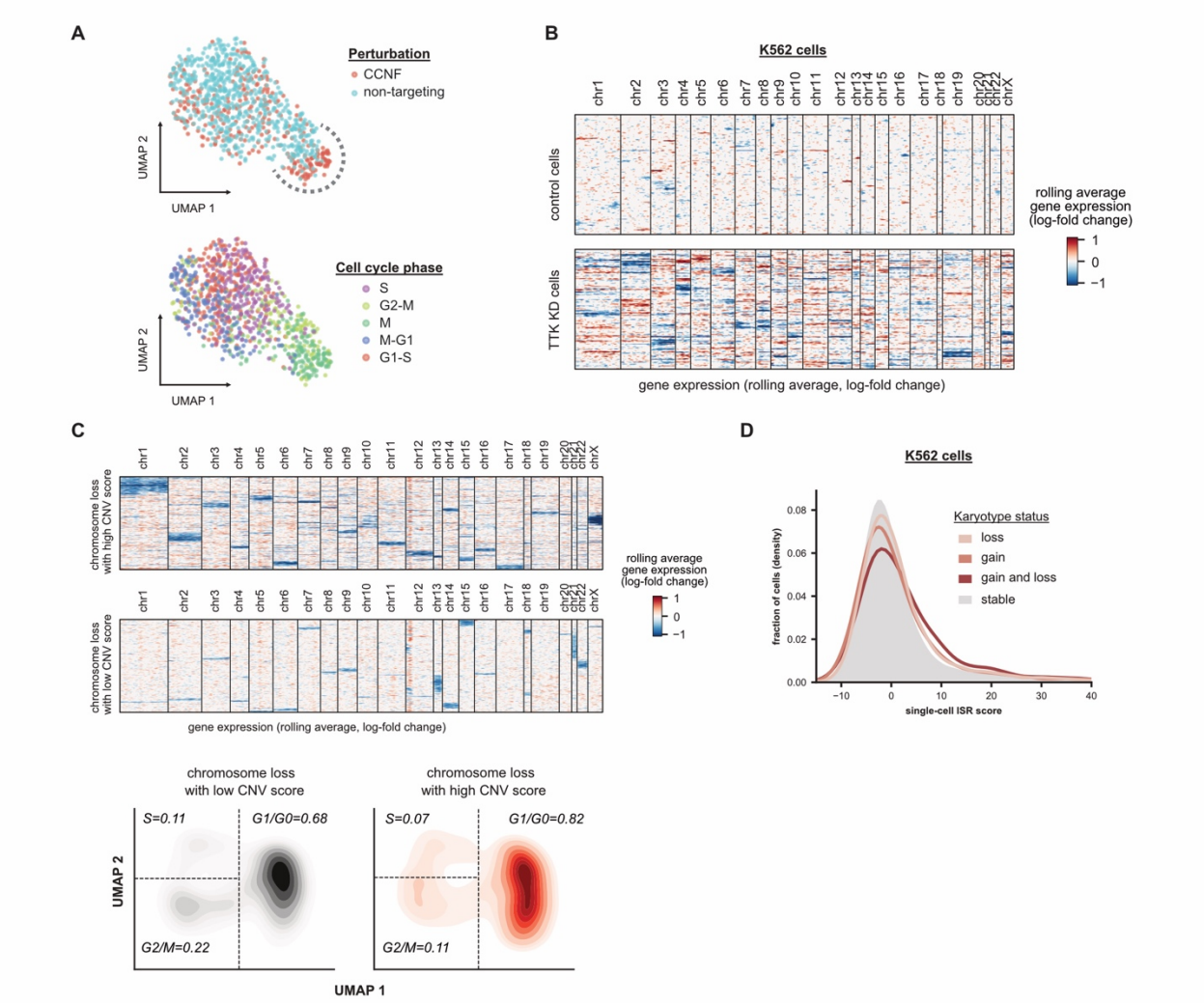
A) Heatmap of results from co-immunoprecipitation of Integrator complex components in Drosophila.
B) Volcano plot of enrichments from co-immunoprecipitation of Integrator complex components with Drosophila C7orf26-homologue CG5274.
C) Co-immunoprecipitation of Integrator complex components in Drosophila. Cell lysates were affinity purified and select Integrator proteins were probed by western blot.



