A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response

Citation: Adamson, Britt et al. "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response." Cell 167, 7 (December 2016): 1867–1882 © 2016 Elsevier Inc

As Published: http://dx.doi.org/10.1016/J.CELL.2016.11.048

Publisher: Elsevier

Persistent URL: http://hdl.handle.net/1721.1/116762

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2017 December 15.

Published in final edited form as:

Cell. 2016 December 15; 167(7): 1867–1882.e21. doi:10.1016/j.cell.2016.11.048.

A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response

Britt Adamson[1,2,3,4,*], **Thomas M. Norman**[1,2,3,4,*], **Marco Jost**[1,2,3,4,5], **Min Y. Cho**[1,2,3,4], **James K. Nuñez**[1,2,3,4], **Yuwen Chen**[1,2,3,4], **Jacqueline E. Villalta**[1,2,3,4], **Luke A. Gilbert**[1,2,3,4], **Max A. Horlbeck**[1,2,3,4], **Marco Y. Hein**[1,2,3,4], **Ryan A. Pak**[1,6], **Andrew N. Gray**[5], **Carol A. Gross**[5,7,8], **Atray Dixit**[9,10], **Oren Parnas**[10,11], **Aviv Regev**[10,12], and **Jonathan S. Weissman**[1,2,3,4,†]

[1]Department of Cellular & Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, CA 94158, USA

[3]California Institute for Quantitative Biomedical Research, University of California, San Francisco, San Francisco, CA 94158, USA

[4]Center for RNA Systems Biology, University of California, San Francisco, San Francisco, CA 94158, USA

[5]Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA 94158, USA

[6]Innovative Genomics Initiative, University of California, Berkeley, Berkeley, CA 94720, USA

[7]Department of Cell and Tissue Biology, University of California, San Francisco, San Francisco, CA 94158, USA

[8]Integrative Program in Quantitative Biology, University of California, San Francisco, San Francisco, CA 94158, USA

[9]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA 02142

[†]Lead contact; Correspondence should be addressed to jonathan.weissman@ucsf.edu.

[11]Present address: The Lautenberg Center for General and Tumor Immunology, The BioMedical Research Institute Israel Canada of the Faculty of Medicine (IMRIC), The Hebrew University Hadassah Medical School, 91120 Jerusalem, Israel

[*]Co-first author

AUTHOR CONTRIBUTIONS

B.A., T.M.N., M.J., and J.S.W. were responsible for the conception, design, and interpretation of the experiments and wrote the manuscript. B.A., T.M.N., M.J., M.Y.C., and J.K.N. conducted experiments. B.A. designed and cloned Perturb-seq vectors, developed the Perturb-seq experimental pipeline, and performed genome-scale screens. T.M.N. developed GBC capture and developed the Perturb-seq analytical pipeline. M.J. designed and cloned all three-guide vectors. T.M.N., J.E.V., and Y.C. prepared sequencing libraries. M.A.H. designed sgRNAs. L.A.G. validated mU6 and hU6 promoters. M.Y.H. built pMH0001. Y.C., M.Y.C., and R.A.P. helped prepare reagents for Perturb-seq experiments. M.J., A.N.G., and C.A.G. conducted bacterial work. A.R. played critical roles in the conceptualization of the project and A.D., O.P., and A.R. were engaged in discussion throughout.

[10]Broad Institute of MIT and Harvard, Cambridge MA 02142, USA

[12]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02140, USA

SUMMARY

Functional genomics efforts face tradeoffs between number of perturbations examined and complexity of phenotypes measured. We bridge this gap with Perturb-seq, which combines droplet-based single-cell RNA-seq with a strategy for barcoding CRISPR-mediated perturbations, allowing many perturbations to be profiled in pooled format. We applied Perturb-seq to dissect the mammalian unfolded protein response (UPR) using single and combinatorial CRISPR perturbations. Two genome-scale CRISPR interference (CRISPRi) screens identified genes whose repression perturbs ER homeostasis. Subjecting ~100 hits to Perturb-seq enabled high-precision functional clustering of genes. Single-cell analyses decoupled the three UPR branches, revealed bifurcated UPR branch activation among cells subject to the same perturbation, and uncovered differential activation of the branches across hits, including an isolated feedback loop between the translocon and IRE1α. These studies provide insight into how the three sensors of ER homeostasis monitor distinct types of stress and highlight the ability of Perturb-seq to dissect complex cellular responses.

Keywords

Single-cell RNA-seq; CRISPR; CRIPSRi; genome-scale screening; unfolded protein response; single-cell genomics; cell-to-cell heterogeneity

INTRODUCTION

Advances in pooled screening have made it possible to readily evaluate mammalian gene function at genome-scale, but to date have relied on simple phenotypic readouts that average properties of a population, such as the expression of a few exogenous reporters or cell viability. These approaches thus cannot distinguish mechanistically distinct perturbations that cause similar responses, or when a bulk phenotype is driven by a subpopulation. These limitations underscore the need for high-content, single-cell screens at genome-scale.

The advent of droplet-based single-cell RNA sequencing (RNA-seq) for profiling gene expression (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2016) has the potential to provide rich phenotypic data at the scale of hundreds of thousands of separately perturbed cells. To build a highly parallel platform for single-cell functional genomics, we paired this technology with our platform for CRISPR-based transcriptional interference (CRISPRi), which mediates gene inactivation with high efficacy and specificity (Qi et al., 2013; Gilbert et al., 2013; Gilbert et al., 2014; Horlbeck et al., 2016). To do this, we developed a robust cell barcoding strategy that encodes the identity of the CRISPR-mediated perturbation in an expressed transcript, which is captured during single-cell RNAseq analyses. This platform, termed “Perturb-seq,” provides a readily implementable and scalable approach for parallel screening with rich phenotypic output from single cells. Moreover, we developed a novel analytical pipeline to parse the massive datasets generated by Perturb-seq, which contain

RNA-seq profiles of tens of thousands of individual cells. This pipeline successfully decomposes the noisy, high-dimensional single-cell data into a handful of more interpretable components, which enables decoupling of the responses to a given perturbation within individual cells and isolation of those responses from confounding effects, such as the cell cycle.

Here, we apply Perturb-seq and its companion analytical pipeline to the systematic analysis of the mammalian unfolded protein response (UPR). The UPR is an integrated endoplasmic reticulum (ER) stress response pathway that is coordinated by three distinct ER transmembrane sensor proteins (IRE1α, ATF6, and PERK). In response to various perturbations, including deleterious changes to protein folding, calcium homeostasis, or membrane integrity, these sensors activate three transcription factors (XBP1, the N-terminal cleavage product of ATF6, and ATF4, respectively) to promote survival or, when ER stress cannot be corrected, trigger cell death pathways (Walter and Ron, 2011). Briefly, IRE1α mediates noncanonical splicing of *XBP1* mRNA to yield expression of the active XBP1 transcription factor (XBP1s). PERK is a kinase that, upon activation, phosphorylates the alpha subunit of the translation initiation factor eIF2 (eIF2α), which suppresses translation generally but paradoxically promotes translation of ATF4. Lastly, ATF6 is targeted to the Golgi where proteolytic cleavage releases a cytosolic transcription factor domain. Once activated, XBP1s, ATF4 and cleaved ATF6 translocate into the nucleus to initiate an integrated, partially co-regulated program of transcription. Considering the diversity of inputs and the complexity of outcome, comprehensive characterization of the UPR in mammalian cells requires both unbiased profiling of the physiological stresses that activate the sensors and delineation of the complex transcriptional phenotypes for each input.

To independently manipulate the three branches of the UPR, we first developed a programmable strategy for simultaneously repressing up to three genes with high efficacy. We then used Perturb-seq with combinatorial repression of the UPR sensor genes to delineate the distinct transcriptional programs of the three branches. Next, we used a two-tiered approach to interrogate the biological systems monitored by the UPR. We identified hundreds of genes that contribute to ER homeostasis from two genome-wide CRISPRi screens and then applied Perturb-seq to interrogate a diverse subset of these genes with single-cell resolution. These experiments allowed us to systematically define functional relationships between genes and to dissect the complex, partially overlapping transcriptional responses to ER stress. Furthermore, analysis of the single cell responses revealed bifurcation of the UPR branches at two levels: among individual cells subject to the same perturbation and at the population level, where differential activation of the three UPR branches occurred across perturbations. The latter includes a dedicated feedback loop that enables a single arm of the UPR (the IRE1α/XBP1 branch) to specifically monitor the integrity of the protein translocation machinery. These data demonstrate the ability of Perturb-seq to provide rich biological insights and systematically dissect complex biological responses.

RESULTS

A robust strategy for pooled profiling of perturbed cells by single-cell RNA-seq

Massively parallel droplet-based approaches for single-cell gene-expression profiling incorporate two indexing strategies that allow pooled RNA-seq data to be deconvolved into single-cell transcriptomes (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2016) (Figure 1A). Briefly, mRNA molecules from individual cells are paired in-droplet with two types of index, a cell barcode (CBC) and a unique molecular identifier (UMI). These indices are affixed to cDNA molecules during reverse transcription and, after pooled RNA-seq library preparation, are read out with mRNA identity by sequencing. The CBC links all sequencing reads from a given cell, and the UMI enables molecular counting of captured mRNA molecules by correcting for duplicates made during PCR. On these platforms, such indexing relies on oligo-dT priming prior to cDNA synthesis and, therefore, captures only polyadenylated RNA transcripts. To enable the recording of other types of information, we built a platform to genetically encode a third type of index on a synthetic polyadenylated transcript (Figures 1A, 1B). This index, which we term a “guide barcode” (GBC), can mark specific cell perturbations (e.g., the identity of a Cas9-targeting single guide RNA, sgRNA) and thus allows complex pools of cells to be interrogated in parallel on existing droplet-based platforms.

To deliver and capture GBCs, we designed the “Perturb-seq vector,” a third generation lentiviral vector that contains two notable features: an RNA polymerase II-driven “GBC expression cassette” and an RNA polymerase III-driven “sgRNA expression cassette” (Figure 1B). The GBC expression cassette carries a 3′ GBC sequence and terminates with a strong polyadenylation signal (BGH pA). Close proximity of the GBC and the BGH pA within this cassette favors faithful transmission of GBC sequences into single-cell RNA-seq libraries, which typically capture only the 3′ ends of transcripts. To prevent the internal BGH pA from disrupting transcription of the lentiviral genome, and therefore transduction competency, the entire expression region was engineered in reverse orientation with respect to the genomic promoter. Finally, to ensure robust GBC capture, we developed a PCR protocol to specifically enrich GBC-containing cDNAs, or “guide-mapping amplicons,” out of single-cell RNA-seq libraries (Figure 1A, 1B). In a pilot experiment, we performed single-cell RNA-sequencing on a pool of individually transduced chronic myeloid leukemia cells (K562) carrying 8 distinct GBCs, analyzing 5,768 cells total (Figure S1A, S1B). For the vast majority of cells, sequencing of guide-mapping amplicons uniquely identified a single dominant GBC with strong enrichment over any competing GBC identity (Figure 1C). Moreover, we observed a median of 45 independent observations per cell of these dominant guide-mapping amplicons (marked by UMIs), allowing us to uniquely infer a single GBC for 92.2% of cells (Figure 1D, 1E). Similar mapping rates were observed in all subsequent experiments (Figure S1). Importantly, our high confidence in GBC calling allowed us to discard information from droplets that fortuitously received more than one cell (Figure 1C).

By including an sgRNA expression cassette in the Perturb-seq vector, we tailored our indexing system to the study of CRISPR-based phenotypes. We confirmed that sgRNA

expression from the Perturb-seq vector was capable of generating robust and homogeneous CRISPRi-mediated gene repression, as activity against genomically integrated GFP (using sgGFP, an sgRNA programmed with the previously validated EGFP-NT2 protospacer (Table S1) (Gilbert et al., 2013)) was robust and comparable to that from a previously validated sgRNA expression construct (95.4% and 96.2% repression of GFP fluorescence, respectively) (Figure 1F, Methods) (Gilbert et al., 2014).

A strategy for multiplexed sgRNA delivery to allow simultaneous genetic perturbations

To systematically delineate IRE1α-, PERK-, and ATF6-controlled transcriptional programs and to expand Perturb-seq to the analysis of higher-order genetic interactions, we sought to design a vector that could mediate robust and homogeneous perturbation of gene combinations in individual cells (Figure 2A). Previous efforts to simultaneously express different sgRNAs (for targeting Cas9) have had limited success achieving uniform genetic perturbations across multiple targets (Kabadi et al., 2014; Nissim et al., 2014). In engineering our vector, we first incorporated three tandem sgRNA expression cassettes (composed of an RNA polymerase III promoter, sgRNA protospacer, and sgRNA constant region) into our Perturb-seq vector (Figure 2A). To minimize intramolecular recombination at repetitive nucleotide sequences during lentiviral transduction (Sack et al., 2016; Smyth et al., 2012), we used three different promoters in this initial three-guide vector (Methods). Test vectors expressing sgGFP from one of the promoters (and negative control sgRNAs from the others) partitioned GFP+ K562 cells with dCas9-KRAB into two subpopulations with either strong GFP depletion (>90%) or no detectable depletion (Figure S2A, S2B, Methods). Such incomplete activity could result from a remaining propensity for recombination between the 93-nt sgRNA constant regions or limiting dCas9-KRAB levels when expressing multiple sgRNAs. To test the latter possibility, we generated GFP+ K562 cells with 10-fold higher dCas9-KRAB levels (cMJ006 cells) (Figure S2C). However, GFP depletion remained bimodal when expressing sgGFP from one of our initial three-guide vectors (Figure 2B, Methods).

To solve this problem, we next engineered two modified sgRNA constant regions (cr2 and cr3) that share at most 20 bases of continuous sequence homology with each other and the original constant region (cr1) (Figures 2C, S2D, Table S2, Methods). These constant regions were functional in bacteria and, when paired with the EGFP-NT2 protospacer and expressed from modified mouse (mU6) and human U6 (hU6) promoters, respectively, mediated GFP depletion in K562 cells that was indistinguishable from that of the Perturb-seq vector or sgGFP expressed from a modified bovine U6 (bU6) promoter (Figures 2C, 2D, S2E, Methods). We then designed our final three-guide Perturb-seq vector with the following sgRNA expression cassettes: the bU6 promoter paired with cr1, mU6 with cr2, and hU6 with cr3 (Figures 2A, S2F). Vectors expressing sgRNAs programmed with EGFP-NT2 from any of the three cassettes in this final design mediated near-uniform and strong depletion of GFP (96–97%), nearly identical to that mediated by the Perturb-seq vector (Figure 2E, Methods). Thus, our final three-guide vector can be faithfully delivered by lentiviral transduction and mediates robust knockdown of targeted genes.

Systematic delineation of the three branches of the UPR using Perturb-seq

With these tools in hand, we applied Perturb-seq to explore the branches of the mammalian UPR (Figures 3A, S1C). Using our three-guide Perturb-seq vector, we introduced sgRNAs targeting each UPR sensor gene in all possible single, double, and triple combinations into K562 cells with dCas9-KRAB (Table S1, Methods). Transduced cells were then pooled, sorted for sgRNA delivery, and, after 5 days of total growth, treated with pharmacological inducers of the UPR: thapsigargin, an ER calcium pump inhibitor, or tunicamycin, an inhibitor of *N*-linked glycosylation. Control cells were treated with DMSO. We sequenced transcriptomes of ~15,000 cells (Figure S1D). Critically, across all conditions, we observed >80% depletion of targeted genes (Figure 2F). Throughout, we refer to this experiment as our "UPR epistasis experiment."

We then devised an analytical approach for finding robust features within the data (Figures 3B, S3A). Single-cell RNA-seq data are rich, but intrinsically noisy and of very high-dimension. However, many genes share common regulation, arguing that cellular behavior is intrinsically low-dimensional (Heimberg et al., 2016). This motivates the use of unsupervised dimensionality reduction methods, describing cellular behavior in terms of tens of components rather than thousands of genes.

To uncover this latent low-dimensional behavior in a way that is robust to noise, we developed low rank independent component analysis (LRICA, Methods). We applied recent advances in sparse matrix theory (Candès et al., 2011; Lin et al., 2010) to decompose the observed gene expression matrix (X) into a low-rank matrix (L), representing the low-dimensional dynamics of the population, and a sparse matrix (S), capturing noise and effects that are highly variable between cells:

$$X = L + S$$

We then identify informative trends in the low-dimensional dynamics by applying independent component analysis (ICA, Methods) to the matrix L. The components aid interpretation in two ways: components that are bimodal define subpopulations and, by asking which genes influence a component, we can identify those driving a behavior.

We applied LRICA to our thapsigargin-treated cells. Four components varied across the different perturbations, including three that tracked the presence of PERK, IRE1α, and ATF6 (Methods, Figure S3B). When projected to two dimensions using *t*-distributed stochastic neighbor embedding (t-sne) (Van Der Maaten, 2014), cells bearing a particular perturbation all grouped together, further validating our triple depletion strategy, and biologically reasonable groups of perturbations clustered (Figure 3B). The same analysis applied to the four components that varied across the cell cycle arranged the cells in a circular pattern ordered by cell cycle phase (Figure 3B). Thus, LRICA identified and decomposed the two largest effects causing variation in the population in an unbiased way and computationally decoupled them from each other.

We did observe an interaction between the two effects, apparent in a "bulge" in Figure 3B. Closer analysis showed this interaction was caused by PERK-dependent cell cycle arrest in

G1 caused by thapsigargin treatment (Figure 3C, 3D) (Hamanaka et al., 2005). One component (right panel of Figure 3C) was bimodal among the cells bearing each perturbation. Defining the cells with that component low as “G1 cells” (cf. middle and right panels of Figure 3C), we looked at the top fifty genes influencing the component (Figure 3E) and noted epistatic interactions between PERK-dependent UPR activation and progression through G1. For some genes the two programs cancel each other out, while for others they act synergistically, as in the thapsigargin-induced expression of *MYC* (Liang et al., 2006), which our data show is most strongly associated with the G1-arrested subpopulation (Figure 3E).

