Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq

Joseph M. Replogle[1-5,†], Reuben A. Saunders[2-5,†], Angela N. Pogson[3-5], Jeffrey A. Hussmann[3-5], Alexander Lenail[4,5], Alina Guna[5], Lauren Mascibroda[6], Eric J. Wagner[6,7], Karen Adelman[8], Gila Lithwick-Yanai[9], Nika Iremadze[9], Florian Oberstrass[9], Doron Lipson[9], Jessica L. Bonnar[3-5], Marco Jost[3,10], Thomas M. Norman[11,*], & Jonathan S. Weissman[3-5,12,13,*]

[1]Medical Scientist Training Program, University of California, San Francisco; San Francisco, CA 94158, USA.
[2]Tetrad Graduate Program, University of California, San Francisco; San Francisco, CA 94158, USA.
[3]Department of Cellular and Molecular Pharmacology, University of California, San Francisco; San Francisco, CA 94158, USA.
[4]Howard Hughes Medical Institute, Massachusetts Institute of Technology; Cambridge, MA 02142, USA.
[5]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology; Cambridge, MA 02142, USA.
[6]Department of Biochemistry & Molecular Biology, The University of Texas Medical Branch at Galveston; Galveston, TX 77555, USA.
[7]Department of Biochemistry & Biophysics, University of Rochester School of Medicine and Dentistry; Rochester, NY 14642, USA.
[8]Department of Biological Chemistry and Molecular Pharmacology, Blavatnik Institute, Harvard Medical School; Boston, MA 02115, USA.
[9]Ultima Genomics; Newark, CA 94560, USA.
[10]Department of Microbiology, Harvard Medical School; Boston, MA 02115, USA.
[11]Program for Computational and Systems Biology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center; New York, NY 10065, USA.
[12]David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology; Cambridge, MA 02142, USA.
[13]Department of Biology, Massachusetts Institute of Technology; Cambridge, MA 02142, USA.

[†]Co-first authors
*Corresponding authors: J.S.W. (weissman@wi.mit.edu) and T.M.N. (normantm@mskcc.org).

Abstract

A central goal of genetics is to define the relationships between genotypes and phenotypes. High-content phenotypic screens such as Perturb-seq (pooled CRISPR-based screens with single-cell RNA-sequencing readouts) enable massively parallel functional genomic mapping but, to date, have been used at limited scales. Here, we perform genome-scale Perturb-seq targeting all expressed genes with CRISPR interference (CRISPRi) across >2.5 million human cells and present a framework to power biological discovery with the resulting genotype-phenotype map. We use transcriptional phenotypes to predict the function of poorly-characterized genes, uncovering new regulators of ribosome biogenesis (including *CCDC86*, *ZNF236*, and *SPATA5L1*), transcription (*C7orf26*), and mitochondrial respiration (*TMEM242*). In addition to assigning gene function, single-cell transcriptional phenotypes allow for in-depth dissection of complex cellular phenomena – from RNA processing to differentiation. We leverage this ability to systematically identify the genetic drivers and consequences of aneuploidy and to discover an unanticipated layer of stress-specific regulation of the mitochondrial genome. Our information-rich genotype-phenotype map reveals a multidimensional portrait of gene function and cellular behavior.

Main Text

Mapping the relationship between genetic changes and their phenotypic consequence is critical to understanding gene and cellular function. This mapping is traditionally carried out in either of two ways. A phenotype-centric, "forward genetic" approach reveals the genetic changes that drive a phenotype of interest. Conversely, a gene-centric, "reverse genetic" approach catalogs the diverse phenotypes caused by a defined genetic change.

Recent technological developments have advanced both forward and reverse genetic efforts (*1*). CRISPR-Cas tools now enable the deletion, mutation, repression, or activation of genes with ease (*2*). In forward genetic screens, CRISPR-Cas systems can be used to generate cells with diverse genetic perturbations. This pool of perturbed cells can then be subjected to a selective pressure, with phenotypes assigned to genetic perturbations by sequencing. Forward genetic screens provide powerful tools for the identification of cancer dependencies, essential cellular machinery, differentiation factors, and suppressors of genetic diseases (*3–6*). In parallel, dramatic improvements in molecular phenotyping now allow for single-cell readouts of epigenetic, transcriptomic, proteomic, and imaging information (*7*). Applied to reverse genetics, single-cell profiling can refine the understanding of how select genetic perturbations affect cell types and cell states.

However, both phenotype-centric and gene-centric approaches suffer conceptual and technical limitations. Pooled forward genetic screens typically use low-dimensional phenotypes (e.g., growth, marker gene expression, drug resistance) for selection. The use of simple phenotypes can conflate genes acting via different mechanisms, requiring extensive follow-up studies to disentangle genetic pathways (*8*). Additionally, in forward genetics, serendipitous discovery is constrained by the prerequisite of selecting phenotypes prior to screening. On the other hand, while reverse genetic approaches enable the study of multidimensional and complex phenotypes, they have typically been restricted in scale to rationally chosen targets, limiting the ability to make systematic comparisons.

Single-cell CRISPR screens present a solution to these problems. These screens simultaneously read out the genetic perturbation and high-dimensional phenotype of individual cells in a pooled screening format, thus combining the throughput of forward genetic screens with the rich phenotypes of reverse genetics. While these approaches initially focused on transcriptomic phenotypes (e.g., Perturb-seq, CROP-seq) (*9–13*), technical advances have enabled their application to epigenetic (*14*), imaging (*15*), or multimodal phenotypes (*16–18*). From these rich

data, it is possible to identify genetic perturbations that cause a specific behavior as well as to catalog the spectrum of phenotypes associated with each genetic perturbation. Despite the promise of single-cell CRISPR screens, their use has generally been limited to studying at most a few hundred genetic perturbations, typically chosen with a bias towards predefined biological questions.

We reasoned that there would be unique value to genome-scale single-cell CRISPR screens. For example, while the number of perturbations scales linearly with experimental cost, the number of pairwise comparisons in a screen—and thus its utility for unsupervised classification of gene function—scales quadratically. Similarly, in large-scale screens, the diversity of perturbations allows one to explore the range of cell states that can be revealed by rich phenotypes. Additionally, as many human genes are well-characterized, these genes serve as natural controls to anchor the interpretation of observations in comprehensive datasets. Finally, genome-scale experiments could help address fundamental biological questions, such as what fraction of genetic changes elicit global transcriptional phenotypes and how transcriptional programs are rewired between cell types, with implications for understanding the organizing principles of cellular systems (*19*).

Here we perform the first genome-scale Perturb-seq screens. We use a compact, multiplexed CRISPR interference (CRISPRi) library to assay thousands of loss-of-function genetic perturbations with single-cell RNA-sequencing (scRNA-seq) in chronic myeloid leukemia (K562) and retinal pigment epithelial (RPE1) cell lines. Leveraging the scale and diversity of these perturbations across >2.5 million cells, we show that Perturb-seq can be used to study numerous complex cellular phenotypes—from RNA splicing to differentiation to chromosomal instability—in a single screen. We demonstrate how the interpretability of scRNA-seq phenotypes enables the discovery of gene function and extensively validate our findings with orthogonal experiments. Finally, we invert our analysis to focus on regulatory networks rather than genetic perturbations and uncover unanticipated stress-specific regulation of the mitochondrial genome. In sum, we use Perturb-seq to reveal a multidimensional portrait of cellular behavior, gene function, and regulatory networks that advances the goal of creating comprehensive genotype-phenotype maps.

Results

A multiplexed CRISPRi strategy for genome-scale Perturb-seq

Perturb-seq uses scRNA-seq to concurrently read out the CRISPR single-guide RNAs (sgRNA) (i.e., genetic perturbation) and transcriptome (i.e., high-dimensional phenotype) of single cells in a pooled format (Fig. 1A). To enable genome-scale Perturb-seq, we considered key parameters that would increase scalability and data quality, such as the genetic perturbation modality and sgRNA library.

Although Perturb-seq is compatible with a range of CRISPR-based perturbations including knockout (*10–12*), knockdown (CRISPRi) (*9*), or activation (CRISPRa) (*20*), we elected to use CRISPRi for several reasons. First, CRISPRi allows the efficacy of the genetic perturbation, knockdown, to be directly measured from scRNA-seq. Exploiting this feature allowed us to target each gene in our library with a single element and empirically exclude unperturbed genes from downstream analysis. Second, CRISPRi tends to yield more homogeneous genetic perturbation than nuclease-based CRISPR knockout, which can generate a subset of cells bearing active in-frame indels (*21*). The relative homogeneity of CRISPRi limits selection for unperturbed cells, especially when studying essential genes. Third, unlike nuclease-based gene knockout, CRISPRi does not lead to activation of the DNA damage response which can alter cell state and transcriptional signatures (*22*).

To improve scalability, we optimized our CRISPRi sgRNA libraries. To maximize CRISPRi efficacy, we used multiplexed CRISPRi libraries in which each construct contains two distinct sgRNAs targeting the same gene (table S1-S3; see *Methods*) (*13*). To avoid low representation of sgRNAs targeting essential genes, we performed preliminary growth screens and, during library synthesis, overrepresented constructs that caused strong growth defects (fig. S1A-D).

Next, we devised a three-pronged Perturb-seq screening approach encompassing multiple timepoints and cell types (Fig. 1A). As a primary cell line, we studied chronic myeloid leukemia (CML) K562 cells engineered to express the CRISPRi effector protein dCas9-KRAB (*23*). In this cell line, we performed two Perturb-seq screens: one targeting all expressed genes sampled at day 8 after transduction (n=9,866 genes; n=10,673 total perturbations; some genes have multiple independent transcripts) and another targeting common essential genes, which was sampled at day

6 after transduction (n=2,057 genes; n=2,176 total perturbations). As a secondary cell line, we used RPE1 cells engineered to express dCas9 fused to a KRAB domain derived from the gene *ZIM3*, which was recently shown to yield improved transcriptional repression compared to the *KOX1*-derived KRAB domain used in previous CRISPRi experiments (*24*). In contrast to K562 cells, RPE1 cells are a non-cancerous retinal pigment epithelial cell line that are hTERT-immortalized, near-euploid, adherent, and p53-positive. In RPE1 cells, we performed a screen targeting common essential genes plus a subset of nonessential genes that produced phenotypes in K562 cells sampled at day 7 after transduction (n=2,393 genes; n=2,549 total perturbations).

We conducted these three screens with 10x Genomics droplet-based 3' scRNA-seq and direct sgRNA capture (*13*). After sequencing and read alignment, we performed sgRNA identification and removed any cells bearing sgRNAs targeting different genes, which are an expected byproduct of lentiviral recombination between sgRNA cassette or doublet encapsulation during scRNA-seq. In total, we obtained >2.5 million high-quality cells with a median coverage of >100 cells per perturbation (fig. S1 E-G; table S4-S6). We observed a median target knockdown of 85.5% in K562 cells and 91.6% in RPE1 cells (Fig. 1B), confirming both the efficacy of our CRISPRi libraries and the fidelity of sgRNA assignment (*13*). The difference in performance between these cell lines was likely due to the use of the optimized *ZIM3*-derived KRAB domain in the RPE1 cells, suggesting that future efforts would benefit from improved CRISPRi efficacy.

A robust computational framework to detect transcriptional phenotypes

The scale of our experiment provided a unique opportunity to ask what fraction of genetic perturbations cause a transcriptional phenotype, a preliminary requirement for inferring gene function. Significant transcriptional phenotypes can take many forms, ranging from altered occupancy of a given cell state to focused changes in the expression level of a small number of target genes. To contend with this diversity, we created a robust framework capable of detecting transcriptional changes between groups of cells in our data. Our experimental design included many control cells bearing diverse non-targeting sgRNAs. These allow for internal *z*-normalization of expression measurements, and we found that this procedure corrected for batch effects that resulted from parallelized scRNA-seq and sequencing (fig. S2). As Perturb-seq captures single-cell genetic perturbation identities in a pooled format, we can use statistical approaches that treat each cell as an independent experimental sample. In general, we chose to use conservative, non-

parametric statistical tests to detect transcriptional changes rather than making specific assumptions about the underlying distribution of gene expression levels.

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

While sensitive, the energy distance test assays global shifts in expression without providing insight into which specific transcripts are altered. To detect individual differentially expressed genes, we applied the Anderson-Darling (AD) test to compare the distribution of expression levels for each gene in cells bearing each genetic perturbation against control cells. Importantly, the AD test is sensitive to transcriptional changes in a subset of cells, enabling us to find differences even when phenotypes have incomplete penetrance. With the AD test, we found 2,935 of 9,608 genetic perturbations targeting a primary transcript (30.5%) compared to 12 of 585 non-targeting controls (2.1%) caused >10 differentially expressed genes in K562 cells. These results were well-correlated between time points and cell types (fig. S3A,B; tables S4-S6) and concordant with the energy distance test (78.7% concordance by Jaccard index).

We then explored features of genetic perturbations that predict a transcriptional phenotype. We found that the strength of the transcriptional response was correlated with the strength of the growth defect (Spearman's rho = –0.51) with 86.6% of essential genetic perturbations (gamma < –0.1) leading to a significant transcriptional response in K562 cells (Fig. 1C; fig. S3C,D). A substantial number of genetic perturbations that cause a transcriptional phenotype nonetheless have a negligible growth phenotype (n=771; fig. S3E), indicating that many genetic perturbations influence cell state but not growth or survival. We also found that highly expressed genes were more likely to produce transcriptional phenotypes (Spearman's rho = 0.42) (fig. S3C).

