Title

POOLED GENETIC SCREENS WITH RICH READOUTS AT SCALE AND IN VIVO

Author

Saunders, Reuben Abraham

Publication Date

2024

Peer reviewed|Thesis/dissertation

Pooled genetic screens with rich readouts at scale and in vivo

by

Reuben Saunders

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Genetics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

Carol Gross

7AFBE9046E0744C...

Chair

DocuSigned by:

Jonathan Weissman

[illegible]57BC64CD...

Signed by:

Bruce Conklin

[illegible]410...

DocuSigned by:

Alexander Johnson

E2BBE6A63C2745B...

Committee Members

לאלה שהגיעוני לזמן הזה

Acknowledgements

I acknowledge several spectacular mentors and colleagues who have contributed to my intellectual development and helped me to reach this point.

Most of all, I acknowledge my advisor Jonathan for his brilliance, encouragement, and support. Jonathan is the only scientist I know who truly understands how the exponential advance of technology impacts biology. Jonathan thinks exponentially. I am generally stuck in the linear regime; if I have ever climb up into his regime, it is because he lent an encouraging hand.

I also acknowledge my previous research mentors, Bob Sauer and David Ron. The rigorous training in biological mechanism I received from them drives my interest in methods for causality screening, such as Perturb-seq and our imaging assay.

I thank Xiaowei Zhuang for welcoming me into her lab for the imaging project. I am extremely excited to have her group as a home base and continue my collaboration with her and her students as a Junior Fellow.

I thank my committee: Carol Gross, Sandy Johnson, and Bruce Conklin. I have felt a bit disconnected from UCSF during my PhD, but they have remained good sources of advice for both scientific and non-scientific issues.

I have had the extraordinarily good fortune of collaborating with Will Allen, who is a friend as well as a mentor. I look forward to going to lab because he is there. I am thrilled to continue our collaboration.

I thank other collaborators in the Weissman Lab and elsewhere, including Karina, Jaspreet, Xingjie, Gayathri, Luke, Katie, Pu, William, Marco, Dan, Angie, Joseph, Jess, Tom, and Richard.

I thank the Hertz foundation for funding and for a wonderful community.

Contributions

Chapter 1 contains conceptual, experimental, analytical, and/or writing contributions from me and from Joseph M. Replogle, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, & Jonathan S. Weissman.

Chapter 2 contains conceptual, experimental, analytical, and/or writing contributions from me and from William E. Allen, Xingjie Pan, Jaspreet Sandhu, Xiaowei Zhuang, and Jonathan Weissman.

POOLED GENETIC SCREENS WITH RICH READOUTS AT SCALE AND IN VIVO

REUBEN SAUNDERS

ABSTRACT

A central goal of organismal biology is to understand how the physiological functions of organs and tissues arise from the coordinated activity of thousands of genes in diverse principal and support cell types. Recent large-scale atlasing efforts have characterized the molecular composition of thousands of distinct cell types across the organs; ongoing work is mapping the spatial organization of these cell types across tissues. Although such efforts provide a foundation resource, a major challenge for the field is to causally understand how cell types, cell states, and multicellular structures and motifs are produced and regulated by the actions of specific genes and molecular pathways. This thesis describes an approach that could enable a systematic and comprehensive understanding of how specific genes control diverse cellular and tissue phenotypes in living animals, under different physiological states. Chapter 1 describes the use of Perturb-seq to create a comprehensive portrait of the transcriptional impact of genetic perturbations at genome-scale in cell culture, enabling the prediction of function for thousands of genes, the investigation of complex cellular phenotypes. Chapter 2 describes two new technologies to take this many genotypes/complex phenotypes approach into animal tissue. The first is a new method for fixed cell Perturb-seq, which drastically simplifies the processing of animal tissues and enables large-scale experiments. The second is an imaging-based method that captures both spatial gene expression and multiplexed immunofluorescence. Through an integrated analysis of imaging and transcriptional phenotypes, we identify novel regulators of hepatocyte zonation, reveal how proteostatic stress pathway activation can broadly affect the expression of secreted proteins, and show how diverse cellular states can produce convergent effects on lipid accumulation. In total,

we provide a new path to characterizing multidimensional genotype-phenotype relationships in living tissues, in a scalable manner that will be broadly applicable across diverse tissues as well as physiological and disease state.

TABLE OF CONTENTS

Chapter 1: Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq *1*

Summary **1**

Keywords **1**

Introduction **2**

Results **5**

Discussion **38**

Chapter 2: Large-scale in vivo pooled genetic screens with highly multiplexed phenotypes in the mouse liver *43*

Introduction **43**

Results **46**

Discussion **79**

REFERENCES *85*

LIST OF FIGURES

Figure 1.1 : Genome-scale Perturb-seq via multiplexed CRISPRi. 8

Figure 1.2: Data-driven inference of gene function from transcriptional phenotypes. 13

Figure 1.3: Perturb-seq discovers a novel gene member and functional submodules of the Integrator complex. 18

Figure 1.4: Summarizing genotype-phenotype relationships with Perturb-seq. 23

Figure 1.5: Exploring acute consequences and genetic drivers of aneuploidy in single-cells. 28

Figure 1.6: Global organization of the transcriptional response to mitochondrial stress. 32

Figure 1.7: Investigating regulation of the mitochondrial genome in stress. 36

Figure 2.1: Pooled in vivo genetic screens through imaging and sequencing.....47

Figure 2.2: Heterogeneity in subcellular morphology by cell type and state 55

Figure 2.3: Large-scale in vivo screening in CRISPR mosaic livers 60

Figure 2.4: Dual-mode in vivo screening with strong phenotypes.....65

Figure 2.5: Mapping in vivo genotype-phenotype landscapes 68

Figure 2.6: Identifying genetic regulators of liver physiology 70

igure 2.7: Exploring liver physiology with dual-mode screening 75

Chapter 1: Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq

A modified version of this chapter is now published in *Cell* as Replogle and Saunders *et al*, 2022.

Summary

A central goal of genetics is to define relationships between genotypes and phenotypes. High-content phenotypic screens such as Perturb-seq (CRISPR-based single-cell RNA-sequencing screens) 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 CRISPRi across >2.5 million human cells and present a framework to power biological discovery from the resulting genotype-phenotype map. We use transcriptional phenotypes to predict gene functions, 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. Leveraging this, we systematically identify genetic drivers and consequences of aneuploidy and discover an unanticipated layer of stress-specific regulation of the mitochondrial genome. Our information-rich genotype-phenotype map reveals a multidimensional portrait of gene and cellular function.

Introduction

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 (Camp et al., 2019). CRISPR-Cas tools now enable the deletion, mutation, repression, or activation of genes with ease (Doench, 2018). 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 (Kramer et al., 2018; Tsherniak et al., 2017; Wang et al., 2021, 2015). In parallel, dramatic improvements in molecular phenotyping now allow for single-cell readouts of epigenetic, transcriptomic, proteomic, and imaging information (Stuart and Satija, 2019). Applied to reverse genetics, single-cell profiling can refine the understanding of how select genetic perturbations affect cell types and cell states.

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

serendipitous discovery is 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) (Adamson et al., 2016; Datlinger et al., 2017; Dixit et al., 2016; Jaitin et al., 2016; Replogle et al., 2020), technical advances have enabled their application to epigenetic (Rubin et al., 2019), imaging (Feldman et al., 2019), or multimodal phenotypes (Frangieh et al., 2021; Mimitou et al., 2019; Papalexi et al., 2021). From these rich data, it is possible to identify genetic perturbations that cause a specific behavior as well as to catalog the spectrum of phenotypes associated with each genetic perturbation. Despite the promise of single-cell CRISPR screens, 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 responses to the same perturbations differ between cell types, with implications for understanding the organizing principles of cellular systems (Tanay and Regev, 2017).

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

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). We sought to exploit and understand the rich information content of transcriptomic phenotypes by studying a comprehensive set of genetic perturbations in a given cell type. To enable genome-scale Perturb-seq, we considered key parameters that would increase scalability and data quality, such as the genetic perturbation modality and sgRNA library.

Although Perturb-seq is compatible with a range of CRISPR-based perturbations including knockout (Datlinger et al., 2017; Dixit et al., 2016; Jaitin et al., 2016), knockdown (CRISPRi) (Adamson et al., 2016), or activation (CRISPRa) (Norman et al., 2019), we elected to use CRISPRi for several reasons. First, compared to gain-of-function CRISPRa perturbations, a higher proportion of loss-of-function CRISPRi perturbations yield phenotypes in growth and chemical-genetic screens, especially for members of large protein complexes (Gilbert et al., 2014; Horlbeck et al., 2016). Second, CRISPRi allows direct measurement of the efficacy of genetic perturbation—knockdown—by scRNA-seq. This is challenging with CRISPR nuclease perturbations due to incomplete nonsense-mediated decay (Dixit et al., 2016) and poor coverage of indels with 3' scRNA-seq methods. Exploiting this feature of CRISPRi allowed us to target each gene in our library with a single element and empirically exclude unperturbed genes from downstream analysis. Third, CRISPRi tends to yield more homogeneous genetic perturbation than nuclease-based CRISPR knockout, which can generate a subset of cells bearing active in-frame indels (Smits et al., 2019). The relative homogeneity of CRISPRi reduces selection for unperturbed cells, especially when studying essential genes. Fourth, unlike nuclease-based gene knockout, CRISPRi

does not lead to activation of the DNA damage response which can alter cell state and transcriptional signatures (Haapaniemi et al., 2018).

To improve scalability, we optimized our CRISPRi sgRNA libraries. To maximize CRISPRi efficacy, we used multiplexed CRISPRi libraries in which each construct contains two distinct sgRNAs targeting the same gene (Replogle et al., 2020). 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.

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 (Gilbert et al., 2014). 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 sampled at day 6 after transduction (n=2,057 genes; n=2,176 total perturbations). These two time points were selected as the phenotypic effects of different genetic perturbations can manifest at variable time points based on technical and biological factors (Bock et al., 2022). We reasoned that assaying an early and intermediate timepoint would allow us to observe strong phenotypes for essential genes that are likely to cause cell death as well for genes with less severe cellular consequences, consistent with kinetic analyses of growth screens (Cross et al., 2016). We also intended to use these two independent datasets to assess experimental reproducibility. To limit costs, we focused the replication screens on essential genes. 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 (Alerasool et al., 2020). In contrast to K562 cells, RPE1 cells are a non-cancerous retinal pigment epithelial cell line that are hTERT-immortalized, near-euploid, adherent, and p53-positive. In RPE1 cells, we performed a screen targeting common essential genes plus a subset of nonessential genes that produced phenotypes in K562 cells sampled at day 7 after transduction (n=2,393 genes; n=2,549 total perturbations). By studying two different cellular models, we hoped to focus our downstream work on conserved cell biological discoveries.

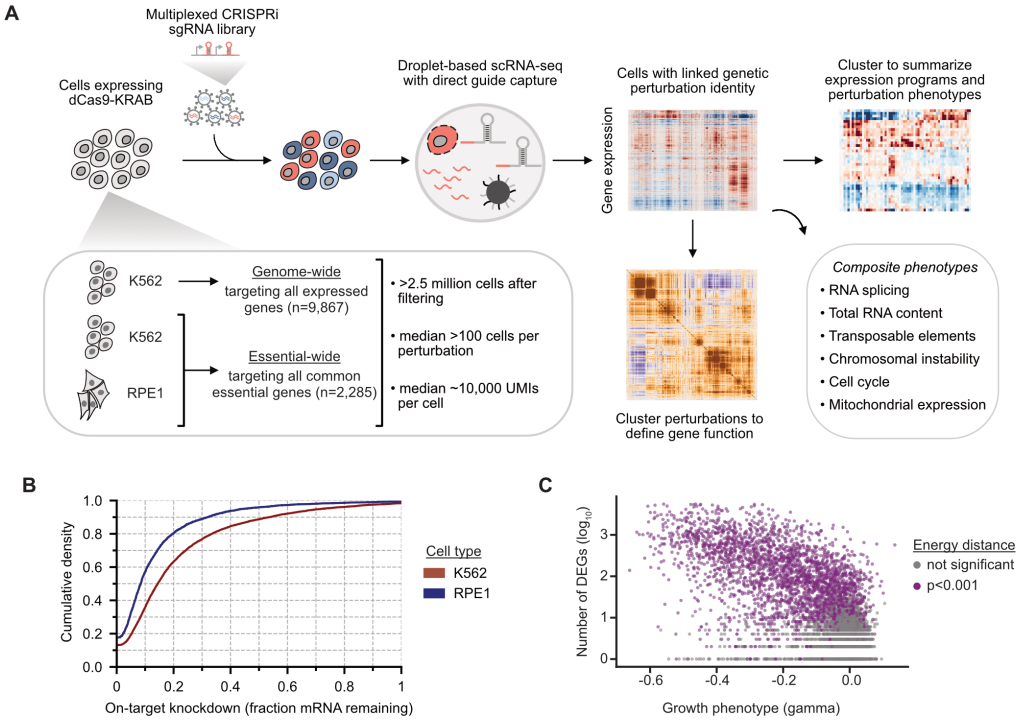

Figure 1.1: Genome-scale Perturb-seq via multiplexed CRISPRi.

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).

We conducted these three screens with 10x Genomics droplet-based 3' scRNA-seq and direct sgRNA capture (Replogle et al., 2020). 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. 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 (Replogle et al., 2020). 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.