We next turned to delineating the three transcriptional programs of the UPR. We identified a set of genes robustly induced by both thapsigargin and tunicamycin treatment and hierarchically clustered them based on their co-expression (Methods). When synthetic bulk RNA-seq profiles (made by averaging all cells containing the same GBC for a given treatment) were ordered according to our clustering, patterns of regulatory control were apparent (Figure 3F). To estimate regulatory overlap, we decomposed the changes across bulk responses using ICA (bottom of Figure 3F, Methods). PERK/ATF4 had the largest regulon in our experiment, with many targets uniquely under its control. ATF6 and IRE1α showed more overlap, consistent with a more common transcriptional regulatory mechanism (Yamamoto et al., 2007). Of the two, IRE1α had more specific targets, notably components of the translocon and translocon auxiliary components (consistent with previous reports (Shoulders et al., 2013)), but ATF6 had stronger activating effects on common targets (Figure 3F). Many genes showed some sensitivity to all branches, particularly a group of very high abundance stress response genes (*HSPA5*, *HERPUD1*, *SDF2L1*). Our experiment thus defined and decoupled the three overlapping branches of the mammalian UPR, both at the bulk level and within single cells.

Genome-scale CRISPRi screens identify genetic perturbations that induce the UPR

We next employed a two-tiered approach to systematically evaluate how UPR transcriptional programs respond to various perturbations. First, we performed two genome-scale CRISPRi screens that identified genes important in maintaining ER homeostasis. For this, we built a K562 cell line (cBA011) that stably carries dCas9-KRAB, an mCherry transcriptional reporter of IRE1α activation (UPRE reporter), and (to control for general effects on gene expression) a constitutively expressed GFP reporter driven by the EF1a promoter (Figure 4A). Importantly, when treated with tunicamycin, these cells demonstrated *XBP1*-dependent mCherry induction (maximally 16-fold), which occurred subsequent to endogenous *XBP1* splicing (Figures 4B, S4A). As expected, we observed no similar induction of GFP.

Using our reporter cell line, we separately screened two genome-scale CRISPRi libraries, our first generation library (CRISPRi-v1), which targets 15,977 genes (20,899 transcriptional start sites, TSSs) with 10 sgRNAs per TSS, and our recently described second-generation library (CRISPRi-v2), which targets 18,905 genes (20,526 TSSs) with 5 sgRNAs per TSS (Figures 4C–D, S4B, S4C, Tables S3–6) (Gilbert et al., 2014; Horlbeck et al., 2016). Briefly, reporter cells (cBA011) transduced with each library were grown for 8 days and then separated into bins according to their ratiometric reporter signal (mCherry/

GFP) by FACS. Cells in the top and bottom thirds of the reporter distribution were collected and processed to measure the frequencies of sgRNAs contained within each, from which we calculated sgRNA and gene-level reporter signal phenotypes. Our CRISPRi-v2 screen identified 397 hit genes with high mCherry/GFP, indicative of UPR activation (Figure 4D, 4E). Importantly, phenotypes were reproducible between replicates and minimal correlation was observed between hit phenotypes and previously calculated gene growth phenotypes (Spearman $R = -0.2$) (Figures S4C, S4D). Of the 141 hits from the CRISPRi-v1 screen, 103 reproduced from screening the CRISPRi-v2 library (Fisher's Exact p-value = 8.97e-138) (Figure S4B).

Among hits from the CRISPRi-v2 screen are well-characterized regulators of protein folding in the ER, most notably *HSPA5*, which encodes the major ER Hsp70 chaperone BiP (Figure 4E). Consistent with results from a similar screen in yeast (Jonikas et al., 2009), our hits featured genes involved in *N*-linked glycosylation, including components of the oligosaccharyltransferase (OST) complex and the dolichol-linked oligosaccharide biosynthesis pathway, ER-associated degradation (ERAD), and protein trafficking. Additionally, genes involved in SRP-mediated protein targeting to the ER were enriched among hits (Fisher's Exact p-value = 2.65e-09). Three out of four subunits of the translocon-associated protein complex (TRAP) scored; and strikingly, among the 7 hits with the strongest phenotypes were all three genes that encode the ER protein-translocation channel or translocon (*SEC61A1*, *SEC61B*, *SEC61G*) (Figures 4D, 4E, S4D). The phenotypes of SRP-targeting factors and the translocon were surprising because recent reports have shown that SRP-mediated recruitment of unspliced *XBP1* (XBP1u) to the ER and IRE1α binding to the translocon are required for maximal XBP1 splicing in response to exogenous stress (Kanda et al., 2016; Plumb et al., 2015). Satisfyingly, targeting of both *ERN1* (IRE1α) and *XBP1* decreased reporter signal in the screen (Figure 4D).

Genes with biological functions not known to be directly related to ER function also scored among hits, some of which are distinct from functional classes seen in the analogous systematic yeast studies (Jonikas et al., 2009). Specifically, sets of genes that control general translation, transcription, and, perhaps most intriguingly, mitochondrial function were enriched among hits (Figures 4E, S4E). While intriguing, these phenotypes alone give us little power to infer mechanisms by which gene repression disrupted ER homeostasis. Additionally, while disruption of these gene functions may impair ER function, it is also possible that such hits represent UPR-independent effects on our reporter system. Individual testing of 257 sgRNAs targeting 152 select hit genes confirmed that a majority induced UPRE reporter signaling; however, some of these sgRNAs, notably ones targeting the mediator transcriptional complex, also reduced GFP levels (Figure 4F).

Perturb-seq of UPR-inducing CRISPRi sub-library reveals functional relationships

Next, to characterize the role of these different gene classes we applied Perturb-seq to a small CRISPRi library of 91 sgRNAs targeting 82 genes, including many of our strongest hits, and 2 negative controls (Figure S4B, Table S1). To test platform scalability, sgRNAs were delivered via pooled transduction using a mixture of separately prepared lentiviruses, and we collected ~65,000 transcriptomes in one large pooled experiment (Figure S1E, S1F,

Methods). Throughout, we refer to this experiment as our “UPR Perturb-seq experiment.” All expected sgRNAs (i.e. GBCs) were detected, with expected and even representation (457 ± 108 cells per sgRNA, mean ± standard deviation).

To explore these data, we first constructed synthetic bulk expression profiles by averaging normalized expression across cells containing each sgRNA (i.e. GBC). Hierarchical clustering of these profiles revealed that sgRNAs targeting the same gene clustered together (Figure S5A). Knockdown was robust, with median 90% depletion of the guide target and similar levels of depletion between sgRNAs with the same target (Figure 5C). Target depletion occurred as a shift in the expression distribution, rather than a bifurcation into perturbed and unperturbed subpopulations (Figure S5B). Indeed, when we computationally split each sgRNA-perturbed subpopulation into most- and least-perturbed (Methods), we observed a median difference in knockdown of 8% (Figure S5C). These findings confirm the ability of CRISPRi to produce uniform knockdown as well as the ability of the barcoding scheme to accurately assign sgRNAs to the appropriate cells. Given the similarity in phenotypes between sgRNAs targeting the same gene, in subsequent analyses we grouped cells by sgRNA target rather than by sgRNA.

The bulk profiles are rich phenotypic fingerprints that identify how different perturbations are related. Hierarchical clustering of profiles revealed gene clusters (boxes on the diagonal in Figure 5A) consistent with known functional and physical interactions, including those composed of genes involved in SRP-mediated protein targeting (*SRP68*/*SRP72* and *SRPRB*/*SRPR*), UFMylation (*UFL1*/*UFM1*/*DDRGK1*), the ubiquitylation reactions of ERAD (*SYVN1*/*SEL1L*), and protein trafficking (*TMED2*/ *TMED10*) (Figure 5A). Perturb-seq can also yield insights at the single-cell level. For example, decomposing the populations by cell-cycle position revealed that perturbation of many aminoacyl tRNA synthetases elicited an accumulation of cells in G2 (Figure 5B).

We next sought to analyze how individual hits effect activation of the different branches of the UPR. We adopted a data-driven strategy and trained random forest regressors to score branch activation using the cells in our UPR epistasis experiment, in which the branches are definitively separated, as training data (Methods). This scoring method performed well and had better accuracy than other metrics (Methods, Figure S5F). Branch activation scores (Figure 5D) showed that hits from the screen activated all three UPR branches with clear correlations in activation among functionally related groups of genes. Intriguingly, different groups elicited differential activation of the three branches. For example, repression of *HSPA5*, which encodes the major ER chaperone BiP, robustly activated all three branches. Repression of aminoacyl tRNA synthetases activated both IRE1α and ATF4 transcriptional programs. Finally, repression of all three subunits of the translocon (*SEC61A1*/*SEC61G*/*SEC61B*) appeared to selectively activate only the IRE1α branch. Comparison with alternate scoring methods and expression of UPR-controlled genes showed good agreement with these calls (Figure S5D, S5E). Thus our data reveal how different genetic perturbations can selectively activate the different branches of the UPR.

Single-cell analysis uncovers a bifurcated response in *HSPA5*-perturbed cells

The above observation raises an immediate question: do the UPR branches also operate independently at the single-cell level? To explore this issue, we examined cells depleted of BiP, where all three branches of the UPR are active.

When compared to unperturbed cells, cells transduced with *HSPA5*-targeting sgRNAs were distinguishable as a distinct population (Figure 6A), and had markedly different patterns of gene expression (~2,100 genes differentially expressed at $P < 0.01$). Using LRICA, we decomposed these differences into 16 independent components. Two of these (IC1 and IC2) varied substantially between control and *HSPA5*-perturbed cells (Figure 6B), and were strongly influenced by UPR-responsive genes. Comparing these hypothesis-free results to the branch activation scores (Figure 6C) showed that our analysis pipeline had independently discovered a subpopulation structure with differential activation of the UPR branches within *HSPA5*-perturbed cells. Indeed, when we ordered the cells by the value of IC1 and examined the expression of UPR-induced genes (as defined in Figure 3F), the trends defining these subpopulations were apparent (Figure 6D).

Of particular note was the switch-like induction of the PERK/ATF4 regulon, revealing that these cells represented a discrete subpopulation. These differences did not reflect levels of BiP depletion, as the subpopulations with IC1 low and high (Figure 6B, 6D) had equal expression of *HSPA5* (Figure 6E). However, the PERK/ATF4-induced subpopulation did have an altered cell cycle, with many cells accumulating in G2 (Figure 6F). These results reveal that the UPR can be executed in markedly different ways within an apparently homogeneous population.

Gene-gene covariance analysis of Perturb-seq data reveals transcriptional regulons

Figure 6D underscores a key point: correlated up- or down-regulation of genes can be a signature of shared regulation. As perturbations elicit coordinated changes, we reasoned that Perturb-seq could help identify related genes (Figure 6G) (Klein et al., 2015).

For example, we identified 200 genes induced in our UPR Perturb-seq experiment (Methods), and when clustered based on co-expression, functional groups appeared, including all three UPR branches (Figures 6H, S6A). Moreover, when we clustered UPR-induced genes (from Figure 3F) using co-expression in either the UPR epistasis experiment or the UPR Perturb-seq experiment, we obtained similar results (cophenetic correlation 0.81, compared to 0.13 when control cells were used) (Figures 6I, S6B, Methods). This similarity suggests that the organization of the UPR is similar between commonly used strong chemical perturbants and the more varied genetic perturbations used here.

We finally investigated a “fishing” strategy to further enhance weak correlations (Figure 6J, 6K, Methods). Our initial analysis (Figure 6H) identified 5 cholesterol biosynthesis genes with correlated expression. When we confined our gene clustering analysis to the ~9,000 cells most perturbed for these genes, we saw strengthened correlations and the emergence of a larger cluster of cholesterol biosynthesis genes grouping together (Figure 6K). Though these demonstrations are not proof, they suggest that correlation information from Perturb-seq may enable automated functional clustering of genes of unknown function.

A homeostatic feedback loop between the translocon and the IRE1α branch of the UPR

Among genes targeted in the UPR Perturb-seq experiment, *SEC61A1*, *SEC61G*, and *SEC61B* were perhaps the most intriguing outliers. Repression of each of these displayed a marked preference for activation of the IRE1α branch with little or no activation of the other branches (Figures 5D, 7A, 7B, S5E, S7A). To confirm that our single-cell data were accurately calling IRE1α activation, we directly probed for *XBP1* splicing. Targeting all three translocon subunits induced *XBP1* splicing at levels consistent with the single-cell data and to a degree at or above that provoked by targeting *HSPA5*, whose depletion induces all three branches of the UPR (Figures 5D, 6C, 6D, 7C, Methods). Additionally, repression of *SEC61A1* and *SEC61B* led to sustained *XBP1* splicing and upregulation of *SSR2*, a translocon auxiliary protein and strongly selective target of IRE1α (Figures 3F, 7D, S7B, Methods). These results were in contrast to transient *XBP1* splicing caused by chemical stress, which diminished on the scale of hours, consistent with previous reports (Figure S4A) (Lin et al., 2007). We note that *SEC61B* appears to share a co-regulated promoter region with *ALG2*, a gene that functions in *N*-linked glycosylation, and as such, we cannot formally separate the effects of repressing these genes (Figure S7B). Nonetheless, the consistent phenotypes from targeting *SEC61A1*, *SEC61B*, and *SEC61G* suggest that translocon depletion elicits selective activation of the IRE1α branch.

To further investigate branch selectivity, we evaluated induction of *CHOP,* also called *DDIT3* and a selective target of PERK/ATF4, after *SEC61A1* and *SEC61B* repression (Figures 3F, 7D, 7E, Methods). Repression of *SEC61B* showed little to no *CHOP* induction. We observed a limited increase in *CHOP* expression in response to *SEC61A1* repression but at lower levels than in cells transduced with an *HSPA5*-targeting sgRNA, and we reason that this could reflect general toxicity. Indeed, *SEC61A1* is an essential gene, perturbation of which, unlike *SEC61B,* caused strong growth phenotypes in both CRISPRi and CRISPR cutting cell viability screens (Figures S4D, S7C) (Gilbert et al., 2014; Wang et al., 2015). An alternative explanation for apparent IRE1α branch selectivity, other than selective activation, is the possibility that general stress caused by translocon loss impairs only the other two branches of the UPR. However, we observed *CHOP* upregulation in response to exogenous ER stress induced by thapsigargin treatment in cells transduced with *SEC61A1*-, *SEC61B*-, or *SEC61G*-targeting sgRNAs (Figure 7E, Methods).

Cumulatively, our data suggest a selective role for the IRE1α branch of the UPR in monitoring translocon availability. Many of the strongest and most selective IRE1α transcriptional targets in the UPR epistasis experiment were translocon subunits and translocon-associated genes (Figure 3F). Conversely, *SEC61A1*, *SEC61G*, and *SEC61B* were among the strongest hits in our unbiased genome-wide screen for IRE1α activation (Figure 4D, 4E) and repression of these genes showed preferential IRE1α pathway activation at the level of single cells (Figure 5D, 7A, 7B, S5E, S7A). Moreover, by RT qPCR analysis, we confirmed reciprocal upregulation of these genes in response to *SEC61A1* or *SEC61B* repression (Figure S7B). These results suggest a model in which IRE1α actively monitors the number of translocons (and perhaps function) and increases them as needed (Figure 7F).

DISCUSSION

We present Perturb-seq, a platform for multiplexed profiling of perturbations with single-cell resolution, and used it to systematically dissect the mammalian UPR. Though we focused on CRISPRi, the same approach can be used to encode a wide range of perturbations, such as CRISPR cutting-mediated loss of function, gene activation, or targeted mutation (Boettcher and McManus, 2015; Komor et al., 2016). We have shown that CRISPRi can give strong, homogeneous, and simultaneous depletion of up to three targets and enables the study of essential genes. As depletion can be observed in the RNA-seq data, performance and quality of GBC identification can be directly assessed. It also has advantages when scaling to high-order combinations relative to CRISPR cutting, as genetic variability during indel formation and non-specific toxicity due to DNA cutting both increase with the number of cut sites (Boettcher and McManus, 2015; Horlbeck et al., 2016; Wang et al., 2015).

Scaling Perturb-seq to genome-scale requires overcoming some obstacles, but none appear intractable. Current techniques (Zheng et al., 2016) already allow RNA to be collected from ~50,000 cells in ~10 min, and our GBCs enable higher loading through computational removal of cell doublets. Cost per cell will decline as technologies mature, and sequencing costs can be mitigated through amplification of select targets (like our guide-mapping amplicons) or depletion of uninteresting high abundance genes (Gu et al., 2016). A more subtle point is that intermolecular provirus recombination during transduction can scramble barcode identities in pooled lentivirus preparations (Sack et al., 2016). We took careful steps to avoid this problem and expect that straightforward protocol alterations will circumvent this issue.

By far the biggest barrier we anticipate is on the analytical side. Perturb-seq generates massive amounts of intrinsically noisy data. We made some progress, using single-cell data to decouple the branches of the UPR, uncover subtle subpopulations within cells of the same type, and infer programs of gene expression using correlated expression. Along with previous successes (Jaitin et al., 2014; Klein et al., 2015; Macosko et al., 2015), and other novel analytical approaches (Dixit et al., co-submitted manuscript), large-scale analyses of single cell behavior should enable systematic understanding of the complex regulatory programs at work within cells.