Supplementary Figure 2.10: Supplementary data related to phenotype relationships.

A) Growth effect of PTPN1 or KDM1A knockdown in K562 cells. Cells were co-transduced with fluorescently labelled sgKDM1A, sgPTPN1, or a non-targeting control guide. Enrichment was determined by flow cytometry relative to uninfected cells in biological triplicate.

B) Comparison of total RNA content with cell cycle state. For single-cells, cell-cycle positioning was inferred by UMAP dimension reduction on differential expression profiles of 199 selected cell-cycle regulated genes. The dimension reduction was performed independently for RPE1 cells (left) and K562 cells (right). Cell cycle occupancy is shown as a scatterplot of UMAP positions of a random subset of 10,000 cells per cell type. Approximate gates between cell cycle phases (G1 or G0; S; G2 or M) are shown as dotted lines. The total RNA content per cell was calculated from the total number of UMIs detected per cell which were z-scored with respect to gemgroup/lane control cells.



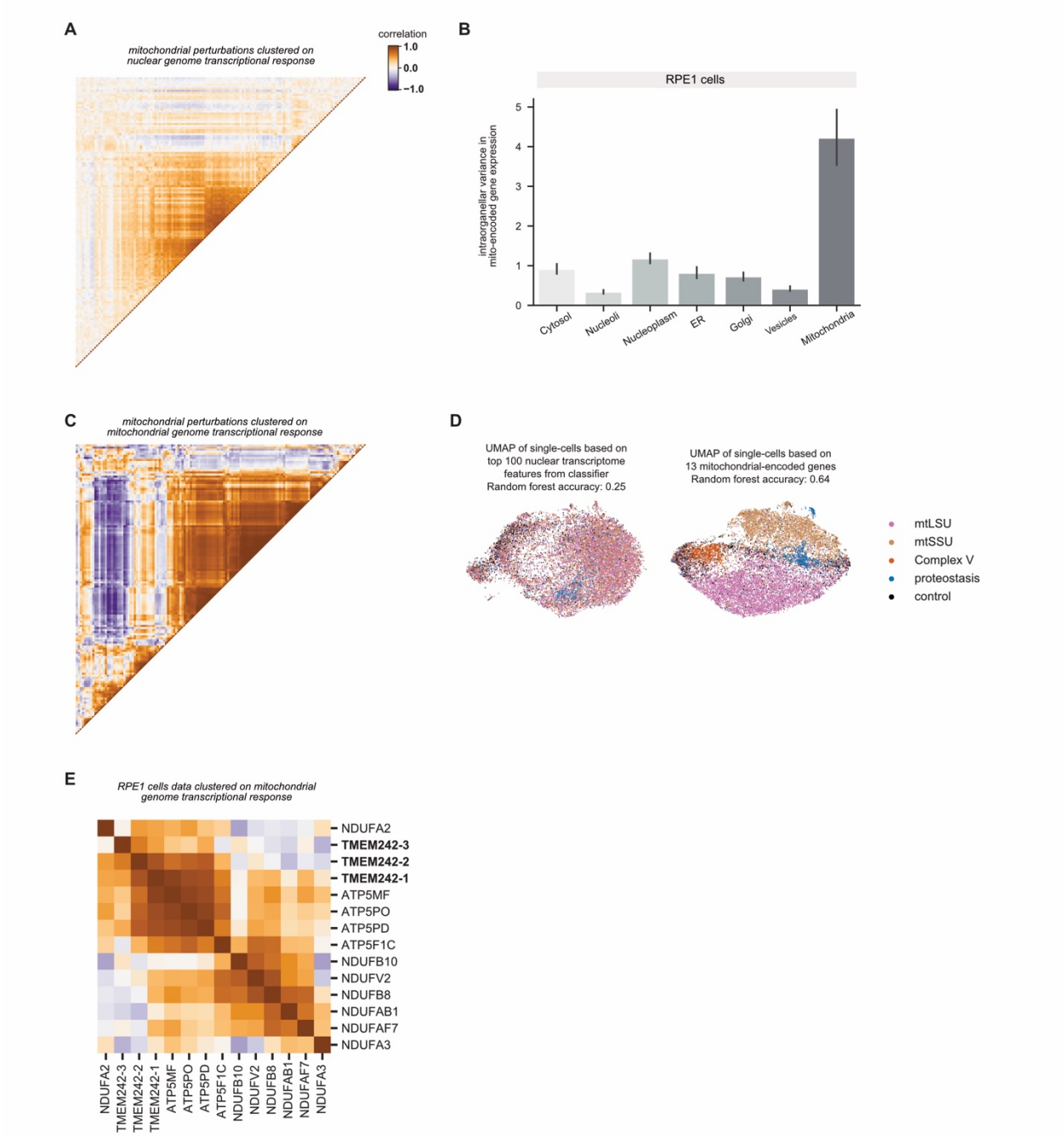
Supplementary Figure 2.11: Supplementary data related to chromosomal instability.