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. 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. 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 and concordant with the energy distance test (78.7% concordance by Jaccard index).

We then explored features of genetic perturbations that predict the likelihood of causing 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). Nonetheless, a substantial number of genetic perturbations that cause a transcriptional phenotype have a negligible growth phenotype (n=771), indicating that many genetic perturbations influence cell state but not growth or survival. We also found that genes whose knockdown caused a strong transcriptional phenotypes were more likely to be highly expressed, to have an annotated subcellular localization, and to be components of core protein complexes such as the ribosome, spliceosome, and RNA polymerase.

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 (Rosenbluh et al., 2017), 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. Taken together, these results present a coherent picture where knockdown of a significant fraction of expressed genes causes a transcriptional response.

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 (Giurgiu et al., 2019). The corresponding expression profile correlations were markedly stronger (median correlation 0.61) than the background distribution of all possible gene pairs (median correlation 0.10) (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 (Szklarczyk et al., 2019). 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).

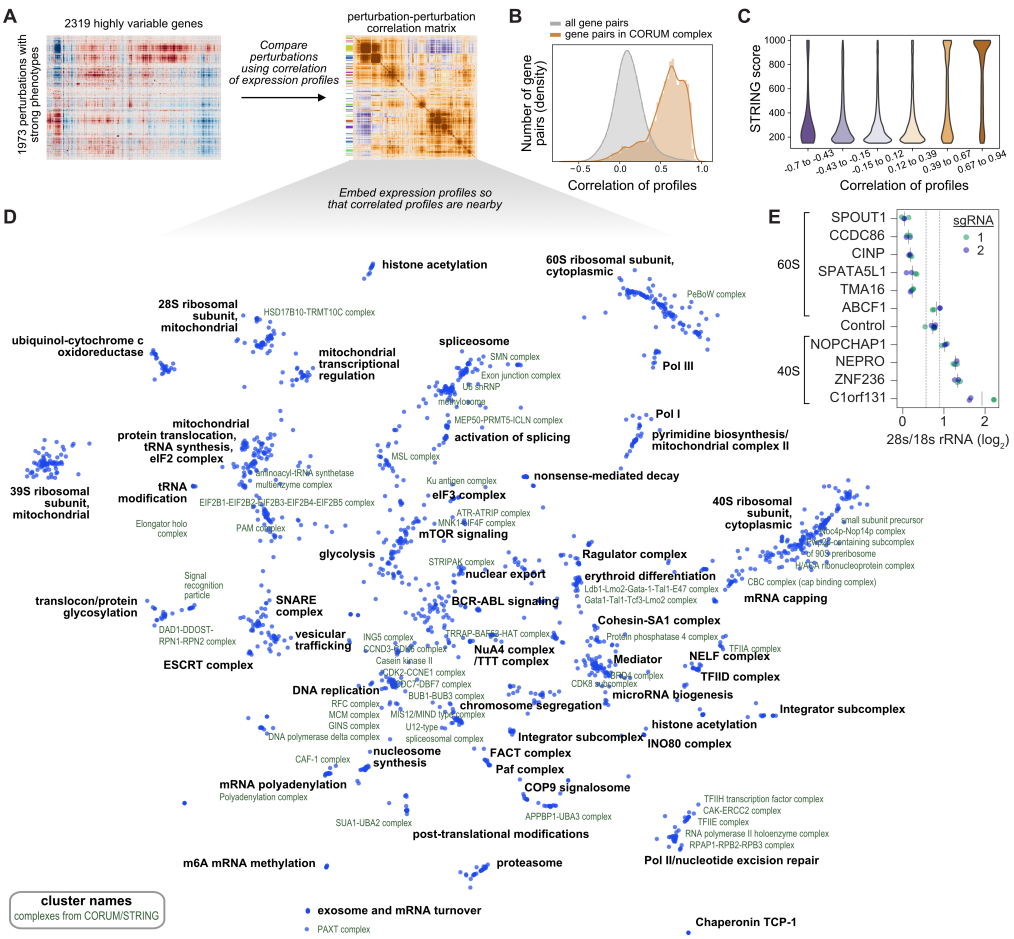

Figure 1.2: Data-driven inference of gene function from transcriptional phenotypes.

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. (Figure caption continued on the next page.)

E) (Figure caption continued from previous page.) CORUM complexes or STRING clusters (green labels) are annotated when involved genes are nearby within the embedding.

F) 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.

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

Next, we compared the similarity of transcriptional phenotypes between the three Perturb-seq datasets we collected. Comparing transcriptional phenotypes in K562 cells sampled at day 8 versus day 6, we found that both the relationships between perturbations (cophenetic correlation = 0.82) and transcriptional phenotypes (median r=0.50) were highly similar. By contrast, when comparing between the K562 and RPE1 datasets, we observed substantial variation in perturbation relationships (cophenetic correlation = 0.37) and transcriptional signatures (median r=0.23). At the level of pathways, knockdown of genes involved in protein export, basal transcription, or mRNA surveillance tended to produce highly similar transcriptional responses. By contrast, KEGG pathways related to “chronic myeloid leukemia” and “cell cycle” were enriched for genes with uncorrelated transcriptional phenotypes between the cell types, consistent with the CML origin of K562 cells and difference in p53 status of the cell lines, respectively. These differences highlight the value in future studies to apply largescale perturb-seq approaches to a range of diverse cell types.

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. 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 (Singh et al., 2021). This validation suggests that many poorly characterized genes can be assigned functional roles through Perturb-seq, although a subset of these relationships might be explained by indirect or off-target effects.

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.

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) (Kirstein et al., 2021). Each of the fourteen core subunits of Integrator was targeted in our experiment, allowing us to systematically compare their transcriptional phenotypes in K562 and RPE1 cells (Fig. 3B). *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 (Fianu et al., 2021; Zheng et al., 2020). The INTS1-2-5-7-8 functional module contained the subunits identified as the structural shoulder and backbone of Integrator. The INTS3-4-9-11 functional module contained the subunits identified as the 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 (Pfleiderer and Galej, 2021; Sabath et al., 2020).

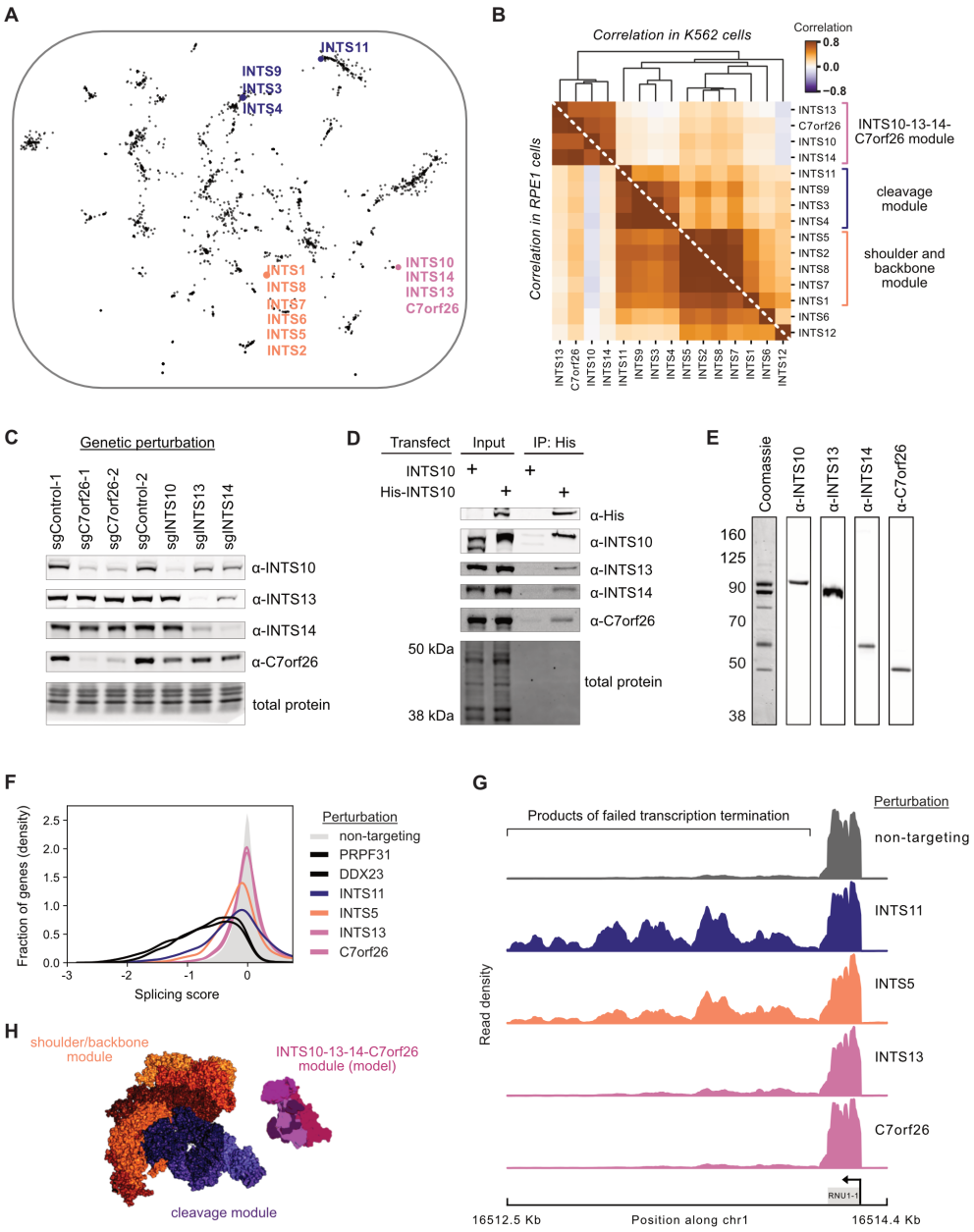

Figure 1.3: Perturb-seq discovers a novel gene member and functional submodules of the Integrator complex.

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. (Figure caption continued on the next page.)

C) (Figure caption continued from the previous page.) Co-depletion of Integrator complex members. Individual Integrator complex members were depleted in CRISPRi K562 cells. Lysates were then probed for other module members by western blot.
D) Co-immunoprecipitation of endogenous C7orf26 with His-INTS10. HEK293T were transfected with His-INTS10 or INTS10. Cell lysates were affinity purified and select Integrator proteins were probed by western blot.
E) Purification of a INTS10-INTS13-INTS14-C7orf26 complex. His-INTS10, INTS13, INTS14, and C7orf26 were overexpressed in Expi293 cells, affinity purified, and separated via SEC. The INTS10-INTS13-INTS14-C7orf26 proteins co-fractionated as visualized by Western blotting.
F) Effects of Integrator modules on splicing from Perturb-seq data. Histogram (kernel density estimate) compares gene-level splicing scores. Splicing scores represent the change in the $\log_2$ ratio of total to unspliced reads for each gene relative to non-targeting control guides. Representative genetic perturbations from Integrator modules as well as the spliceosome are shown colored by module.
G) Density of PRO-seq reads at the snRNA *RNU1-1* locus mapping actively engaged RNA polymerase II. For each perturbation, densities are shown relative to the maximum read count in the locus.
H) Structure of the Integrator complex colored by functional modules revealed by Perturb-seq. The endonuclease (blue) and shoulder/backbone (orange) modules were obtained from the cryo-EM structure (Zheng et al., 2020). The model of the newly discovered 10-13-14-C7orf26 module was built by docking the crystal structure of INTS13-INTS14 (Sabath et al., 2020) with an AlphaFold multimeric model of INTS10 and C7orf26.

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), consistent with previous reports (Boeing et al., 2016; Malovannaya et al., 2010). 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). 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 (Pan et al., 2022; Wainberg et al., 2021). 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, 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.

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, consistent with concurrent studies (Fig. 3H) (Funk et al., 2021; Pan et al., 2022; Wainberg et al., 2021). 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.

While clustering can organize genetic perturbations into pathways or complexes, it does not reveal the functional consequences of perturbations. To globally summarize the genotype-phenotype relationships in our dataset, we: (i) clustered genes into expression programs based on their co-regulation across perturbations; (ii) clustered perturbations with strong phenotypes based on their transcriptional profiles (as described above); and (iii) computed the average activity of each gene expression program within each perturbation cluster (Fig. 4A,B). This map uncovered many known gene expression programs associated with genetic perturbations, including

upregulation of proteasomal subunits due to proteosome dysfunction (Radhakrishnan et al., 2010), activation of NFkB signaling upon loss of ESCRT proteins (Mamińska et al., 2016), downregulation of growth-related genes in response to essential genetic perturbations, and upregulation of the cholesterol biosynthesis pathway in response to defects in vesicular trafficking (Luo et al., 2020). Beyond these large-scale relationships, we could also score the effects of individual genetic perturbations on different expression programs. For example, our analysis delineated the canonical branches of the cellular stress response into the independently regulated Unfolded Protein Response (UPR) and Integrated Stress Response (ISR) (Fig. 4C) (Adamson et al., 2016). The ISR was highly activated by loss of mitochondrial proteins, aminoacyl-tRNA synthetases, and translation initiation factors, whereas the UPR was activated by loss of ER-resident chaperones and translocation machinery. Collectively, this analysis establishes the ability of Perturb-seq to learn regulatory circuits by leveraging the variability of responses across perturbations.