Our experiments also provide insights into how the mammalian UPR senses and responds to the diverse challenges faced by the ER. A central question is why metazoan cells have evolved three independent and mechanistically distinct sensors of protein misfolding. As expected from previous work (Acosta-Alvear et al., 2007; Han et al., 2013; Lee et al., 2003; Shoulders et al., 2013), epistasis analysis using combinatorial depletions of PERK, ATF6, and IRE1α revealed both distinct and overlapping programs of gene expression. One of our main observations is that these branches nevertheless can operate independently, both at the bulk and single-cell levels.

Our genome-wide screens identified diverse genetic perturbations that activate IRE1α signaling, including some categories not expected from analogous yeast screens (Jonikas et al., 2009). Subjecting these hits to Perturb-seq showed that the screen in fact captured all

three branches of the UPR, and that genes with similar functional roles induced the UPR in similar ways. The remarkable bifurcation in behavior we observed in cells depleted of BiP illustrated the utility of single-cell data: bulk RNA-seq would in this case describe a state that no cell actually occupies. As all cells were treated identically, the cause of such marked differences remains in question.

Perhaps the most intriguing example of branch specificity was our observation that depletion of translocon subunits led to selective activation of the IRE1α branch, which is notable in light of recent studies suggesting that IRE1α, unlike ATF6 or PERK, acts in physical association with the translocon (Plumb et al., 2015). Given that we, in agreement with others (Shoulders et al., 2013), observed regulation of translocon expression to be uniquely under IRE1α control, this suggests a feedback model in which IRE1α monitors the state of translocation. Isolated IRE1α induction would enable repair to or upregulation of the translocation machinery without broader UPR induction, potentially forestalling responses such as cell death.

Our study of the mammalian UPR serves as a blueprint for the study of complex and overlapping transcriptional networks, in which a primary genome-wide screen serves as the input to more detailed analysis via Perturb-seq. Our success here and the parallel success in understanding dendritic cell activation (Dixit et al., co-submitted manuscript) speak well to the potential of the Perturb-seq approach to become a standard strategy for understanding regulatory interactions in the cell.

STAR METHODS

Key Resources Table

See separate file.

Contact for Reagent and Resource Sharing

Requests for further information and resources may be directed to Jonathan S. Weissman (Jonathan.Weissman@ucsf.edu).

Method Details

Plasmid design and construction—The “Perturb-seq vector” backbone (pBA439, Addgene, Cat#85967) was derived from a previously described CRISPRi vector (herein referred to as the “original sgRNA expression vector”) (Addgene, Cat#60955). To construct pBA439, the mU6-sgRNA-EF1a-PURO-BFP region from this parental vector and a BGH polyadenylation sequence amplified by PCR from pcDNA3.1(+) (Invitrogen, V790-20) were inserted in reverse origination between the XbaI and EcoRI sites of the parental. A random 18-nt barcode was then inserted between the BFP and BGH polyA sequences (using subsequently disrupted EcoRI and AvrII sites) by Gibson assembly to construct the “Perturb-seq GBC library” (pBA571, Addgene, Cat#85968). This library was prepared with an estimated barcode diversity of >100,000 essentially as previously described (Kampmann et al., 2014). Guide RNA protospacer sequences were individually cloned into both the original sgRNA expression vector and the pBA571 library (between the BstXI and BlpI sites) by

ligation. Each vector was then verified by Sanger sequencing of the protospacer and, if applicable, its corresponding barcode. Final Perturb-seq vectors containing barcodes that introduced the conserved polyadenylation signal AATAAA were discarded. To construct pMH0001 (Addgene, Cat#85969), a minimal ubiquitous chromatin opening element (UCOE) (Müller-Kuller et al., 2015) was inserted upstream of the SFFV promoter in the lentiviral dCas9-KRAB expression vector (pHR-SFFV-dCas9-BFP-KRAB, Addgene, Cat#46911). Throughout this manuscript, the term dCas9-KRAB is frequently used to indicate the dCas9-BFP-KRAB construct and corresponding fusion protein.

Three-guide expression vectors were assembled by a two-step cloning procedure (Figure S2F). First, complementary oligonucleotides (Integrated DNA Technologies) containing the protospacer sequence and ligation overhangs were annealed and ligated into BstXI/BlpI-digested “one-guide Perturb-seq vector” backbones (pMJ114, Addgene, Cat#85995; pMJ179, Addgene, Cat#85996; pMJ117, Addgene, Cat#85997). These one-guide Perturb-seq vectors each contained specific primer binding sites flanking the sgRNA expression cassette for PCR amplification. Three-guide expression cassettes were then assembled from PCR-amplified single cassettes and inserted into HpaI/XhoI-digested pBA571 (Perturb-seq GBC library) by a single four-piece Gibson assembly step. Resulting vectors were clonally isolated and then sequence verified as described above. Our initial three-guide expression vectors (“initial three-guide vectors”) were assembled from one-guide expression cassettes that contained a modified mouse U6 promoter (mU6), a modified human U6 promoter (hU6), and a modified human 7SK promoter (h7SK). These were ordered hU6, mU6, h7SK (5’ to 3’ relative to lentiviral transcription). However, we found that the h7SK promoter generally performed poorly in the context of our Perturb-seq vector design (Figure S2A, S2B). Therefore, various U6 promoter sequences were tested for use in our final three-guide vector design (“final three-guide vector” or “final three-guide Perturb-seq vector”) (Figures 2A, S2E). For testing of U6 promoters, U6 promoters from cow (bU6-2, GenBank DQ150531 and bU6-3, GenBank DQ150532), sheep (sU6-1, GenBank HM641427 and sU6-2, GenBank HM641426), buffalo (buU6, GenBank JN417659), and pig (pU6, GenBank EU520423) spanning ~300–500 bp upstream of the TSS, modified to contain a BstXI site at the TSS, and fused to both the EGFP-NT2 (Table S1) protospacer and the original constant region (cr1) (Table S2) (Gilbert et al., 2013; Gilbert et al., 2014) were obtained as synthetic DNA segments (Integrated DNA technologies). These were inserted into HpaI/XhoI-digested pBA439 by Gibson assembly. The modified bovine U6-2 promoter (bU6) was used instead of h7SK in our final three-guide vector design (Figures 2A, S2F). For testing of constant region variants in K562 cells, constant region variants fused to the EGFP-NT2 protospacer or a negative control protospacer were PCR-amplified and inserted into BstXI/XhoI-digested pBA439 or one-guide Perturb-seq vectors by Gibson assembly.

For final three-guide Perturb-seq vectors targeting the UPR branches, the bU6, mU6, and hU6 cassettes (containing the cr1, cr2, and cr3 constant regions, respectively) were designed to either express an sgRNA targeting *ATF6*, *EIF2AK3* (PERK), or *ERN1* (IRE1α), respectively, or a non-targeting negative control sgRNA. The following protospacer sequences were used: *ATF6*-targeting, gGGGATCTGAGAATGTACCA; *EIF2AK3*-targeting, gCGGGCTGAGACGTGGCCAG; *ERN1*-targeting, gAGAACTGACTAGGCAGCGG; non-targeting sgRNA in bU6 cassette,

gACGACTAGTTAGGCGTGTA; non-targeting sgRNA in mU6 cassette, gGCCAAACGTGCCCTGACGG; non-targeting sgRNA in hU6 cassette, gCCTTGGCTAAACCGCTCCC (Table S1).

The UPRE reporter was built into a backbone for lentiviral expression that has been previously described (Addgene, Cat#44012). This parental vector was digested with AgeI and religated to remove unwanted functional cassettes, and the UPRE promoter region or EF1a promoter were inserted between the BamHI and XhoI site of the resulting product. The UPRE promoter region contains 5 UPR elements (UPREs, 5′-TGACGTGG-3′) upstream of the *c-fos* minimal promoter (−53 to +45 of the human *c-fos* promoter) (Wang et al., 2000). Lastly, mCherry (mCh) and sfGFP were cloned adjacent to UPRE and EF1a promoters, respectively (into an HpaI site). The resulting vectors are pBA407 (UPRE-mCh-Ubc-Neo, Addgene, Cat#85970) and pBA409 (EF1a-sfGFP-Ubc-Neo, Addgene, Cat#85971).

Cell culture, DNA transfections, viral production, and construction of reporter cell lines—K562 cells were grown in 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. HEK293T cells were grown in Dulbecco's modified eagle medium (DMEM) in 10% FBS, 100 units/mL penicillin and 100 μg/mL streptomycin. Cells were treated with tunicamycin or thapsigargin (Sigma, T9033) solubilized in DMSO. Lentivirus was produced by transfecting HEK293T with standard packaging vectors using *Trans*IT®-LTI Transfection Reagent (Mirus, MIR 2306). Viral supernatant was harvested at least ~2 days after transfection and filtered through a PVDF syringe filter and/or frozen prior to infection.

To construct the UPRE reporter cell line, K562 cells stably expressing dCas9-KRAB (Gilbert et al., 2014), originally constructed from K562 cells obtained from ATCC 536 (RRID:CVCL_0004), were stably transduced with pBA407 and selected in media supplemented with 500 μg/mL Geneticin (Gibco, 10131-035). The clonal line cBA010 was then selected by limiting dilution. cBA011 is a derivative of cBA010 containing pBA409. cBA011 was made by stable transduction and selection of GFP positive cells using fluorescence activated cell sorting on a BD FACSAria2. Separately, the GFP+ K562 dCas9-KRAB cell line (also referred to as GFP+ K562 with dCas9-KRAB) was constructed by infecting K562 cells stably expressing dCas9-KRAB with a Murine Stem Cell Virus (MSCV) retrovirus that carries GFP under the control of the SV40 promoter. MSCV retrovirus was produced by transfecting amphotropic Phoenix packaging cell lines with standard packaging vectors. K562 cells stably expressing GFP were then sorted to purity by flow cytometry using a BD FACSAria2. These cells were generated for testing CRISPRi-mediated gene depletion from new sgRNA expression vectors (described below), and use of this cell line is denoted in figures with the label "low dCas9-K562." To construct the GFP+ K562 UCOE-dCas9-KRAB cell line (cMJ006), GFP+ K562 dCas9-KRAB cells were transduced with pMH0001 at a multiplicity of infection of ~3. Use of this cell line for testing CRISPRi-mediated gene depletion is denoted in figures with the label "high dCas9-K562." Transduced cells were sorted for BFP expression (top 33%) by flow cytometry on a BD FACSAria2. BFP fluorescence was monitored for several generations and found to be stable.

Design and cloning of constant region variants for testing in *E. coli*—Bases in the original sgRNA constant region (cr1, see Table S2) were selected for mutation by inspection of the crystal structure of Cas9 bound to guide RNA and target DNA (PDB ID code 4OO8 (Nishimasu et al., 2014)) (Figures 2C, S2D). Bases that did not form direct contacts with Cas9 or with other nucleotides of the constant region were deemed amenable for mutation. If applicable, sequence conservation patterns of the base in crRNAs/tracrRNAs of *Streptococcus* species were used to determine the type of mutation. In this fashion, 15 constant region variants with mutations in different parts of the constant region were designed (Figure 2C, Table S2). The most diverse constant region variants cr2 and cr3 were designed by combining multiple individual mutations (Figure 2C, Table S2).

To rapidly assess the activity of the variant constant regions, the variants were fused to an mRFP-targeting protospacer (mRFP-NT1, sequence AACTTTCAGTTTAGCGGTCT) (Qi et al., 2013) and tested in an *E. coli* CRISPRi reporter strain for knockdown of mRFP (described below). To eliminate variability from copy number variation, sgRNA sequences were cloned into a plasmid for site-specific integration into the *E. coli* genome at *attL* and expressed from single copy from an IPTG-inducible $P_{LlacO-1}$ promoter. To construct the integrating sgRNA expression plasmid, an sgRNA expression cassette was PCR-amplified from pgRNA-bacteria (Addgene, Cat#44251), modified to be flanked by strong synthetic terminators, and inserted into pCAH63 (Haldimann and Wanner, 2001) at the ClaI/NheI sites. The constitutive promoter from pgRNA-bacteria was replaced with the IPTG-inducible $P_{LlacO-1}$ promoter, generating pCs-550r. Then, pCs-550r was further modified to include the constant region used in mammalian CRISPRi (cr1) (Gilbert et al., 2014), PCR-amplified with an mRFP-targeting protospacer and inserted into pCs-550r at the SpeI and KpnI sites to generate pMJ020. Finally, constant region variants 1–15 as well as cr2 and cr3 were cloned into pMJ020 by inverse PCR with mutations encoded in primer overhangs, by site-directed mutagenesis following standard procedures, or by insertion of a synthetic DNA segment encoding the constant region (Integrated DNA Technologies) into SpeI/KpnI-digested pMJ020 by Gibson assembly.

Construction of *E. coli* CRISPRi reporter strain and testing of constant region variants—The *E. coli* CRISPRi reporter strain was constructed by sequential insertion of a construct for IPTG-inducible expression of dCas9, a construct for constitutive expression of mRFP, and a construct for IPTG-inducible guide RNA expression (described above) into the *E. coli* genome. First, a *lacIq*-t0-$P_{LlacO-1}$-*dCas9* cassette (*lacIq* for strong expression of the Lac repressor; t0, a transcription terminator; $P_{LlacO-1}$-*dCas9*, for IPTG-inducible expression of *S. pyogenes* D10A/H840A *Cas9* (dCas9)) was inserted into the chromosome of *E. coli* BW25113 at +19 *attL* via lambda Red recombinase-mediated recombineering following established protocols. Then, a *nfsA::mRFP-kan* cassette for expression of mRFP from the J23119 promoter, a strong synthetic constitutive promoter from the Anderson promoter collection (http://parts.igem.org/Promoters/Catalog/Anderson), was inserted into an *E. coli* MG1655-derived strain by lambda Red recombinase-mediated recombineering as described previously (Qi et al., 2013), and moved from the MG1655-derived strain into the dCas9-expressing BW25113 strain by P1 transduction and selection on kanamycin following established protocols. Plasmids for expression of mRFP-NT1 with the different constant

region variants were integrated into the dCas9- and mRFP-expressing strain at *attL* using the helper plasmid pINT-ts (Haldimann and Wanner, 2001), selecting for chloramphenicol resistance.

Single colonies of strains with the integrated guide RNA expression plasmids were inoculated into LB and grown overnight in deep 96-well blocks at 37 °C with shaking at 900 rpm. Stationary-phase cultures were back-diluted 1:30 and grown into mid-exponential phase, at which point they were back-diluted 1:10000 into LB with 1 mM IPTG for induction of sgRNA and dCas9 expression. Induced cultures were grown at 37 °C with shaking until $OD_{600\ nm}$ reached ~0.4–0.7 (~5 hrs), at which point they were diluted 1:30 in PBS in a 96-well plate. RFP fluorescence was recorded on a LSR-II flow cytometer (BD Biosciences) equipped with a 96-well high-throughput sampler. Each experiment was carried out using three individual colonies for each constant region variant. RFP levels were normalized to those of a strain expressing a non-targeting sgRNA. Almost all constant region variants including cr2 and cr3 retained strong CRISPRi activity as indicated by a 97–99% reduction in mRFP levels in these assays suggesting that the introduced mutations do not disrupt sgRNA:Cas9 binding (Figure 2C).

Testing of sgRNA expression vectors in K562 cells—Vectors for sgRNA expression were transduced into GFP+ K562 dCas9-KRAB cells or GFP+ K562 UCOE-dCas9-KRAB cells (cMJ006) (both described above) at an MOI of 0.1–0.5. For all experiments using GFP + K562 UCOE-dCas9-KRAB, transduced cells were allowed to recover for 2 days, then selected to purity using 2 μg/mL puromycin for 3 days, and allowed to recover for another 2 days before GFP levels were recorded by flow cytometry on a LSR-II flow cytometer (BD Biosciences). For experiments involving only GFP+ K562 dCas9-KRAB cells, cells were grown out for 8–11 days after transduction and GFP levels were recorded by flow cytometry, using BFP expression to gate for transduced cells. Flow cytometry data were analyzed using FlowCytometryTools v0.4.5 (http://eyurtsev.github.io/FlowCytometryTools/). For plotting, flow cytometry events were normalized to population size and the histograms were smoothened by kernel density estimation. For estimating knockdowns, GFP levels from normal (GFP-) K562 cells were subtracted. Experimental details relevant to specific figures in the main text are included below. Similar experimental details related to supplemental figures can be found in the corresponding supplemental figure legends.

Related to Figure 1F: GFP+ K562 dCas9-KRAB cells were transduced with the indicated sgRNA expression vectors carrying either sgGFP (programmed with the GFP-targeting protospacer EGFP-NT2) or a negative control. GFP expression was evaluated 11 days later. Untransduced GFP-K562 cells were also evaluated to determine background florescence. Data are representative of three independent experiments.

Related to Figure 2B: GFP+ K562 UCOE-dCas9-KRAB cells (cMJ006, described above) were transduced with the indicated sgRNA expression vectors and evaluated for GFP expression after 7 days. Data are representative of two independent experiments.

Related to Figure 2D: GFP+ K562 dCas9-KRAB cells were transduced with the indicated sgRNA expression vectors and evaluated for GFP expression after 10 days. In this

experiment, we compared GFP depletion from 4 different sgRNA expression vectors using sgRNAs programmed with the EGFP-NT2 protospacer and fused to 3 different constant region variants (cr1, cr2, and cr3). These were the Perturb-seq vector (Figure 1B) with sgGFP (EGFP-NT2_cr1), a one-guide vector (described above, Figure S2F) with sgGFP under control of bU6, a one-guide vector with EGFP-NT2_cr2 under control of mU6, and a one-guide vector with EGFP-NT2_cr3 under control of hU6. Data are representative of two independent experiments.