A) Heatmap of chromosomal copy number inference from Perturb-seq data. For all genes (expressed >0.05 UMI per cell), the log-fold change in expression is calculated with respect to the average of non-targeting control cells, and genes are ordered along the genome. A weighted moving average of 100 genes is used infer copy number changes (columns) in single-cells (rows) with noise and median filtering. 199 TTK knockdown K562 cells and 199 randomly sampled non-targeting control K562 cells are shown (data from K562 essential-wide day 6 dataset). Cells are ordered by average linkage hierarchical clustering based on correlation of chromosomal copy number profiles.

B) Comparison of cell cycle effects by magnitude of karyotypic abnormality in RPE1 cells. RPE1 cells with at least one chromosomal loss (defined as evidence of chromosomal loss for >80% of the chromosomal length) were stratified into high, medium, and low degree of karyotypic abnormality based on their CNV score. 500 randomly-sampled high and low CNV score cells were visualized in a heatmap of chromosomal copy number inference. Below, the cell cycle occupancy of high and low CNV cells is shown for 1000 randomly-sampled cells. For single-cells, cell-cycle positioning was inferred by UMAP dimension reduction on differential expression profiles of 199 selected cell-cycle regulated genes. Cell cycle occupancy is shown as a 2D kernel density estimate of a random subset of 1000 cells per karyotypic status. Approximate gates between cell cycle phases (G1 or G0;

S; G2 or M) are shown as dotted lines, and the fraction of cells in each cell cycle phase are indicated.

C) Effect of chromosomal instability (CIN) on activation of the Integrated Stress Response (ISR). Histogram (kernel density estimate) compares the ISR score versus CIN status in K562 cells (day 6). CIN status is defined as evidence of gain or loss of chromosomal copy number for >80% of the chromosomal length, with 290,432 stable cells, 11,100 cells bearing chromosomal loss, 5,852 cells bearing chromosomal gain, and 4,541 cells bearing gain and loss of chromosomes. ISR score is defined as the sum of z-normalized expression of ISR marker genes where increased values indicate stronger ISR activation.



Supplementary Figure 2.12: Supplementary data related to mitochondrial genome regulation.

A) Clustering mitochondrial perturbations by nuclear transcriptional response. CRISPRi enables knockdown of nuclear-encoded genes whose protein products are targeted to mitochondria (mitochondrial perturbations). Mitochondrial perturbations were annotated by MitoCarta3.0 and subset to those with a strong transcriptional phenotype (n=140 mitochondrial perturbations). Gene expression profiles were restricted to nuclear encoded genes (including 99% of mitochondrial proteins). The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of mitochondrial perturbations in RPE1 cells. Genetic perturbations are ordered by HDBSCAN with a correlation metric.