Considering that some of our genetic perturbations did not yield strong on-target knockdown, our estimate of the fraction of genetic perturbations that cause a transcriptional phenotype is likely to be a lower bound. While a fraction of phenotypes may result from off-target effects, an advantage of Perturb-seq is the ability to directly detect potential off-target activities

such as the knockdown of neighboring genes. Consistent with earlier studies (*25*), we found that ~7.5% of perturbations caused significant knockdown of a neighboring gene in K562 cells, but neighbor gene knockdown was not enriched in genetic perturbations with a negligible growth defect that produced a transcriptional phenotype (fig. S4). Taken together, these results present a coherent picture where knockdown of a significant fraction of expressed genes causes a transcriptional response.

Annotating gene function from transcriptional phenotypes

Previous Perturb-seq screens have focused on targeted sets of genetic perturbations that are often related biologically, such as genes identified in forward genetic screens. Our large-scale screen targeting all expressed genes in K562 cells presented a unique opportunity to assess how well transcriptional phenotypes can resolve gene function when used in an unbiased manner.

We focused on a subset of 1,973 perturbations that had strong transcriptional phenotypes (>50 differentially expressed genes by AD test) (Fig. 2A). Because related perturbations could have different magnitudes of effect, we used the correlation between mean expression profiles as a scale-invariant metric of similarity.

To assess the extent to which correlated mean expression profiles between genetic perturbations indicated common function, we compared our results to two curated sources of biological relationships. First, among the 1,973 targeted genes, there were 327 protein complexes from the CORUM 3.0 database with at least two thirds of the complex members present, representing 14,165 confirmed protein-protein interactions (*26*). The corresponding expression profile correlations were markedly stronger (median correlation 0.61) than the background distribution of all possible gene pairs (median correlation 0.10) (Fig. 2B). Second, we compared the correlation between genetic perturbations to the STRING database of known and predicted protein-protein interactions, which had scores for 243,558 of the possible gene-gene relationships within our dataset (*27*). High STRING scores, reflecting high-confidence interacting proteins, were also strongly associated with high expression correlations (Fig. 2C).

We next performed an unbiased search for global structure to group similar perturbations within the dataset. We identified 64 discrete clusters based on strong intra-cluster correlations and annotated their function using CORUM, STRING, and manual searches. To visualize the dataset,

we constructed a minimum distortion embedding that places genes with correlated expression profiles close to each other in the plane and labeled the location of gene clusters (Fig. 2D).

Both the clusters and the embedding showed clear organization by biological function spanning a diverse array of different processes including: chromatin modification; transcription; mRNA splicing, capping, polyadenylation, and turnover; nonsense-mediated decay; translation; post-translational modification, trafficking, and degradation of proteins; central metabolism; mitochondrial transcription and translation; DNA replication; cell division; microRNA biogenesis; and major signaling pathways active in K562 cells such as BCR-ABL and mTOR (table S7). We further annotated the embedding visualization by labeling CORUM complexes and STRING clusters whose members were placed in nearby positions, revealing structure at finer resolution such as identifying the SMN complex, exon junction complex, U6 snRNP, and methylosome within the spliceosome and the association of ribosome biogenesis factors with the 40S ribosomal subunit.

In our dataset, we identified many poorly annotated genes whose perturbation led to similar transcriptional responses to genes of known function, naturally predicting a role for the uncharacterized genes. To orthogonally test a subset of these predictions, we selected ten poorly annotated genes whose perturbation response correlated (r>0.6) with subunits and biogenesis factors of either the large or small subunit of the cytosolic ribosome, which formed distinct clusters in our data (fig. S5A). This included genes that had no previous association with ribosome biogenesis (*CCDC86*, *CINP*, *SPATA5L1*, *ZNF236*, *C1orf131*) as well as genes that had not been associated with functional defects in a particular subunit (*SPOUT1*, *TMA16*, *NOPCHAP1*, *ABCF1*, and *NEPRO*). We used CRISPRi to target these genes in K562 cells and looked for evidence of ribosome biogenesis defects by assessing the ratio of 28S to 18S rRNA by Bioanalyzer electrophoresis. Knockdown of nine of the ten candidate factors led to substantial defects in ribosome biogenesis, with the exception of *ABCF1* (Fig. 2E). In every case, the affected ribosomal subunit corresponded to the Perturb-seq clustering across two independent sgRNAs. While this study was in progress, another group used cryo-EM to identify C1orf131 as a core structural component of the pre-A1 small subunit processome, complementing our functional evidence (*28*). This validation suggests that many poorly characterized genes can be assigned functional roles through Perturb-seq, although a subset of these relationships might be explained by indirect or off-target effects (fig. S5B,C).

In total, these results show that transcriptional phenotypes revealed by Perturb-seq have utility beyond studying gene regulation or transcriptional programs, and can serve as valuable tools for resolving and interrogating many central processes in cell biology.

Delineating functional modules of the Integrator complex

In general, perturbations to members of known protein complexes produced similar transcriptional phenotypes in our dataset. Therefore, we were surprised by the wide spectrum of transcriptional responses to knockdown of subunits of Integrator, a metazoan-specific essential nuclear complex with roles in small nuclear RNA (snRNA) biogenesis and transcription termination at paused RNA polymerase II (Fig. 3A) (*29*). Each of the fourteen core subunits of Integrator was targeted in our experiment, allowing us to systematically compare their transcriptional phenotypes in K562 and RPE1 cells (Fig. 3B; fig. S6A). *INTS1*, *INTS2*, *INTS5*, *INTS7*, and *INTS8* formed a tight cluster which weakly correlated with *INTS6* and *INTS12*. Separately, *INTS3*, *INTS4*, *INTS9*, and *INTS11* clustered together alongside splicing regulators involved in snRNP assembly and the tri-snRNP. Finally, *INTS10*, *INTS13*, and *INTS14* formed another discrete cluster together with *C7orf26*, an uncharacterized gene.

These distinct functional modules mirror the architecture of the Integrator complex observed in recent structures (*30*, *31*). The INTS1-2-5-7-8 functional module contained the subunits identified as the structural shoulder and backbone of Integrator. The INTS3-4-9-11 functional module contained the subunits identified as the structural cleavage module (as well as INTS3 which was not resolved). While INTS10, INTS13, and INTS14 were not resolved in the recent cryo-EM Integrator structures, these subunits have been identified as a stable biochemical subcomplex (*32*, *33*).

Integrator is an essential and well-studied complex, so we were intrigued by the robust clustering of the uncharacterized gene *C7orf26* with Integrator subunits 10, 13, and 14. To explore this, we tested whether loss of C7orf26 impacted the abundance of Integrator subunits. CRISPRi-based depletion of C7orf26 destabilized INTS10 in K562 cells, confirming either a regulatory or protein-level relationship (Fig. 3C). Next, we checked for a biochemical interaction between these proteins. Pulldown of His-INTS10 from cell lysates recovered endogenous C7orf26 alongside INTS13 and INTS14 (Fig. 3D). Additionally, overexpression of C7orf26 with INTS10, INTS13, and INTS14 enabled the purification of a stable INTS10-13-14-C7orf26 complex by size-

exclusion chromatography (Fig. 3E; fig. S6B and fig. S7). We also detected a physical interaction between the Drosophila C7orf26 orthologue and fly Integrator in S2 cells and observed co-essentiality between C7orf26 and INTS10, INTS13, INTS14 in the Cancer Dependency Map, suggesting that this relationship is conserved across species and cell types (fig. S8). Together, these results suggest that C7orf26 is a core subunit of a novel INTS10-13-14-C7orf26 Integrator module.

We sought to better understand the distinct transcriptional phenotypes induced by knockdown of INTS10-13-14-C7orf26 compared to the shoulder/backbone and cleavage modules. Comparison of the genes differentially expressed between modules did not reveal function in an obvious way (fig. S6C,D), perhaps owing to the late time point assayed in our experiment. We next explored the canonical role of Integrator in snRNA biogenesis. As mature snRNAs are not captured in 3' scRNAs-seq, we monitored changes in global splicing as a proxy for snRNA biogenesis defects. In our Perturb-seq data, we quantified changes in splicing by comparing the ratio of intronic (unspliced) to exonic (spliced) reads for each gene. Validating our approach, depletion of known splicing factors as well as subunits of the cleavage and shoulder/backbone modules led to gross splicing defects (Fig. 3F). By contrast, depletion of subunits of the INTS10-13-14-C7orf26 module did not cause a substantial splicing defect. To directly test the effect of the INTS10-13-14-C7orf26 module on snRNA biogenesis, we used PRO-seq to probe the positioning of active RNA-polymerase. These data confirmed that extended knockdown of the cleavage and backbone/shoulder modules, but not *INTS10*, *INTS13*, or *C7orf26*, caused a dramatic increase in transcriptional readthrough past the 3' cleavage site of snRNAs (Fig. 3G). In addition, the PRO-seq data confirmed that loss of the INTS10-13-14-C7orf26 module causes a transcriptional phenotype distinct from other modules (fig. S6E).

In sum, our results show that INTS10-13-14-C7orf26 represents a functionally and biochemically distinct module of the Integrator complex, and we propose that *C7orf26* be renamed *INTS15* for future studies (Fig. 3H). Although Integrator has been subjected to extensive structural analyses, it has been difficult to resolve the INTS10-13-14 components in relation to the rest of the complex. Inclusion of C7orf26 may facilitate future structural efforts. Broadly, this example highlights the utility of high-dimensional functional phenotypes for the unsupervised classification of protein complex subunits into functional modules.

340 ***Data-driven definition of transcriptional programs***

341

342 While clustering can organize genetic perturbations into pathways or complexes, it does
343 not reveal the functional consequences of perturbations. To globally summarize the genotype-
344 phenotype relationships in our dataset, we: (i) clustered genes into expression programs based on
345 their co-regulation across perturbations; (ii) clustered perturbations with strong phenotypes based
346 on their transcriptional profiles (as described above); and (iii) computed the average activity of
347 each gene expression program within each perturbation cluster (Fig. 4A,B; fig. S9A; table S7,S8;
348 see *Methods*). This map uncovered many known gene expression programs associated with genetic
349 perturbations, including upregulation of proteasomal subunits due to proteosome dysfunction (*34*),
350 activation of NFkB signaling upon loss of ESCRT proteins (*35*), downregulation of growth-related
351 genes in response to essential genetic perturbations, and upregulation of the cholesterol
352 biosynthesis pathway in response to defects in vesicular trafficking (*36*). Beyond these large-scale
353 relationships, we could also score the effects of individual genetic perturbations on different
354 expression programs. For example, our analysis delineated the canonical branches of the cellular
355 stress response into the independently regulated Unfolded Protein Response (UPR) and Integrated
356 Stress Response (ISR) (Fig. 4C) (*9*). The ISR was highly activated by loss of mitochondrial
357 proteins, aminoacyl-tRNA synthetases, and translation initiation factors, whereas the UPR was
358 activated by loss of ER-resident chaperones and translocation machinery. Collectively, this
359 analysis establishes the ability of Perturb-seq to learn regulatory circuits by leveraging the
360 variability of responses across perturbations.

361 Interestingly, our unbiased clustering uncovered many perturbations that drove the
362 expression of markers of erythroid or myeloid differentiation, consistent with the known
363 multilineage potential of K562 cells (Fig. 4D) (*37*). The scale of our experiment allowed us to
364 comprehensively search for genes whose modulation promotes cellular differentiation, an
365 application of major interest in both developmental and cancer biology. As expected, loss of
366 central regulators of erythropoiesis (*GATA1*, *LDB1*, *LMO2*, and *KDM1A*) caused myeloid
367 differentiation, whereas knockdown of *BCR-ABL* and its downstream adaptor *GAB2* induced
368 erythroid differentiation (*38*). Surprisingly, loss of a number of common essential genes (i.e.,
369 essential across cell lines in the Cancer Dependency Map) also caused expression of either myeloid
370 (e.g., Integrator subcomplex) or erythroid (e.g., NuRD complex, DNA replication machinery)
371 markers. Next, we investigated the differentiation effect of selectively essential genes, which could

be promising targets for differentiation therapy, analogous to ongoing efforts for KDM1A (*39*, *40*). We observed that loss of *PTPN1*, a tyrosine phosphatase selectively essential in K562 cells, drove myeloid differentiation. While inhibitors of PTPN1 have been developed for use in diabetes and certain cancers (*41*), to our knowledge they have not been tested as a differentiation therapy. Remarkably, in targeted experiments, we found that knockdown of *PTPN1* and *KDM1A* in combination caused a substantial increase in differentiation and growth defect compared to either single genetic perturbation, suggesting that these targets act via different cellular mechanisms (Fig. 4E; fig. S9B). These results highlight the utility of rich phenotypes for understanding differentiation as well as nominating promising therapeutic targets or combinations.

Hypothesis-driven study of composite phenotypes

We next recognized that our scRNA-seq readout could be used to study phenotypes that integrate data from across the transcriptome and, therefore, would be difficult to study in traditional forward genetic screens. Examples of these "composite phenotypes" include total cellular RNA content and the fraction of RNA derived from transposable elements (TE). We found numerous composite phenotypes under strong genetic control, with highly reproducible effects across screen replicates and cell types (Fig. 4F). In the specific case of TE regulation, two major classes of perturbations increased the fraction of TE RNA by affecting broad classes of elements including Alu, L1, and MIR (Fig. 4G; fig. S9C). First, loss of subunits of the exosome led to a substantial increase in the fraction of TE RNA, suggesting that transcripts deriving from TEs might be preferentially degraded. Second, loss of the CPSF cleavage and polyadenylation complex and parts of the Integrator complex produced a similar phenotype, suggesting that many of the TE RNAs observed in K562 cells may be derived from failure of normal transcription termination.