Related to Figure 2E: GFP+ K562 UCOE-dCas9-KRAB cells were transduced with the indicated sgRNA expression constructs and evaluated as in Figure 2B. The Perturb-seq vector trace is the same as in Figure 2B; other traces are from distinct samples processed alongside. Here we compared GFP depletion from 4 different sgRNA expression vectors using sgRNAs programmed with the EGFP-NT2 protospacer and fused to 3 different constant regions variants (cr1, cr2, and cr3). These were the Perturb-seq vector with sgGFP and 3 final three-guide Perturb-seq vectors expressing an EGFP-NT2 programmed sgRNA from the indicated promoter/position with two different control sgRNAs expressed from the other promoters/positions. We also evaluated a three-guide Perturb-seq vector expressing three control sgRNAs as a negative control. Data are representative of two independent experiments.

Perturb-seq screening—For schematics of Perturb-seq experiments, see Figure S1A, S1C, S1E. Viruses were individually packaged (using sequence-verified lentiviral Perturb-seq vectors or final three-guide Perturb-seq vectors) and harvested in preparation for Perturb-seq screening. Individual packaging of the lentivirus and pooling at the step of virus or cells was done to avoid intermolecular recombination of proviral genomes and to ensure maintenance of paired barcode-sgRNA coupling (Sack et al., 2016). For the “pilot experiment” (schematic in Figure S1A, data represented in Figures 1C–E, S1B) cBA010 cells were individually spinfected with virus (at 33°C for 2 hours at 1000×*g*) in media supplemented with 8 μg/mL polybrene; 5 hours post spinfection, virus was removed by centrifugation and cells were resuspended in fresh media. Three days later, a transduction efficiency of 20–30%, as determined by percentage of BFP positive (BFP+) cells, was measured by flow cytometry and cells were pooled with equal numbers of sgRNA-containing (BFP+) cells, except cells transduced with a negative control sgRNA were included in the pool at 3-fold coverage. Pooled cells were then grown in the presence of puromycin (3 μg/mL) for 5 additional days. Seven days post transduction cells were sorted on a BD FACSAria2 to near purity and eight days post transduction the sorted cells were separated into droplet emulsion using the Chromium™ Single Cell 3′ Solution according to manufacturer’s instructions (10X Genomics).

For the “UPR epistasis experiment” (schematic in Figure S1C, data represented in Figures 3, 6H, 6I, S1D, S3B, S5D, S5F, S6), seven three-guide vectors (“final three-guide Perturb-seq vector” design) targeting every possible combination of ATF6, ERN1 (IRE1α), and EIF2AK3 (PERK) as well as two independent final three-guide Perturb-seq vectors with three negative control sgRNAs and different barcodes were individually packaged into lentiviruses. Freshly produced (i.e. not frozen) lentiviruses were then spinfected into cBA010 cells (at 33°C for 2 hours at 1000×*g*) in media supplemented with 8 μg/mL

polybrene. The virus was removed by centrifugation and cells were resuspended in fresh media. Three days after infection, transduction efficiencies of 5–10% were measured by flow cytometry. Cells were combined into a pool with equal numbers of transduced (BFP+) cells for each vector (resulting in 2-fold excess of negative control vectors) and the combined cells were then sorted on a BD FACSAria2 to near purity. To limit heterogenous effects of cell microenvironments caused by cell settling, the sorted cells were grown with continuous agitation on an orbital shaker. Five days after infection, the pooled and sorted cells were split into three populations, which were treated as follows: 1) DMSO control treatment for 6 hours; 2) treatment with 4 μg/mL tunicamycin for 6 hours; and 3) treatment with 100 nM thapsigargin for 4 hours. At the end of the treatment, the cells were separated into droplet emulsion using the Chromium™ Single Cell 3′ Solution according to manufacturer's instructions (10X Genomics). Cells loaded onto the device were 90.4%, 87.9%, and 85.3% viable for the different treatment conditions, respectively.

For the large-scale Perturb-seq screen of UPR-inducing sgRNAs (the "UPR Perturb-seq experiment;" schematic in Figure S1E, data represented in Figures 5, 6, 7A, 7B, S1F, S5A–E, S6, S7A), viruses were individually titered by test infections into cBA011 cells and then pooled. To account for varied effects on cell viability across the sgRNA sub-library and minimize cell number difference at final evaluation, pooling titers were determined by the percentage of BFP+ cells remaining 6 days post transduction. Two negative control sgRNAs were included, NegCtrl-2 and NegCtrl-3. NegCtrl-2 and select sgRNAs (those encoded by pDS002, pDS017, pDS026, pDS032, pDS033, pDS052, pDS088, pDS091, pDS160, pDS186; see Table S1) were included at higher representation within the lentivirus pool, 8-fold and 2-fold, respectively. The lentivirus library pool was then used to infect cBA010 cells (performed by spinfection at 33°C for 3 hours at 1000×*g*) so that a single pooled cell population with all perturbations would be carried though subsequent steps. Post centrifugation, cells were immediately removed from virus and transferred to a spinner flask for growth in fresh media. Three days later, a transduction efficiency of 15% was measured by flow cytometry and BFP+ cells were sorted to near purity on a BD FACSAria2. To limit heterogenous effects of cell microenvironments caused by cell settling, the sorted cells were grown with continuous agitation on an orbital shaker. Approximately 7 days post transduction, cells were separated into droplet emulsion using the Chromium™ Single Cell 3′ Solution across two separate runs totaling 10 lanes on the device according to manufacturer's instructions (10X Genomics). Cells loaded onto the device were 92% BFP+ and 93–94% viable, as determined by flow cytometry.

For all Perturb-seq experiments, single-cell RNA-seq libraries were prepared according to the Single Cell 3′ Reagent Kits User Guide (10X Genomics). However, this protocol produces libraries that are not compatible with analysis on the HiSeq 4000 Sequencing System (Illumina) due to the presence of unique byproducts. To remove this issue, we implemented a short, post-preparation library cleanup protocol. Specifically, 120–200 ng of library material was split into parallel PCR reactions containing 0.3 μM each of the Illumina P5 and P7 primers, and amplified using Kapa HiFi ReadyMix according to the following protocol: (1) 95°C for 80 seconds, (2) 98°C for 20 seconds, then 65°C for 30 seconds, then 72°C for 20 seconds (6 cycles), (3) 72°C for 1 minute. PCR products were then SPRI-purified at 1X ratio, re-pooled during elution, and then fragments of length 350–525 bp were

selected using the BluePippin (Sage Science). For the UPR epistasis experiment, the library for each drug condition was sequenced using two HiSeq 4000 lanes. For the UPR Perturb-seq experiment, each of the 10 Chromium libraries was sequenced using 1.5 HiSeq 4000 lanes (one dedicated lane each plus half of a lane shared with another library). Our initial pilot experiment was sequenced using a single HiSeq 2500 Rapid Run.

Specific amplification of guide barcodes—Parallel PCR reactions were constructed containing 30 ng of final library as template, 0.6 μM PTMN050-P7 (CAAGCAGAAGACGGCATACGAGAT), and 0.6 μM barcoded PTMN051 (AATGATACGGCGACCACCGAGATCTACAC [ILLUMINA S513–S522 INDEX] TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACCTCCCTAGCAAACTGGG GCACAAG), and amplified using Kapa HiFi ReadyMix according to the following PCR protocol: (1) 95°C for 3 minutes, (2) 98°C for 15 seconds, then 70°C for 10 seconds (14–16 cycles). Reactions were re-pooled during 0.8X SPRI selection, and then fragments of length 350–425 were selected using the BluePippin. Guide barcode libraries were sequenced either as spike-ins alongside the parent RNA-seq libraries (pilot experiment and UPR epistasis experiment) or using half of a separate HiSeq 2500 rapid run (UPR Perturb-seq experiment).

Genome-scale CRISPRi screening—Reporter screens were conducted using protocols similar to those previously described (PMID:) (Gilbert et al., 2014; Horlbeck et al., 2016; Sidrauski et al., 2015). The CRISPRi-v1 (Addgene, Cat#62217) or the compact (5 sgRNA/gene) CRISPRi-v2 (Addgene, Cat#83969) sgRNA libraries were transduced into cBA011 cells at an MOI < 1 (BFP+ cell percentages were ~45% and 26%, respectively). For the CRISPRi-v1 screen, cells were grown in spinner flasks for 2 days without selection, followed by 3 days of selection with 1 μg/mL puromycin. Screen replicates were split post infection and carried separately throughout the remainder of the experiment. One replicate arm of the CRISPRi-v1 screen was carried with media supplemented with 88–150nM ISRIB throughout, although differences observed between the replicates were negligible (Table S3). For the CRISPRi-v2 screen, cells were grown in spinner flasks for 2 days without selection, followed by 5 days of selection with 1–3 μg/mL puromycin. Screen replicates were split into separate spinner flasks on day 3. For both screens, cells were separated into those with the highest (~28–33%) and lowest (~30–35%) mCherry/GFP ratio 8 days post transduction by fluorescence-activated cell sorting (FACS). Cell pellets were frozen after collection. Approximately 23–30 million cells were collected per bin during screening of the CRISPRi-v1 library (a representation of ~450) and 19–22 million cells per bin for CRISPRi-v2 (a representation of ~600). Practically, the more compact CRISPRi-v2 library allowed us to maintain higher screen representation through the flow cytometer with similar sorting times. Genomic DNA was isolated from frozen cells and the sgRNA-encoded regions were enriched, amplified, and prepared for sequencing. CRISPRi-v2 samples were sequenced with greater coverage.

Sequenced protospacer sequences were aligned and data were processed as described (Gilbert et al., 2014; Horlbeck et al., 2016) with custom Python scripts (ScreenProcessing, available at https://github.com/mhorlbeck/ScreenProcessing). Reporter phenotypes for library sgRNAs were calculated as the $\log_2$ enrichment of sgRNA sequences identified

within the high mCherry/GFP cells over the low mCherry/GFP cells (Table S6). Phenotypes for each transcription start site ("Gene reporter phenotypes") were then calculated as the average reporter phenotype of the 3 sgRNAs with the strongest phenotype by absolute value (most active sgRNAs). Mann-Whitney test p-values were calculated by comparing all sgRNAs targeting a given TSS to the full set of negative control sgRNAs. For data presented in Figures 4D, 4E, S4B and S4D, genes with multiple targeted TSSs were collapsed such that only the TSS with the lowest p-value was used (Table S4). Screen hits were defined as those genes (or separately those TSSs from all targeted, Table S5) with a discriminant score, defined as the absolute value of a gene reporter phenotype over the standard deviation of all gene reporter phenotypes multiplied by the $\log_{10}$ of the Mann-Whitney p-value for each candidate, greater than 7. Growth screen data in Figure S4D and S7C has been reported elsewhere (Horlbeck et al., 2016), except in Figure S7C data from a second, unreported screen was also used. This second screen was conducted in parallel to the first and as described (Horlbeck et al., 2016). Gene ontology analysis was conducting using select databases (GOTERM_BP_FAT, GOTERM_CC_FAT, GOTERM_MF_FAT, KEGG_PATHWAY) and hits (calculated from all TSSs, Table S5) with a phenotype of greater than 1 using DAVID Bioinformatic Resources 6.8 Beta (https://davidd.ncifcrf.gov/) (Huang et al., 2009). Biological classifications reported in Figure 4E and 4F were manually assembled from the literature and using resources from the HUGO Gene Nomenclature Committee (www.genenames.org), AmiGO, the GO Consortium's annotation and ontology toolkit (Carbon et al., 2009) (http://amigo.geneontology.org), DAVID Bioinformatic Resources (https://david.ncifcrf.gov) (Huang et al., 2009) (Table S7).

Individual evaluation of sgRNA reporter phenotypes—Viruses were individually packaged, harvested, and frozen (described above). UPRE reporter cells (cBA011) were separately transduced with targeting sgRNAs and negative controls. In parallel, parental K562 cells with dCas9-KRAB (Gilbert et al., 2014) were transduced with negative controls. Medians of mCherry (from the UPRE reporter) and GFP (from the constitutive EF1a reporter) expression were recorded periodically and 8 days post-transduction for both transduced (BFP+) and untransduced (BFP−) cells in each cell population assayed using an LSR-II flow cytometer (BD Biosciences) equipped with a 96-well high-throughput sampler. EF1a and UPRE signals were calculated for each sgRNA by subtracting an average background signal (median from control K562 dCas9-KRAB cells without reporter constructs) from these measurements and normalizing the resulting difference calculated from guide-containing cells (as determined by BFP fluorescence) to that from corresponding untransduced cells. Data from wells with fewer than 500 transduced or untransduced cells or with lower than expected BFP signal (3 standard deviations below the mean of BFP medians from all other wells) were systematically discarded from further analysis. For experiments where a flow cytometer reading was taken on the second day post transduction, data was also filtered for a minimum day 2 viability. Data were collected across 4 separate experiments and data without a minimum of 2 experimental replicates were discarded.

RT-qPCR and semi-quantitative PCR for *XBP1* mRNA splicing—Cells were harvested and total RNA was isolated using TRIzol® Reagent (ThermoFisher Scientific, 15596-018) and Phase Lock Gel tubes (VWR, 10052-170) or NucleoSpin® RNA

(Macherey-Nagel, 740955.50) essentially according to manufacturers' instructions. RNA prepared by TRIzol® extraction was treated with TURBO™ DNase (ThermoFisher Scientific). RNA was converted to cDNA using SuperScript® II or SuperScript® III Reverse Transcriptase (ThermoFisher Scientific) under standard conditions with oligo(dT) primers or random hexamers with or without RNaseOUT™ Recombinant Ribonuclease Inhibitor (ThermoFisher Scientific). Quantitative PCR reactions were prepared with 1X master mix containing 1X Colorless GoTaq® Reaction Buffer (Promega, M792A), MgCl2 (0.7 mM), dNTPs (0.2 mM each), primers (0.75 μM each), and 1000X SYBR Green with GoTaq® DNA polymerase (Promega, M830B) in 22 μL reactions. Reactions were run on a LightCycler® 480 Instrument (Roche). Semi-quantitative *XBP1*-specific PCR reactions were prepared with 2 μL of cDNA diluted 1:10 using a master mix containing 0.9X Colorless GoTaq® Reaction Buffer (Promega, M792A), dNTPs (0.23 mM each), primers (0.45 μM each) with GoTaq® DNA polymerase (Promega, M830B) in 22.1 μL reactions. These reactions were run on a standard thermocycler program with 30 second at 60.5°C for annealing and 28 cycles. PCR products were visualized on 8% TBE gels. Primers used were against *XBP1* (DAA_Hs_XBP1_A_RT_L: AGCTTTTACGAGAGAAAACTCAT; DAA_Hs_XBP1_B_RT_R: ACTGGGTCCAAGTTGTCCAG), *ACTB* (oBA74: GCTACGAGCTGCCTGACG, oBA75: GGCTGGAAGAGTGCCTCA), *CHOP* (oBA249: AGAACCAGGAAACGGAAACAGA, oBA250: TCTCCTTCATGCGCTGCTTT) (Oslowski and Urano, 2011), *SEC61A1* (oBA360: TGCAAAGCAGCTGAAGGA, oBA361: ATGCACAGCCCACCAAAG), *SSR2* (oBA364: TTCACCTCGGCAACAATTACT, oBA365: GGTGCACTGGTAGAGCCAAT), *SEC61B* (oBA366: GCTCTCCCAGCAAAGCAGT, oBA367: CCCACAGCTGGCATTTTT), *SEC61G* (oBA368: TTGTGAAATTGATCCATATTCCTATT, oBA369: AGATGAAAAACTCTCTTCCAAAATG), and *ALG2* (oBA372: ACCTTCCTTAAAAGCCACCAT, oBA373: TGTAAATGCTTCAGGGGAAAA). Experimental details relevant to specific figures in the main text are included below. Similar experimental details related to Figure S7B can be found in the corresponding supplemental figure legend.

Related to Figure 7C and 7E: cBA010 K562 cells (described above) were transduced with the indicated sgRNAs and after 2 days, carried in the presence of puromycin. Six days post transduction, cells were treated with 0.5 μM thapsigargin for 1.5 hours (or left untreated) and collected for RT-qPCR and semi-quantitative PCR to visualize *XBP1* mRNA splicing (described above). In this experiment sgRNAs were expressed from the original sgRNA expression vector (Addgene, Cat#60955).

Related to Figure 7D: cBA011 K562 cells (described above) were transduced and sorted for expression of the indicated sgRNAs. These were then collected on the indicated days post transduction for RT-qPCR and semi-quantitative PCR to visualize *XBP1* mRNA splicing (described above). In this experiment sgRNAs were expressed from the original sgRNA expression vector (Addgene, Cat#60955).

Quantification and Statistical Analysis

We will first provide an overview of the methods used, and then describe their specific application to each figure.

Pipeline overview—All analysis was performed in Python, using a combination of Numpy, Pandas, scikit-learn, and a custom-made Perturb-seq library. The general outline is presented in Figure S3A, and we will outline the steps below.

Sequencing—Reads from 10X single-cell RNA-seq experiments were aligned and collapsed to unique molecular identifier (UMI) counts using 10X's cellranger software (version 1.1, except for the pilot experiment in Figure 1 where version 1.0 was used). The result is a large digital expression matrix with cell barcodes as rows and gene identities as columns.