B) Comparing variability in the mitochondrial transcriptome by perturbation localization. The mitochondrial genome encodes 13 protein-coding genes. Genetic perturbations were grouped based on localization of their protein products as determined by the Human Protein Atlas. For each of these 13 mitochondrial-encoded genes, the variance in pseudobulk z-normalized expression profiles was calculated between all perturbations with the same localization. Barplots represent the average across genes with 95% confidence interval obtained by bootstrapping.

C) Clustering mitochondrial perturbations by mitochondrial transcriptional response. Mitochondrial perturbations were annotated by MitoCarta3.0 and subset to those with a strong transcriptional phenotype as above (n=140 mitochondrial perturbations). Gene expression profiles were restricted to the 13 mitochondrial-encoded genes. The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of mitochondrial perturbations in RPE1 cells. Genetic perturbations are ordered by HDBSCAN with a correlation metric.

D) Comparison of predictive accuracy of nuclear versus mitochondrial genome response. To assess the specificity of the nuclear and mitochondrial genome regulation, random forest classifiers were trained on gene expression profiles to classify cells as having perturbation to one of four mitochondrial complexes ("mtLSU" which corresponds to components of the mitochondrial large ribosomal subunit; "mtSSU" which corresponds to components of the mitochondrial small ribosomal subunit; "Complex V" which corresponds to components of ATP synthase; "proteostasis" which corresponds to essential proteostatic machinery; "control" which corresponds to non-targeting control cells). In K562 cells (day 8), the top 100 gene features led to an accuracy of 25% for the nuclear-encoded genes, versus 64% for the 13 mitochondrial-encoded genes. As a visual guide, the plot displays the UMAP embedding of single cells colored by perturbed mitochondrial complex based on sgRNA assignment.

E) Clustering of TMEM242 genetic perturbation based on the mitochondrial transcriptome. Genetic perturbations to members of ATP synthase and Complex I of the respiratory chain were compared to knockdown of TMEM242, a mitochondrial gene of unknown function. Gene expression profiles were restricted to the 13 mitochondrial-encoded genes. The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of mitochondrial perturbations in RPE1 cells. Multiple independent sgRNA pairs targeting TMEM242 were used. Genetic perturbations are ordered by HDBSCAN with a correlation metric.

REFERENCES

Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, Villalta JE, Gilbert LA, Horlbeck MA, Hein MY, Pak RA, Gray AN, Gross CA, Dixit A, Parnas O, Regev A, Weissman JS. 2016. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* 167:1867-1882.e21. doi:10.1016/j.cell.2016.11.048

Alerasool N, Segal D, Lee H, Taipale M. 2020. An efficient KRAB domain for CRISPRi applications in human cells. *Nat Methods* 17:1093–1096. doi:10.1038/s41592-020-0966-x

Allen JF. 2017. The CoRR hypothesis for genes in organelles. *J Theor Biol* 434:50–57. doi:10.1016/j.jtbi.2017.04.008

Ben-David U, Amon A. 2020. Context is everything: aneuploidy in cancer. *Nat Rev Genet* 21:44–62. doi:10.1038/s41576-019-0171-x

Carroll J, He J, Ding S, Fearnley IM, Walker JE. 2021. TMEM70 and TMEM242 help to assemble the rotor ring of human ATP synthase and interact with assembly factors for complex I. *P Natl Acad Sci Usa* 118:e2100558118. doi:10.1073/pnas.2100558118

Cleary B, Cong L, Cheung A, Lander ES, Regev A. 2017. Efficient Generation of Transcriptomic Profiles by Random Composite Measurements. *Cell* 171:1424-1436.e18. doi:10.1016/j.cell.2017.10.023

D'Angiolella V, Donato V, Vijayakumar S, Saraf A, Florens L, Washburn MP, Dynlacht B, Pagano M. 2010. SCF(Cyclin F) controls centrosome homeostasis and mitotic fidelity through CP110 degradation. *Nature* 466:138–42. doi:10.1038/nature09140