Turning to total RNA content (Fig. 4H), we found that loss of many essential regulators of S-phase and mitosis increased the RNA content of cells. This is consistent with the observation that cells tend to increase their size, and thus their RNA content, as they progress through the cell cycle (fig. S9D), so perturbations that arrest cells in later cell cycle stages yield increased total RNA content on average. By contrast, loss of essential transcriptional machinery, including general transcription factors, the Mediator complex, and transcription elongation factors, decreased total RNA content. In sum, these analyses show that genome-scale Perturb-seq enables

hypothesis-driven exploration of complex cellular features that are challenging to study through other means.

Exploring genetic drivers and consequences of aneuploidy in single-cells

As Perturb-seq is a single-cell assay, it enables the study of cell-to-cell heterogeneity in response to genetic perturbations. We reasoned that systematically exploring sources of heterogeneity could reveal insights into phenotypes that are missed in bulk or averaged measurements.

To assess the penetrance of perturbation-induced phenotypes, we first applied SVD-based leverage scores as a metric of single-cell phenotypic magnitude (see *Methods*). In this formulation, leverage scores quantify how outlying each perturbed cell's transcriptome is relative to non-targeting control cells without assuming that perturbations drive a single axis of variation. Supporting this approach, we found that mean leverage scores for each genetic perturbation were correlated with the number of differentially expressed genes (fig. S10A, Spearman's rho = 0.71), and reproducible across the day 6 and day 8 K562 experiments (fig. S10B, Spearman's rho = 0.79). To quantify the degree of heterogeneity in response to genetic perturbations, we then scored perturbations by the variation in single-cell leverage scores (Fig. 5A; see *Methods*). Comparing leverage scores across subunits of large essential complexes, we observed evidence for both biological (e.g., subcomplex function or dosage imbalance) and technical (e.g., selection to escape toxic perturbations) sources of phenotypic variation in response to genetic perturbations (fig. S10C-F).

Intriguingly, many genes implicated in chromosome segregation were among the top drivers of heterogeneity, including *TTK*, *SPC25*, and *DSN1* (Fig. 5B) (*42*). We hypothesized that the extreme transcriptional variability caused by these genetic perturbations might result from acute changes in the copy number of individual chromosomes due to mitotic mis-segregation. To explore this, we used inferCNV (*43*) to estimate single-cell DNA copy number along the genome by quantifying the change in moving average gene expression compared to control cells. Consistent with our hypothesis, knockdown of *TTK*, a core component of the spindle assembly checkpoint (*44*), led to dramatic changes in estimated DNA copy number in both intrinsically aneuploid K562 and near euploid RPE1 cells (Fig. 5C; fig. S11A). Specifically, in RPE1 cells, we found that 61/80 (76%) of *TTK* knockdown cells had evidence of karyotypic changes compared to

274/13140 (2%) of unperturbed cells. Notably, *TTK* knockdown cells bore highly variable karyotypes due to the stochastic gain or loss of chromosomes, accounting for the phenotypic heterogeneity observed in these cells (Fig. 5C).

An important advantage of the rich data provided by Perturb-seq is the ability to dissect not only perturbation-phenotype associations but also relationships between cellular phenotypes. We were curious how chromosomal instability (CIN) would affect cell cycle progression in euploid, p53-positive RPE1 cells versus constitutively aneuploid, p53-deficient K562 cells. Expanding our analysis to all cells in our experiment independent of genetic perturbation, we found that RPE1 cells with abnormal karyotypes tended to arrest in G1 or G0 of the cell cycle (G1 or G0 fraction 0.68 for abnormal karyotype vs. 0.44 for stable karyotype), while K562 cells with altered karyotypes had less significant shifts in cell cycle occupancy (Fig. 5D,E). Within the population of RPE1 cells bearing a chromosomal loss, the likelihood of cell cycle arrest directly depended on the magnitude of karyotypic abnormality (fig. S11B). Additionally, we observed that cells with the most severe karyotypic changes—those bearing both chromosomal gains and losses—had marked upregulation of the ISR (Fig. 5F and fig. S11C). These results are consistent with models in which cell cycle checkpoints are activated by the secondary consequences of aneuploidy (e.g., DNA damage or proteostatic stress) rather than changes in chromosome number *per se* (*45*, *46*).

Finally, we looked across all perturbations to systematically identify genetic drivers of CIN. We assigned a score to each perturbation based on the average magnitude of induced karyotypic abnormalities. Validating our approach, we found that CIN scores were strongly correlated across K562 and RPE1 cell lines (r=0.69) and identified many known regulators of chromosomal segregation, including components of the spindle assembly checkpoint, centromere, and NDC80 complex (Fig. 5G). Remarkably, we uncovered CIN regulators with diverse cellular roles, from cytoskeletal components to DNA repair machinery (Fig. 5H; table S4-S6). While many of these genes have previously been associated with chromosomal instability through targeted studies, the scale and single-cell resolution of Perturb-seq allowed us to identify numerous genetic drivers of CIN in a single experiment. This analysis also shows the potential of single-cell CRISPR screens to dissect phenotypes that were not predefined endpoints of the experiment.

Discovery of stress-specific regulation of the mitochondrial genome

Mitochondria arose from the engulfment and endosymbiotic evolution of an ancestral alphaproteobacterium by the precursor to eukaryotic cells (*47*). While the vast majority (~99%) of mitochondrially-localized proteins are encoded in the nuclear genome, mitochondria contain a small (~16.6 kilobase) remnant of their ancestral genome encoding 2 rRNAs, 22 tRNAs, and 13 protein-coding genes in humans. An open question is how the nuclear and mitochondrial genomes coordinate their expression to cope with mitochondrial stress (*48*). The scale of our experiment provided a unique opportunity to investigate this question.

We began by comparing the nuclear transcriptional responses to CRISPRi-based depletion of nuclear-encoded mitochondrial genes (i.e., mitochondrial perturbations). We found that mitochondrial perturbations elicited relatively homogeneous nuclear transcriptional responses, illustrated by well-correlated transcriptional phenotypes across mitochondrial perturbations (Fig. 6A and fig. S12A). While there was some variation in the magnitude of transcriptional responses (e.g., proteostatic injury drove an especially strong ISR activation), nuclear transcriptional responses generally failed to discriminate genetic perturbations by function. Although this result was broadly consistent with recent literature that has highlighted the role of the ISR as response to mitochondrial stress (*49–53*), the lack of functional specificity of the transcriptional response was puzzling in light of: (i) the multifaceted roles of mitochondria in diverse processes such as respiration, intermediary metabolism, iron-sulfur cluster biogenesis, and apoptosis and (ii) the high-resolution separation of cytosolic perturbations by transcriptional response in our data described above.

In contrast to the nuclear transcriptional response, we observed that the expression of mitochondrially encoded genes was highly variable between different mitochondrial perturbations (Fig. 6B; fig. S12B,C,D). When we clustered mitochondrial perturbations based solely on expression levels of the 13 mitochondrially encoded genes, a remarkably intricate and coherent pattern emerged: the clustering separated perturbations to Complex I, Complex IV, Complex III, Complex V (ATP synthase), the mitochondrial large ribosomal subunit, the mitochondrial small ribosomal subunit, chaperones/import machinery, and RNA processing factors (Fig. 6C; fig. S12E). To quantitatively support this observation, we trained a random forest classifier to distinguish cells with perturbations to different mitochondrial complexes and found that the mitochondrial transcriptome was far more predictive than the nuclear transcriptome

(mitochondrial accuracy 0.64; nuclear accuracy: 0.25) (fig. S12F). We then visualized the gene expression signatures of a subset of representative perturbations (Fig. 6D). The coregulation of mitochondrial genes tended to reflect function, with the exception of the bicistronic mRNAs *ND4L/ND4* and *ATP8/ATP6* (*54*). However, we did not identify a simple logic to explain the connection between genetic perturbations to their observed transcriptional consequences. While previous studies have described distinct regulation of the mitochondrial genome in response to specific perturbations [notably, related to loss of complex III and complex IV assembly factors (*55*, *56*)], our data generalize this phenomenon to a comprehensive set of stressors.

Next, we wanted to shed light on the mechanistic basis for this unappreciated complexity of mitochondrial genome responses. Given its singular origin, the mitochondrial genome is expressed by unique processes (Fig. 7A) (*57*). Mitochondrially encoded genes are transcribed as part of three polycistronic transcripts punctuated by tRNAs. These transcripts are then processed into rRNAs and mRNAs by tRNA excision, and individual mRNAs can be polyadenylated, expressed, or degraded. This complex system limits the potential for distinct transcriptional control but presents multiple opportunities for post-transcriptional regulation. To identify modes of perturbation-elicited differential expression, we examined the distribution of Perturb-seq reads along the mitochondrial genome (Fig. 7B). As our scRNA-seq used poly-A selection, most reads aligned to the 3' ends of mRNAs. To validate the utility of this position-based analysis, we confirmed that knockdown of known regulators of mitochondrial transcription (*TEFM*) and RNA degradation (*PNPT1*) led to major shifts in the position of reads along the mitochondrial genome. By contrast, many of the perturbation-specific responses discovered in the present study appeared to cause shifts in the relative abundance of mRNAs rather than gross shifts in positional alignments. To determine whether the observed mitochondrial genome responses reflected regulation of the total level of mitochondrial mRNAs or specific regulation of mRNA polyadenylation, we performed a bulk RNA-sequencing experiment with no poly-A selection. We observed perturbation-specific changes in the level of total RNA similar to those measured by scRNA-seq (cophenetic correlation r=0.79; Fig. 7C). Given the complexity of the observed responses, we propose that there are likely to be multiple mechanisms that impact the levels of the various mitochondrially encoded transcripts in response to different stressors.

Finally, we asked whether we could use the detailed clustering produced by the mitochondrial genome to predict gene function. Knockdown of an unannotated gene, *TMEM242*, produced a transcriptional signature resembling loss of ATP synthase in both K562 and RPE1 cells

(Fig. 7D; fig. S12G). Supporting this relationship, the top five co-essential genes with *TMEM242* were components of ATP synthase in the Cancer Dependency Map. Using a Seahorse assay, we further confirmed that basal respiration was decreased in *TMEM242* knockdown cells (Fig. 7E). While this work was in progress, another group used a biochemical approach to show that *TMEM242* physically interacts with ATP synthase subunits and regulates ATP synthase complex assembly in cells (*58*). Together, these experiments discover a novel factor required for ATP synthase activity and point to the precision of mitochondrial genome regulation.

Discussion

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

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

We argue that these screens are particularly powerful because of the intrinsic interpretability of comprehensive genotype-phenotype maps, enabling in-depth dissection of the functional consequences of genetic perturbations that impinge on many distinct aspects of cell biology. Of particular note is the ability to use the information-rich readouts to study complex, composite phenotypes, which are difficult to measure by other modalities. These composite phenotypes can be created in a data-driven (e.g., deriving transcriptional programs) or hypothesis-

563 driven manner (e.g., measuring intron/exon ratios to study splicing), resulting in an enormous
564 breadth of measured phenotypes. In the case of scRNA-seq, we show that it measures not only
565 features such as differential gene expression and the activity of critical transcriptional programs,
566 but also RNA splicing and processing, expression of transposable elements, differentiation,
567 transcriptional heterogeneity, cell cycle progression, and chromosomal instability. Once a
568 phenotype is defined, the genotype-phenotype map can be used to explore its genetic
569 underpinnings, in a manner analogous to a forward genetic screen, as well as its relationship to
570 other cellular phenotypes.

571 An illustrative example of this process is our study of chromosomal instability. Based on
572 an initial observation of heterogeneous responses to specific perturbations, we suspected that some
573 cells carried genetically-induced chromosomal gains or losses. In a hypothesis-driven manner, we
574 then used our rich phenotypic data to discover a large collection of perturbations—which were
575 only loosely connected by clustering on average transcriptional phenotypes—that promote
576 chromosomal instability. Importantly, the single-cell nature of our Perturb-seq data also allowed
577 us to explore the relationship between karyotypic changes and other phenotypes, including cell
578 cycle progression and stress induction. While aneuploidy is an important hallmark of most cancers,
579 it has not been easy to study with traditional genetic screens as it requires both a single-cell and
580 multimodal readout. In future work, this platform could be used to investigate interactions between
581 genetic perturbations and specific karyotypes, karyotype-dependent stress responses, or the
582 temporal evolution of karyotypes (*59*).

583 Genetic perturbations can push cells into extreme states that are not observed in
584 unperturbed cells. Because composite phenotypes can be generated and explored without being
585 preregistered at the time of data collection, rich genotype-phenotype maps provide a powerful
586 resource for the discovery of new cellular behaviors. Using this ability, we discovered a remarkable
587 degree of stress-specific changes in the expression of mitochondrially encoded transcripts. It was
588 only possible to appreciate the functional specificity of this regulation by pairing a defined set of
589 mitochondrial perturbations with a high-dimensional readout. This discovery suggests a
590 framework to explain how cells cope with diverse insults to mitochondria: a general nuclear
591 response is layered over perturbation-specific changes in the expression level of mitochondrially
592 encoded genes (Fig. 7F). Building on this observation, we can ask new questions about the
593 mitochondrial stress response. The transcriptional changes we observed may reflect adaptive
594 responses or, alternatively, complex patterns of dysfunction owing to disruption of the intricate

system of mitochondrial gene expression. Understanding how and in what contexts this regulation is adaptive may have important implications for diseases associated with mitochondrial stress. An intriguing additional question is whether individual mitochondria are able to regulate their expression autonomously. Combined with the nuanced responses observed here, this would support and substantially extend the "co-location for redox regulation" (CoRR) hypothesis which holds that the endosymbiotically derived mitochondrial genome has been retained through evolution to enable localized regulation of mitochondrial gene expression (*60*).