Perturbation identity mapping—Specifically amplified guide barcode libraries were created as described above and either sequenced as spike-ins or independently. The specific amplification strategy we used (Figure 1A, 1B) preserved the $3'$ end of the transcript (and thus the CBC and UMI of a given captured molecule) and introduced an Illumina read 1 primer upstream of the GBC sequence. These reads were aligned using bowtie (flags: -v2 -q -m1) to a library of expected GBC sequences. We then collapsed all reads with common CBC, UMI, and read identity (as some reads were not mapped by bowtie due to low quality scores) to produce a table consisting of possible guide identities for each cell, and the number of reads and molecules attributing a given guide identity to that cell. We defined the coverage of a given proposed identity as the number of reads divided by the number of UMIs. The distribution of coverages was always bimodal (Figure 1C). We defined a proposed identity as having good coverage if it: (1) was in the upper mode of the coverage distribution (defined by a threshold) (2) was attested to by at least 50 raw reads and (3) was attested to by at least 3 UMIs. Any cell that had only a single identity that met these criteria was assigned that perturbation (sgRNA) identity. Any cell that had two or more identities meeting these criteria was assigned as a multiple (either a multiple infection, PCR artifact, or a multiple encapsulation during emulsion generation). Any cell that had no identities meeting these criteria was assigned as unidentifiable.

Expression normalization—To normalize for differences in sequencing capture and coverage across emulsion droplets, we rescaled all cells to have the median number of total UMIs (i.e. each row of the raw digital expression matrix is normalized to the same sum). Expression of each gene was then *z*-normalized with respect to the mean and standard deviation of that gene in the control (unperturbed) population:

$$x_{\text{normalized}} = \frac{x - \mu_{\text{control}}}{\sigma_{\text{control}}}$$

This normalization means that control cells always have mean normalized expression of 0 for all genes and standard deviation 1, so that the units of expression are "standard deviations above/below the control distribution."

In the UPR epistasis experiment, the control population was the DMSO-treated cells. In the UPR Perturb-seq experiment, they were the cells containing the NegCtrl-2 guide. In the UPR Perturb-seq experiment, the mixed population was run in ten separate pools that were treated independently during library preparation (corresponding to lanes on the 10X Chromium instrument and on the Illumina sequencer). To avoid any lane-dependent batch effects, cells were normalized with respect to control cells within the same lane.

Low cell count/inviable cell removal—While developing LRICA method described below, we observed that all experiments always contained two subpopulations that were peculiar in that they contained roughly equal membership from all perturbations. Further investigation showed that these were a group of cells with systematically lower total UMI counts (visible as a small second mode in the distribution of total UMIs per cell) and a group of cells that contained markers of activation of apoptotic programs. We attributed the first population to inefficient reverse transcription occurring in a small number of emulsion droplets, and the second to inviable cells (which we knew were present at low frequency in the cells used in the 10X experiments). Though LRICA always isolated these in an unbiased way, we generally excluded them from analysis. The low UMI count cells were simply removed using a threshold. To remove the apoptotic cells, we trained a random forest regressor (described in more detail below in the section on UPR branch activation scoring below) to recognize them using the cells in our UPR epistasis experiment as training data. Apoptosis scores were assigned between 0 and 1 using this method to all cells within the population.

Identification of differentially expressed genes—The end result of the previous steps is a normalized gene expression matrix where each cell has been assigned a perturbation identity. In general, we were interested in analyzing differences between populations, and used two distinct strategies for isolating interesting genes.

Kolmogorov-Smirnov test/metric: The Kolmogorov-Smirnov test is a nonparametric test for equality of probability distributions based on a metric defined on their cumulative distribution functions (CDFs). Specifically, if $F_{\text{perturbed}}$ and F_{control} are the CDFs for a given gene in the perturbed and control distribution, the test statistic is

$$D = \sup_x |F_{\text{perturbed}}(x) - F_{\text{control}}(x)|$$

This can be assigned a p-value in a standard way. However, the large scale of single-cell data means that many genes were often significantly perturbed without being interestingly perturbed, simply because of small differences detected by great sampling depth. Thus in some cases we placed a direct threshold on the test statistic D itself, which ensured that changes were both significant (in the statistical sense) and also of reasonable magnitude, as it is valid metric on the space of CDFs.

Random forest classifier: An advantage of Perturb-seq is that cell populations are known, which means that supervised learning methods can be brought to bear. Our strategy here was

motivated by the idea that a gene is likely important for a given perturbation if its expression level can be used to accurately predict that perturbation's identity. This idea is particularly useful when many perturbations are being compared, as what you want then are the genes that best distinguish all of the perturbations from each other. To leverage this idea, we used random forest classifiers. Given a set of perturbations, we would train a random forest classifier to predict perturbation identity using a subset of genes. Specifically, we used the implementation of extremely randomized trees implemented in scikit-learn, generally with 1000 trees in the forest. We performed a two-stage fitting process for a given number of desired features N_{genes}. First, we set aside 20% of the cells. The remaining 80% were used to train a random forest classifier (usually with 1000 estimators) to predict the perturbation identity using the normalized expression profile for each cell as the set of features. (With some threshold on gene expression level to restrict the number of possible features; we usually restricted attention for example to genes present at at least 0.5 UMI/cell on average.) The random forest assigns importances to features during training based on their predictive value. We would then take the top N_{genes} sorted by importance as the set of most informative genes. To evaluate how informative these genes were, we would then retrain the classifier using only these genes, and predict the perturbation present in the 20% of cells we had initially set aside. For sets of perturbations with large differences, we routinely saw accuracies of 80–90%. The genes chosen by the random forest essentially always showed marked differences by the Kolmogorov-Smirnov approach outlined above, and the forests had the advantage that they scaled to an arbitrary number of perturbations, and that the selected genes were known to vary informatively across perturbations instead of simply having a difference in distribution.

Low rank ICA—Single-cell data are intrinsically very noisy, either due to real biological variation or problems with capture efficiency. To try to separate out this noise and robustly identify larger trends within the data, we developed a simple two-step approach called low rank ICA (LRICA). The first step consists of isolating a low rank approximation of the dynamics within the experiment. To do this, we used Robust PCA (Candès et al., 2011), which seeks a decomposition of the form

$$X = L + S$$

where X is the normalized expression matrix, L is a low rank matrix, and S is a sparse matrix (most entries are zero). Specifically, Robust PCA solves the optimization problem

$$\min_{\mathrm{L,S}} \|L\|_* + \lambda \|S\|_1 \quad \text{subject to} \quad X = L + S$$

where $\|\cdot\|_*$ is the nuclear norm (sum of singular values) and $\|\cdot\|_1$ is the sum of the absolute values of the entries of the matrix. These constraints naturally induce L to be low rank, and S to be sparse. In implementations, we used the augmented Lagrangian multiplier method (Lin et al., 2010), which was fast and efficient.

We should note that our interpretation of this optimization problem is slightly different from that seen in some other instances, where S is regarded as capturing noise corrupting the "true" dynamics seen in L. In single-cell data the "noise" may actually be biological in origin, but our primary intent is to isolate the low rank approximation L, which is effectively a smoothed version of the population's dynamics that leaves major trends intact. The advantage of the decomposition of course is that the S matrix is still available afterward, and it may in fact carry useful information about highly stochastic processes within the population.

Our next goal was to isolate the major trends within the low rank dynamics of the population. To do this we applied independent components analysis (ICA). ICA posits a model in which the expression of a given gene (y_j) can be decomposed as a linear sum of various effects (s_1 to s_n) that are statistically independent of each other:

$$y_j = a_j s_1 + a_{j2} s_2 + \cdots + a_{jn} s_n$$

Solving this problem is beyond the scope of this section, but our interest lies primarily in the vector version of this formula,

$$\boldsymbol{y} = A\boldsymbol{s}$$

in which a cell's expression profile **y** (over all genes) is viewed as a linear sum of independent effects, and the equivalent matrix version

$$Y = AS$$

in which we decompose all of the dynamics of the cells within our population (the columns of Y) into sums of independent components (ICs). The matrix A above is called the mixing matrix, and in our context describes which genes contribute to which effects. A key difference in this case from principal components analysis is that the **s** components are derived in a way to make them as statistically independent as possible, rather than uncorrelated. Once the matrix A is estimated, we can then "unmix" the dynamics of each cell in the population by applying the inverse operation (denoted here by W) to its expression profile:

$$\boldsymbol{s} = W\boldsymbol{y}$$

This yields a low-dimensional description of what each cell is doing in terms of the independent factors given by **s**.

In our case we apply ICA to the low rank matrix L, i.e. $Y = L^T$ above. Thus we try to separate the population's low rank dynamics into independent factors. As the ICA minimization problem posed in the strongest form cannot practically be solved, different algorithms will give somewhat different answers based on the tradeoffs they make. After

trying several methods, we settled on the ProDenICA algorithm (Friedman et al., 2001), which we found to frequently give the highest quality components.

In general we applied low rank ICA in two ways. First, it can be used to partition cells into subpopulations. Strong trends often lead to independent components that are bimodal, so simply thresholding the value of a component is a means of clustering. We note however that an advantage of this method of subpopulation identification is that it can also identify continuous trends, rather than enforcing discrete categories that may not exist like in other methods of clustering. Secondly, the mixing matrix A is very informative, as it determines the extent to which each gene contributes to a given component. This can be useful both in understanding what the component is measuring (if the most heavily weighted genes have a clear common function) and in identifying groups of genes that are co-expressed in an unbiased way.

Interpretation of independent components does have some caveats. First, they have no natural sign (so an “enriched” effect may appear as a low value of an independent component) or scale: thus there is no natural order where the first IC is somehow more informative than the next, consistent with the fact that they are meant to represent independent effects. We do note that one pragmatic solution is to order the components by the L_2 norm of the corresponding column in the mixing matrix, which tends to place the most interesting components first.

t-sne visualization—To obtain two-dimensional projections of the population’s dynamics, we first reduce the dimensionality of the low rank matrix L using classical PCA (with the number of components determined from a scree plot), and then further reduce these components via *t*-distributed stochastic neighbor embedding (t-sne). We occasionally directly visualize the ICs in this way as well, but because they lack intrinsic scale like principal components, dominant effects can be crowded out by minor ones.

Hierarchical clustering of genes—Several of the analyses in the paper use single-cell co-expression information to cluster genes. For a given list of genes, we perform this clustering by first calculating the gene-gene correlation matrix ρ over all cells in the population. This is then converted to a dissimilarity matrix π via the transformation $\pi = \sqrt{2(1-\rho)}$. The dissimilarity matrix is then clustered using Ward’s method. For visualization purposes, we then apply the optimal leaf ordering algorithm in MATLAB. This reorders the leaves in the dendrogram by flipping tree branches to maximize the similarity between adjacent leaves, but without dividing any branches (i.e. the clustering is unchanged, but the dendrogram ordering is in some sense optimal). We then reorder the columns and rows of the correlation matrix via the resulting ordering, so that groups of genes with correlated expression appear as blocks along the diagonal.

Cell cycle position—We used an approach previously described, in which the expression of sets of experimentally-derived genes specific for each cell cycle phase is used for each cell to score cell cycle phase (Macosko et al., 2015).

Average expression profiles—We often create synthetic bulk profiles for different populations. These are created by averaging the normalized expression profile of each cell within that population together.

Analytical steps for each figure

We now describe the analysis behind each figure in the paper, with references as necessary to the above sections.

Single-cell analyses in Figure 3—We formed a population consisting of cells treated with 100 nM thapsigargin in each of our 8 genetic backgrounds, along with DMSO-treated control cells (containing three non-targeting sgRNAs), totaling 5334 cells. As outlined in the "Low cell count/inviable cell removal" section, we removed cells with substantially lower than average UMI counts or that scored strongly for inviability markers from analysis, as these groups partitioned away from the rest of the population in preliminary analyses. 4541 cells remained after these filters. For each perturbation, we then looked for genes that were differentially expressed relative to the control, as described in the "Identification of differentially expressed genes." We made a list of all genes that had a mean expression of at least 0.5 UMI per cell in the population and for which the Kolmogorov-Smirnov test statistic $D > 0.15$ in at least one perturbation. This led to a group of 1,711 differentially expressed genes. We formed a reduced gene expression matrix containing only these genes, and performed low rank ICA to reduce the population's dynamics therein to 16 ICs (Figure S3B). We examined the raw trends in the population by reducing the low rank matrix to 16 components via PCA (16 components) and then to two dimensions via t-sne, revealing a general breakdown by perturbation and by cell cycle within each perturbation (Figure S3B). We then looked for ICs whose average value varied either across the perturbation, or across the cell cycle position. For each category, four components showed clear trends at the average level and in the t-sne plots (Figure S3B). For example, several of the components clearly showed the expected epistasis patterns for PERK, ATF6, and IRE1α (Figure S3B). The plots made in Figure 3B of the main text were then made by furthering reducing only the ICs that varied across perturbation (IC1 – IC4 in Figure S3B) or across the cell cycle (IC5 – IC8 in Figure S3B) to two dimensions using t-sne. (i.e., we constructed matrices with cells as rows and the given ICs as columns and reduced those matrices to two dimensions with t-sne.)

To make the plots in Figure 3C, we then subsampled our population to only look at cells treated with thapsigargin with or without depletion of PERK, and the DMSO-treated control (2042 cells in total). We applied the same methodology as above, though with 12 ICs instead of 16. The "G1 cell" IC described in the main text was bimodal within each subpopulation (see inset in right panel of Figure 3C), but with varying distances between the two modes (note that the IC takes a substantially lower value in the +Tg population than in any of the others, Figure 3C). We split each population based on a population-specific threshold that separated the two modes. The cell cycle position histograms were made as described above. To make Figure 3E, we took the 25 genes that most positively influenced the IC and the 25 genes that most negatively influenced the IC (by sorting the mixing matrix column for that IC by coefficient value) and then clustered them based on co-expression as described in the

"Hierarchical clustering of genes" section. The meaning of each cluster was discerned by the pattern of up- and down-regulation observed within.

Note in the raw sequencing data the tunicamycin-treated cells have gemgroup 1 (as a BAM tag), the thapsigargin-treated cells have gemgroup 2, and the DMSO-treated cells have gemgroup 3.

Branch epistasis analysis in Figure 3F—We created two populations: (1) consisting of cells treated with 100 nM thapsigargin in each of our 8 genetic backgrounds, along with DMSO-treated control cells, or (2) consisting of cells treated with 4 μg/mL tunicamycin in each of our 8 genetic backgrounds, along with DMSO-treated control cells. To identify informative differentially regulated genes, we used the random forest classifier method described in the "Identification of differentially expressed genes" section, limiting the random forest to pick 100 genes for each of the two populations. We then combined these two lists and discarded any duplicate genes. We created average profiles of expression of these genes for each of the nine conditions present in the two populations, as visualized in Figure 3F. The average epistatic phenotype of a gene can then be viewed as a 9-vector in either the thapsigargin- or tunicamycin-treated populations. We discarded any genes where the correlation between these two conditions was less than 0.9, as we were only interested in factors that showed the same regulation in response to both condtions. The end result was the 104 genes presented in Figure 3F. These were then clustered based on their co-expression pattern as described in the "Hierarchical clustering of genes" section, with the exception that Spearman correlation was used instead of Pearson correlation (to emphasize the large shifts in expression across the population). Rough meanings were ascribed to clusters based on the average pattern of gene expression across perturbations, but we emphasize that many targets show some degree of cross-regulation. To assess this in an unbiased way, we constructed a matrix consisting of the average expression of the 104 assayed genes across the 17 unique conditions present in the experiment, and reduced it to four independent components using FastICA. Three of the components clearly corresponded to ATF6, IRE1α, and PERK perturbations, as they showed banded patterns in the reduced matrix matching the pattern of epistasis for those regulators seen in Figure 3F (e.g. the PERK component was high in all conditions where PERK was present, and low everywhere else). The fourth component was low in the DMSO and all tunicamycin-treated conditions, and high in the thapsigargin-treated condition, so we discarded it as representing the difference between chemical perturbations. The panel at the bottom of Figure 3F plots the mixing matrix coefficients for each gene in the indicated component, and thus determines how much that gene affects that component's value.

Genome-wide CRISPRi screens in Figure 4—Analysis of the screen is described above along-side the experimental details above.

Clustering of guides and perturbations in Figure 5—We first split our large UPR Perturb-seq population into subpopulations based on guide identity and created average expression profiles (see "Average expression profiles" section) of all genes with mean representation >1 UMI per cell. We calculated the perturbation-perturbation correlation matrix between all average expression profiles, and then clustered it using the same

methodology described in the “Hierarchical clustering of genes.” The ordering is seen in Figure S5A. Because guides targeting the same gene behaved similarly in this analysis, in subsequent analyses we instead split the population into subpopulations based on guide target (thus merging subpopulations that had different guides that targeted the same gene). We clustered these profiles using the same criteria, and optimally ordered the resulting dendrogram and correlation matrix (as described in “Hierarchical clustering of genes”) to produce Figure 5A.

Assessing knockdown homogeneity in Figure 5—Most guide targets were too low abundance to interrogate directly at single-cell resolution. We first directly visualized the shift in guide target expression induced by the guide, comparing the distribution of expression in control cells to cells perturbed for a given target (Figure S5B). We calculated mean knockdown per guide (Figure 5C), and assigned 95% confidence intervals to our estimates via bootstrapping.

We also attempted to assess to what extent knockdown varied throughout the population based on phenotype. To do this, we needed an unbiased means of assessing deviation in behavior from the control cells. We leveraged a method called OneClassSVM, which is a means of novelty detection. Given a set of training exemplars, a OneClassSVM learns an estimate of how those points are distributed (potentially in a high-dimensional space). When given new observations, the OneClassSVM then estimates how likely it is that those observations came from the same distribution as the training set, or if they are outliers (potentially novel). In our case we trained the OneClassSVM using control cells, and thus scored the extent to which perturbed cells scored as outliers, or if they fell within the expected range of behavior for unperturbed cells. Specifically, for each guide target, we performed the following algorithm:

1. Form a population of all cells perturbed for that target, and an equal number of randomly sampled control cells.
2. Find all genes that are expressed at an average level of 0.5 UMI per cell or higher and that are differentially expressed between control and perturbed cells by the Kolmogorov-Smirnov test (as described in “Identification of differentially expressed genes”) at $P < 0.01$.
3. Form a reduced gene expression matrix consisting only of the differentially expressed genes. Create a low-dimensional picture of the dynamics within the population by reducing this matrix to 8 dimensions via PCA.
4. To form an estimate of “normal” behavior, train a OneClassSVM model to estimate the support of the control cells in this 8-dimensional space. The model was trained assuming a contamination rate with outliers of 5%.
5. Score each cell in the perturbed population using the OneClassSVM model to estimate the extent it deviates from control behavior.