Interestingly, our unbiased clustering uncovered many perturbations that drove the expression of markers of erythroid or myeloid differentiation, consistent with the known multilineage potential of K562 cells (Fig. 4D) (Leary et al., 1987). The scale of our experiment allowed us to comprehensively search for genes whose modulation promotes cellular differentiation, an application of major interest in both developmental and cancer biology. As expected, loss of central regulators of erythropoiesis (*GATA1*, *LDB1*, *LMO2*, and *KDM1A*) caused myeloid differentiation, whereas knockdown of *BCR-ABL* and its downstream adaptor *GAB2* induced erythroid differentiation (Orkin and Zon, 2008).

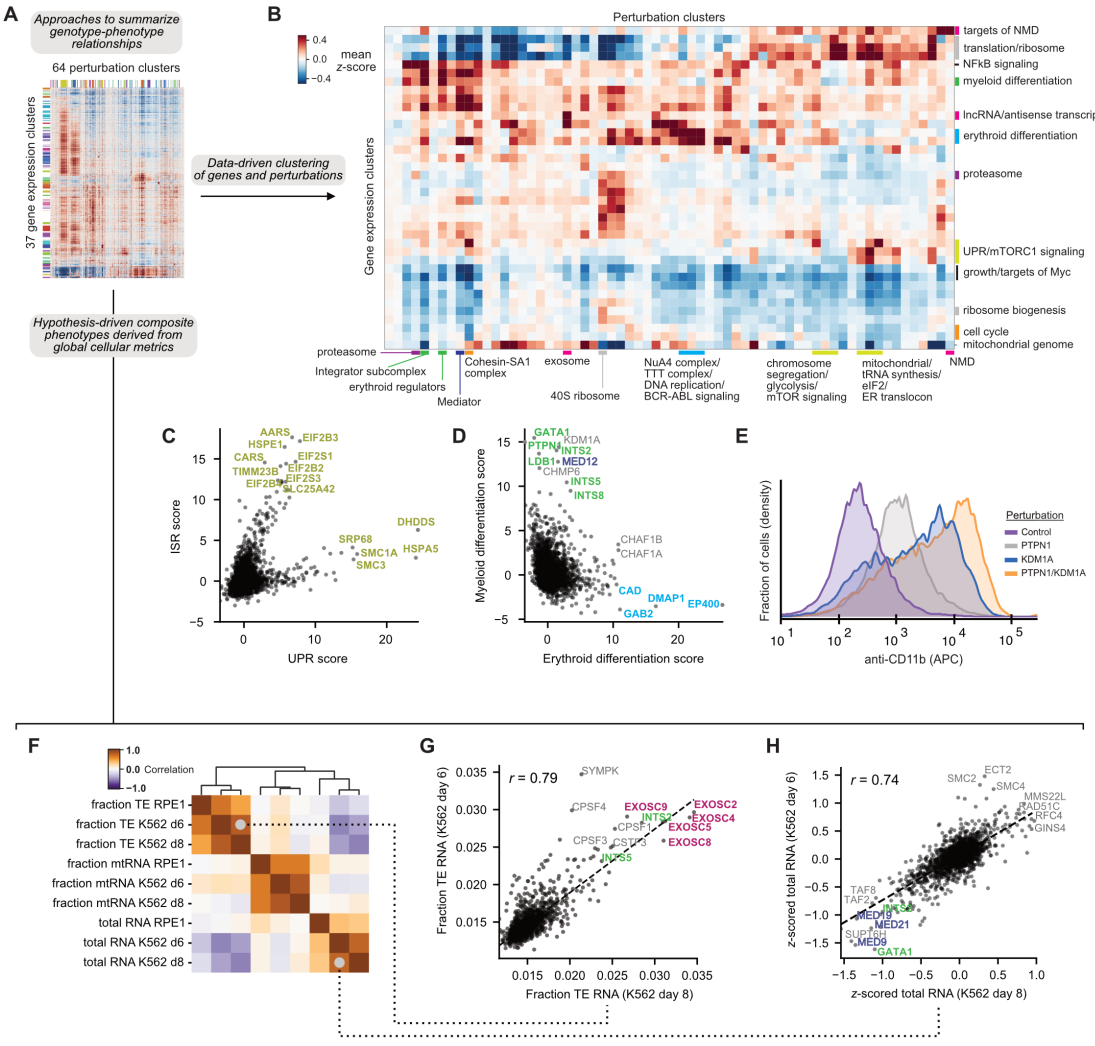

Figure 1.4: Summarizing genotype-phenotype relationships with Perturb-seq.

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. (Figure caption continued on the next page.)

E) (Figure caption continued from the previous page.) 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.

Surprisingly, loss of a number of common essential genes (i.e., essential across cell lines in the Cancer Dependency Map) also caused expression of either myeloid (e.g., Integrator subcomplex) or erythroid (e.g., NuRD complex, DNA replication machinery) markers. Next, we investigated the differentiation effect of selectively essential genes, which could be promising targets for differentiation therapy, analogous to ongoing efforts for KDM1A (Maes et al., 2018; Yu et al., 2021). We observed that loss of *PTPN1*, a tyrosine phosphatase selectively essential in K562 cells, drove myeloid differentiation. While inhibitors of PTPN1 have been developed for use in diabetes and certain cancers (Sharma et al., 2020), to our knowledge they have not been tested as a differentiation therapy. Remarkably, in targeted experiments, we found that knockdown of *PTPN1* and *KDM1A* in combination caused a substantial increase in differentiation and growth defect compared to either single genetic perturbation, suggesting that these targets act via different cellular mechanisms (Fig. 4E). These results highlight the utility of rich phenotypes for understanding differentiation as well as nominating promising therapeutic targets or combinations.

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). 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, 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.

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. 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 (Spearman's rho = 0.71), and reproducible across the day 6 and day 8 K562 experiments (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). 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.

Intriguingly, many genes implicated in chromosome segregation were among the top drivers of heterogeneity, including *TTK*, *SPC25*, and *DSN1* (Fig. 5B) (Musacchio and Salmon, 2007). 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 (Patel et al., 2014) to estimate single-cell DNA copy number along the genome by quantifying the change in moving average gene expression compared to control cells. Consistent with our hypothesis, knockdown of *TTK*, a core component of the spindle assembly checkpoint (Jelluma et al., 2008), led to dramatic changes in estimated DNA copy number in both intrinsically aneuploid K562 and near euploid RPE1 cells (Fig. 5C). 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).

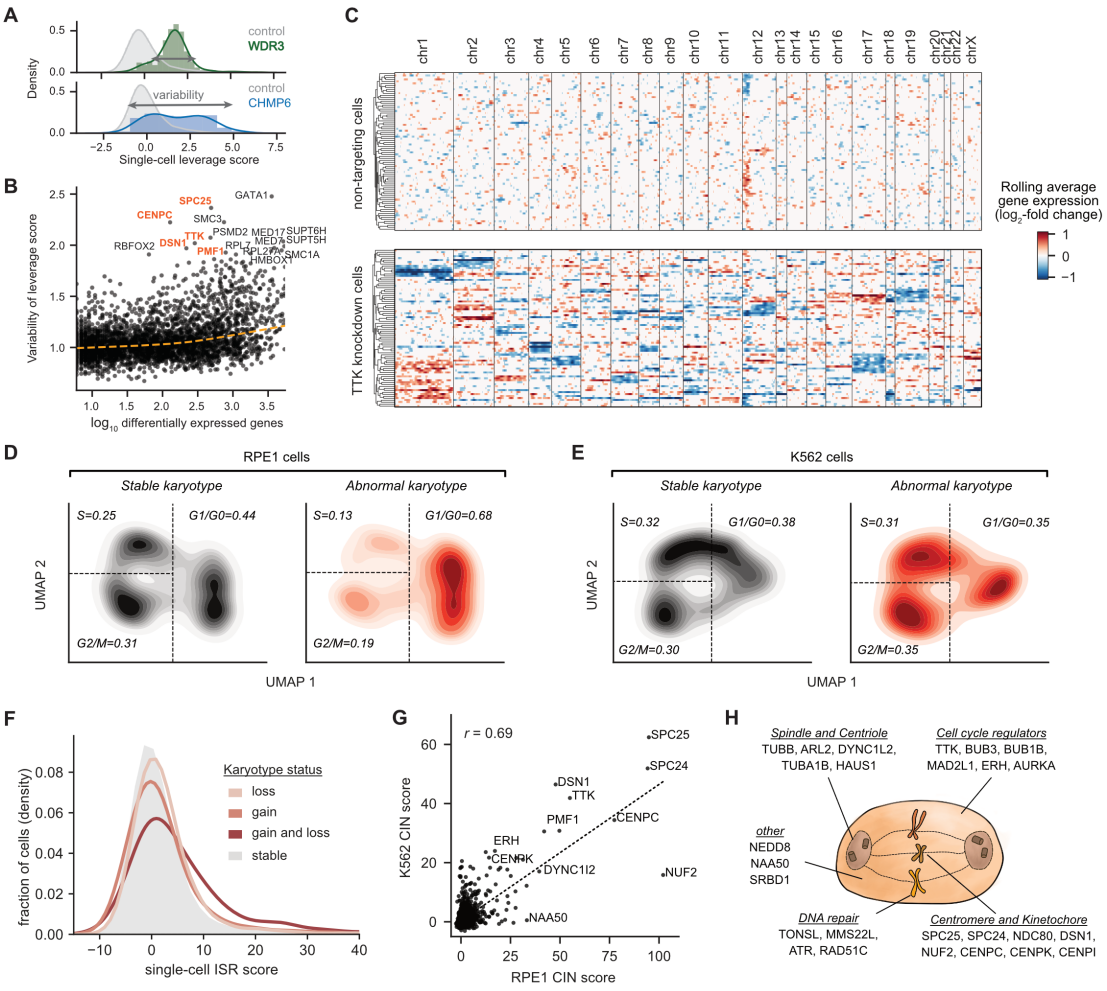

Figure 1.5: Exploring acute consequences and genetic drivers of aneuploidy in single-cells.

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. (Figure caption continued on the next page.)

D) (Figure caption continued from the previous page.) 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.

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

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

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

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. 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). These results are consistent with models in which cell cycle checkpoints are activated by the secondary consequences of aneuploidy (e.g., DNA damage or proteostatic stress) rather than changes in chromosome number *per se* (Santaguida and Amon, 2015; Santaguida et al., 2017).

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). 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.

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

We began by comparing the nuclear transcriptional responses to CRISPRi-based depletion of nuclear-encoded mitochondrial genes (i.e., mitochondrial perturbations). We found that mitochondrial perturbations elicited relatively homogeneous nuclear transcriptional responses, illustrated by well-correlated transcriptional phenotypes across mitochondrial perturbations (Fig. 6A). 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.

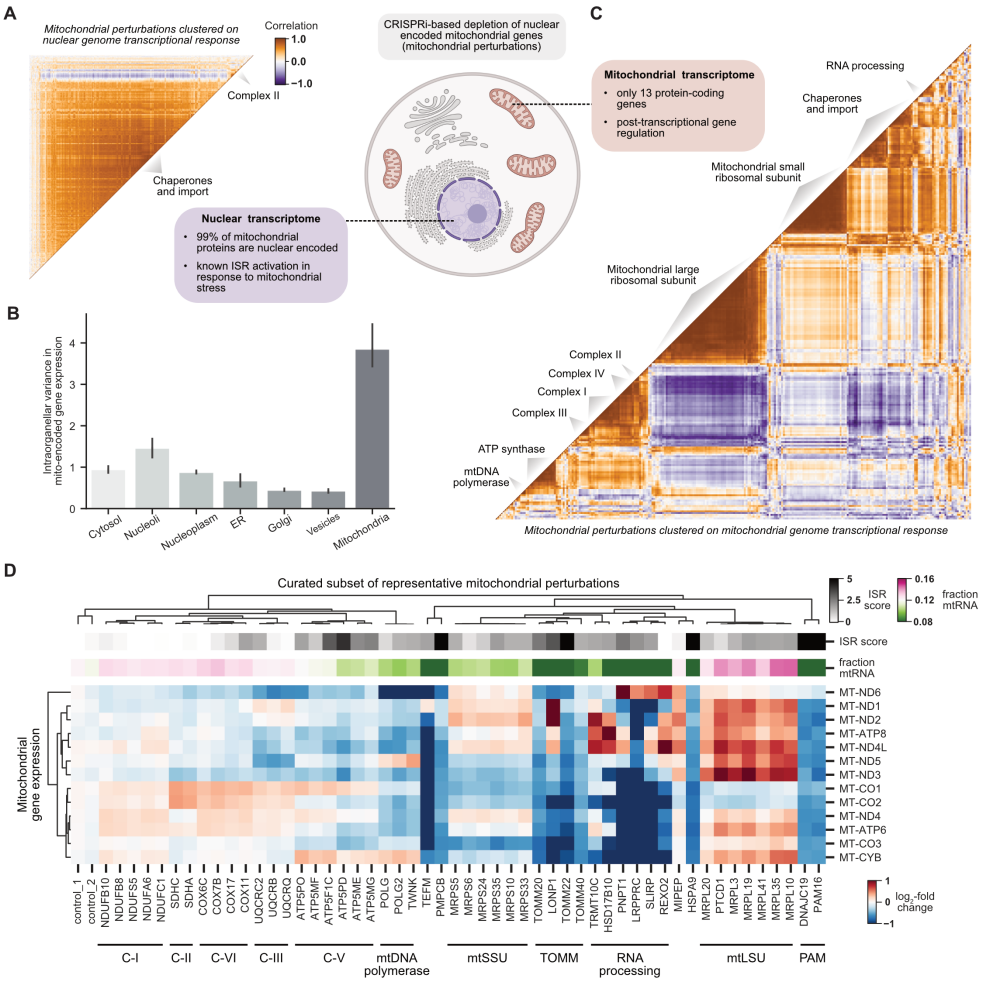

Figure 1.6: Global organization of the transcriptional response to mitochondrial stress.