A final theme emerging from our work is the flexibility of single-cell CRISPR screens compared to other functional genomic approaches. Because these screens extract rich information from each cell in a pooled format, they require only a fraction of the number of cells used by other approaches and thus are well suited to the study of iPSC-derived cells and *in vivo* samples. As technologies for single-cell, multimodal phenotyping advance, single-cell screens will continue to become more powerful. At present, the major limitation of single-cell CRISPR screens is cost. Careful experimental designs, such as multiplexed libraries or compressed sensing (*61*), together with advances in single-cell phenotyping (*62*, *63*) and DNA sequencing promise to greatly increase the scale of these experiments. To this point, we concluded our work by sequencing our genome-scale K562 libraries on a lower-cost, ultra-high throughput sequencing platform developed by Ultima Genomics, generating results equivalent to those sequenced on Illumina instruments (fig. S13).

In sum, our study presents a blueprint for the construction and analysis of rich genotype-phenotype maps to serve as a driving force for the systematic exploration of genetic and cellular function.

Materials and Methods

A complete description of our Material and Methods is found in the Supplementary Material online. This includes methods experimental methods related to Perturb-seq screens and functional experiments, as well as computational methods detailing all data analysis.

References and Notes

1. J. G. Camp, R. Platt, B. Treutlein, Mapping human cell phenotypes to genotypes with single-cell genomics. *Sci New York N Y*. 365, 1401–1405 (2019).

2. J. G. Doench, Am I ready for CRISPR? A user's guide to genetic screens. *Nat Rev Genet*. 19, 67–80 (2018).

3. N. J. Kramer, M. S. Haney, D. W. Morgens, A. Jovičić, J. Couthouis, A. Li, J. Ousey, R. Ma, G. Bieri, C. K. Tsui, Y. Shi, N. T. Hertz, M. Tessier-Lavigne, J. K. Ichida, M. C. Bassik, A. D. Gitler, CRISPR–Cas9 screens in human cells and primary neurons identify modifiers of C9ORF72 dipeptide-repeat-protein toxicity. *Nat Genet*. 50, 603–612 (2018).

4. A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, R. M. Meyers, L. Ali, A. Goodale, Y. Lee, G. Jiang, J. Hsiao, W. F. J. Gerath, S. Howell, E. Merkel, M. Ghandi, L. A. Garraway, D. E. Root, T. R. Golub, J. S. Boehm, W. C. Hahn, Defining a Cancer Dependency Map. *Cell*. 170, 564-576.e16 (2017).

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

6. E. Wang, H. Zhou, B. Nadorp, G. Cayanan, X. Chen, A. H. Yeaton, S. Nomikou, M. T. Witkowski, S. Narang, A. Kloetgen, P. Thandapani, N. Ravn-Boess, A. Tsirigos, I. Aifantis, Surface antigen-guided CRISPR screens identify regulators of myeloid leukemia differentiation. *Cell Stem Cell*. 28, 718-731.e6 (2021).

7. T. Stuart, R. Satija, Integrative single-cell analysis. *Nat Rev Genet*. 20, 257–272 (2019).

645 8. L. Przybyla, L. A. Gilbert, A new era in functional genomics screens. *Nat Rev Genet*, 1–15
646 (2021).

647 9. B. Adamson, T. M. Norman, M. Jost, M. Y. Cho, J. K. Nuñez, Y. Chen, J. E. Villalta, L. A.
648 Gilbert, M. A. Horlbeck, M. Y. Hein, R. A. Pak, A. N. Gray, C. A. Gross, A. Dixit, O. Parnas, A.
649 Regev, J. S. Weissman, A Multiplexed Single-Cell CRISPR Screening Platform Enables
650 Systematic Dissection of the Unfolded Protein Response. *Cell*. 167, 1867-1882.e21 (2016).

651 10. A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D.
652 Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman,
653 N. Friedman, A. Regev, Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell
654 RNA Profiling of Pooled Genetic Screens. *Cell*. 167, 1853-1866.e17 (2016).

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

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

661 13. J. M. Replogle, T. M. Norman, A. Xu, J. A. Hussmann, J. Chen, J. Z. Cogan, E. J. Meer, J.
662 M. Terry, D. P. Riordan, N. Srinivas, I. T. Fiddes, J. G. Arthur, L. J. Alvarado, K. A. Pfeiffer, T.
663 S. Mikkelsen, J. S. Weissman, B. Adamson, Combinatorial single-cell CRISPR screens by direct
664 guide RNA capture and targeted sequencing. *Nat Biotechnol*. 38, 954–961 (2020).

665 14. A. J. Rubin, K. R. Parker, A. T. Satpathy, Y. Qi, B. Wu, A. J. Ong, M. R. Mumbach, A. L. Ji,
666 D. S. Kim, S. W. Cho, B. J. Zarnegar, W. J. Greenleaf, H. Y. Chang, P. A. Khavari, Coupled
667 Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory
668 Networks. *Cell*. 176, 361-376.e17 (2019).

669 15. D. Feldman, A. Singh, J. L. Schmid-Burgk, R. J. Carlson, A. Mezger, A. J. Garrity, F. Zhang,
670 P. C. Blainey, Optical Pooled Screens in Human Cells. *Cell*. 179, 787-799.e17 (2019).

671 16. E. P. Mimitou, A. Cheng, A. Montalbano, S. Hao, M. Stoeckius, M. Legut, T. Roush, A.
672 Herrera, E. Papalexi, Z. Ouyang, R. Satija, N. E. Sanjana, S. B. Koralov, P. Smibert, Multiplexed
673 detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat*
674 *Methods*. 16, 409–412 (2019).

675 17. C. J. Frangieh, J. C. Melms, P. I. Thakore, K. R. Geiger-Schuller, P. Ho, A. M. Luoma, B.
676 Cleary, L. Jerby-Arnon, S. Malu, M. S. Cuoco, M. Zhao, C. R. Ager, M. Rogava, L. Hovey, A.
677 Rotem, C. Bernatchez, K. W. Wucherpfennig, B. E. Johnson, O. Rozenblatt-Rosen, D.
678 Schadendorf, A. Regev, B. Izar, Multimodal pooled Perturb-CITE-seq screens in patient models
679 define mechanisms of cancer immune evasion. *Nat Genet*. 53, 332–341 (2021).

680 18. E. Papalexi, E. P. Mimitou, A. W. Butler, S. Foster, B. Bracken, W. M. Mauck, H.-H.
681 Wessels, Y. Hao, B. Z. Yeung, P. Smibert, R. Satija, Characterizing the molecular regulation of
682 inhibitory immune checkpoints with multimodal single-cell screens. *Nat Genet*. 53, 322–331
683 (2021).

684 19. A. Tanay, A. Regev, Scaling single-cell genomics from phenomenology to mechanism.
685 *Nature*. 541, 331–338 (2017).

686 20. T. M. Norman, M. A. Horlbeck, J. M. Replogle, A. Y. Ge, A. Xu, M. Jost, L. A. Gilbert, J. S.
687 Weissman, Exploring genetic interaction manifolds constructed from rich single-cell phenotypes.
688 *Science*. 365, 786–793 (2019).

689 21. A. H. Smits, F. Ziebell, G. Joberty, N. Zinn, W. F. Mueller, S. Clauder-Münster, D.
690 Eberhard, M. F. Savitski, P. Grandi, P. Jakob, A.-M. Michon, H. Sun, K. Tessmer, T.
691 Bürckstümmer, M. Bantscheff, L. M. Steinmetz, G. Drewes, W. Huber, Biological plasticity
692 rescues target activity in CRISPR knock outs. *Nat Methods*. 16, 1087–1093 (2019).

693 22. E. Haapaniemi, S. Botla, J. Persson, B. Schmierer, J. Taipale, CRISPR–Cas9 genome editing
694 induces a p53-mediated DNA damage response. *Nat Med*. 24, 927–930 (2018).

695 23. L. A. Gilbert, M. A. Horlbeck, B. Adamson, J. E. Villalta, Y. Chen, E. H. Whitehead, C.
696 Guimaraes, B. Panning, H. L. Ploegh, M. C. Bassik, L. S. Qi, M. Kampmann, J. S. Weissman,

697 Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*. 159, 647–
698 661 (2014).

699 24. N. Alerasool, D. Segal, H. Lee, M. Taipale, An efficient KRAB domain for CRISPRi
700 applications in human cells. *Nat Methods*. 17, 1093–1096 (2020).

701 25. J. Rosenbluh, H. Xu, W. Harrington, S. Gill, X. Wang, F. Vazquez, D. E. Root, A. Tsherniak,
702 W. C. Hahn, Complementary information derived from CRISPR Cas9 mediated gene deletion
703 and suppression. *Nat Commun*. 8, 15403 (2017).

704 26. M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C.
705 Montrone, A. Ruepp, CORUM: the comprehensive resource of mammalian protein complexes—
706 2019. *Nucleic Acids Res*. 47, D559–D563 (2019).

707 27. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic,
708 N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, C. von Mering, STRING v11: protein–
709 protein association networks with increased coverage, supporting functional discovery in
710 genome-wide experimental datasets. *Nucleic Acids Res*. 47, D607–D613 (2019).

711 28. S. Singh, A. V. Broeck, L. Miller, M. Chaker-Margot, S. Klinge, Nucleolar maturation of the
712 human small subunit processome. *Science*. 373, eabj5338 (2021).

713 29. N. Kirstein, H. G. D. Santos, E. Blumenthal, R. Shiekhattar, The Integrator complex at the
714 crossroad of coding and noncoding RNA. *Curr Opin Cell Biol*. 70, 37–43 (2021).

715 30. I. Fianu, Y. Chen, C. Dienemann, O. Dybkov, A. Linden, H. Urlaub, P. Cramer, Structural
716 basis of Integrator-mediated transcription regulation. *Science*. 374, 883–887 (2021).

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

720 32. M. M. Pfleiderer, W. P. Galej, Structure of the catalytic core of the Integrator complex. *Mol*
721 *Cell*. 81, 1246-1259.e8 (2021).

722 33. K. Sabath, M. L. Stäubli, S. Marti, A. Leitner, M. Moes, S. Jonas, INTS10–INTS13–INTS14
723 form a functional module of Integrator that binds nucleic acids and the cleavage module. *Nat*
724 *Commun*. 11, 3422 (2020).

725 34. S. K. Radhakrishnan, C. S. Lee, P. Young, A. Beskow, J. Y. Chan, R. J. Deshaies,
726 Transcription Factor Nrf1 Mediates the Proteasome Recovery Pathway after Proteasome
727 Inhibition in Mammalian Cells. *Mol Cell*. 38, 17–28 (2010).

728 35. A. Mamińska, A. Bartosik, M. Banach-Orłowska, I. Pilecka, K. Jastrzębski, D. Zdżalik-
729 Bielecka, I. Castanon, M. Poulain, C. Neyen, L. Wolińska-Nizioł, A. Toruń, E. Szymańska, A.
730 Kowalczyk, K. Piwocka, A. Simonsen, H. Stenmark, M. Fürthauer, M. González-Gaitán, M.
731 Miaczynska, ESCRT proteins restrict constitutive NF-κB signaling by trafficking cytokine
732 receptors. *Sci Signal*. 9, ra8–ra8 (2016).

733 36. J. Luo, H. Yang, B.-L. Song, Mechanisms and regulation of cholesterol homeostasis. *Nat Rev*
734 *Mol Cell Bio*. 21, 225–245 (2020).

735 37. J. F. Leary, B. M. Ohlsson-Wilhelm, R. Giuliano, S. Labella, B. Farley, P. T. Rowley,
736 Multipotent human hematopoietic cell line K562: Lineage-specific constitutive and inducible
737 antigens. *Leukemia Res*. 11, 807–815 (1987).

738 38. S. H. Orkin, L. I. Zon, Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell*.
739 132, 631–644 (2008).

740 39. L. Yu, G. Myers, C.-J. Ku, E. Schneider, Y. Wang, S. A. Singh, N. Jearawiriyapaisarn, A.
741 White, T. Moriguchi, R. Khoriaty, M. Yamamoto, M. G. Rosenfeld, J. Pedron, J. H. Bushweller,
742 K.-C. Lim, J. D. Engel, An erythroid-to-myeloid cell fate conversion is elicited by LSD1
743 inactivation. *Blood*. 138, 1691–1704 (2021).

744 40. T. Maes, C. Mascaró, I. Tirapu, A. Estiarte, F. Ciceri, S. Lunardi, N. Guibourt, A. Perdones,
745 M. M. P. Lufino, T. C. P. Somervaille, D. H. Wiseman, C. Duy, A. Melnick, C. Willekens, A.
746 Ortega, M. Martinell, N. Valls, G. Kurz, M. Fyfe, J. C. Castro-Palomino, C. Buesa, ORY-1001, a
747 Potent and Selective Covalent KDM1A Inhibitor, for the Treatment of Acute Leukemia. *Cancer*
748 *Cell*. 33, 495-511.e12 (2018).