Datlinger P, Rendeiro AF, Boenke T, Senekowitsch M, Krausgruber T, Barreca D, Bock C. 2021. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat Methods* 18:635–642. doi:10.1038/s41592-021-01153-z

Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. 2017. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* 14:297–301. doi:10.1038/nmeth.4177

Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, Adamson B, Norman TM, Lander ES, Weissman JS, Friedman N, Regev A. 2016. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167:1853-1866.e17. doi:10.1016/j.cell.2016.11.038

Doench JG. 2018. Am I ready for CRISPR? A user's guide to genetic screens. *Nat Rev Genet* 19:67–80. doi:10.1038/nrg.2017.97

Feldman D, Singh A, Schmid-Burgk JL, Carlson RJ, Mezger A, Garrity AJ, Zhang F, Blainey PC. 2019. Optical Pooled Screens in Human Cells. *Cell* 179:787-799.e17. doi:10.1016/j.cell.2019.09.016

Fessler E, Eckl E-M, Schmitt S, Mancilla IA, Meyer-Bender MF, Hanf M, Philippou-Massier J, Krebs S, Zischka H, Jae LT. 2020. A pathway coordinated by DELE1 relays mitochondrial stress to the cytosol. *Nature* 579:433–437. doi:10.1038/s41586-020-2076-4

Fianu I, Chen Y, Dienemann C, Dybkov O, Linden A, Urlaub H, Cramer P. 2021. Structural basis of Integrator-mediated transcription regulation. *Science* 374:883–887. doi:10.1126/science.abk0154

Frangieh CJ, Melms JC, Thakore PI, Geiger-Schuller KR, Ho P, Luoma AM, Cleary B, Jerby-Arnon L, Malu S, Cuoco MS, Zhao M, Ager CR, Rogava M, Hovey L, Rotem A, Bernatchez C, Wucherpfennig KW, Johnson BE, Rozenblatt-Rosen O, Schadendorf D, Regev A, Izar B. 2021. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nat Genet* 53:332–341. doi:10.1038/s41588-021-00779-1

Friedman JR, Nunnari J. 2014. Mitochondrial form and function. *Nature* 505:335–343. doi:10.1038/nature12985

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

Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Ruepp A. 2019. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res* 47:D559–D563. doi:10.1093/nar/gky973

Guo X, Aviles G, Liu Y, Tian R, Unger BA, Lin Y-HT, Wiita AP, Xu K, Correia MA, Kampmann M. 2020. Mitochondrial stress is relayed to the cytosol by an OMA1-DELE1-HRI pathway. *Nature* 579:427–432. doi:10.1038/s41586-020-2078-2

Haapaniemi E, Botla S, Persson J, Schmierer B, Taipale J. 2018. CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. *Nat Med* 24:927–930. doi:10.1038/s41591-018-0049-z

Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M, Weissman JS. 2016. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* 5:e19760. doi:10.7554/elife.19760

Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, Salame TM, Tanay A, Oudenaarden A van, Amit I. 2016. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* 167:1883-1896.e15. doi:10.1016/j.cell.2016.11.039

Jelluma N, Brenkman AB, Broek NJF van den, Cruijsen CWA, Osch MHJ van, Lens SMA, Medema RH, Kops GJPL. 2008. Mps1 Phosphorylates Borealin to Control Aurora B Activity and Chromosome Alignment. *Cell* 132:233–246. doi:10.1016/j.cell.2007.11.046

Jost M, Chen Y, Gilbert LA, Horlbeck MA, Krenning L, Menchon G, Rai A, Cho MY, Stern JJ, Prota AE, Kampmann M, Akhmanova A, Steinmetz MO, Tanenbaum ME, Weissman JS. 2017. Combined CRISPRi/a-Based Chemical Genetic Screens Reveal that Rigosertib Is a Microtubule-Destabilizing Agent. *Mol Cell* 68:210-223.e6. doi:10.1016/j.molcel.2017.09.012