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. (Figure caption continued on the next page.)

C) (Figure caption continued from the previous page.) 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.

Although this result was broadly consistent with recent literature that has highlighted the role of the ISR as response to mitochondrial stress (Fessler et al., 2020; Guo et al., 2020; Mick et al., 2020; Münch and Harper, 2016; Quirós et al., 2017), the lack of functional specificity of the transcriptional response was puzzling in light of: (i) the multifaceted roles of mitochondria in diverse processes such as respiration, intermediary metabolism, iron-sulfur cluster biogenesis, and apoptosis and (ii) the high-resolution separation of cytosolic perturbations by transcriptional response in 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). 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). 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). 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* (Mercer et al., 2011). 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 (Richter-Dennerlein et al., 2016; Salvatori et al., 2020)], 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) (Kummer and Ban, 2021). 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

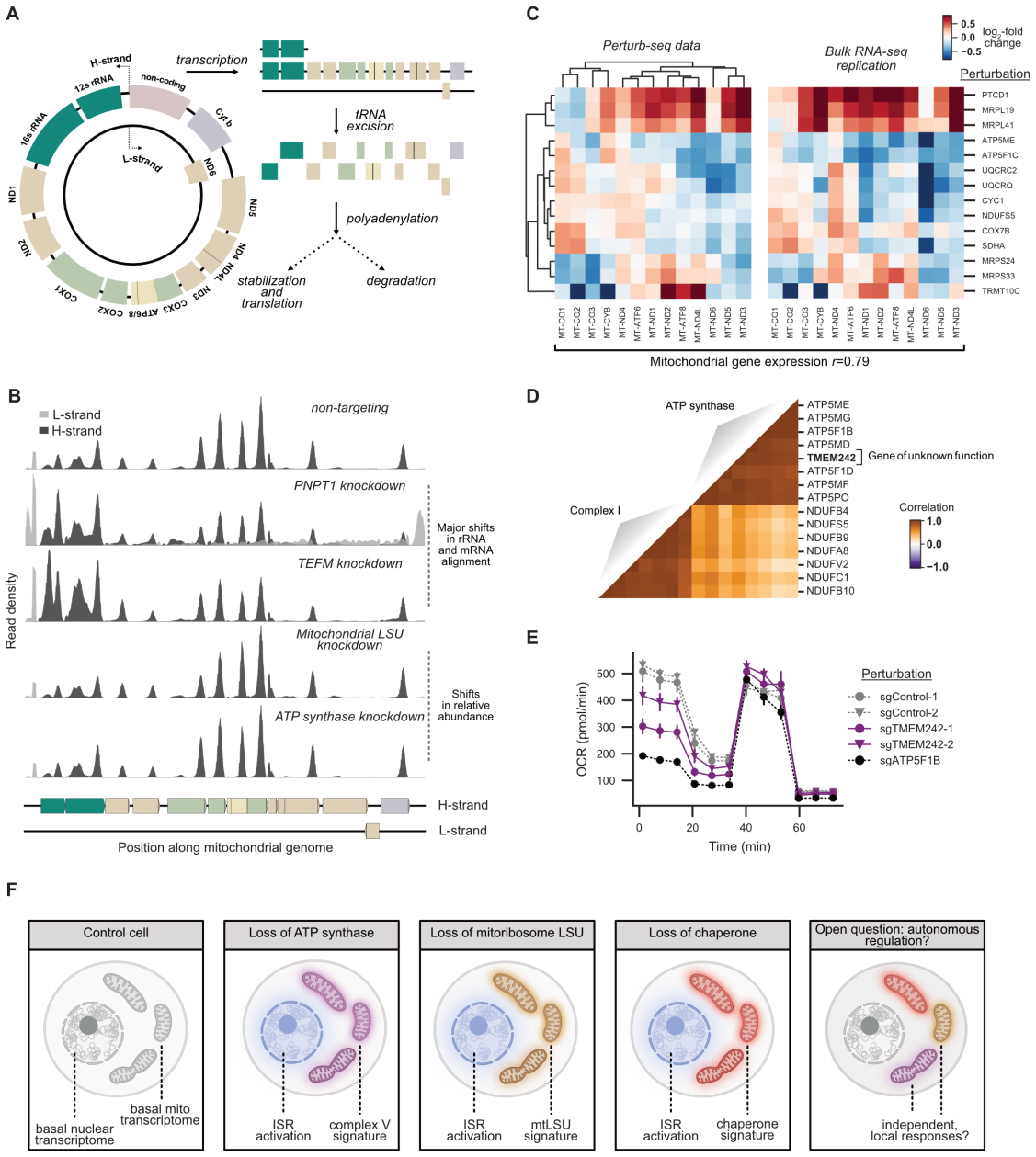

Figure 1.7: Investigating regulation of the mitochondrial genome in stress.

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. (Figure caption continued on the next page.)

B) (Figure caption continued from the previous page.) 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.

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). 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 (Carroll et al., 2021). Together, these experiments discover a novel factor required for ATP synthase activity and point to the precision of mitochondrial genome regulation.

Discussion

Single-cell CRISPR screens represent an emerging tool to generate rich genotype-phenotype maps. However, to date, their use has been limited to the study of preselected genes focused on discrete, predefined biological questions. Here, we perform genome-scale single-cell CRISPR screens using Perturb-seq and demonstrate how these screens enable data-driven dissection of a 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-driven manner (e.g., measuring intron/exon ratios to study splicing), resulting in an enormous breadth of measured phenotypes. In the case of scRNA-seq, we show that it measures not only features such as differential gene expression and the activity of critical transcriptional programs, but also RNA splicing and processing, expression of transposable elements, differentiation, transcriptional heterogeneity, cell cycle progression, and chromosomal instability. Once a phenotype is defined, the genotype-phenotype map can be used to explore its genetic underpinnings, in a manner analogous to a forward genetic screen, as well as its relationship to other cellular phenotypes.

An illustrative example of this process is our study of chromosomal instability. Based on an initial observation of heterogeneous responses to specific perturbations, we suspected that some

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

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

is adaptive may have important implications for diseases associated with mitochondrial stress. An intriguing additional question is whether individual mitochondria are able to regulate their expression autonomously. Combined with the nuanced responses observed here, this would support and substantially extend the "co-location for redox regulation" (CoRR) hypothesis which holds that the endosymbiotically derived mitochondrial genome has been retained through evolution to enable localized regulation of mitochondrial gene expression (Allen, 2017).

A final theme emerging from our work is the 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 (Cleary et al., 2017), together with advances in single-cell phenotyping (Datlinger et al., 2021; Martin et al., 2021) and DNA sequencing promise to greatly increase the scale of these experiments. To this point, we concluded our work by sequencing our genome-scale K562 libraries on a lower-cost, ultra-high throughput sequencing platform developed by Ultima Genomics, generating results equivalent to those sequenced on Illumina instruments.

In sum, our study presents a blueprint for the construction and analysis of rich genotype-phenotype maps to serve as a driving force for the systematic exploration of genetic and cellular function. Our data is available in interactive and raw formats at https://weissman.wi.mit.edu/perturbseq/.

Declaration of 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. GLY, NI, FO, and DL are employees and shareholders of Ultima Genomics. 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 at https://weissman.wi.mit.edu/perturbseq/. 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.

Chapter 2: Large-scale in vivo pooled genetic screens with highly multiplexed phenotypes in the mouse liver

Introduction

A central goal of organismal biology is to understand how the physiological functions of organs and tissues arise from the coordinated activity of thousands of genes in diverse principal and support cell types. Recent large-scale atlasing efforts have characterized the molecular composition of thousands of distinct cell types across the organs; ongoing work is mapping the spatial organization of these cell types across tissues. Although such efforts provide a foundation resource, a major challenge for the field is to causally understand how cell types, cell states, and multicellular structures and motifs are produced and regulated by the actions of specific genes and molecular pathways. However, we currently lack approaches that would enable a systematic and comprehensive understanding of how specific genes control diverse cellular and tissue phenotypes in living animals, under different physiological states.

The physiological functions of organs and tissues are the product of the coordinated action of thousands of genes in many distinct principal and support cell types. These functions are dynamic, as cellular states can change over time in response to fluctuating physiological conditions, damage, developmental cues, and disease. Recent large-scale atlasing efforts have characterized the molecular composition of thousands of distinct cell types across different organs; ongoing work is mapping the spatial organization of these cell types across tissues. Although such efforts provide a foundational resource for mapping the diverse cellular composition of different tissues, a major challenge for the field is to causally understand how cell types, cell states, and multicellular structures and motifs are produced and regulated by the actions of specific genes and molecular pathways. However, we currently lack approaches that would enable a systematic and

comprehensive understanding of how specific genes control diverse cellular and tissue phenotypes in living animals, under different physiological states.

Cellular and tissue state can be manifest across multiple dimensions, ranging from gene expression to protein phosphorylation to cellular morphology to the organization of cells in niches and neighborhoods. Historically, microscopy has been a leading approach for understanding the function of genes in cells and organisms. Forward genetic screens in multiple systems, from yeast to flies to zebrafish, rely heavily on visual identification of phenotypically abnormal organisms. High resolution microscopy of cellular processes have been crucial to innumerable discoveries, ranging from cell division to the mechanisms of autophagy. Recent advances in computational capacity and machine learning methods have dramatically advanced the capabilities to quantify and explore imaging-based phenotypes. In parallel, new developments in molecular profiling provide complementary pictures of cellular state (such as transcriptional state, translation, and chromatin), in a manner amenable to high-throughput single cell characterization in tissues.

Coupling these tools for deeply phenotyping cells to advances in targeted genetic perturbations is markedly expanding our ability to map genotype-phenotype relationships at a large scale. In particular, Perturb-seq – the combination of CRISPR screening and single-cell RNA-seq – enables the dissectiion of molecular regulators of core cellular processes such as the unfolded protein response and RNA splicing, as well as the unsupervised classification of gene function and the principled discovery of cellular responses when applied at the genome-wide scale. Several studies have used Perturb-seq to profile the effects of a subset (typicaly dozens to a few hundred) of disease-relevant genes, chosen through genetic association studies, to perturb in the brain, in the skin, and in the immune system. In parallel, advances in multiplexed molecular measurements through microscopy have enabled new forms of pooled optical screening, using both MERFISH

and *in situ* sequencing to read out genotypes (i.e. the perturbed gene). *In situ* sequencing-based approaches have been applied at a genome-wide scale in cultured cells. Although these imaging approaches have recently been expanded to include multiple modalities including immunofluorescence and spatial gene expression of limited predefined set of genes (cite preprints) for cultured or transplanted cells, there remains a critical tradeoff between *in situ* optical and morphological measurement and transcriptome-wide transcriptional phenotyping especially when studying perturbation in an in vivo physiological setting.

The ideal approach to comprehensively map the relationship between genetic perturbations and their phenotypic consequences would bridge multiple modalities to obtain a holistic picture of cellular and tissue state under different genotypes in living tissue. Such an approach would allow one relate changes in genome-wide cellular transcriptional state under different perturbations with the effects of those perturbations on specific morphological and multicellular features. Transcriptionally-defined cellular states could be linked to interpretable phenotypes measured through microscopy. By performing these perturbations in different cell types in their native tissue context, the function of different genes can be interrogated in their native physiological context, under both homeostatic and pathological conditions.

Here, we report the development of an approach for performing large-scale pooled genetic screens in native, living tissue with rich phenotypes, as well as its application to map the function of hundreds of genes in the liver under multiple physiological states. Through this approach, we are able to study the effects of many genetic perturbations on diverse cellular phenotypes in the liver, at the levels of transcriptional state, subcellular morphology, and tissue organization. To do so, we develop new methods for fixed-cell Perturb-seq as well as imaging-based pooled perturbation screening in heavily fixed tissue, enabling joint analysis of the same perturbations in

the same tissue. For the latter, we apply multiplexed subcellular protein and RNA imaging in combination with imaging-based spatial transcriptomics, using *in situ* enzymatic probe amplification to read out both endogenous RNA and short barcodes for genotyping. Through an integrated analysis of imaging and transcriptional phenotypes, we identify novel regulators of hepatocyte zonation, reveal how proteostatic stress pathway activation can broadly affect the expression of secreted proteins, and show how diverse cellular states can produce convergent effects on lipid accumulation. In total, we provide a new path to characterizing multidimensional genotype-phenotype relationships in living tissues, in a scalable manner that will be broadly applicable across diverse tissues as well as physiological and disease states.