749 41. B. Sharma, L. Xie, F. Yang, W. Wang, Q. Zhou, M. Xiang, S. Zhou, W. Lv, Y. Jia, L.
750 Pokhrel, J. Shen, Q. Xiao, L. Gao, W. Deng, Recent advance on PTP1B inhibitors and their
751 biomedical applications. *Eur J Med Chem*. 199, 112376 (2020).

752 42. A. Musacchio, E. D. Salmon, The spindle-assembly checkpoint in space and time. *Nat Rev*
753 *Mol Cell Bio*. 8, 379–393 (2007).

754 43. A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P.
755 Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà,
756 A. Regev, B. E. Bernstein, Single-cell RNA-seq highlights intratumoral heterogeneity in primary
757 glioblastoma. *Science*. 344, 1396–1401 (2014).

758 44. N. Jelluma, A. B. Brenkman, N. J. F. van den Broek, C. W. A. Cruijsen, M. H. J. van Osch,
759 S. M. A. Lens, R. H. Medema, G. J. P. L. Kops, Mps1 Phosphorylates Borealin to Control
760 Aurora B Activity and Chromosome Alignment. *Cell*. 132, 233–246 (2008).

761 45. S. Santaguida, A. Richardson, D. R. Iyer, O. M'Saad, L. Zasadil, K. A. Knouse, Y. L. Wong,
762 N. Rhind, A. Desai, A. Amon, Chromosome Mis-segregation Generates Cell-Cycle-Arrested
763 Cells with Complex Karyotypes that Are Eliminated by the Immune System. *Dev Cell*. 41, 638-
764 651.e5 (2017).

765 46. S. Santaguida, A. Amon, Short- and long-term effects of chromosome mis-segregation and
766 aneuploidy. *Nat Rev Mol Cell Bio*. 16, 473–485 (2015).

767 47. J. R. Friedman, J. Nunnari, Mitochondrial form and function. *Nature*. 505, 335–343 (2014).

768 48. P. M. Quirós, A. Mottis, J. Auwerx, Mitonuclear communication in homeostasis and stress.
769 *Nat Rev Mol Cell Biology*. 17, 213–26 (2016).

770 49. E. Mick, D. V. Titov, O. S. Skinner, R. Sharma, A. A. Jourdain, V. K. Mootha, Distinct
771 mitochondrial defects trigger the integrated stress response depending on the metabolic state of
772 the cell. *Elife*. 9, e49178 (2020).

773 50. P. M. Quirós, M. A. Prado, N. Zamboni, D. D'Amico, R. W. Williams, D. Finley, S. P. Gygi,
774 J. Auwerx, Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress
775 response in mammals. *J Cell Biol*. 216, 2027–2045 (2017).

776 51. C. Münch, J. W. Harper, Mitochondrial unfolded protein response controls matrix pre-RNA
777 processing and translation. *Nature*. 534, 710–713 (2016).

778 52. E. Fessler, E.-M. Eckl, S. Schmitt, I. A. Mancilla, M. F. Meyer-Bender, M. Hanf, J.
779 Philippou-Massier, S. Krebs, H. Zischka, L. T. Jae, A pathway coordinated by DELE1 relays
780 mitochondrial stress to the cytosol. *Nature*. 579, 433–437 (2020).

781 53. X. Guo, G. Aviles, Y. Liu, R. Tian, B. A. Unger, Y.-H. T. Lin, A. P. Wiita, K. Xu, M. A.
782 Correia, M. Kampmann, Mitochondrial stress is relayed to the cytosol by an OMA1-DELE1-HRI
783 pathway. *Nature*. 579, 427–432 (2020).

784 54. T. R. Mercer, S. Neph, M. E. Dinger, J. Crawford, M. A. Smith, A.-M. J. Shearwood, E.
785 Haugen, C. P. Bracken, O. Rackham, J. A. Stamatoyannopoulos, A. Filipovska, J. S. Mattick,
786 The Human Mitochondrial Transcriptome. *Cell*. 146, 645–658 (2011).

787 55. R. Richter-Dennerlein, S. Oeljeklaus, I. Lorenzi, C. Ronsör, B. Bareth, A. B. Schendzielorz,
788 C. Wang, B. Warscheid, P. Rehling, S. Dennerlein, Mitochondrial Protein Synthesis Adapts to
789 Influx of Nuclear-Encoded Protein. *Cell*. 167, 471-483.e10 (2016).

790 56. R. Salvatori, K. Kehrein, A. P. Singh, W. Aftab, B. V. Möller-Hergt, I. Forne, A. Imhof, M.
791 Ott, Molecular Wiring of a Mitochondrial Translational Feedback Loop. *Mol Cell*. 77, 887-
792 900.e5 (2020).

793 57. E. Kummer, N. Ban, Mechanisms and regulation of protein synthesis in mitochondria. *Nat*
794 *Rev Mol Cell Bio*. 22, 307–325 (2021).

795 58. J. Carroll, J. He, S. Ding, I. M. Fearnley, J. E. Walker, TMEM70 and TMEM242 help to
796 assemble the rotor ring of human ATP synthase and interact with assembly factors for complex I.
797 *P Natl Acad Sci Usa*. 118, e2100558118 (2021).

798 59. U. Ben-David, A. Amon, Context is everything: aneuploidy in cancer. *Nat Rev Genet*. 21,
799 44–62 (2020).

800 60. J. F. Allen, The CoRR hypothesis for genes in organelles. *J Theor Biol*. 434, 50–57 (2017).

801 61. B. Cleary, L. Cong, A. Cheung, E. S. Lander, A. Regev, Efficient Generation of
802 Transcriptomic Profiles by Random Composite Measurements. *Cell*. 171, 1424-1436.e18 (2017).

803 62. P. Datlinger, A. F. Rendeiro, T. Boenke, M. Senekowitsch, T. Krausgruber, D. Barreca, C.
804 Bock, Ultra-high-throughput single-cell RNA sequencing and perturbation screening with
805 combinatorial fluidic indexing. *Nat Methods*. 18, 635–642 (2021).

806 63. B. K. Martin, C. Qiu, E. Nichols, M. Phung, R. Green-Gladden, S. Srivatsan, R. Blecher-
807 Gonen, B. J. Beliveau, C. Trapnell, J. Cao, J. Shendure, An optimized protocol for single cell
808 transcriptional profiling by combinatorial indexing. *Arxiv* (2021).

809 64. M. Jost, Y. Chen, L. A. Gilbert, M. A. Horlbeck, L. Krenning, G. Menchon, A. Rai, M. Y.
810 Cho, J. J. Stern, A. E. Prota, M. Kampmann, A. Akhmanova, M. O. Steinmetz, M. E.
811 Tanenbaum, J. S. Weissman, Combined CRISPRi/a-Based Chemical Genetic Screens Reveal
812 that Rigosertib Is a Microtubule-Destabilizing Agent. *Mol Cell*. 68, 210-223.e6 (2017).

813 65. S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale,
814 T. R. Hughes, M. T. Weirauch, The Human Transcription Factors. *Cell*. 172, 650–665 (2018).

815 66. M. A. Horlbeck, L. A. Gilbert, J. E. Villalta, B. Adamson, R. A. Pak, Y. Chen, A. P. Fields,
816 C. Y. Park, J. E. Corn, M. Kampmann, J. S. Weissman, Compact and highly active next-
817 generation libraries for CRISPR-mediated gene repression and activation. *Elife*. 5, e19760
818 (2016).

819 67. N. D. Elrod, T. Henriques, K.-L. Huang, D. C. Tatomer, J. E. Wilusz, E. J. Wagner, K.
820 Adelman, The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-
821 Coding Genes. *Mol Cell*. 76, 738-752.e7 (2019).

822 68. K.-L. Huang, D. Jee, C. B. Stein, N. D. Elrod, T. Henriques, L. G. Mascibroda, D. Baillat, W.
823 K. Russell, K. Adelman, E. J. Wagner, Integrator Recruits Protein Phosphatase 2A to Prevent
824 Pause Release and Facilitate Transcription Termination. *Mol Cell*. 80, 345-358.e9 (2020).

825 69. A. P. Anderson, X. Luo, W. Russell, Y. W. Yin, Oxidative damage diminishes mitochondrial
826 DNA polymerase replication fidelity. *Nucleic Acids Res*. 48, 817–829 (2020).

827 70. K. A. Reimer, C. A. Mimoso, K. Adelman, K. M. Neugebauer, Co-transcriptional splicing
828 regulates 3′ end cleavage during mammalian erythropoiesis. *Mol Cell*. 81, 998-1012.e7 (2021).

829 71. Z. Xie, A. Bailey, M. V. Kuleshov, D. J. B. Clarke, J. E. Evangelista, S. L. Jenkins, A.
830 Lachmann, M. L. Wojciechowicz, E. Kropiwnicki, K. M. Jagodnik, M. Jeon, A. Ma'ayan, Gene
831 Set Knowledge Discovery with Enrichr. *Curr Protoc*. 1, e90 (2021).

832 72. J. He, I. A. Babarinde, L. Sun, S. Xu, R. Chen, J. Shi, Y. Wei, Y. Li, G. Ma, Q. Zhuang, A.
833 P. Hutchins, J. Chen, Identifying transposable element expression dynamics and heterogeneity
834 during development at the single-cell level with a processing pipeline scTE. *Nat Commun*. 12,
835 1456 (2021).

836 73. P. Ma, M. W. Mahoney, B. Yu, A Statistical Perspective on Algorithmic Leveraging. *Arxiv*
837 (2013).

838 74. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-
839 cell gene expression data. *Nat Biotechnol*. 33, 495–502 (2015).

840

Acknowledgements: We thank S. Vazquez, L. Gilbert, F. Urnov, C. Cotta-Ramusino, V. Sankaran, K. Loh, W. Allen, B. Do, P. Hsu, F. Diehl, C. Jan, J. Replogle, D. Phizicky, and all members of the Weissman and Norman labs for helpful discussions. We thank Jorge Dinis and the Innovative Genomics Institute for critical administrative and scientific support. We also thank: the UCSF Center for Advanced Technology, Eric Chow, and Delsy Martinez for sequencing and access to FACS machines; the Whitehead Institute Flow Cytometry Core and Kathy Daniels for access to FACS machines; the Whitehead Institute Genome Technology Core for support with bulk RNA-seq library preparation and rRNA bioanalyzers; and the Harvard Nascent Transcriptomics Core for support with PRO-seq library preparation.

Author contributions: JMR, RAS, TMN, and JSW were responsible for the conception, design, and interpretation of the experiments and wrote the manuscript. JMR led the Perturb-seq screens. JMR and TMN led the Perturb-seq data analysis. RAS led the functional studies. ANP generated and validated the ZIM3 CRISPRi RPE1 cell line and optimized Perturb-seq protocols across cell lines. JAH assisted with data interpretation and analysis. AL designed the interactive web visualization. AG produced preliminary data for Perturb-seq across cell lines. EJW and LM performed and supervised Drosophila Integrator biochemistry. KA supervised PRO-seq library preparation. GY, NI, FO, and DL sequenced libraries on the Ultima Genomics platform. JLB and MJ cloned pJB108 and validated the ZIM3 KRAB domain. JMR and RAS helped obtain funding for experiments. All authors provided feedback on the manuscript.

Funding: Research reported in this publication was supported by:
Defense Advanced Research Projects Agency (DARPA) grant HR0011-19-2-0007 (JSW)
National Institutes of Health (NIH) Centers of Excellence in Genomic Science (CEGS) (JSW)
Howard Hughes Medical Institute (JSW)
Chan Zuckerberg Initiative (JSW)
NIH grant 1DP2 GM140925-01 (TMN)
NIH grant R00-GM130964 (MJ)
NIH grant R01-GM134539 (EJW, KA)
Fannie and John Hertz Foundation Fellowship (RAS)
NSF Graduate Research Fellowship (RAS)
NIH grant F31-NS115380 (JMR)

Competing interests: JMR consults for Maze Therapeutics and is a consultant for and equity holder in Waypoint Bio. RAS consults for Maze Therapeutics. KA is a consultant for Syros Pharmaceuticals, is on the SAB of CAMP4 Therapeutics, and received research funding from Novartis not related to this work. MJ consults for Maze Therapeutics and Gate Bioscience. TMN consults for Maze Therapeutics. JSW declares outside interest in 5 AM Venture, Amgen, Chroma Medicine, KSQ Therapeutics, Maze Therapeutics, Tenaya Therapeutics, Tessera Therapeutics and Third Rock Ventures. The Regents of the University of California with RAS, TMN, MJ, and JSW as inventors have filed patent applications related to CRISPRi/a screening and Perturb-seq.

Data and materials availability: Raw sequencing data will be deposited into SRA. An interactive data browser including processed, downloadable single-cell and pseudobulk populations will be made available upon publication. Our previously published analytic framework for Perturb-seq analysis is available at https://github.com/thomasmaxwellnorman/Perturbseq_GI. Scripts for guide assignment are available at https://github.com/josephreplogle/guide_calling. Additional code related to specific analyses will be made available on github upon publication.