These scores generally assigned most or all of the perturbed cells outlier status, except in guides where very few genes were perturbed to begin with (bottom panel of Figure 5D). Ordering the cells by score, we split each perturbed cell population into top third and bottom

third (i.e. the most and least perturbed cells) and assessed the difference in average knockdown in each of these populations (Figure S5C), with a difference of ~8% on average.

We also reported the number of differentially expressed genes measured above in the bottom panel of Figure 5D.

Scoring branch activation in Figure 5D—As outlined in the main text, we adopted a data-driven strategy to score activation of each of the UPR branches using the UPR epistasis experiment as training data. To do this, we assigned the label “ATF6 active”, “IRE1 active”, or “PERK active” to each cell in the UPR epistasis experiment based on whether a given branch was present (i.e. sensor gene not repressed) and induced (tunicamycin or thapsigargin had been added). For example, cells treated with thapsigargin and IRE1α-repressed would have ATF6 and PERK active, but not IRE1α. We converted these labels to scores of 0 (inactive) and 1 (active) and then trained three random forest regressors to predict activation of each branch. The training strategy was the same as outlined in the “Identification of differentially expressed genes” section: each cell was regarded as a training data point, with the normalized expression of every gene of mean > 1 UMI initially regarded as a possible feature for predicting branch activation. In training, 20% of the data was always set aside to use for performance testing, and we generally observed correlation coefficients of 0.8 or higher between predicted and actual scores. Each regressor was constrained to use the top 25 genes for predicting branch activation, as we found no performance improvement when more genes were included. The genes isolated as most important by the three regressors for scoring activation of the three branches all appear in the epistasis analysis in Figure 3F.

To validate performance, we compared this approach to scoring based on two other strategies:

1. *Gene list approach:* A list of hand-picked branch-specific genes were chosen from Figure 3F, and a score was defined as the sum of the normalized expression of those genes.
2. *ICA approach*: To allow for more complicated logic than simple sums, we applied the ICA decomposition seen in Figure 3F to each cell’s normalized expression profile and computed the value of each IC to produce a score for the expression of each branch.

With each scoring system, we normalized scores by subtracting the median of the DMSO-treated control cells and thresholded all cells with negative scores to zero. We then assessed the overlap of score distributions between cells expected to have a given branch active or inactive. As the random forests performed well in separating active and inactive branches in this analysis, we used them as our primary scoring method (Figure S5F).

The branch scores seen in Figure 5D are thus the result of applying the random forest regressor scoring system to each cell in the UPR Perturb-seq experiment, and then averaging the results within cells knocked down for the same gene. Note that because the regressors were trained using normalized expression data (see “Expression normalization” section), scoring is independent of sequencing depth. The average scores assigned by the ICA method agree well (cf. Figures 5D, S5E).

Single-cell analysis in Figure 6—We formed a population of cells containing either of two guides targeting *HSPA5*, or the NegCtrl-3 guide. In total, this consisted of 646 control cells and 1002 perturbed cells. We then removed all cells that had apoptosis scores greater than 0.85 (on a scale of 0 to 1, see "Low cell count/inviable cell removal" section), leaving 620 control cells and 969 perturbed cells. We found all genes that had mean abundance >0.5 UMI per cell and that were differentially expressed between the two populations by Kolmogorov-Smirnov test ($P < 0.01$), resulting in ~2,100 genes. We formed a reduced gene expression matrix consisting only of these genes and applied low rank ICA to reduce the population's dynamics therein to 12 ICs. The t-sne plots were made by reducing the low rank matrix to 16 components using PCA and then applying t-sne (see "t-sne visualization" section). Branch activation scores in Figure 6C were assigned as described above in the "Scoring branch activation in Figure 5D" section.

Two ICs varied substantially in average value between the control and perturbed cells (Figure 6B). The first, IC1, had a two-phase distribution in which all control cells and the majority of *HSPA5*-perturbed cells fell in the large lower peak, and a subpopulation of *HSPA5*-perturbed cells fell into a long tail of higher values (Figure 6B). We defined the sgHSPA5 IC1 HIGH cells to be the ones that fell within this tail (Figure 6B). Figure 6D shows the normalized expression of genes found in our epistasis analysis (Figure 3F) as columns, and the *HSPA5*-perturbed cells as rows, ordered by increasing IC1. Figure 6E was created by averaging the expression of *HSPA5* within the subpopulations defined in Figure 6B. Figure 6F was created using the cell cycle positions called in the "Cell cycle position" section.

Gene clustering analysis in Figure 6H—We first needed an unbiased approach to find programs of gene expression induced in the UPR Perturb-seq experiment. To do this we separated the population into control cells (containing our two control guides) and perturbed cells (containing any targeting guide). We constructed average expression profiles (see "Average expression profiles" section) of each, and then restricted our analysis to genes of mean expression > 0.5 UMI per cell on average in the perturbed population, and whose normalized expression was > 0.5. (Control cells by definition have mean normalized expression 0 for all genes, see "Expression normalization" section.) We then used a random forest classifier approach to select 200 of these induced genes that varied informatively across all of the perturbations in the Perturb-seq experiment (see "Identification of differentially expressed genes" section). The genes were then clustered based on their co-expression throughout the population, with the dendrogram leaves optimally reordered (see "Hierarchical clustering of genes" section). Our assumption was that many of these "induced genes" were involved in the unfolded protein response. We evaluated UPR dependence by examining the expression pattern of the induced genes within thapsigargin- and tunicamycin-treated cells (Figure S6A). We also assigned identities to some other clusters based on clear functional connections (as seen in Figure 6H).

Comparison of clustering of UPR genes in Figure 6I—As many UPR genes fell out of the previous analysis, we wanted to evaluate the ability to go the opposite direction, and cluster known interactions. We thus reexamined the list of UPR-regulated genes found in

Figure 3F. We separated the UPR Perturb-seq population into control cells (containing our two control guides) and perturbed cells (containing any targeting guide). We constructed average expression profiles (see "Average expression profiles" section) of each, and then restricted our analysis to the UPR-regulated genes that showed the same pattern of induction or repression in the perturbed cell population as they did in the cells treated with thapsigargin in the UPR epistasis experiment that had all branches of the UPR intact (i.e. with no knockdowns). We then performed hierarchical clustering of these genes (see "Hierarchical clustering of genes" section) using co-expression information from either (1) all cells in the UPR epistasis experiment, (2) all cells in the UPR Perturb-seq, and (3) only control cells in the UPR Perturb-seq experiment. We assessed the similarity among clusterings using the cophenetic correlation coefficient, i.e. the correlation coefficient between dendrogram distances taken over all possible pairs of genes. Closeness in cophenetic correlation thus implies that the dendrograms tend to place the same genes close to each other. The figure is meant only as a visual aid, as the cophenetic correlation carries information beyond the linear order. The genes were roughly grouped based on their epistasis pattern in the UPR epistasis experiment (as in Figure 3F), and then color was preserved as they were shuffled by the other two clusterings.

Enrichment of cholesterol genes in Figure 6K—Our unbiased analysis in Figure 6H contained a cluster of genes involved in cholesterol biosynthesis: *ACAT2*, *FDPS*, *FADS1*, *INSIG1*, *TMEM97*. We made a "cholesterol score" by summing the normalized expression of this group of genes in each cell, and then created a subpopulation containing (1) cells with cholesterol scores at or above the 95% of the control cell population and (2) control cells. This gave ~9,000 cells. Within this subpopulation, we then correlated the cholesterol score with the normalized expression of all genes with mean > 0.25 UMI per cell. We then selected all genes that had a correlation of 0.15 or higher with the cholesterol score for further analysis. We clustered the genes by co-expression within the population (see "Hierarchical clustering of genes" section), and then selected a group of 23 genes that clustered together with the original five and that appeared as a distinct block on the diagonal of the gene-gene correlation matrix. To demonstrate the improvement in correlation obtained by this "fishing" approach, we compared correlation matrices composed of these 23 genes and 23 random genes of similar average abundance between our enriched population, and control cells (seen in Figure 6K). Finally, we used Enrichr (Kuleshov et al., 2016) to obtain Reactome annotations and Encode SREBP binding state. Note that some of the genes that don't have annotations nevertheless are almost certainly cholesterol-related, such as the lncRNA RP11-660L16 which is directly next to *DHC7R*. SREBP binding data from Encode corresponds to the "SREBF1_HepG2_hg19" data set. "Reactome cholesterol synthesis" corresponds to the "Cholesterol biosynthesis_Homo sapiens_R-HSA-191273" data set.

Single-cell analysis in Figure 7—We formed populations of cells containing guides targeting either *SEC61A1* or *SEC61B*, along with cells containing the NegCtrl-3 guide, and that had apoptosis scores < 0.85. In total there were 620 control cells, 1381 *SEC61B*-perturbed cells, and 946 *SEC61A1*-perturbed cells. We found all genes that had mean abundance >0.5 UMI per cell and that were differentially expressed between the two populations by Kolmogorov-Smirnov test setting a threshold of $D > 0.15$ for *SEC61A1*, and

$D > 0.1$ for *SEC61B*, which is a weaker perturbation (see "Identification of differentially expressed genes" section). The different thresholds were chosen largely for esthetic reasons: lowering the threshold with *SEC61A1*, which is a strong perturbation, resulted in the inclusion of a number of cell cycle genes that caused the control population to fragment into subpopulations by cell cycle phase, which we felt was distracting. In each case we formed a reduced gene expression matrix consisting only of differentially expressed genes, then applied robust PCA (see "Low rank ICA" section) to these matrices, and then visualized the cells using t-sne plots generated using the first 16 principal components (see "t-sne visualization" section). Branch activation scores in Figure 7A, 7B, S7A were assigned as described above in the "Scoring branch activation in Figure 5D" section.

Data and Software Availability

Custom Python scripts for analysis of genome-scale CRISPRi screens is available at https://github.com/mhorlbeck/ScreenProcessing. The accession number for the sequencing data reported in this paper is GEO: GSE90546.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank E. Chow, D. Bogdanoff, and S. Elmes (UCSF Center for Advanced Technology) and N. Rapicavoli (10X Genomics) for technical expertise and assistance. We thank C.M. Gallagher, D. Acosta-Alvear, J. Peters, and M. Silvis for sharing of protocols and reagents. We thank L.M. Sack and S.J. Elledge for sharing of unpublished results. UCOE sequence was a gift from G. Sienski. We thank members of the J.S.W. lab for input and advice. This work was funded by National Institutes of Health Grants P50 GM102706, U01 CA168370, R01 DA036858 (all to J.S.W.), R01 GM102790, R35 GM118061 (to C.A.G.), and F32 GM116331 (to M.J.). J.S.W. is a Howard Hughes Medical Institute Investigator. T.M.N. is a fellow and B.A. is an HHMI fellow of the Damon Runyon Cancer Research Foundation (B.A. DRG-[2182-14], T.M.N. DRG-[2211-15]). L.A.G. is supported by NIH/NCI Pathway to Independence Award K99 CA204602. M.Y.H. is an EMBO postdoctoral fellow (EMBO ALTF 1193-2015, co-funded by the European Commission FP7, Marie Curie Actions, LTFCOFUND2013, GA-2013-609409). A.D. is supported by an NDSEG Fellowship, and A.R. is supported by the Klarman Cell Observatory, NHGRI (P50 HG006193), and HHMI.

References

Acosta-Alvear D, Zhou Y, Blais A, Tsikitis M, Lents NH, Arias C, Lennon CJ, Kluger Y, Dynlacht BD. XBP1 controls diverse cell type- and condition-specific transcriptional regulatory networks. Mol Cell. 2007; 27:53–66. [PubMed: 17612490]

Boettcher M, McManus MT. Choosing the right tool for the job: RNAi, TALEN, or CRISPR. Mol Cell. 2015; 58:575–585. [PubMed: 26000843]

Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? Journal of the ACM (JACM). 2011; 58:11.

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009; 25:288–289. [PubMed: 19033274]

Friedman, J., Hastie, T., Tibshirani, R. Springer series in statistics. Springer; Berlin: 2001. The elements of statistical learning.

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC. Genome-scale CRISPR-mediated control of gene repression and activation. Cell. 2014; 159:647–661. [PubMed: 25307932]

Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell. 2013; 154:442–451. [PubMed: 23849981]

Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, DeRisi JL. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol. 2016; 17:41. [PubMed: 26944702]

Haldimann A, Wanner BL. Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. J Bacteriol. 2001; 183:6384–6393. [PubMed: 11591683]

Hamanaka RB, Bennett BS, Cullinan SB, Diehl JA. PERK and GCN2 Contribute to eIF2α Phosphorylation and Cell Cycle Arrest after Activation of the Unfolded Protein Response Pathway. Mol Biol Cell. 2005; 16:5493–5501. [PubMed: 16176978]

Han J, Back SH, Hur J, Lin Y, Gildersleeve R, Shan J, Yuan CL, Krokowski D, Wang S, Hatzoglou M, et al. ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death. Nat Cell Biol. 2013; 15:481–490. [PubMed: 23624402]

Heimberg G, Bhatnagar R, El-Samad H, Thomson M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. Cell Syst. 2016; 2:239–250. [PubMed: 27135536]

Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. eLife. 2016; 5:e19760. [PubMed: 27661255]

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4:44–57. [PubMed: 19131956]

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014; 343:776–779. [PubMed: 24531970]

Jonikas MC, Collins SR, Denic V, Oh E, Quan EM, Schmid V, Weibezahn J, Schwappach B, Walter P, Weissman JS, Schuldiner M. Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. Science. 2009; 323:1693–1697. [PubMed: 19325107]

Kabadi AM, Ousterout DG, Hilton IB, Gersbach CA. Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. Nucleic Acids Res. 2014; 42:e147. [PubMed: 25122746]

Kampmann M, Bassik MC, Weissman JS. Functional genomics platform for pooled screening and generation of mammalian genetic interaction maps. Nat Protoc. 2014; 9:1825–1847. [PubMed: 24992097]

Kanda S, Yanagitani K, Yokota Y, Esaki Y, Kohno K. Autonomous translational pausing is required for XBP1u mRNA recruitment to the ER via the SRP pathway. Proc Natl Acad Sci U S A. 2016; 113:E5895.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161:1187–1201. [PubMed: 26000487]

Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature. 2016; 533:420–424. [PubMed: 27096365]

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016; 44:90.

Lee A, Iwakoshi NN, Glimcher LH. XBP-1 regulates a subset of endoplasmic reticulum resident chaperone genes in the unfolded protein response. Mol Cell Biol. 2003; 23:7448–7459. [PubMed: 14559994]

Liang S, Zhang W, McGrath BC, Zhang P, Cavener DR. PERK (eIF2alpha kinase) is required to activate the stress-activated MAPKs and induce the expression of immediate-early genes upon disruption of ER calcium homoeostasis. Biochem J. 2006; 393:201–209. [PubMed: 16124869]

Lin JH, Li H, Yasumura D, Cohen HR, Zhang C, Panning B, Shokat KM, Lavail MM, Walter P. IRE1 signaling affects cell fate during the unfolded protein response. Science. 2007; 318:944–949. [PubMed: 17991856]

Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. 2010 arXiv Preprint arXiv:1009.5055.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015; 161:1202–1214. [PubMed: 26000488]

Müller-Kuller U, Ackermann M, Kolodziej S, Brendel C, Fritsch J, Lachmann N, Kunkel H, Lausen J, Schambach A, Moritz T, Grez M. A minimal ubiquitous chromatin opening element (UCOE) effectively prevents silencing of juxtaposed heterologous promoters by epigenetic remodeling in multipotent and pluripotent stem cells. Nucl Acids Res. 2015:gkv019.

Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell. 2014; 156:935–949. [PubMed: 24529477]

Nissim L, Perli SD, Fridkin A, Perez-Pinera P, Lu TK. Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. Mol Cell. 2014; 54:698–710. [PubMed: 24837679]

Oslowski CM, Urano F. Measuring ER stress and the unfolded protein response using mammalian tissue culture system. Meth Enzymol. 2011; 490:71. [PubMed: 21266244]

Plumb R, Zhang Z, Appathurai S, Mariappan M. A functional link between the co-translational protein translocation pathway and the UPR. Elife. 2015; 4

Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013; 152:1173–1183. [PubMed: 23452860]

Sack LM, Davoli T, Xu Q, Li MZ, Elledge SJ. Sources of Error in Mammalian Genetic Screens. G3 (Bethesda). 2016; 6:2781–2790. [PubMed: 27402361]

Shoulders MD, Ryno LM, Genereux JC, Moresco JJ, Tu PG, Wu C, Yates JR, Su AI, Kelly JW, Wiseman RL. Stress-independent activation of XBP1s and/or ATF6 reveals three functionally diverse ER proteostasis environments. Cell Rep. 2013; 3:1279–1292. [PubMed: 23583182]

Sidrauski C, Tsai JC, Kampmann M, Hearn BR, Vedantham P, Jaishankar P, Sokabe M, Mendez AS, Newton BW, Tang EL, et al. Pharmacological dimerization and activation of the exchange factor eIF2B antagonizes the integrated stress response. Elife. 2015; 4:e07314. [PubMed: 25875391]

Smyth RP, Davenport MP, Mak J. The origin of genetic diversity in HIV-1. Virus Res. 2012; 169:415–429. [PubMed: 22728444]

Van Der Maaten L. Accelerating t-SNE using tree-based algorithms. Journal of Machine Learning Research. 2014; 15:3221–3245.