Results

Our goal was to develop an approach that would allow us to profile the effects of hundreds to thousands of genetic perturbations at subcellular resolution in different cell types in the tissue of a living mouse under defined physiological conditions. For each perturbation, we aimed to capture multiple distinct dimensions of cellular state to resolve a more comprehensive picture of gene function. To this end, building on top of recent work establishing ability to deliver genome-wide *in vivo* CRISPR libraries to the mouse liver, we established an approach where hepatocytes were infected in a mosaic with lentivirus delivering CRISPR guides (Fig. 1A). After perfusion with paraformaldehyde to rapidly lock cellular phenotypes into place, perturbed cells were then interrogated using either sequencing- or imaging-based approaches to capture multiple aspects of cellular state.

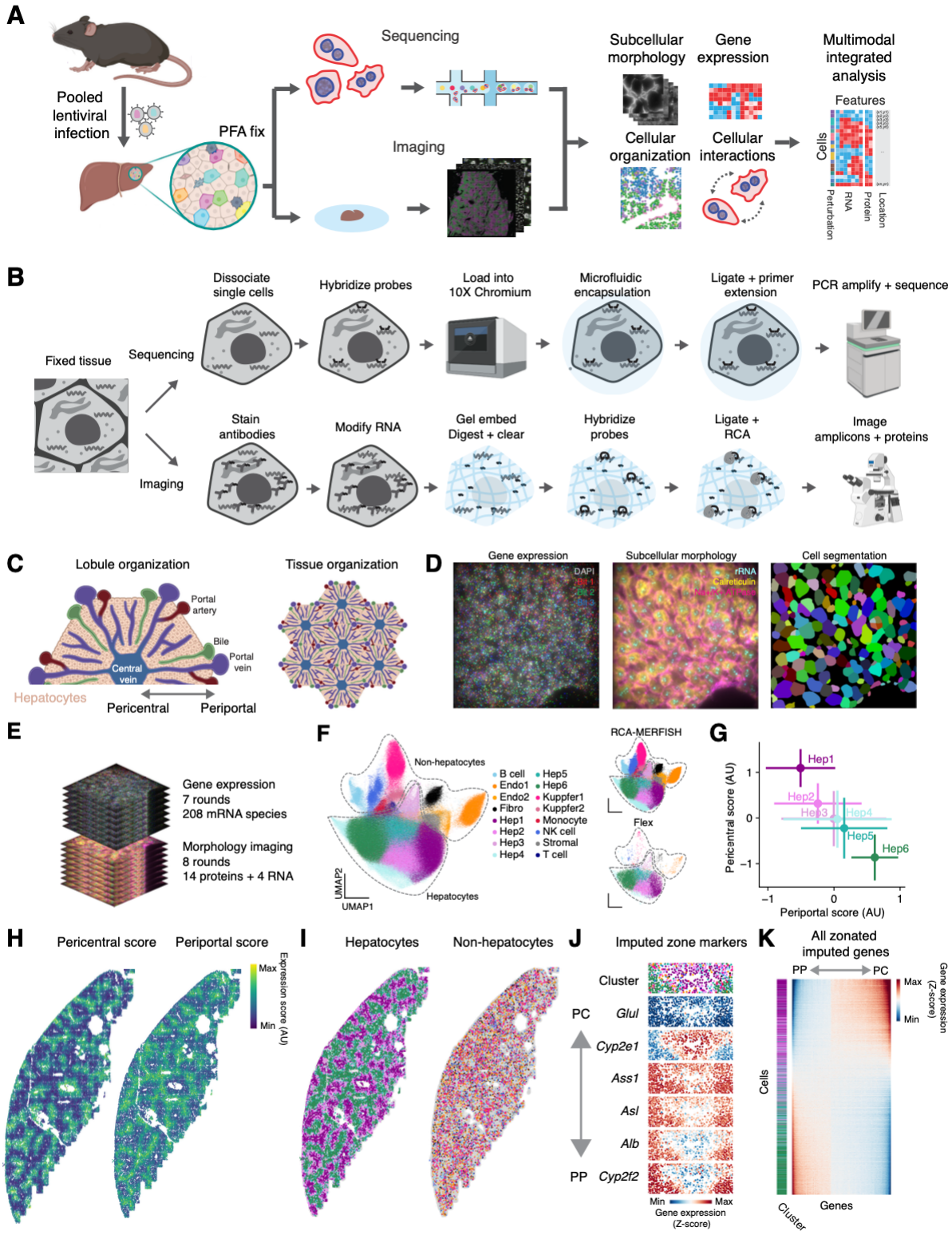

Figure 2.1: Pooled in vivo genetic screens through imaging and sequencing

A) Multimodal mapping of genetic and physiological perturbations *in vivo*. (Figure legend continued on the next page.)

B) (Figure legend continued from the previous page.) Multiple genotype and phenotype readout methods using hybridization-based sequencing and imaging.
C) Diagram of liver organization the lobule and tissue scale, showing major pericentral to periportal axis. Image courtesy of BioRender.
D) Multimodal measurement of gene expression (left) and subcellular morphology (center), with machine learning-based segmentation of cells (right).
E) Multimodal readout of RNA and protein through sequential hybridization and imaging.
F) Left: Integrated UMAP plot of liver cell types between RCA-MERFISH and Flex single-cell RNA-seq, identified through unsupervised clustering. Right detail: same UMAP showing cells from either RCA-MERFISH or Flex.
G) Periportal vs pericentral gene expression score for hepatocyte subtypes.
H) Spatial distribution of periportal and pericentral scores in hepatocytes.
I) Spatial organization of hepatocyte subtypes (left) and non-hepatocyte cell types (right).
J) Spatial organization of imputed hepatocyte zone markers radially organized around a central vein.
K) Genome-wide imputed gene expression of individual hepatocytes (colored by subtype), sorted by pericentral score for that cell.

From single cell sequencing, we could resolve the effects of perturbations on genome-wide gene expression profiles to obtain rich insights into regulators of cell state and cell type. From imaging, we would be able to measure different spatial phenotypes, including the subcellular morphology, localization, or quantity of proteins, the cellular organization of the tissue, and the interactions between different cells. By linking cells measured through different modalities via their perturbation identity, we could then perform integrated multimodal analyses to examine the multifarious effects of the same perturbation on different aspects of cellular and tissue function.

Measuring phenotypes through both sequencing and imaging required the development of improved readout modalities that were compatible with heavy tissue fixation and enabled multimodal spatial readout of both protein and RNA. For sequencing-based phenotyping, we built on top of the 10X Genomics Flex technology, which uses microfluidic encapsulation of dissociated single cells that have been hybridized with a transcriptome-wide library of split probes against different RNA species (Fig. 1B, top). Probes that hybridize adjacent to each other are then ligated in the droplet and given a cellular barcode for sequencing and expression quantification. We extended this approach to also target sgRNA, using probes against crRNA targeting sequence as a perturbation barcode.

Creating an imaging approach for genotyping and phenotyping intact tissue sections required more extensive technical development. In developing this approach, we had several goals: 1) To create a multimodal protocol that allows for the readout of both RNA and protein from tissue fixed in a manner that is standard for high resolution immunofluorescence, in a single sample preparation on an automated microscope. 2) To amplify RNA FISH signals, both in order to increase the throughput of data collection and to enable the readout of short barcode sequences.

3) To produce stable samples that could be stored and imaged later, such that many samples could be prepared in parallel and then imaged sequentially.

Our approach – RCA-MERFISH – builds on top of many advances over the past several years in methods for spatial transcriptomics and proteomics. We began by building on the highly scalable and specific approach of MERFISH of encoding gene identity as error-correcting binary codes that are read out through multiple rounds of hybridization of fluorescent oligonucleotide 'bits'. Inspired by recent advances in highly specific rolling circle amplification (RCA)-based readout of gene expression, such as BOLORAMIS, HybISS, STARmap, we targeted both endogenous mRNA and barcode RNA with padlock probes that are ligated *in situ*, with probes that directly contained sequences for readout of a MERFISH code over multiple rounds of hybridization. To overcome limitations on performing *in situ* enzymatic reactions in heavily crosslinked tissue, we adopted approaches from the hydrogel-based RNA-embedding and tissue clearing in MERFISH and embedding of RCA-amplicons in STARmap and ExSeq. Finally, we combined this approach with multiplexed immunostaining of oligo-conjugated antibodies.

The final protocol consists of a series of steps to remove the proteins and lipids from the tissue, while preserving the location of proteins and RNA in an acrylamide gel matrix in a manner that can be read out through multiple rounds of hybridization of short fluorescent oligonucleotides following by stripping (Fig. 1B). Briefly, after perfusion and cryopreservation, we cut thin (10 μm) sections of tissue onto silanized coverslips. We then decrosslinked the tissue and stained it with a pool of oligo-conjugated antibodies targeting different subcellular organelles, membrane, and phosphorylated states of signaling proteins. After antibody staining, we modified all RNA and DNA in the tissue with an alkylating agent containing an acrydite motif (MelphaX), and embedded the sample in a thin acrylamide gel. We then digested the proteins and washed away lipids to both

clear the tissue and make it accessible to oligo probes and enzymes. After clearing, we hybridized a library of padlock probes targeting both endogenous RNA and barcodes. After stringent washing, we then ligated the remaining padlock probes bound to RNA and performed rolling circle amplification. The final protocol was the result of extensive optimization of decrosslinking conditions, oligo modifications, additives, and digestion conditions to identify optimal conditions that allow for amplification while being compatible with immunostaining of oligo-conjugated antibodies in heavily fixed tissue.

Each individual RCA-MERFISH probe consists of two binding arms that must hybridize adjacent to each other in order to be ligated, as well as the reverse complement sequences of 4-6 readout probes that are used to encode the identity of the targeted RNA species. Each mRNA species was targeted by eight probes tiling the transcript. These probes are relatively long (~140-170 nt) and must be full-length in order to ligate properly, so the synthesis of thousands of probes at the necessary nanomole scale is cost-prohibitive with traditional phosphoramidite synthesis. To overcome this limitation, we developed a method for synthesizing ligatable probes from low-cost femtomolar pools of long oligos synthesized in array. After testing multiple different options, we settled on an approach that uses limited cycle PCR and then RCA to amplify the initial pool, followed by Type IIS restriction enzyme digestion to liberate the full length, phosphorylated probes with variable binding regions on either end.

As a proof of principle, we applied this imaging approach in combination with single-cell transcriptomics to map the spatial organization of the murine liver. The liver is a classic system for studying cell biology, enabling discoveries over the past century ranging from the isolation of different organelles to uncovering mechanisms of mTor signaling. For our purposes, the liver provided an excellent test system, as the principal cell type – hepatocytes – are relatively

homogeneous, have a well-defined spatial organization, and are readily infected with virus for perturbation delivery. At the anatomical level, the liver has long been understood to consist of distinct cellular 'zones' radially organized with respect to the central and portal veins (Fig. 1C, left). Periportal and pericentral hepatocytes have distinct physiological functions, as well as gene expression. This periportal to pericentral distinction is repeated in multiple units around different central veins in a crystalline manner throughout the organ (Fig. 1C, right).

After optimization, our multimodal approach allowed us to measure endogenous gene expression with high specificity while imaging multiple subcellular structures simultaneously. We resolved hundreds of amplicons per cell while simultaneously imaging multiple proteins and RNA species. In total, we measured 208 genes across 7 rounds of 3 color imaging, with a total of 21 bits, using a HW4, MHD4 codebook, altogether resolving hundreds of amplicons per cell (Fig. 1D, left). This gene panel was primarily focused on hepatocyte state but also included probes to label other cell types. After these RNA were detected through combinatorial MERFISH readout, we sequentially read out 14 proteins with oligo-conjugated antibodies (Fig. 1E), in addition to 4 abundant RNA species that exhibited specific patterns of subcellular localization with RNA FISH (Fig. 1D, middle). One antibody was against the membrane marker Na/K+ ATPase, which we used for deep-learning based segmentation (Fig. 1D, right).

Unsupervised clustering revealed a number of hepatocyte subtypes and non-hepatocyte cell types. We classified different cell types as well as subtypes of hepatocytes through a joint clustering between RCA-MERFISH and single-cell RNA-seq data (Fig. 1F). We focused our analysis on hepatocytes, which were enriched by our FACS gating strategy. We computed a periportal and pericentral gene expression score based on known marker genes and found that the hepatocyte clusters fell onto a zonated continuum from Hep1 to Hep6, with Hep1 as the most

pericentral and Hep6 as the most periportal (Fig. 1G). Visualizing these zonation scores and hepatocyte subtypes across the tissue revealed the expected patchwork localization of periportal and pericentral cells (Fig. 1H, 1I), whereas non-hepatocytes had less obvious spatial arrangement (Fig. 1I). We imputed the expression of the full transcriptome for each cell measured with RCA-MERFISH from the scRNA-seq data and were able to resolve the expected pattern of zonated gene expression for imputed genes (Fig. 1J). Finally, as expected, we found that a large fraction of the transcriptome in hepatocytes varied along the periportal to pericentral axis (Fig. 1K).

To capture diverse aspects of cellular state through high-resolution microscopy, we devised a panel of 18 different protein and RNA targets that covered a variety of subcellular organelles, morphological features such as membrane, and phosphorylated states of proteins. The proteins were measured through oligo-conjugated antibodies stained in a pool, and the RNA species were labeled using RNA FISH probes. Targeted proteins included both organellar and structural markers such as anti-GAPDH (cytoplasm), anti-calreticulin (endoplasmic reticulum), anti-perilipin (lipid droplets), anti-Na+/K+ ATPase (membrane), anti-CathB (lysosome), and anti-Tomm20 (mitochondria), as well as cell state-dependent markers such as the mTOR marker anti-phospo-S6 ribosomal protein. The targeted RNA species included the highly abundant Albumin mRNA, pre-rRNA in the nucleolus, mitochondrial RNA (mtRNA), as well as total polyadenylated RNA (Fig 2A,B).