Supplementary Materials:

Materials and Methods

Figs. S1 to S13

Tables S1 to S9

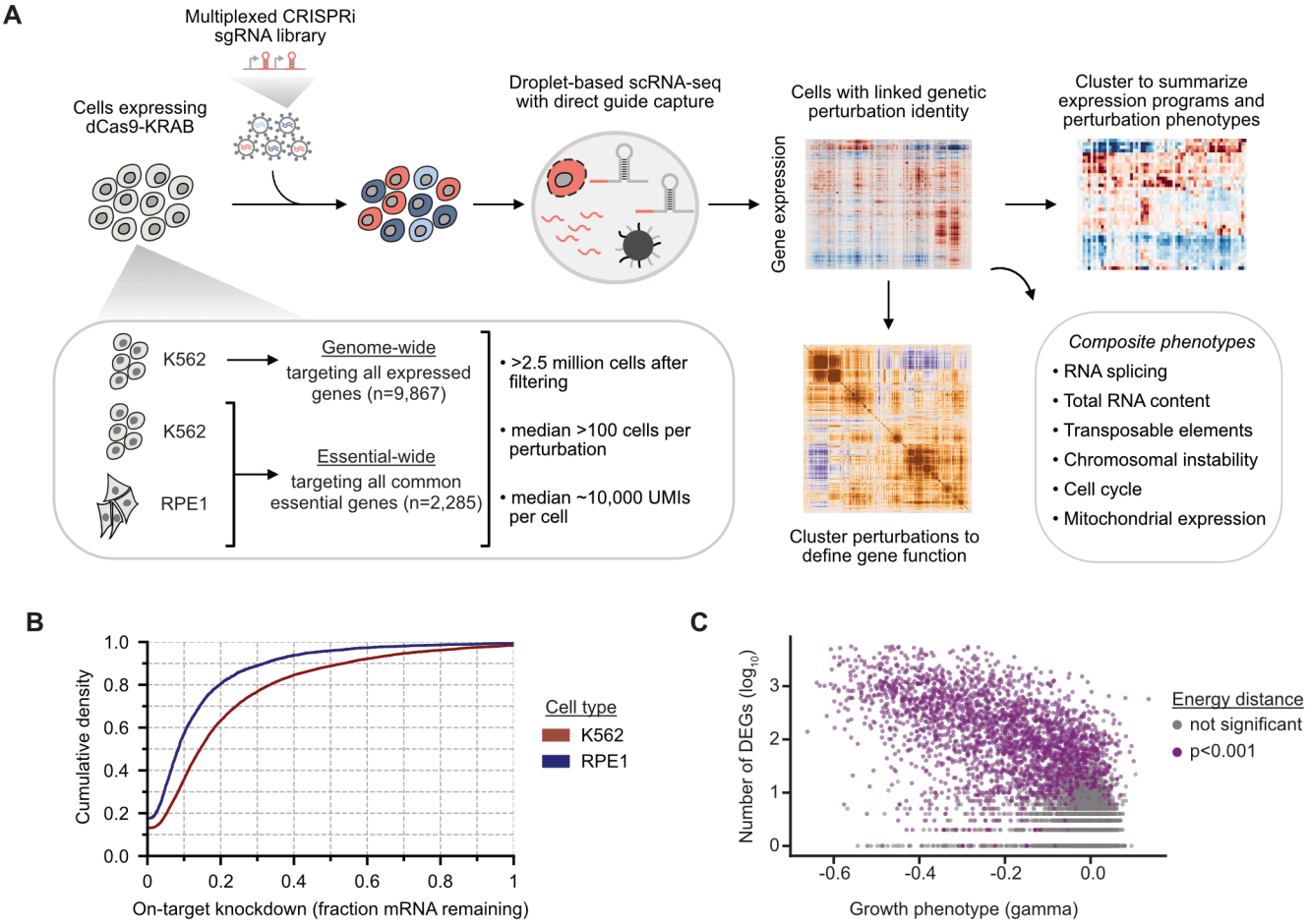

Figure 1: Genome-scale Perturb-seq via multiplexed CRISPRi. (A) Schematic experimental strategy. A multiplexed CRISPRi sgRNA library was used to knock down all expressed genes (in K562 cells) or all common essential genes (in RPE1 and K562 cells). Cells were transcriptionally profiled using droplet-based single-cell RNA-sequencing, with genetic perturbations assigned to cells by direct capture and sequencing of sgRNAs. **(B)** On-target knockdown statistics. Cumulative density plot of on-target knockdown, for n=9,464 target genes in K562 cells (red) and n=2,333 target genes in RPE1 cells (blue). **(C)** Comparing growth phenotype versus the number of differentially expressed genes (DEGs) for each multiplexed guide pair in K562 cells. Growth phenotypes are reported as the $\log_2$ guide enrichment per cell doubling (gamma). DEGs were determined using a two-sample Anderson-Darling test compared against non-targeting guides, and a pseudocount of 1 was added to the number of DEGs before $\log_{10}$ transformation. Dots are colored by Energy distance as either permutation significant (purple) or not significant (grey). The growth phenotype and number of DEGs are anticorrelated (Spearman's rho=-0.51).

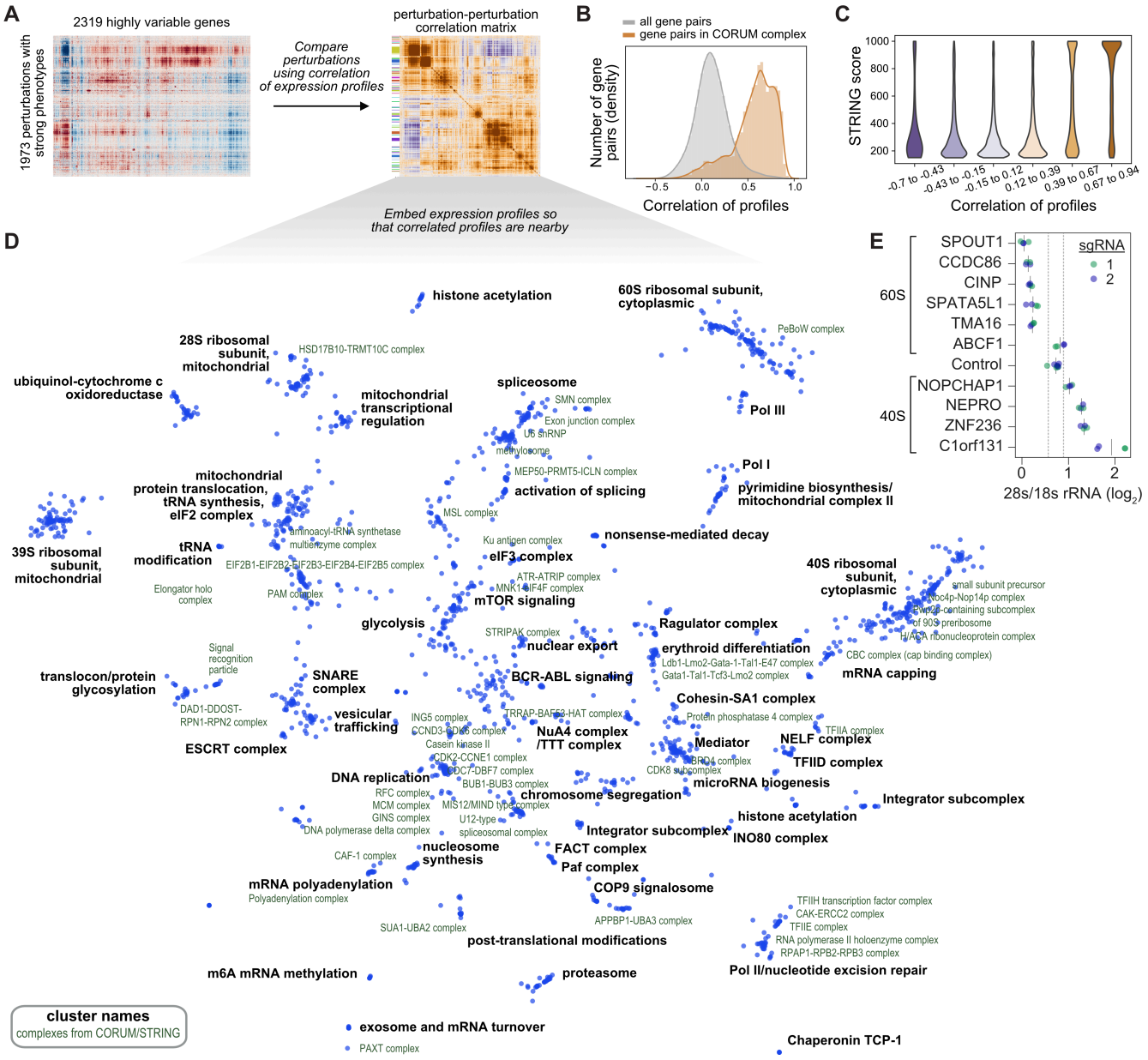

Figure 2: Data-driven inference of gene function from transcriptional phenotypes. (A) Schematic of analysis. To examine the ability of transcriptional phenotypes to assign gene function, we analyzed 1973 genetic perturbations that elicited strong responses. Perturbations were compared and clustered using the correlation of gene expression across 2319 highly variable genes. **(B)** Expression profile correlations among genes in curated complexes. 327 protein complexes from the CORUM3.0 database have at least two thirds of complex subunits within the dataset. Plot compares the distribution of pairwise expression profile correlations among genes in complexes vs. all possible gene-gene pairs. **(C)** Comparing expression profile correlations to predicted protein-protein interactions from STRING. 243,558 gene-gene relationships within the dataset are scored within STRING. The relationships were sorted into 6 equally spaced bins based on expression profile correlation. Plot shows kernel density estimates of STRING scores within each bin. **(D)** Minimum distortion embedding of dataset. Each dot represents a genetic perturbation, arranged so that perturbations with correlated expression profiles are nearby in the two dimensional embedding. Manual annotations (black labels) of

cluster function are placed near the median location of genes within the cluster. CORUM complexes or STRING clusters (green labels) are annotated when involved genes are nearby within the embedding. **(E)** Quantification of 28s to 18s rRNA ratio. Poorly characterized genes with Perturb-seq predicted roles in ribosome biogenesis were targeted by CRISPRi. The 28s/18s rRNA ratio was measured by Bioanalyzer electrophoresis in biological duplicate with two distinct sgRNAs per gene (green and blue; solid grey lines represent mean). Dotted grey lines represent two standard deviations above and below the mean of non-targeting controls.

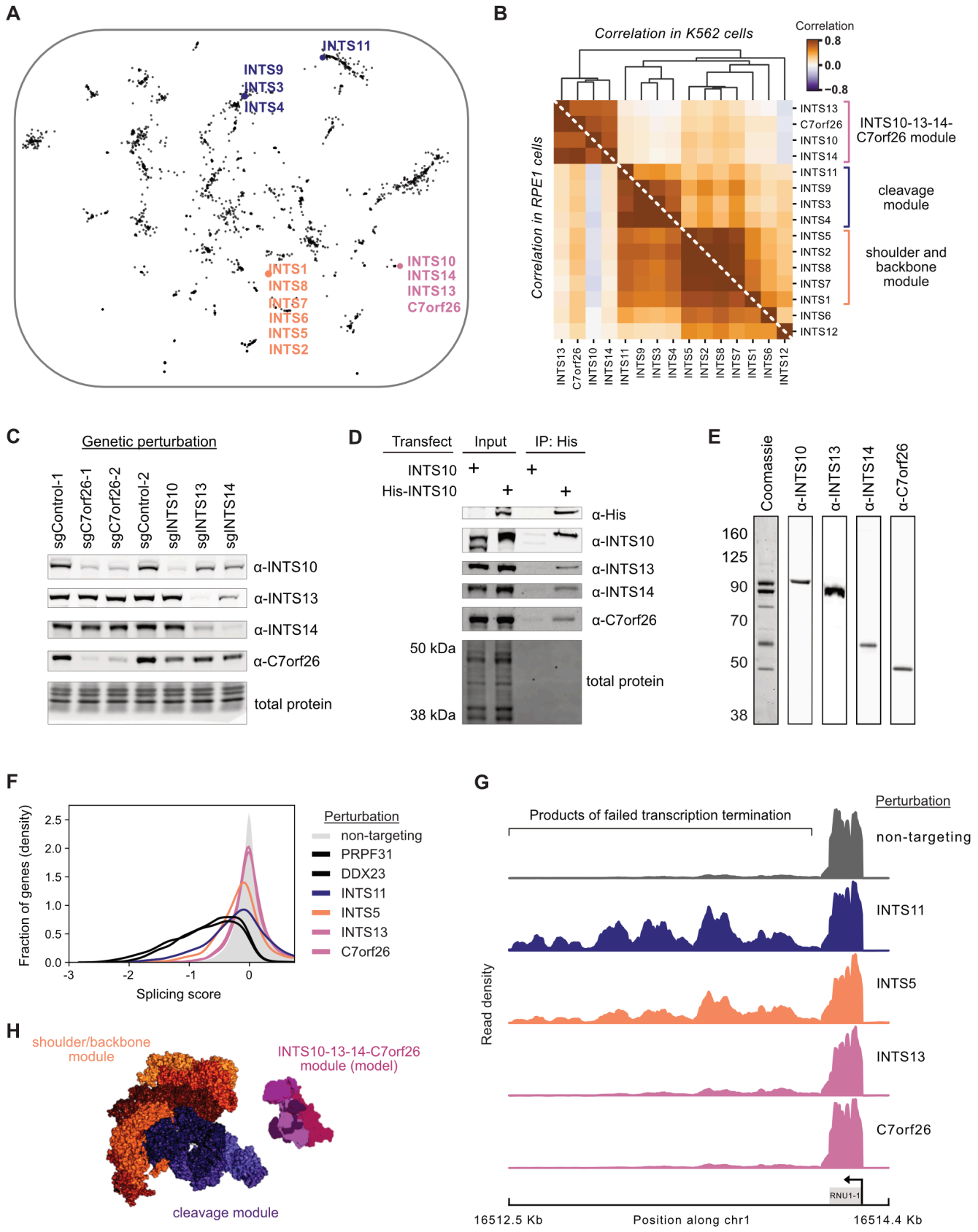

Figure 3. (legend on next page)