Walter P, Ron D. The unfolded protein response: from stress pathway to homeostatic regulation. Science. 2011; 334:1081–1086. [PubMed: 22116877]

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. Identification and characterization of essential genes in the human genome. Science. 2015; 350:1096–1101. [PubMed: 26472758]

Wang Y, Shen J, Arenzana N, Tirasophon W, Kaufman RJ, Prywes R. Activation of ATF6 and an ATF6 DNA binding site by the endoplasmic reticulum stress response. J Biol Chem. 2000; 275:27013–27020. [PubMed: 10856300]

Yamamoto K, Sato T, Matsui T, Sato M, Okada T, Yoshida H, Harada A, Mori K. Transcriptional induction of mammalian ER quality control proteins is mediated by single or combined action of ATF6α and XBP1. Developmental Cell. 2007; 13:365–376. [PubMed: 17765680]

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. bioRxiv. 2016

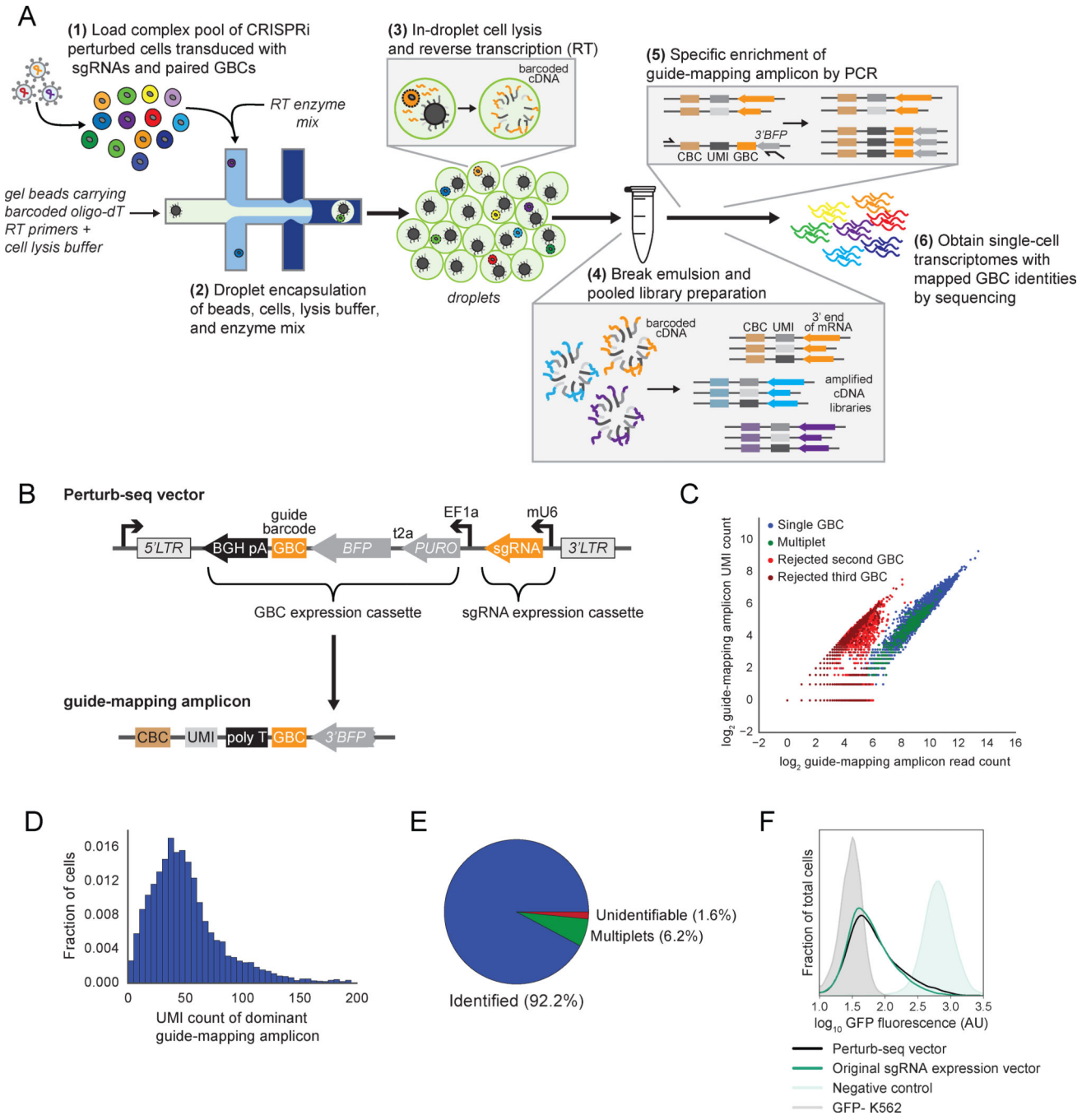


Figure 1. A robust strategy for genetic screens using single-cell gene expression profiling

(A) Schematic of the Perturb-seq platform. CBC, cell barcode (index unique to each bead). UMI, unique molecular identifier (index unique to each bead oligo). GBC, guide barcode (index unique to each sgRNA).

(B) Schematic of the Perturb-seq vector and guide-mapping amplicon.

(C) Performance of GBC capture. Top 3 possible GBCs for each CBC. CBC identity was assigned to sgRNA identity when a single GBC dominated (blue dots) and any lower

abundance GBCs were rejected (red dots). CBC was identified as a “multiplet” when a second or third GBC also had good coverage (green dots). Compare with (D,E).
(D) Distribution of captured UMIs from dominant guide-mapping amplicons.
(E) Performance of perturbation (sgRNA) identification. Data also represented in Figure S1B.
(F) Kernel density estimates of normalized flow cytometry counts representing GFP expression and knockdown achieved from the indicated sgRNA expression constructs.
See also Figure S1.

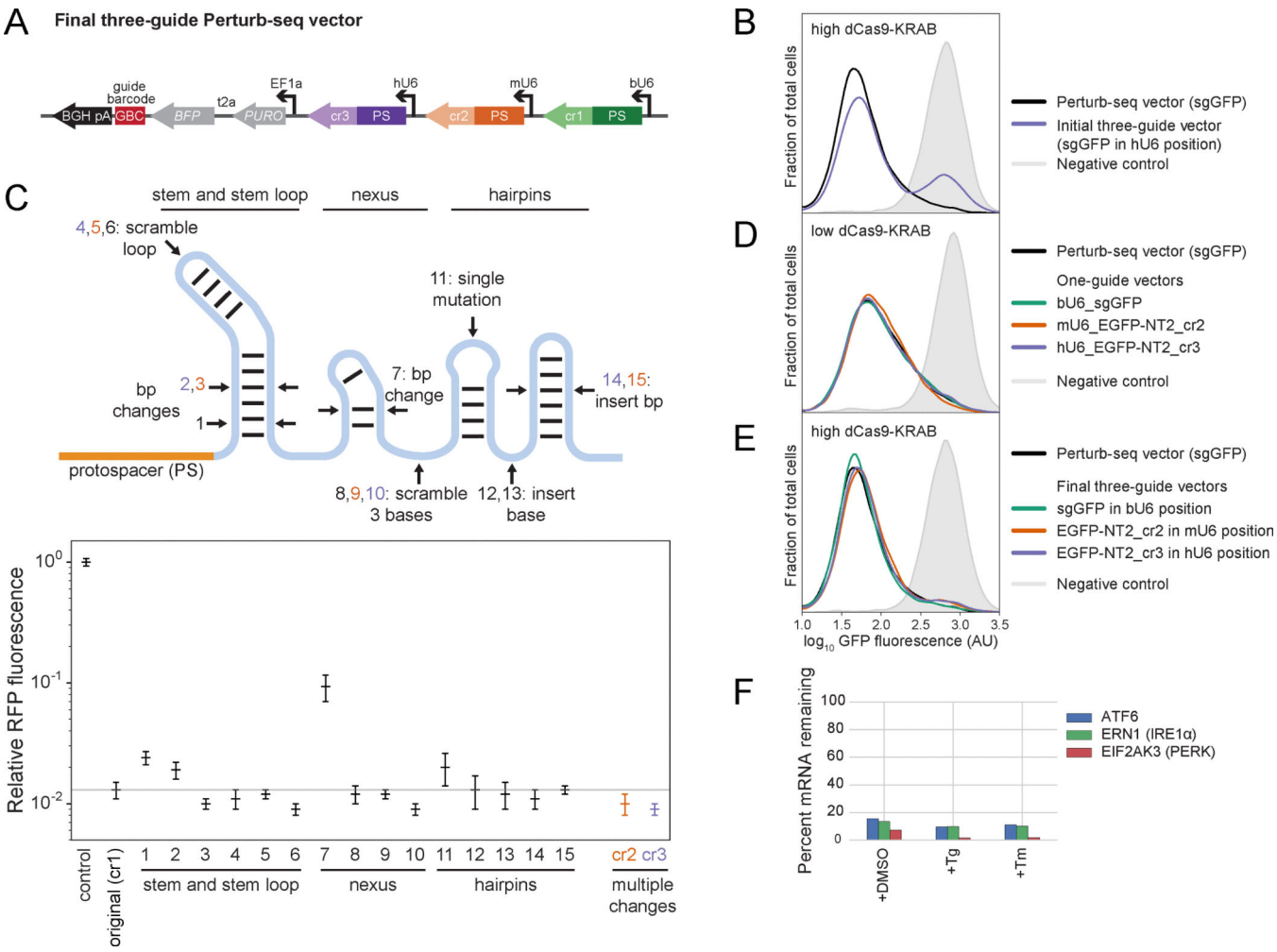


Figure 2. Strategy for multiplexed delivery of CRISPR sgRNAs in a single expression vector

(A) Schematic of the final three-guide Perturb-seq vector. “PS” denotes protospacer.

(B) Kernel density estimates of normalized flow cytometry counts representing GFP expression and knockdown achieved from the indicated sgRNA expression constructs.

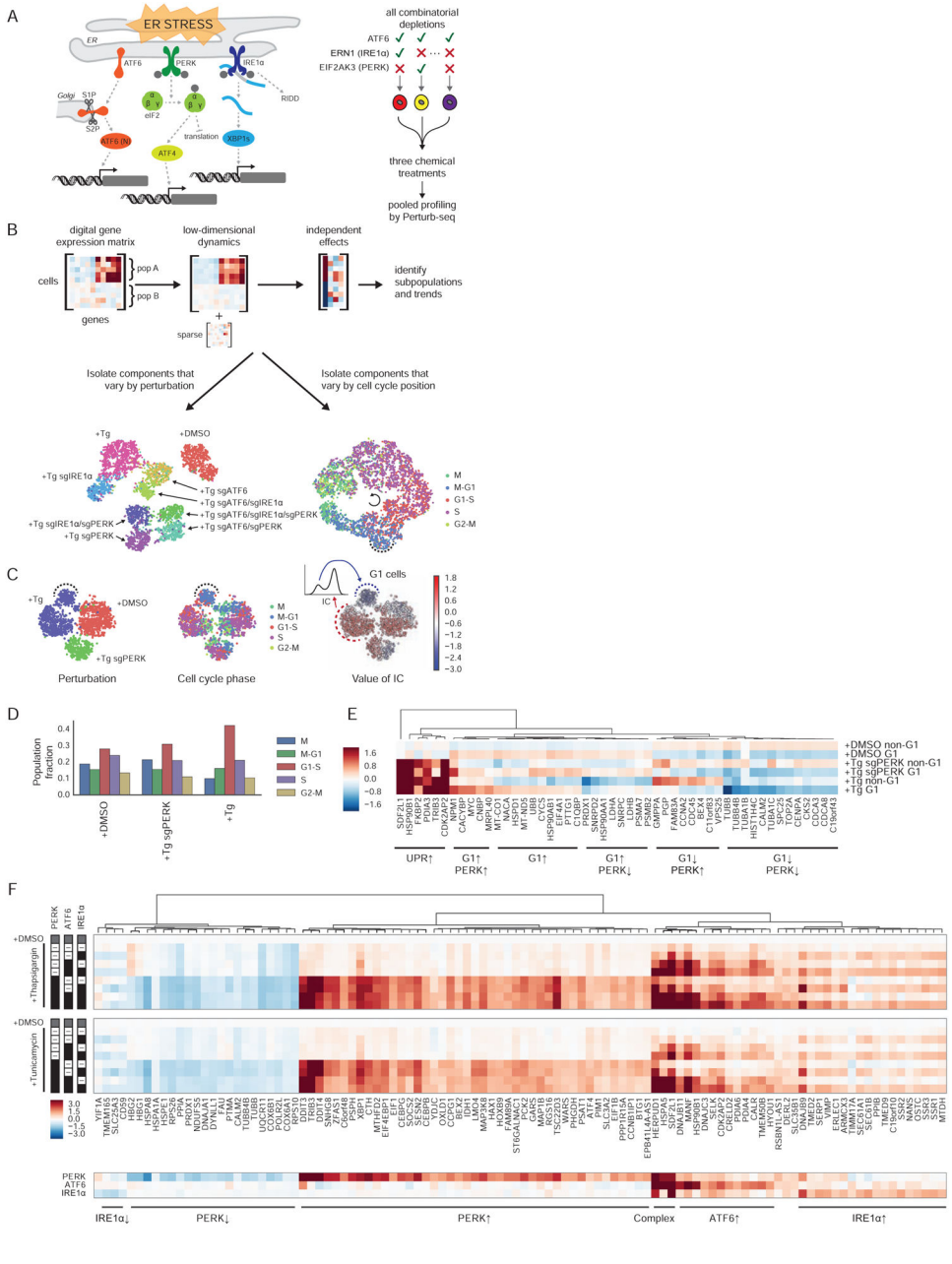
(C) Top: Schematic of sgRNA constant region with indicated changes. Orange, cr2 changes. Purple, cr3 changes. Bottom: Relative RFP from an *E. coli* CRISPRi reporter strain expressing an sgRNA with the indicated constant region variant and an mRFP-targeting protospacer. Data represent mean fluorescence of replicates normalized to negative control sgRNA ± standard deviations (n = 3).

(D) Kernel density estimates of normalized flow cytometry counts representing GFP expression and knockdown achieved from the indicated sgRNA expression constructs. For details on one-guide vectors see Figure S2F and Methods.

(E) Kernel density estimates of normalized flow cytometry counts representing GFP expression and knockdown achieved from the indicated sgRNA expression constructs. Data for the Perturb-seq vector is the same as in panel (B).

(F) Average percent mRNA remaining after simultaneous gene repression of *ERN1* (IRE1α), *EIF2AK3* (PERK), and *ATF6* using a final three-guide Perturb-seq vector determined via Perturb-seq.

See also Figure S2.



α!

Figure 3. Defining the three arms of the unfolded protein response using Perturb-seq

(A) Schematics of the unfolded protein response (UPR) and Perturb-seq UPR epistasis experiment.

(B) Unbiased identification and decoupling of single-cell behaviors via low rank independent component analysis (LRICA) in UPR epistasis experiment. Gene expression in cells (dots) is reduced to components identifying major trends in the population. Plots show t-sne projections of components that vary across genetic perturbations and chemical treatments (bottom left) or cell cycle position (bottom right). Tg, thapsigargin. DMSO-treated control cells (+DMSO) contain non-targeting control sgRNAs (throughout Figure 3).

(C) Plots (t-sne) of perturbation subpopulations (indicated GBC/treatment pairs: +DMSO and Tg-treated cells with or without PERK) from UPR epistasis experiment. LRICA identified a component (IC) that is bimodal within each of these subpopulations and marks G1 cells.

(D) Cell cycle composition of perturbation subpopulations from panel (C).

(E) Perturbation subpopulations from panel (C) were further divided into G1 and non-G1 cells based on IC value. Heatmap displays normalized expression of the 50 genes that most influenced IC, exposing both synergistic and antagonistic interactions.

(F) Genetic interactions among the three branches of the UPR. Top: Heatmap displays average expression profiles of 104 genes that strongly varied within the UPR epistasis experiment for each perturbation (i.e. indicated GBC/treatment pairs). Genes were clustered by their expression pattern within the entire population (i.e. all cells in all conditions). These patterns determine the branch specificity of each gene. Bottom: Unbiased decomposition of the total response into three components obtained via ICA.

See also Figure S3.