To analyze this highly multiplexed cellular imaging data, we built on top of recent advances in self-supervised learning methods for dimensionality reduction. We reduced each image channel in each cell from a high-resolution z-stack to a 128x128 pixel matrix and trained a VQ-VAE network to reduce the dimensionality of that protein or RNA further, to a 512-dimensional vector (Fig. 2C). The dimensionality reduction was constrained by an auxiliary task that sought to use the

low-dimensional embedding to discriminate different transcriptionally-defined cell types or conditions, in order to find an embedding that reflected true biological variation in the sample. We visualized these embeddings with UMAP and found that embeddings representing images of morphologically similar targets were grouped together (Fig. 2D); we also observed this through quantification of mutual information between channels (Fig. 2E).

We then examined the tissue-scale spatial organization of variation in these embeddings. In one approach, we took each 512 dimensional embedding for a protein across all cells and visualized the top few principal components. While some protein features exhibited zonated spatial patterns similar to those observed in hepatocyte gene expression, others captured morphological variation that did not appear related to spatial zonation (Fig. 2F). In general, we observed significant spatial heterogeneity in the morphological feature embeddings that appeared unrelated to the dominant periportal to pericentral axis of transcriptomic variation.

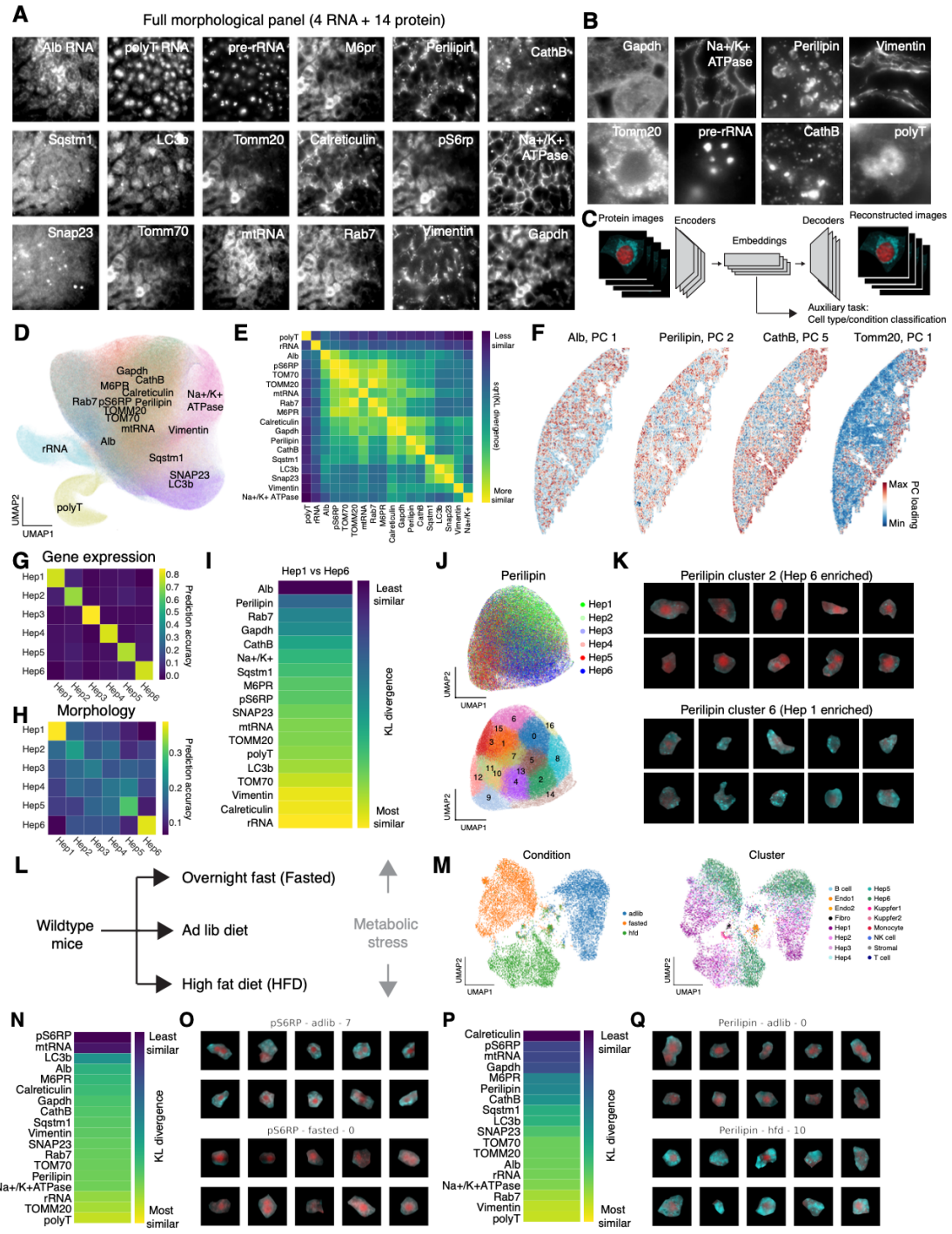

Figure 2.2: Heterogeneity in subcellular morphology by cell type and state

A) Example images of the full subcellular morphology panel, comprising 4 RNA species and 14 proteins.
B) Details of subset of subcellular morphologies.
C) Diagram of deep learning autoencoder to reduce dimensionality and featurize subcellular morphology images. (Figure legend continued on the next page.)

D) (Figure legend continued from the previous page.) UMAP plot of individual subcellular morphology image embeddings, colored by channel identity. The text indicating the morphology channel identity is located at or near (to avoid overlapping text) the centroid of images of the channel in two-dimensional UMAP space.
E) Similarity of embeddings of subcellular morphological channels, quantified by Kullback-Leibler (KL) divergence.
F) Example tissue-scale spatial organizations of a single principal component used to quantify variation in subcellular morphology embeddings. Each cell in the liver section is colored by its value in the listed principal component.
G) Confusion matrix representing from the imaging transcriptomic data.
H) Confusion matrix representing from the morphological data.
I) Heat map showing mutual information between each morphological channel at hepatocyte subtype identity.
J) UMAP representation of anti-Perilipin morphological embeddings, colored by hepatocyte subtype identity (top) or embedding Leiden cluster (bottom).
K) Unbiased sampling of hepatocytes from Perilipin embedding clusters 2 and 6.
L) Diagram of the diet experiment.
M) UMAP representation of gene expression in the diet experiment, colored by condition (left) or cell type (right).
N) Heat map showing the capacity of morphological channel embeddings to discriminate between hepatocytes from *ad lib* and fasted mice.
O) Unbiased sampling of hepatocytes from anti-p-S6 RP embedding cluster 7 (from the *ad lib* condition) and cluster 0 (from the fasted condition).
P) Heat map showing the capacity of morphological channel embeddings to discriminate between hepatocytes from *ad lib* and HFD mice.
Q) Unbiased sampling of hepatocytes from anti-perilipin embedding cluster 0 (from the *ad lib* condition) and cluster 10 (from the HFD condition).

To compare the relative information contained in protein embeddings versus transcriptome in more detail, we attempted to quantify the relative information that was contained in the single-cell transcriptional profiles versus the morphological feature embeddings in the same cells. To do so, we trained a classifier to predict the subtype identity of individual hepatocytes – from Hep1 to Hep6 – on held out cells, using either the 208 genes measured through RCA-MERFISH or the concatenated 18 x 512 dimensional feature embeddings. We found that whereas the transcriptional profiles clearly discriminated all subtypes of hepatocytes (Fig. 2G), the feature embeddings only clearly distinguished the most periportal and pericentral subtypes Hep1 and Hep6 (Fig. 2H). By contrast, the other intermediate subtypes were less clearly distinguished, suggesting a subcellular morphological gradient between the two extremes.

To understand the differences in subcellular morphology between the most periportal and pericentral hepatocyte subtypes, we identified the morphological feature embeddings that best discriminated between Hep1 and Hep6 cells (Fig 2I). Albumin FISH and anti-perilipin were the most dissimilar, consistent with the literature understanding of zonated Albumin expression and lipid storage and metabolism. We visualized anti-perilipin embeddings with UMAP and found that embeddings representing cells with different hepatocyte subtype identity were separated (Fig. 2J); we conducted unsupervised clustering of the embeddings and found that Hep6 cells were enriched in anti-perilipin embedding cluster 2, which has very few lipid droplets, whereas Hep1 cells were enriched in anti-perilipin embedding cluster 6, which has many lipid droplets (Fig. 2K).

To understand the features of liver cellular state that respond to fluctuating physiological conditions and metabolic stress, we overnight-fasted mice or fed them a high-fat diet (HFD) for a month and fixed tissue samples for analysis by single cell sequencing and imaging (Fig 2L). These physiological perturbations caused large changes in gene expression, especially in hepatocytes

(Fig 2M). We embedded the morphological images with our VQ-VAE model and looked for image channels whose embeddings best discriminated between the physiological conditions. anti-phospho-S6 ribosomal protein separated hepatocytes between the fasted and *ad lib* samples, consistent with S6 ribosomal protein phosphorylation by mTOR in response to nutritional status (Fig 2N,O). Perilipin differed in cells from the *ad lib* and HFD samples (Fig 2P,Q). We note that stains such as anti-calreticulin, anti-pS6RP, and anti-Gapdh were also quite different between the *ad lib* and HFD samples, apparently due to negative space in the staining caused by large lipid droplets.

In total, these experiments demonstrated that our multiplexed subcellular morphology readout approach through multiplexed immunofluorescence and FISH, combined with deep learning analysis, was able to resolve significant heterogeneity in subcellular morphology across a variety of imaging channels. Furthermore, we could relate this heterogeneity to distinctions in transcriptionally-defined cell types through multimodal analysis, as well as across samples to changes in the animal's physiological state.

Having established an integrated sequencing and imaging methodology for studying the molecular and cellular organization of tissue at multiple scales, we combined this approach with large-scale pooled *in vivo* CRISPR screening to map the effect of genetic knockout perturbations on genome-wide transcriptional state as well as subcellular staining intensity and morphology. We targeted ~200 genes with two independent sgRNAs and included many negative control sgRNAs. The genes chosen for perturbation were involved in diverse aspects of hepatocyte and liver physiology, including metabolism, intercellular and intracellular signaling, transcription and translation, and protein trafficking and secretion. We also targeted several hepatocyte genes with uncertain functions.

We introduced barcoded genetic perturbations that could be read out through either imaging or sequencing into the liver through lentiviral delivery. Building on the CROP-seq lentiviral vector design to mitigate the effects of recombination during lentivirus production, we engineered a vector that expresses a fluorescent protein (mTurquoise) as well as a 180-mer barcode for imaging from a hepatocyte-specific HBG Pol II promoter while simultaneously expressing an sgRNA from a Pol III promoter (Fig. 3A). Each sgRNA and its associated 180-mer barcode were synthesized on a single DNA fragment and then cloned into the vector in a pool. An intervening region composed of one of three alternative sgRNA constant regions was introduced between the barcode and sgRNA to reduce the effects of barcode swapping through recombination of the constant intervening sequence.

We then used this vector to introduce a mosaic of genetic perturbations into the liver of Cas9 transgenic mice. We produced high-titre pools of CROP-seq lentivirus that were concentrated through ultracentrifugation, and introduced the virus systemically into CAG-LSL-Cas9 mice at P1 through injection into the temporal vein (Fig. 3B). We targeted a relatively low MOI (10-30%) such that most infected cells had a single perturbation. We then allowed the mice to grow to adulthood and used AAV containing the Cre recombinase driven by a hepatocyte-specific TBG promoter.