Kirstein N, Santos HGD, Blumenthal E, Shiekhattar R. 2021. The Integrator complex at the crossroad of coding and noncoding RNA. *Curr Opin Cell Biol* 70:37–43. doi:10.1016/j.ceb.2020.11.003

Kramer NJ, Haney MS, Morgens DW, Jovičić A, Couthouis J, Li A, Ousey J, Ma R, Bieri G, Tsui CK, Shi Y, Hertz NT, Tessier-Lavigne M, Ichida JK, Bassik MC, Gitler AD. 2018. CRISPR–Cas9 screens in human cells and primary neurons identify modifiers of C9ORF72 dipeptide-repeat-protein toxicity. *Nat Genet* 50:603–612. doi:10.1038/s41588-018-0070-7

Kummer E, Ban N. 2021. Mechanisms and regulation of protein synthesis in mitochondria. *Nat Rev Mol Cell Bio* 22:307–325. doi:10.1038/s41580-021-00332-2

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* 172:650–665. doi:10.1016/j.cell.2018.01.029

Leary JF, Ohlsson-Wilhelm BM, Giuliano R, Labella S, Farley B, Rowley PT. 1987. Multipotent human hematopoietic cell line K562: Lineage-specific constitutive and inducible antigens. *Leukemia Res* 11:807–815. doi:10.1016/0145-2126(87)90065-8

Luo J, Yang H, Song B-L. 2020. Mechanisms and regulation of cholesterol homeostasis. *Nat Rev Mol Cell Bio* 21:225–245. doi:10.1038/s41580-019-0190-7

Ma P, Mahoney MW, Yu B. 2013. A Statistical Perspective on Algorithmic Leveraging. *Arxiv*.

Maes T, Mascaró C, Tirapu I, Estiarte A, Ciceri F, Lunardi S, Guibourt N, Perdones A, Lufino MMP, Somervaille TCP, Wiseman DH, Duy C, Melnick A, Willekens C, Ortega A, Martinell M, Valls N, Kurz G, Fyfe M, Castro-Palomino JC, Buesa C. 2018. ORY-1001, a Potent and Selective Covalent KDM1A Inhibitor, for the Treatment of Acute Leukemia. *Cancer Cell* 33:495-511.e12. doi:10.1016/j.ccell.2018.02.002

Mamińska A, Bartosik A, Banach-Orłowska M, Pilecka I, Jastrzębski K, Zdżalik-Bielecka D, Castanon I, Poulain M, Neyen C, Wolińska-Nizioł L, Toruń A, Szymańska E, Kowalczyk A, Piwocka K, Simonsen A, Stenmark H, Fürthauer M, González-Gaitán M, Miaczynska M. 2016.

ESCRT proteins restrict constitutive NF-κB signaling by trafficking cytokine receptors. *Sci Signal* 9:ra8–ra8. doi:10.1126/scisignal.aad0848

Martin BK, Qiu C, Nichols E, Phung M, Green-Gladden R, Srivatsan S, Blecher-Gonen R, Beliveau BJ, Trapnell C, Cao J, Shendure J. 2021. An optimized protocol for single cell transcriptional profiling by combinatorial indexing. *Arxiv*.

Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood A-MJ, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA, Filipovska A, Mattick JS. 2011. The Human Mitochondrial Transcriptome. *Cell* 146:645–658. doi:10.1016/j.cell.2011.06.051

Mick E, Titov DV, Skinner OS, Sharma R, Jourdain AA, Mootha VK. 2020. Distinct mitochondrial defects trigger the integrated stress response depending on the metabolic state of the cell. *Elife* 9:e49178. doi:10.7554/elife.49178

Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, Legut M, Roush T, Herrera A, Papalexi E, Ouyang Z, Satija R, Sanjana NE, Koralov SB, Smibert P. 2019. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat Methods* 16:409–412. doi:10.1038/s41592-019-0392-0

Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket S, Yuan F, Chen S, Leung HM, Villoria J, Rogel N, Burgin G, Tsankov A, Waghray A, Slyper M, Waldmann J, Nguyen L, Dionne D, Rozenblatt-Rosen O, Tata PR, Mou H, Shivaraju M, Bihler H, Mense M, Tearney GJ, Rowe SM, Engelhardt JF, Regev A, Rajagopal J. 2018. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* 560:319–324. doi:10.1038/s41586-018-0393-7