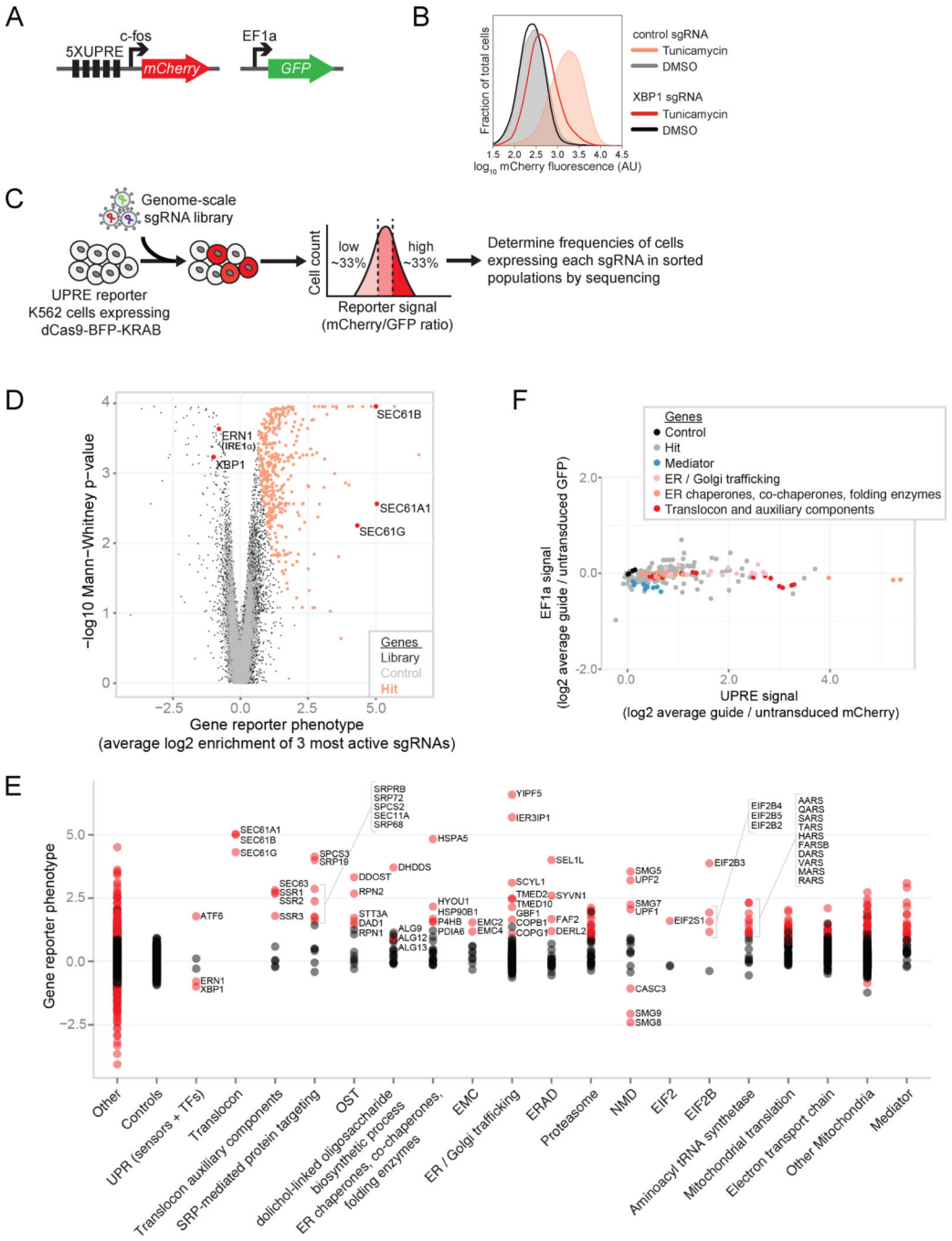


Figure 4. Genome-scale CRISPRi screening to identify gene depletion events that induce the IRE1α branch of the UPR

(A) Schematic of UPRE and constitutive EF1a reporter cassettes.

(B) K562 reporter (cBA011) cells were transduced with the indicated sgRNAs and treated with 2 μg/mL tunicamycin or DMSO after 4 days. Approximately 12 hr later, these cells were evaluated by flow cytometry. Data are representative of two independent experiments.

(C) Schematic of CRISPRi screens.

(D) Volcano plot of gene reporter phenotypes and p-values from CRISPRi-v2 screen. Gray indicates data generated from negative control sgRNAs. Pink indicates screen hits.

(E) Gene reporter phenotypes from CRISPRi-v2 screen (as in D) by functional category. Red indicates screen hits. See also Table S7.

(F) Comparison of UPRE and EF1a signals from K562 reporter (cBA011) cells transduced with 257 sgRNAs targeting 152 hit genes from the CRISPRi-v2 screen and 3 distinct negative controls. Data represent $\log_2$ averages of background-adjusted fluorescence medians (normalized to untransduced cells) collected from four separate experiments (n = 2–7 replicates).

See also Figure S4.

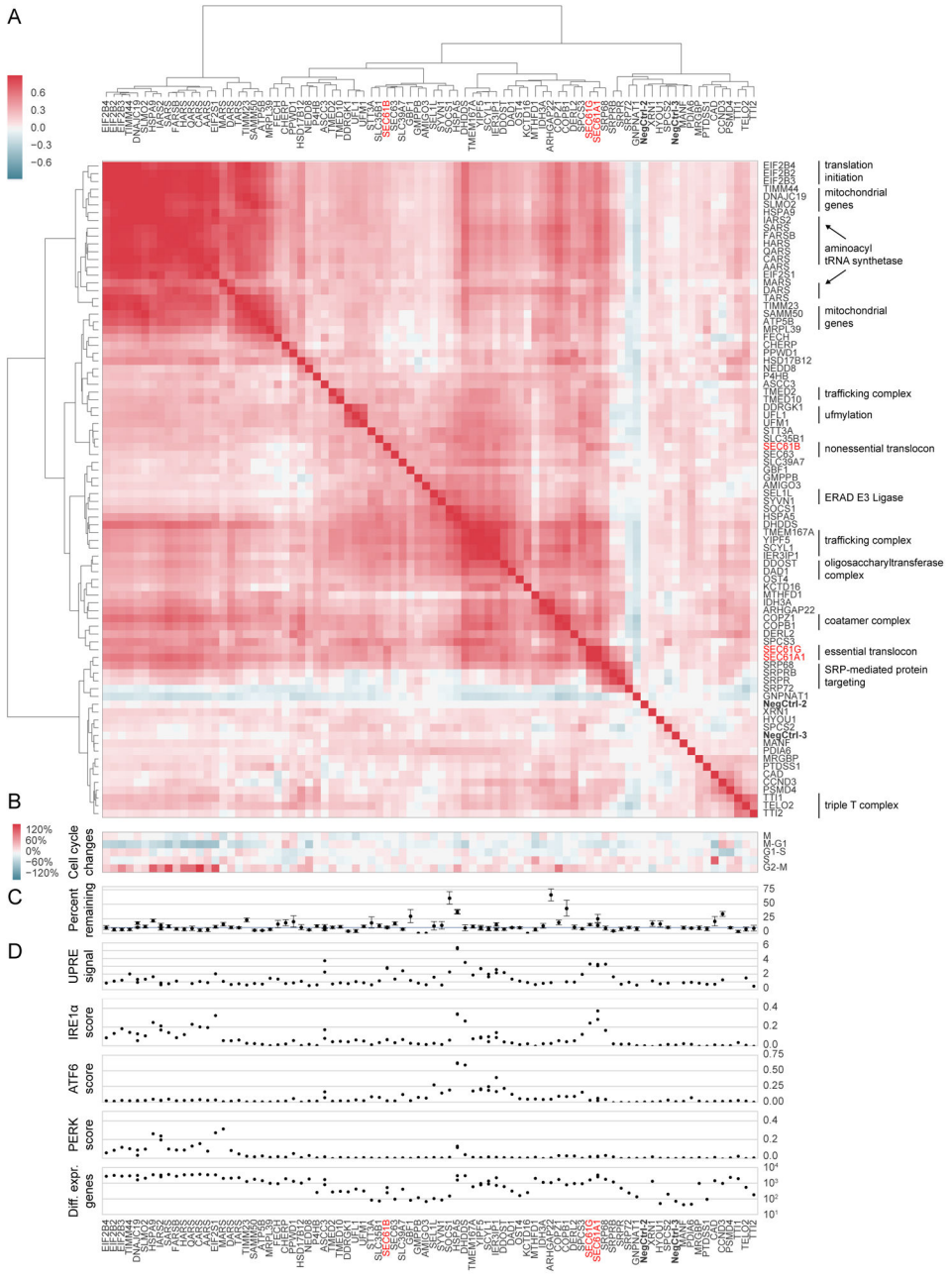


Figure 5. A large-scale Perturb-seq experiment interrogating ER homeostasis

(A) Functional clustering of genes from UPR Perturb-seq experiment. Heatmap displays correlations between hierarchically clustered average expression profiles from all cells bearing sgRNAs targeting the same gene (identified by GBCs). Functional annotations are indicated.

(B) Change in cell cycle composition induced by indicated genetic perturbations (identified by GBC) relative to control (NegCtrl-2) cells.

(C) Average percent target mRNA remaining from each subpopulation (identified by GBC). Genes targeted by multiple sgRNAs have multiple, possibly overlapping dots. Error bars are 95% CI estimated by bootstrapping.

(D) Individually evaluated UPRE signal phenotypes (data for hit genes also represented in Figure 4F) and scores measuring activation of the three UPR branches for each genetic perturbation. Final panel represents the $\log_{10}$ number of genes differentially expressed relative to control cells measured by the Kolmogorov-Smirnov test at $P < 0.01$.

See also Figure S5.

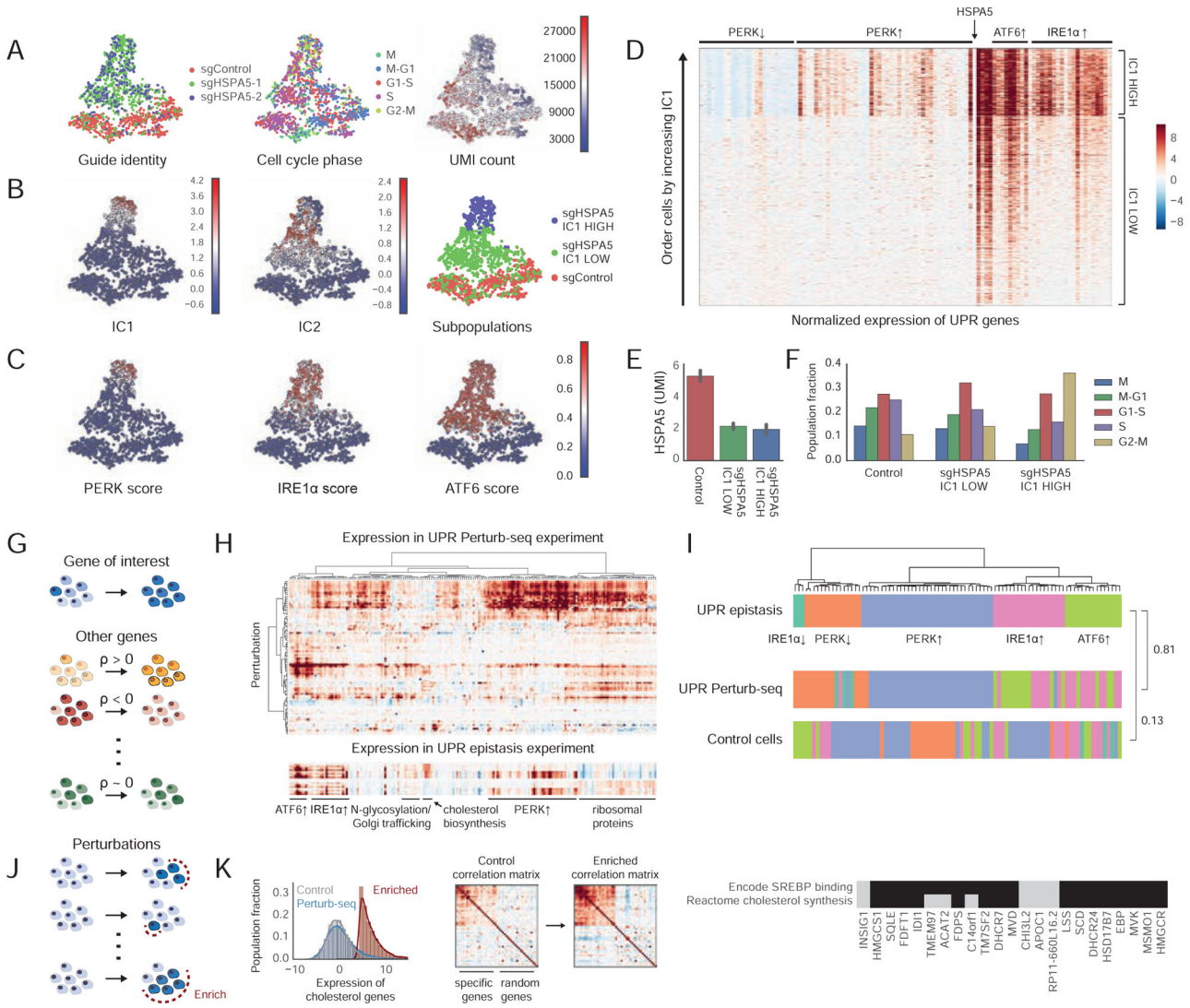


Figure 6. Single-cell information reveals a bifurcated UPR within a population and allows unbiased discovery of UPR-controlled genes

(A) Single-cell projections (t-sne) of sgRNA identity, cell cycle position, and UMI count per cell in *HSPA5*-perturbed and control cells (containing the NegCtrl-3 guide). We note that the *HSPA5*-targeting sgRNAs indicated differ by only 1-nt (Table S1).

(B) LRICA analysis of *HSPA5*-perturbed cells identifies two subpopulation-defining independent components. Right panel: subpopulations defined by thresholding IC1.

(C) Branch activation scores in *HSPA5*-perturbed cells.

(D) Normalized expression of UPR genes in *HSPA5*-perturbed cells. Each row is a cell, ordered by increasing IC1, and each column is a gene in the same order as Figure 3F.

(E) Mean expression of *HSPA5* across subpopulations. Error bars are 95% CI.

(F) Cell cycle composition of *HSPA5*-perturbed cells.

(G) Strategy for using correlated expression to identify functionally related genes.

(H) Unbiased identification of induced gene expression programs. Top: Normalized expression of 200 genes with significantly altered expression in UPR Perturb-seq experiment

clustered based on co-expression. Bottom: Normalized expression in UPR epistasis experiment, to assess UPR dependence. Full version in Figure S6A.
(I) UPR-responsive genes with altered expression in the UPR Perturb-seq experiment clustered by co-expression in the UPR epistasis experiment, the UPR Perturb-seq experiment, and control cells. Cophenetic correlation coefficients between dendrograms along with a visual guide to the movement of major groups included. Full version in Figure S6B.
(J) Strategy for enriching cells perturbed for a trait of interest.
(K). Within cells enriched for a set of bait cholesterol biosynthesis genes, a group of genes clustered with the bait genes and had more correlated expression than in control cells. Reactome annotations and SREBP binding data for the group included (right panel).
See also Figure S6.

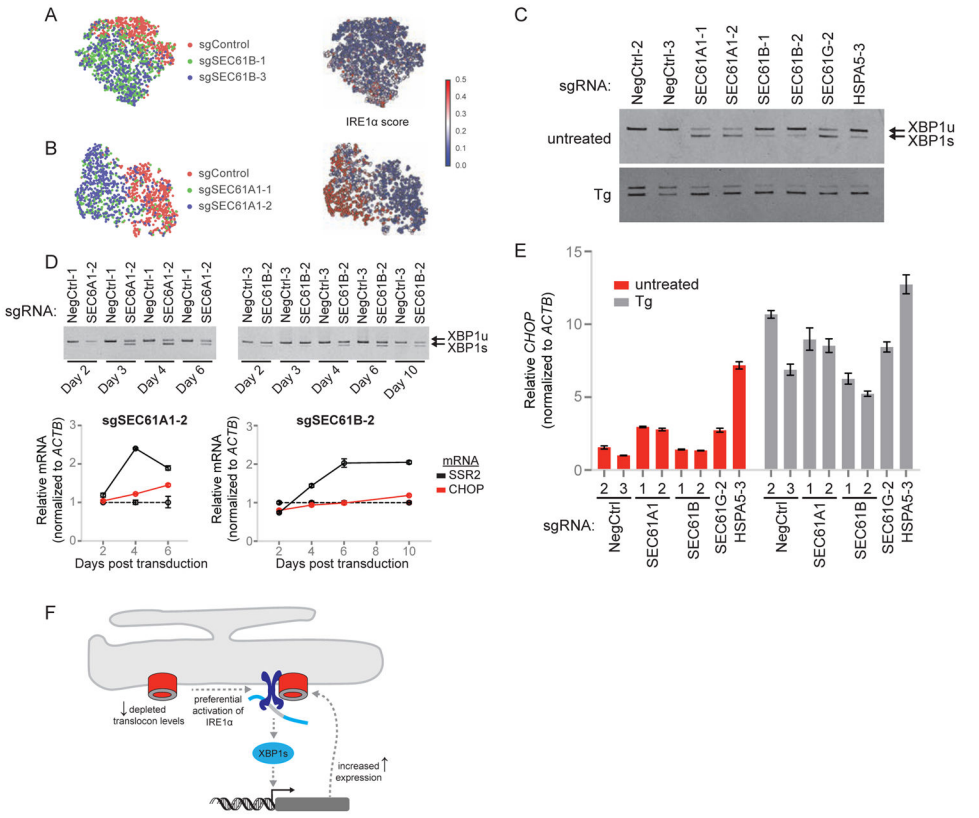


Figure 7. Translocon Gene Repression Preferentially Activates IRE1α UPR Signaling

(A) Single-cell analysis of *SEC61B*-perturbed cells in UPR Perturb-seq experiment. Control cells contain the NegCtrl-3 guide.

(B) Analysis of *SEC61A1*-perturbed cells (as in A).

(C) *XBP1* mRNA splicing from cells transduced with the indicated sgRNAs and treated ± thapsigargin (0.5 μM Tg for 1.5 hr).

(D) *XBP1* mRNA splicing (top) and *SSR2* and *CHOP* mRNA expression (bottom) from cells transduced with the indicated sgRNAs. Graphical data represent means relative to *ACTB* mRNA and normalized to cells transfected with NegCtrl-1 sgRNA (dotted lines) ± standard error of technical replicates (n = 3).

(E) Relative *CHOP* mRNA in cells described in (C). Data represent means relative to *ACTB* mRNA and normalized to cells transfected with NegCtrl-3 sgRNA ± standard error of technical replicates (n = 3).

(F) Model of translocon feedback signaling through IRE1α.

See also Figure S7.