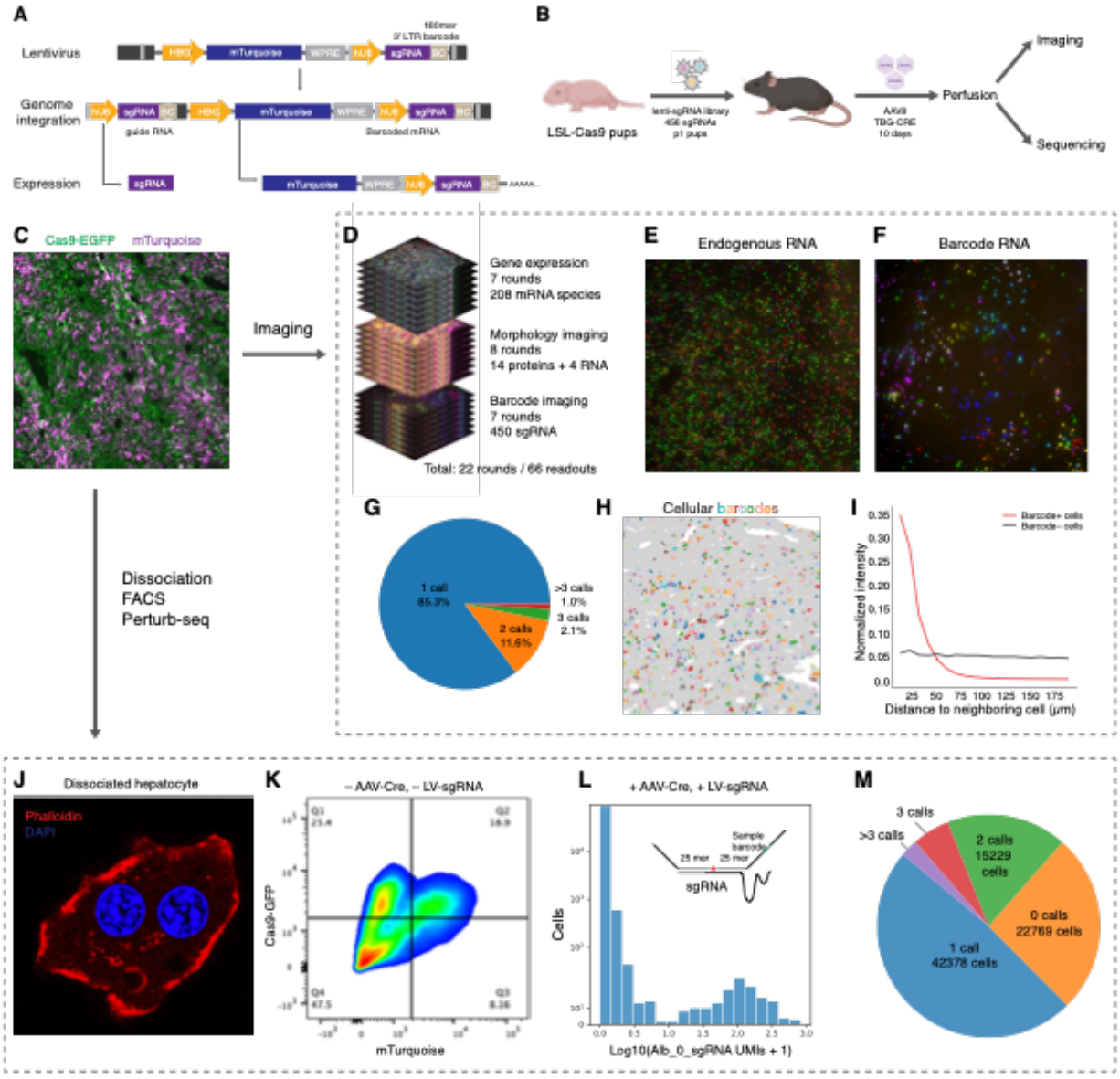

Figure 2.3: Large-scale in vivo screening in CRISPR mosaic livers

A) Schematic of lentiviral vector used for dual-mode mosaic screens.
B) Diagram of CRISPR perturbation experiment.
C) Fluorescence micrograph of PFA-perfused, lentivirus- and AAV-transduced liver tissue.
D) Multimodal readout of RNA, protein, and perturbation barcode through sequential hybridization and imaging.
E) Representative fluorescence micrograph of bits M1, M2, and M3 from an mRNA RCA-MERFISH library.
F) Representative fluorescence micrograph of bits P1, P2, and P3 from of perturbation RCA-MERFISH library.
G) Pie chart representing the number of sgRNA calls in each cell in the imaging. (Figure legend continued on the next page.)

H) (Figure legend continued from the previous page.) Representative tissue section, with perturbed cells colored by perturbation identity.
I) Local density of guides with the same sgRNA
J) Fluorescence micrograph of a representative hepatocyte recovered from PFA-perfused liver.
K) Representative flow cytometry data from dissociated PFA-perfused, lentivirus- and AAV-transduced liver tissue
L) Histogram representing counts data from custom hybridization probes targeting the Alb_0 sgRNA. The y axis is semi-log.
M) Pie chart representing the number of sgRNA calls in each cell in the Perturb-seq dataset.

After waiting 10 days to allow Cas9 to express and perturb the targeted gene, the mice were perfused and the livers were cryopreserved for analysis. Lentivirus introduced at P1 expressed mTurquoise in mosaic distribution throughout the liver of adult mice, whereas EGFP expressed with Cas9 was uniformly expressed in presumptive hepatocytes following Cre recombination (Fig. 3C). We then mapped the effects of these perturbations on hepatocytes using a combination of imaging and sequencing.

For imaging-based phenotype and genotype measurement, we extended our approach to RNA and morphological imaging through the inclusion of padlock probes targeting the perturbation-specific barcode sequences, which are orthogonal to the mouse genome. These perturbation-specific probes specify the identity of a perturbation through a highly scalable HD6, HW4 code; they bear separate bits and codes from mRNA probes and oligo-antibodies, but are imaged sequentially through seven additional rounds of imaging on the same automated microscope (Fig 3D,E,F).

We found decoded barcode molecules within a subset of cells. We established a count threshold to confidently identify cells with each perturbation and found that most perturbed cells had a single perturbation, consistent with the low infection MOI (Fig 3G). We observed relatively homogeneous infection throughout the tissue and within lobules. We found small clusters of cells sharing the same perturbation, such that for a cell expressing any given barcode there was an enrichment of other cells within 50 μm expressing the same barcode (Fig. 3H, 3I); these might reflect local proliferation of single infected hepatocytes during postnatal growth after transduction.

We developed a fixed-cell Perturb-seq method to identify the transcriptome-wide consequences of genetic perturbations. In addition to compatibility with imaging, we reasoned that a fixed cell protocol could overcome many of the challenges that have limited the scale and

practicality of *in vivo* Perturb-seq studies, such as sample degradation during tissue dissociation and FACS enrichment, especially when processing many tissues in parallel or recovering millions of cells.

Dissociation of fixed 100 μm sections from the same livers we imaged yielded intact cells with morphologies similar to those observed in tissue (Fig 3J). We used FACS to enrich for sgRNA-transduced (mTurquoise+) and Cas9-active (GFP+) cells (Fig 3K) and then performed fixed-cell single-cell RNA-seq using the hybridization-based 10X Flex approach. To determine the identity of expressed sgRNA alongside the mRNA transcriptome, we spike in a pool of custom probes that directly targeted the protospacers of the sgRNA library (right hand side probe) and the constant region of the sgRNA (left hand side probe). We observed very few UMIs for most sgRNA probes in most cells (mostly 0). In most cells, one or two sgRNA probes had much higher UMI counts (Fig 3L). We used this distribution to establish a threshold and confidently identify the sgRNAs present in each cell. Most perturbed cells had a single sgRNA above threshold (Fig 3M); some had multiple called sgRNAs, consistent with approximately Poisson-distributed transduction of the lentivirus-accessible cells in the liver.

We confirmed the accuracy of our cell calling in Perturb-seq and Perturb-RCA-MERFISH by inspecting for depletion of sgRNA targets. Fortuitously, sgRNAs targeting Albumin, an abundant secreted protein whose mRNA provides ~10% of the mRNA UMIs in hepatocytes, caused strong depletion of Alb mRNA related to control sgRNAs, presumably by nonsense-mediated decay (NMD). With Perturb-seq, we quantified *Alb* mRNA depletion directly (Fig 4A) and observed minimal overlap between the distributions of *Alb* expression in sgAlb and sgControl cells, consistent with a negligible false positive rate of sgRNA calling. We observed similar

presumed NMD for several other targets. We also conducted a fixed-cell Perturb-seq experiment in cultured CRISPRi K562 cells to confirm the accuracy and specificity of the method.

We included FISH probes against *Alb* in our Perturb-RCA-MERFISH screens. We quantified raw *Alb* FISH intensity in sgAlb and sgControl cells and observed much lower intensity in sgAlb cells, albeit with noise due to background staining and out-of-focus light from adjacent cells (Fig 4E, 4F). We also included an oligo-antibody to Gapdh and observed much lower multiplex immunofluorescence intensity in sgGapdh cells (Fig 4G, 4H).

Confident in our sgRNA assignments, we explored the broader phenotypic consequences of each genetic perturbation on the single cell transcriptome and immunofluorescent staining of hepatocytes. To simplify first-pass analysis of the imaging data, we quantified the mean intensity of each immunofluorescent stain in each cell.

We used an energy-distance permutation test with stringent multiple testing correction to determine whether perturbations caused a difference in the distribution of (1) global transcriptional states and (2) overall immunofluorescent states, relative to cells with control sgRNAs. In the Perturb-seq experiment, 109/406 targeting sgRNAs had a significant impact on the distribution of transcriptional states (measured as the first 20 PCs of mRNA expression) (Fig 4B); at that significance threshold, 0/50 control sgRNAs had a significant transcriptional impact. In the imaging experiment, 84/318 targeting sgRNAs and 3/50 control sgRNAs had a significant impact on the distribution of imaging intensities (Fig 4I).

To verify the activity and specificity of our perturbations, we compared pairs of sgRNAs targeting the same gene. We generated pseudo-bulk sgRNA-phenotype maps and found that pairs of active sgRNAs targeting the same gene tended to cause both highly correlated transcriptional phenotypes and highly correlated immunofluorescence imaging phenotypes (Fig 4C,4J).

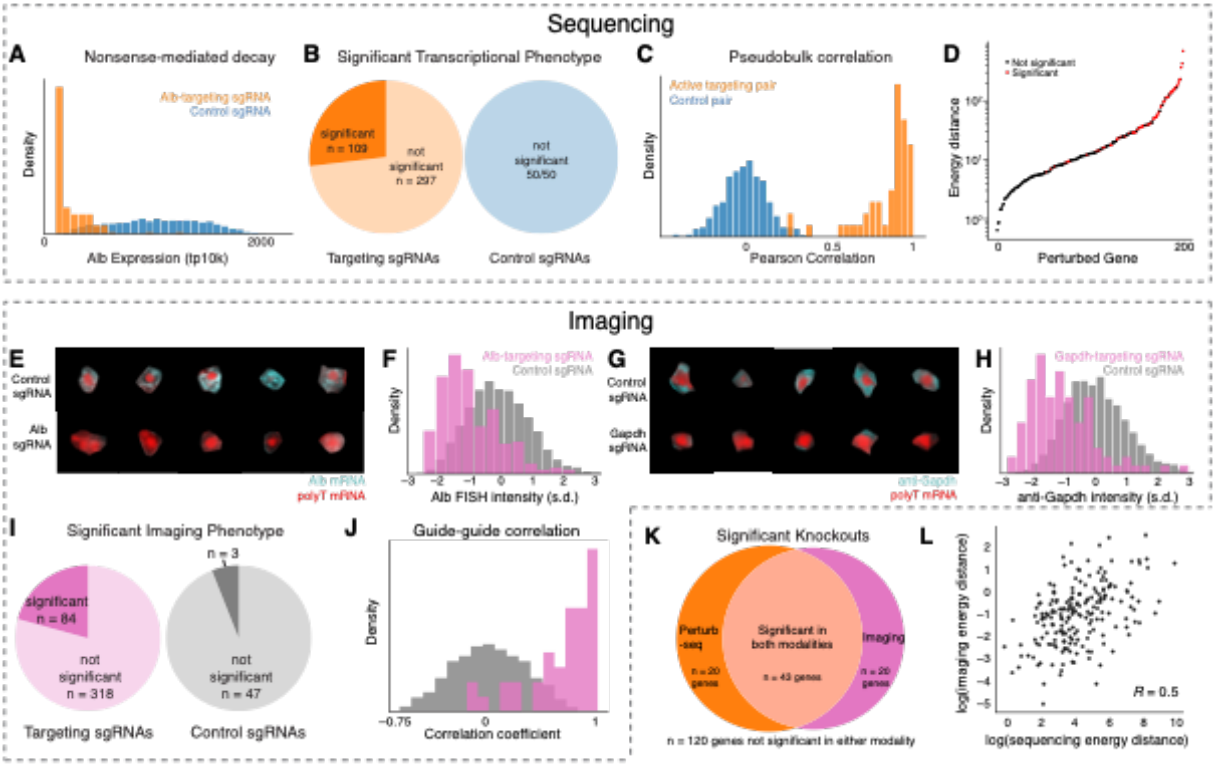

Figure 2.4: Dual-mode in vivo screening with strong phenotypes

A) Histogram comparing Albumin expression in called cells with a control sgRNA and called cells with Albumin-targeting sgRNAs, from the perturb-seq dataset.
B) Pie charts showing the number of targeting (left) and non-targeting (right) sgRNAs that caused a significant transcriptional phenotype, as measured by a Holm-Sidak-corrected energy distance permutation test ($p < 0.05$), in the perturb-seq dataset.
C) Histogram comparing the correlation of pseudo-bulk transcriptional changes associated with pairs of active targeting sgRNAs targeting the same gene, to pairs of control sgRNAs, in the perturb-seq dataset.
D) Ranking knockouts according to the energy distance vs control cells, in the perturb-seq dataset.
E) Unbiased sampling of cells with control sgRNAs and sgRNAs targeting *Alb*. The fluorescence micrographs show Alb and polyT FISH channels.
F) Histogram comparing Albumin expression in called cells with a control sgRNA and called cells with Albumin-targeting sgRNAs.
G) Unbiased sampling of cells with control sgRNAs and sgRNAs targeting *Gapdh*. The fluorescence micrographs show anti-GAPDH and polyT FISH channels.
H) Histogram comparing Albumin expression in called cells with a control sgRNA and called cells with Gapdh-targeting sgRNAs, from the imaging dataset.
I) Pie charts showing the number of targeting (left) and non-targeting (right) sgRNAs that caused a significant transcriptional phenotype, as measured by a Holm-Sidak-corrected energy distance permutation test ($p < 0.05$), in the perturb-seq dataset. (Figure legend continued on the next page).