Münch C, Harper JW. 2016. Mitochondrial unfolded protein response controls matrix pre-RNA processing and translation. *Nature* 534:710–713. doi:10.1038/nature18302

Musacchio A, Salmon ED. 2007. The spindle-assembly checkpoint in space and time. *Nat Rev Mol Cell Bio* 8:379–393. doi:10.1038/nrm2163

Norman TM, Horlbeck MA, Replogle JM, Ge AY, Xu A, Jost M, Gilbert LA, Weissman JS. 2019. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* 365:786–793. doi:10.1126/science.aax4438

Orkin SH, Zon LI. 2008. Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell* 132:631–644. doi:10.1016/j.cell.2008.01.025

Papalexi E, Mimitou EP, Butler AW, Foster S, Bracken B, Mauck WM, Wessels H-H, Hao Y, Yeung BZ, Smibert P, Satija R. 2021. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat Genet* 53:322–331. doi:10.1038/s41588-021-00778-2

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A, Bernstein BE. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344:1396–1401. doi:10.1126/science.1254257

Pfleiderer MM, Galej WP. 2021. Structure of the catalytic core of the Integrator complex. *Mol Cell* 81:1246-1259.e8. doi:10.1016/j.molcel.2021.01.005

Plasschaert LW, Žilionis R, Choo-Wing R, Savova V, Knehr J, Roma G, Klein AM, Jaffe AB. 2018. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* 560:377–381. doi:10.1038/s41586-018-0394-6

Przybyla L, Gilbert LA. 2021. A new era in functional genomics screens. *Nat Rev Genet* 1–15. doi:10.1038/s41576-021-00409-w

Quirós PM, Mottis A, Auwerx J. 2016. Mitonuclear communication in homeostasis and stress. *Nat Rev Mol Cell Biology* 17:213–26. doi:10.1038/nrm.2016.23

Quirós PM, Prado MA, Zamboni N, D'Amico D, Williams RW, Finley D, Gygi SP, Auwerx J. 2017. Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals. *J Cell Biol* 216:2027–2045. doi:10.1083/jcb.201702058

Radhakrishnan SK, Lee CS, Young P, Beskow A, Chan JY, Deshaies RJ. 2010. Transcription Factor Nrf1 Mediates the Proteasome Recovery Pathway after Proteasome Inhibition in Mammalian Cells. *Mol Cell* 38:17–28. doi:10.1016/j.molcel.2010.02.029

Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Cogan JZ, Meer EJ, Terry JM, Riordan DP, Srinivas N, Fiddes IT, Arthur JG, Alvarado LJ, Pfeiffer KA, Mikkelsen TS, Weissman JS, Adamson B. 2020. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat Biotechnol* 38:954–961. doi:10.1038/s41587-020-0470-y

Richter-Dennerlein R, Oeljeklaus S, Lorenzi I, Ronsör C, Bareth B, Schendzielorz AB, Wang C, Warscheid B, Rehling P, Dennerlein S. 2016. Mitochondrial Protein Synthesis Adapts to Influx of Nuclear-Encoded Protein. *Cell* 167:471-483.e10. doi:10.1016/j.cell.2016.09.003

Rubin AJ, Parker KR, Satpathy AT, Qi Y, Wu B, Ong AJ, Mumbach MR, Ji AL, Kim DS, Cho SW, Zarnegar BJ, Greenleaf WJ, Chang HY, Khavari PA. 2019. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* 176:361-376.e17. doi:10.1016/j.cell.2018.11.022

Sabath K, Stäubli ML, Marti S, Leitner A, Moes M, Jonas S. 2020. INTS10–INTS13–INTS14 form a functional module of Integrator that binds nucleic acids and the cleavage module. *Nat Commun* 11:3422. doi:10.1038/s41467-020-17232-2