Figure 3: Perturb-seq discovers a novel gene member and functional submodules of the Integrator complex. (A) Location of known Integrator complex members in the minimum distortion embedding. **(B)** Relationship between Integrator complex members and *C7orf26* in K562 cells (top) and RPE1 cells (bottom). The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of Integrator complex members. Genetic perturbations are ordered by average linkage hierarchical clustering based on correlation in K562 cells. Functional modules suggested by the clustering are highlighted. **(C)** Co-depletion of Integrator complex members. Individual Integrator complex members were depleted in CRISPRi K562 cells. Lysates were then probed for other module members by western blot. **(D)** Co-immunoprecipitation of endogenous C7orf26 with His-INTS10. HEK293T were transfected with His-INTS10 or INTS10. Cell lysates were affinity purified and select Integrator proteins were probed by western blot. **(E)** Purification of a INTS10-INTS13-INTS14-C7orf26 complex. His-INTS10, INTS13, INTS14, and C7orf26 were overexpressed in Expi293 cells, affinity purified, and separated via SEC. The INTS10-INTS13-INTS14-C7orf26 proteins co-fractionated as visualized by Western blotting. **(F)** Effects of Integrator modules on splicing from Perturb-seq data. Histogram (kernel density estimate) compares gene-level splicing scores. Splicing scores represent the change in the $\log_2$ ratio of total to unspliced reads for each gene relative to non-targeting control guides. Representative genetic perturbations from Integrator modules as well as the spliceosome are shown colored by module. **(G)** Density of PRO-seq reads at the snRNA *RNU1-1* locus mapping actively engaged RNA polymerase II. For each perturbation, densities are shown relative to the maximum read count in the locus. **(H)** Structure of the Integrator complex colored by functional modules revealed by Perturb-seq. The endonuclease (blue) and shoulder/backbone (orange) modules were obtained from the cryo-EM structure (*31*). The model of the newly discovered 10-13-14-C7orf26 module was built by docking the crystal structure of INTS13-INTS14 (*33*) with an AlphaFold multimeric model of INTS10 and C7orf26.

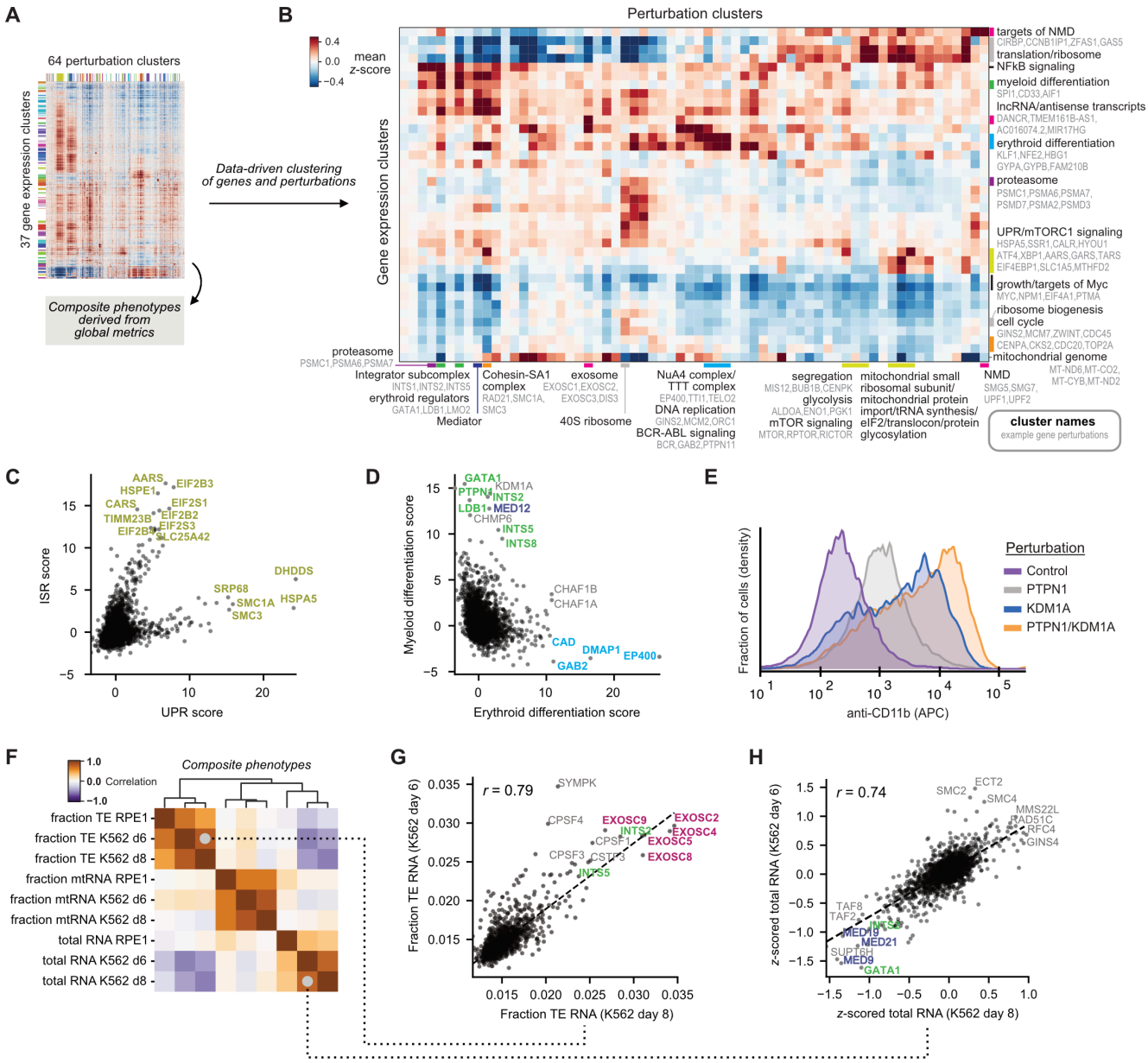

Figure 4: Summarizing genotype-phenotype relationships with Perturb-seq. (A) Schematic of analysis. To produce a high-level summary of genotype-phenotype relationships in K562 cells, 1973 genetic perturbations that elicited strong responses and 2319 highly variable genes were clustered using HDBSCAN after nonlinear embedding. Alternatively, composite phenotypes were derived from global metrics in a hypothesis-driven manner. **(B)** Heatmap of the high-level genotype-phenotype map. The heatmap represents the mean z-scored expression for gene expression and perturbation clusters. For a subset of clusters, clustered are labelled with manual annotations (black labels) of cluster function along with example genes within the cluster (light gray labels). **(C)** Comparison of ISR and UPR scores for genetic perturbations. Scores were recovered from unbiased clustering of genes by genetic dependency, and manually annotated. **(D)** Comparison of erythroid and myeloid differentiation scores for genetic perturbations. Scores were recovered from unbiased clustering of genes by genetic dependency, and manually annotated. Genetic perturbations are colored to reflect cluster identity. **(E)** Expression of CD11b/ITGAM in

K562 cells upon knockdown of *PTPN1* or *KDM1A*. CD11b was labelled by cell surface staining with anti-CD11b antibody and measured by flow cytometry. **(F)** Correlation of composite phenotypes across time points and cell types. Composite phenotypes were defined in a hypothesis-driven manner. Fraction TE (repetitive and transposable element) represents the number of non-intronic reads mapped to TEs over total, averaged over all cells bearing each perturbation (both collapsed on UMIs). Fraction mtRNA represents the mean number of reads mapped to mitochondrial genome protein-coding genes over total. Total RNA represents the mean total RNA content (number of UMIs). **(G)** Comparison of TE expression across time points. The mean fraction TE reads per perturbation is highly correlated across time points in K562 cells (r=0.79). Genetic perturbations are colored to reflect cluster identity. **(H)** Comparison of total RNA content across time points. Total RNA content per perturbation is highly correlated across time points in K562 cells (r=0.74). Genetic perturbations are colored to reflect cluster identity.

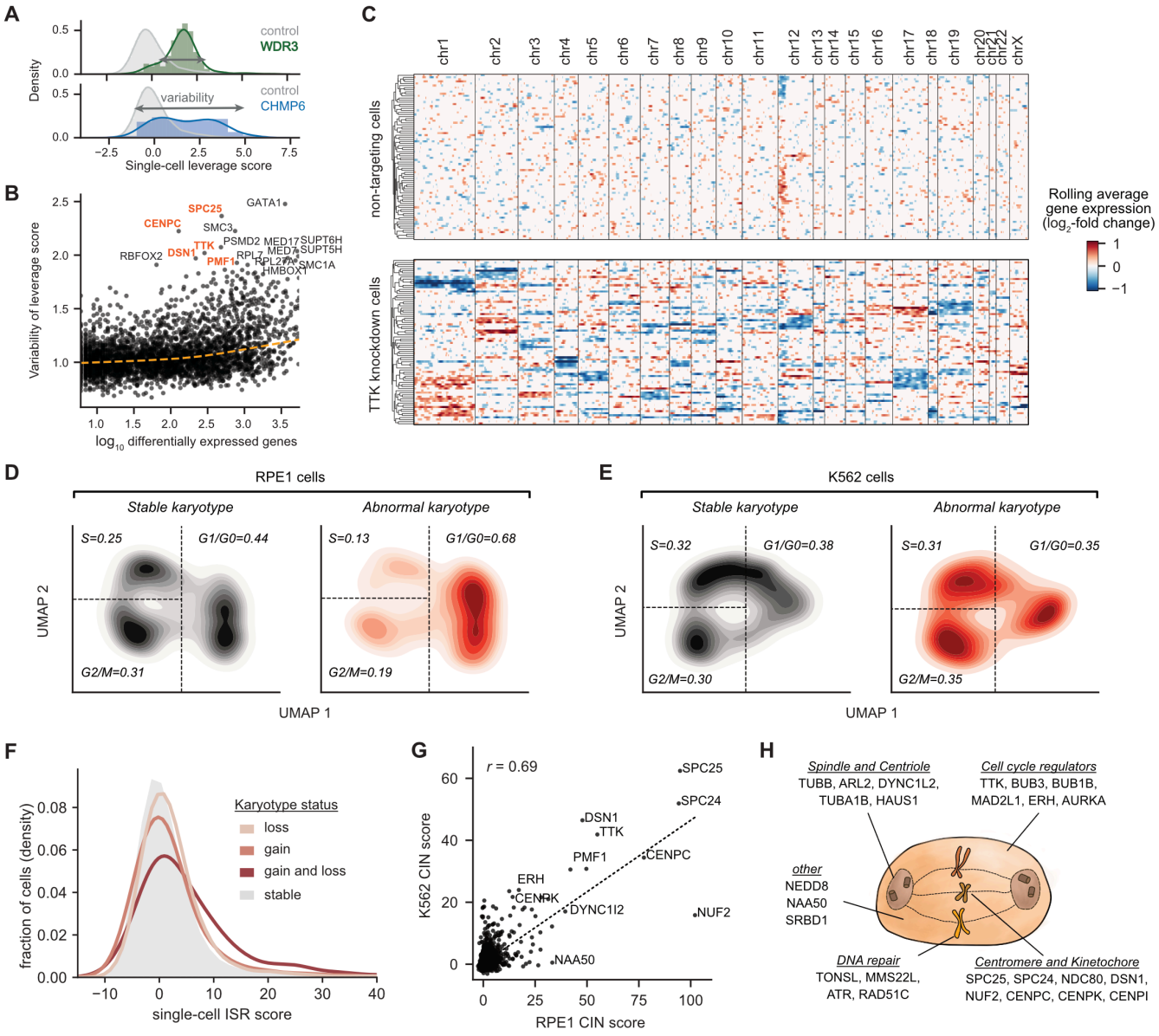

Figure 5: Exploring acute consequences and genetic drivers of aneuploidy in single-cells. (A) Schematic of heterogeneity statistic. Single-cell leverage scores quantify how outlying each cell is relative to non-targeting control cells by PCA. For each perturbation, heterogeneity of single-cell phenotypes is quantified as the standard deviation of leverage scores. **(B)** Identifying heterogeneous perturbations. Known regulators of chromosome segregation were among the perturbations with the highest single-cell heterogeneity (high variability of leverage scores), especially compared to their number of differentially expressed genes (based on Anderson-Darling test). **(C)** Heatmap of chromosomal copy number inference from Perturb-seq data. For all genes (expressed >0.05 UMI per cell), the log-fold change in expression is calculated with respect to the average of non-targeting control cells, and genes are ordered along the genome. A weighted moving average of 100 genes is used infer copy number changes (columns) in single-cells (rows) with noise and median filtering. 80 *TTK* knockdown RPE1 cells and 80 randomly sampled non-targeting control RPE1 cells are shown. Cells are ordered by average linkage hierarchical clustering based on correlation of chromosomal copy number profiles. **(D, E)** Comparison of cell

cycle occupancy upon acute karyotypic changes. Abnormal karyotypic cells were defined as having ≥1 chromosome with evidence of changes in chromosomal copy number for >80% of the chromosomal length. For single-cells, cell-cycle positioning was inferred by UMAP dimension reduction on differential expression profiles of 199 selected cell-cycle regulated genes. Cell cycle occupancy is shown as a 2D kernel density estimate of a random subset of 1000 cells per karyotypic status. Approximate gates between cell cycle phases (G1 or G0; S; G2 or M) are shown as dotted lines, and the fraction of cells in each cell cycle phase are indicated. **(F)** Effect of chromosomal instability (CIN) on activation of the Integrated Stress Response (ISR). Histogram (kernel density estimate) compares the ISR score versus CIN status in RPE1 cells. CIN status is defined as evidence of gain or loss of chromosomal copy number for >80% of the chromosomal length, with 240,768 stable cells, 5,522 cells bearing chromosomal loss, 1987 cells bearing chromosomal gain, and 904 cells bearing gain and loss of chromosomes. ISR score is defined as the sum of z-normalized expression of ISR marker genes where increased values indicate stronger ISR activation. **(G)** Comparison of the effect of genetic perturbations on the CIN score across cell types. For each genetic perturbation, the CIN score is calculated as the mean single-cell sum of squared CIN values, *z*-normalized relative to non-targeting control perturbations. The CIN score is correlated across cell types (r=0.69). **(H)** Schematic of a subset of genetic perturbations that drive CIN. CIN drivers play diverse roles in mitosis, cell cycle regulation, and DNA repair.