Salvatori R, Kehrein K, Singh AP, Aftab W, Möller-Hergt BV, Forne I, Imhof A, Ott M. 2020. Molecular Wiring of a Mitochondrial Translational Feedback Loop. *Mol Cell* 77:887-900.e5. doi:10.1016/j.molcel.2019.11.019

Santaguida S, Amon A. 2015. Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat Rev Mol Cell Bio* 16:473–485. doi:10.1038/nrm4025

Santaguida S, Richardson A, Iyer DR, M'Saad O, Zasadil L, Knouse KA, Wong YL, Rhind N, Desai A, Amon A. 2017. Chromosome Mis-segregation Generates Cell-Cycle-Arrested Cells with Complex Karyotypes that Are Eliminated by the Immune System. *Dev Cell* 41:638-651.e5. doi:10.1016/j.devcel.2017.05.022

Sharma B, Xie L, Yang F, Wang W, Zhou Q, Xiang M, Zhou S, Lv W, Jia Y, Pokhrel L, Shen J, Xiao Q, Gao L, Deng W. 2020. Recent advance on PTP1B inhibitors and their biomedical applications. *Eur J Med Chem* 199:112376. doi:10.1016/j.ejmech.2020.112376

Singh S, Broeck AV, Miller L, Chaker-Margot M, Klinge S. 2021. Nucleolar maturation of the human small subunit processome. *Science* 373:eabj5338. doi:10.1126/science.abj5338

Smits AH, Ziebell F, Joberty G, Zinn N, Mueller WF, Clauder-Münster S, Eberhard D, Savitski MF, Grandi P, Jakob P, Michon A-M, Sun H, Tessmer K, Bürckstümmer T, Bantscheff M, Steinmetz LM, Drewes G, Huber W. 2019. Biological plasticity rescues target activity in CRISPR knock outs. *Nat Methods* 16:1087–1093. doi:10.1038/s41592-019-0614-5

Stuart T, Satija R. 2019. Integrative single-cell analysis. *Nat Rev Genet* 20:257–272. doi:10.1038/s41576-019-0093-7

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering C von. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47:D607–D613. doi:10.1093/nar/gky1131

Tanay A, Regev A. 2017. Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541:331–338. doi:10.1038/nature21350

Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J, Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn WC. 2017. Defining a Cancer Dependency Map. *Cell* 170:564-576.e16. doi:10.1016/j.cell.2017.06.010

Wang E, Zhou H, Nadorp B, Cayanan G, Chen X, Yeaton AH, Nomikou S, Witkowski MT, Narang S, Kloetgen A, Thandapani P, Ravn-Boess N, Tsirigos A, Aifantis I. 2021. Surface antigen-guided CRISPR screens identify regulators of myeloid leukemia differentiation. *Cell Stem Cell* 28:718-731.e6. doi:10.1016/j.stem.2020.12.005

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015. Identification and characterization of essential genes in the human genome. *Science* 350:1096–1101. doi:10.1126/science.aac7041

Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, Ma'ayan A. 2021. Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* 1:e90. doi:10.1002/cpz1.90

Yu L, Myers G, Ku C-J, Schneider E, Wang Y, Singh SA, Jearawiriyapaisarn N, White A, Moriguchi T, Khoriaty R, Yamamoto M, Rosenfeld MG, Pedron J, Bushweller JH, Lim K-C, Engel JD. 2021. An erythroid-to-myeloid cell fate conversion is elicited by LSD1 inactivation. *Blood* 138:1691–1704. doi:10.1182/blood.2021011682

Zheng H, Qi Y, Hu S, Cao X, Xu C, Yin Z, Chen X, Li Y, Liu W, Li J, Wang J, Wei G, Liang K, Chen FX, Xu Y. 2020. Identification of Integrator-PP2A complex (INTAC), an RNA polymerase II phosphatase. *Sci New York N Y* 370. doi:10.1126/science.abb5872

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Joseph Replogle
22CB4E51C1AC445...
Author Signature

12/2/2021
Date