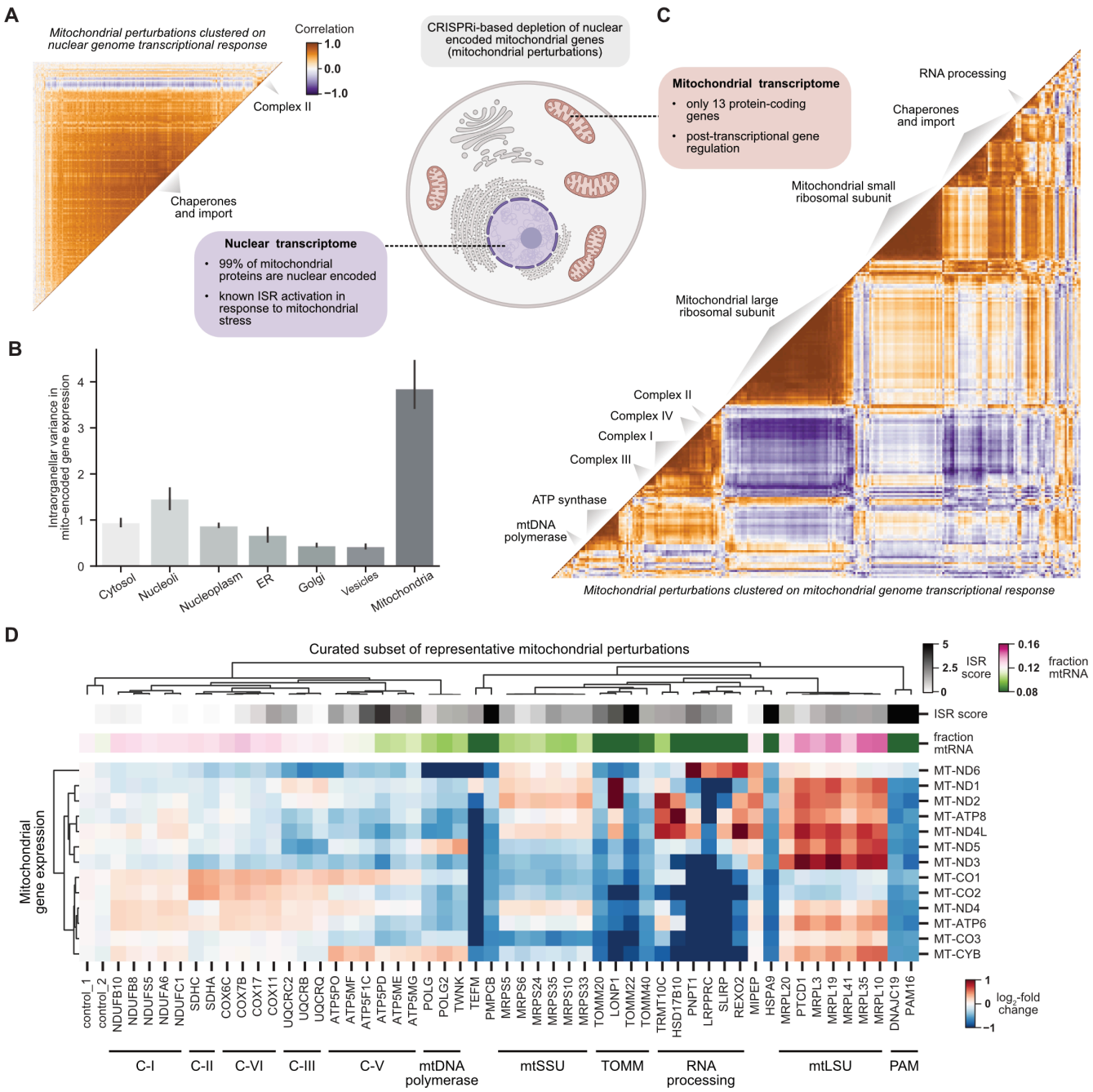

Figure 6: Global organization of the transcriptional response to mitochondrial stress. (A) Clustering mitochondrial perturbations by nuclear transcriptional response. CRISPRi enables knockdown of nuclear-encoded genes whose protein products are targeted to mitochondria (mitochondrial perturbations). Mitochondrial perturbations were annotated by MitoCarta3.0 and subset to those with a strong transcriptional phenotype (n=268 mitochondrial perturbations). Gene expression profiles were restricted to nuclear encoded genes which includes 99% of mitochondrial proteins. The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of mitochondrial perturbations in K562 cells. Genetic perturbations were clustered by HDBSCAN with a correlation metric. **(B)** Comparing variability in the mitochondrial transcriptome by perturbation localization. The mitochondrial genome encodes 13 protein-coding

genes. Genetic perturbations were grouped based on localization of their protein products as determined by the Human Protein Atlas. For each of these 13 mitochondrially encoded genes, the variance in pseudobulk z-normalized expression profiles was calculated between all perturbations with the same localization. Barplots represent the average across genes with 95% confidence interval obtained by bootstrapping. **(C)** Clustering mitochondrial perturbations by mitochondrial transcriptional response. Mitochondrial perturbations were annotated by MitoCarta3.0 and subset to those with a strong transcriptional phenotype as above (n=268 mitochondrial perturbations). Gene expression profiles were restricted to the 13 mitochondrial-encoded genes. The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of mitochondrial perturbations in K562 cells. Genetic perturbations are clustered by HDBSCAN with a correlation metric. Clusters were manually annotated. **(D)** Heatmap visualizing the mitochondrial genome transcriptional response to diverse mitochondrial stressors. The expression (log_2 fold-change relative to non-targeting controls) of the 13 mitochondrially encoded genes is shown for a subset of perturbations representative of different mitochondrial complexes or function. Neither the ISR score nor mean fraction of mitochondrial RNA (mtRNA) would allow for high-resolution clustering by function as provided by the mitochondrial genome response. Genetic perturbations and genes are ordered by average linkage hierarchical clustering with a correlation metric.

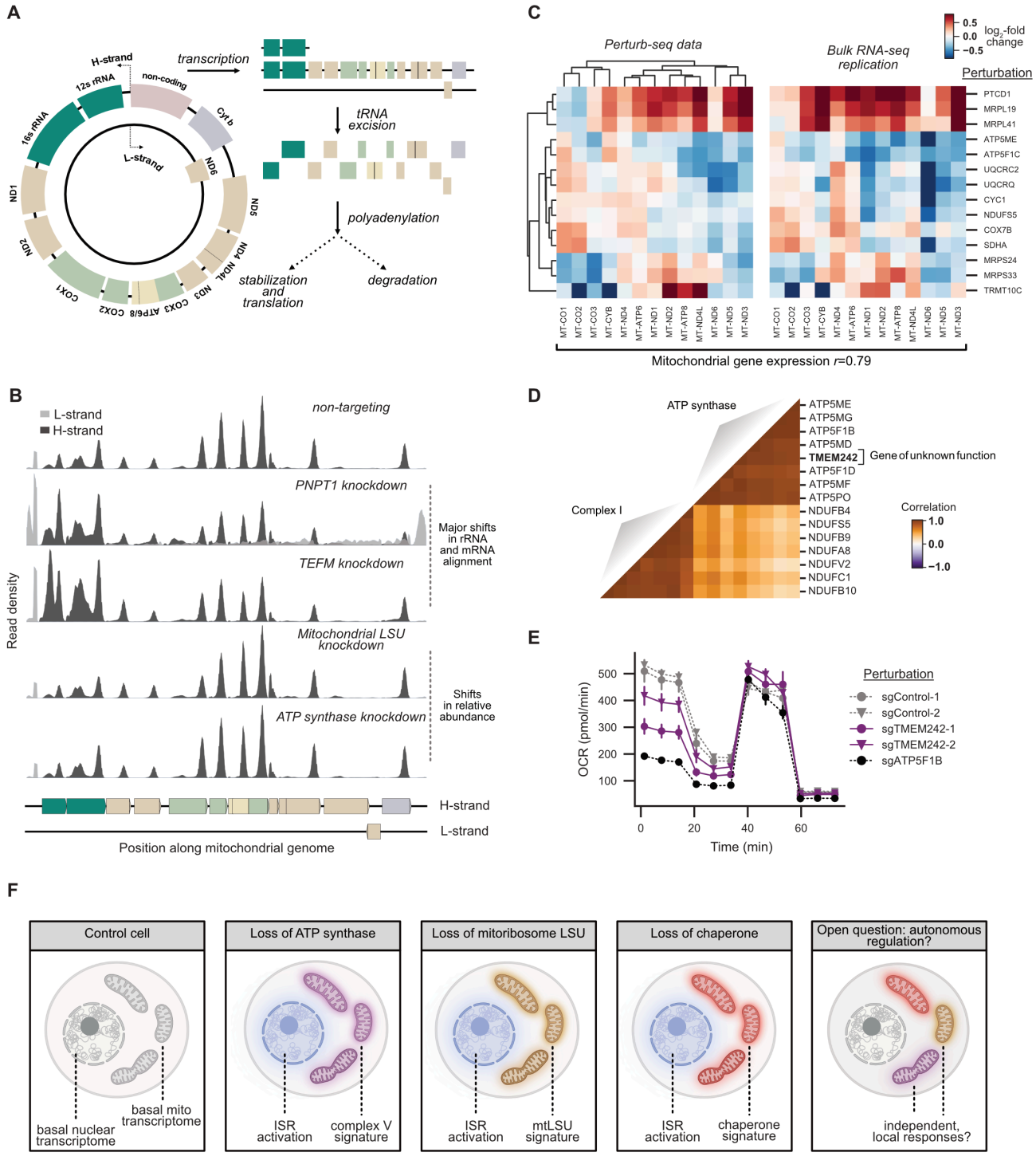

Figure 7: Investigating regulation of the mitochondrial genome in stress. (A) Schematic of the mitochondrial transcriptome. Each human cell contains many copies of the circular 16.6 kb mitochondrial genome distributed throughout the mitochondrial network. The human mitochondrial genome encodes 2 rRNAs, 22 tRNAs, and 13 protein-coding genes. Both the heavy (H) and light (L) strand of the genome are transcribed as polycistronic transcripts punctuated by tRNAs. Excision of tRNAs from transcripts generates nascent mRNA precursors (colored by complex membership). mRNA precursors can then be polyadenylated, stabilized, or

degraded. **(B)** Density of Perturb-seq reads along the mitochondrial genome from select genetic perturbations. Reads are aligned to both the H-strand (dark grey) and L-strand (light grey). For each perturbation, densities are shown relative to the maximum read count in the locus. **(C)** Comparison of mitochondrial gene expression profiles between Perturb-seq and bulk RNA-seq. Heatmap displays $\log_2$-fold changes in expression of the 13 mitochondrial encoded genes (columns) for genetic perturbations (rows) in Perturb-seq and bulk RNA-seq data collected from K562 cells. Bulk RNA-seq was conducted to analyze total RNA (including non-polyadenylated RNA), with data representing the average of biological replicates. Genetic perturbations and genes are ordered by average linkage hierarchical clustering with a Euclidean distance metric. The profiles are strongly correlated ($r=0.79$, $p<10^{-39}$). **(D)** Clustering of *TMEM242* genetic perturbation based on the mitochondrial transcriptome. Genetic perturbations to members of ATP synthase and Complex I of the respiratory chain were compared to knockdown of *TMEM242*, a mitochondrial gene of unknown function. Gene expression profiles were restricted to the 13 mitochondrially encoded genes. The heatmap displays the Pearson correlation between pseudobulk *z*-normalized gene expression profiles of mitochondrial perturbations in K562 cells. Genetic perturbations are ordered by HDBSCAN with a correlation metric. **(E)** Effect of *TMEM242* knockdown on mitochondrial respiration. A Seahorse analyzer was used to monitor oxygen consumption rate (OCR). The Mito Stress Test consists of sequential addition of oligomycin (an ATP synthase inhibitor that enables measurement of ATP-productive respiration), FCCP (an uncoupling agent that enables measurement of maximal respiratory capacity), and a mixture of rotenone and antimycin A (inhibitors of Complex I and Complex III, respectively, that enable measurement of non-mitochondrial respiration). Data is presented as average ± SEM, n=6. **(F)** Schematic diagram of mitochondrial stress response.