J) (Figure legend continued from the previous page.) Histogram comparing the correlation of pseudo-bulk transcriptional changes associated with pairs of active targeting sgRNAs targeting the same gene, to pairs of control sgRNAs, in the imaging dataset.
K) Ranking knockouts according to the energy distance vs control cells, in the imaging dataset.
L) Scatterplot comparing the energy distance vs control cells for each knockout in the imaging and Perturb-seq datasets.

We inspected the perturbations that had the largest impact on the distribution of global phenotypic states, reflected in energy distance from cells with control sgRNAs (Fig 4D). In the Perturb-seq experiment, many genes with functions essential for cellular growth and survival, such as Atp2a2, Aars, Sf3b6, and Psmc4, caused large impacts to global transcriptional state upon acute knockout. The knockout of signaling genes such as Vhl, Prkar1a, and Insig1 also had large transcriptional impacts. Many of the same genes also had large impacts on the global immunofluorescence state (Fig 4K). Overall, there was a substantial correlation between the magnitude of the knockout phenotypes in the Perturb-seq and imaging assays (Fig 4L, $r = 0.5$).

We identified the differentially expressed genes associated with each sgRNA. Many targeting sgRNAs caused hundreds to thousands of genes to be differentially expressed. We also identified the immunofluorescent channels that were significantly impacted by each perturbation. Most active targeting perturbations hit a relatively small subset of markers, but some perturbations such as the essential nuclear genes Sf3b6, Sbno1, Polr1a, and Kin had a significant impact on impacted most channels, presumably reflecting the pleiotropic consequences of disrupted transcription and RNA processing.

We performed unbiased clustering of the pseudobulk genotype-phenotype map to identify perturbations that cause similar transcriptional phenotypes and to discover mRNAs and imaging channels that co-vary across perturbations (Fig 5A). The perturbation clustermap derived from perturb-seq recovered dozens of interpretable blocks of perturbed gene function, including a large block of ribosomal proteins and ribosome biogenesis genes and a block of nuclear mRNA processing genes that retained substantial substructure (Fig 5B). We also identified dozens of blocks of mRNAs that were co-expressed across perturbations; we visualized the co-expression with a two-dimensional embedding that places co-expressed mRNAs nearby (Fig 5C).

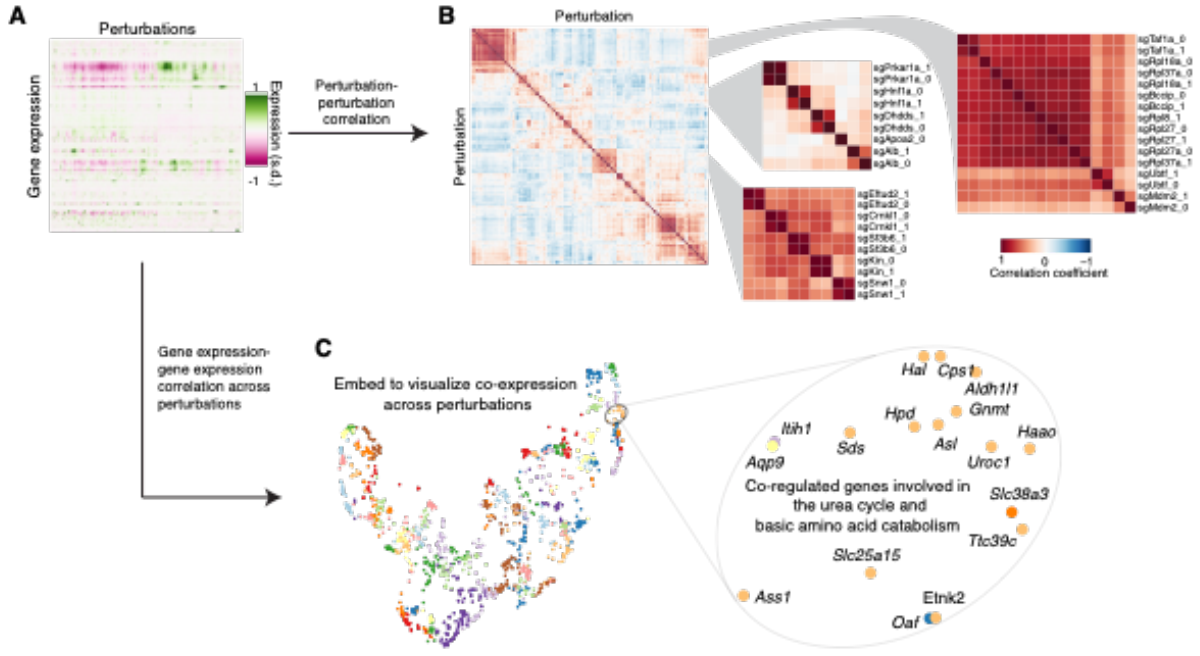

Figure 2.5: Mapping* in vivo *genotype-phenotype landscapes

A) Heat map representation of pseudobulk transcriptional changes associated with each sgRNA, relative to the mean in cells with control sgRNAs, in the Perturb-seq dataset. The color represents Z-normalized expression change and is clipped at -1 and 1 for visual emphasis.
B) Clustermap reflecting the perturbation-perturbation correlation of pseudobulk transcriptional changes associated with each active, targeting sgRNA, in the Perturb-seq dataset. The color represents pearson correlation.
C) Minimal distortion embedding where each dot represents an mRNA expressed in hepatocytes. mRNAs that are co-expressed across the genetic perturbations are placed in proximity.

We surveyed the multimodal genotype-phenotype map to identify key regulators of *in vivo* liver physiology. For first-pass exploration of the imaging data, we ranked genetic perturbations by their impact on the intensity of each immunofluorescent stain and identified perturbations that had a significant impact on stain intensity. For example, sgRNAs targeting GAPDH had the strongest negative impact on anti-GAPDH staining intensity (Fig 6A) and sgRNAs targeting Alb had the strongest negative impact on Albumin FISH. sgRNAs targeting the polymerase II machinery (Polr2l, Gpn1) or the core hepatocyte transcription factor *Hnf1a*, which directly binds the Alb promoter, also had a strong negative impact on Alb mRNA FISH, as did sgRNAs against the ERAD factor *Sel1l*.

We identified the genetic regulators of extremely essential processes, suggesting that the ten days we waited between AAV-CRE infection and tissue perfusion captured the acute impacts of perturbations. For example, targeting the RNA polymerase I machinery (Polr1a) and pol I transcription factors (Taf1a, Rrn3, Ubtf) had the strongest negative impact on pre-rRNA FISH intensity (Fig 6C). Intriguingly, targeting the nuclear speckle component Sf3b6 caused an increase in pre-rRNA FISH intensity, suggesting interplay between subnuclear structures in hepatocytes *in vivo*.

We also identified the *in vivo* regulators of cellular signaling pathways that play important roles in liver physiology. Phosphorylation of ribosomal protein S6 is a core function of the mTor signaling pathway. Knockout of mTor and the mTor complex chaperone Cdc37 caused the strongest decrease in anti-phospho S6 ribosomal protein intensity (Fig 6D); knockout of the Mtor inhibitors Tsc1 and Tsc2 and the growth inhibitor Pten caused the strongest increase in anti-phospho S6 ribosomal protein intensity.

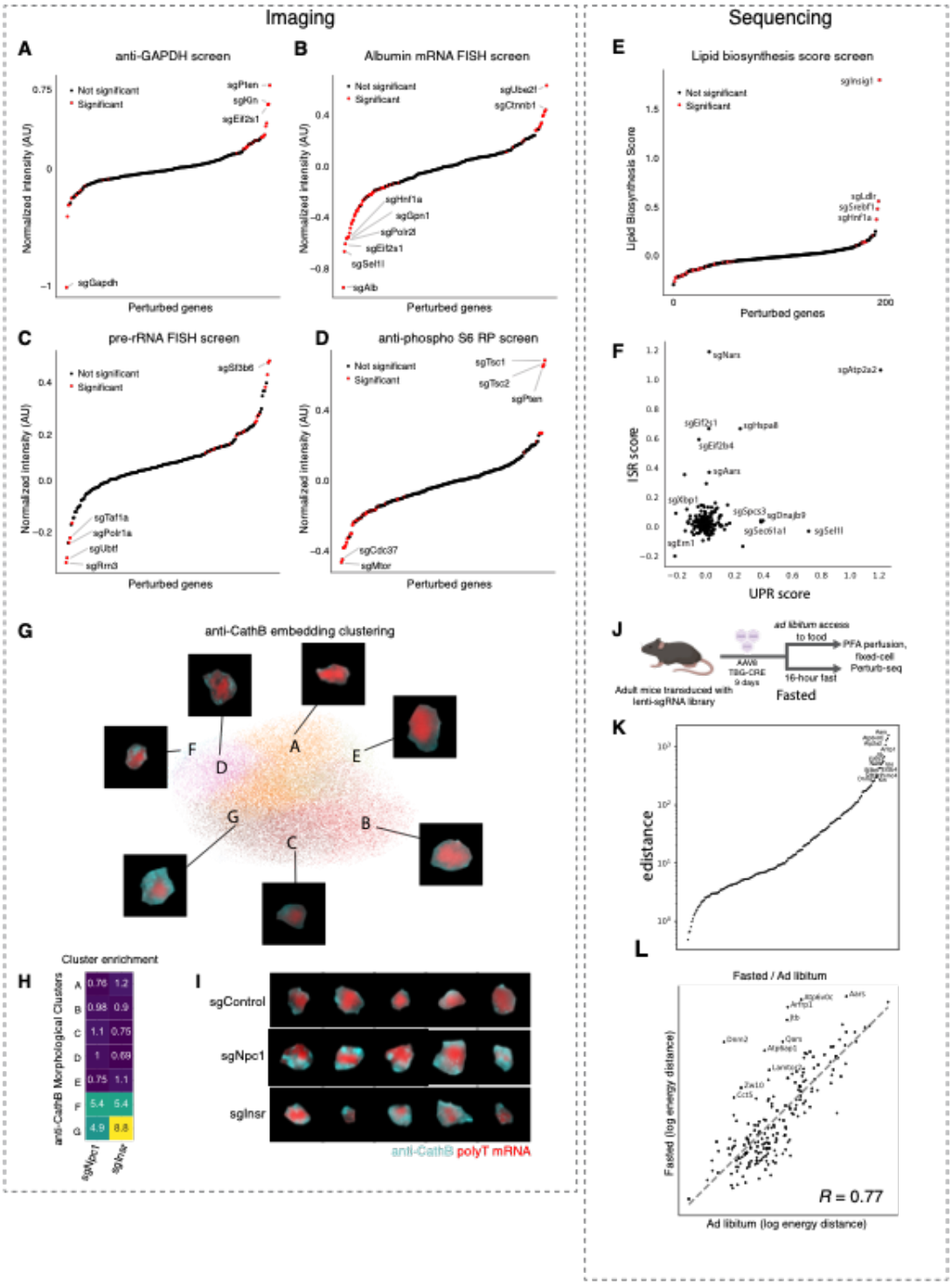

Figure 2.6: Identifying genetic regulators of liver physiology

A) Ranking of perturbed genes by their impact on anti-GAPDH intensity in the imaging experiment. Red reflects FDR-corrected significance. (Figure legend continued on the next page.)

B) (Figure legend continued from the previous page.) Ranking of perturbed genes by their impact on *Albumin* FISH intensity in the imaging experiment. Red reflects FDR-corrected significance.
C) Ranking of perturbed genes by their impact on anti-Phospho S6 ribosomal protein intensity in the imaging experiment. Red reflects FDR-corrected significance.
D) Ranking of perturbed genes by their impact on pre-rRNA FISH intensity in the imaging experiment. Red reflects FDR-corrected significance.
E) Ranking of perturbed genes by their impact on a literature-defined set of lipid and cholesterol biosynthesis genes in the perturb-seq experiment. Red reflects FDR-corrected significance.
F) Scatterplot showing the impact of perturbed genes on literature-defined sets of Unfolded Protein Response (UPR) and Integrated Stress Response (ISR) genes.
G) A Leiden-clustered UMAP representation of CathB embeddings from the imaging experiment. Sampled cells from each cluster are shown as insets.
H) Heatmap representing enrichment in each of the clusters in NPC1 and Insr knockout cells, relative to cells with control perturbations.
I) An unbiased sampling of cells with control sgRNAs,NPC1 sgRNAs, and Insr sgRNAs, showing fluorescence micrographs the anti-CathB channel alongside polyA FISH.
J) Diagram of the diet + genetic perturbation Perturb-seq experiment
K) Ranking knockouts according to the energy distance vs control cells, in the fasted perturb-seq dataset.
L) Scatterplot comparing the energy distance vs control cells for each knockout in the *ad lib* and fasted Perturb-seq datasets.

We conducted a conceptually similar analysis of the transcriptional data and identified the *in vivo* drivers of gene expression modules with crucial functions in hepatocytes. For example, we ranked genetic perturbations by their impact on a core set of lipid biosynthesis genes (Fig 6E). Knockout of the lipid biosynthesis inhibitor Insig1 had the strongest positive effect on lipid biosynthesis gene expression; knockout of the Ldl receptor and Srebf1 also had strong positive impacts, possibly due to compensatory expression.

We also identified perturbations that activated stress response pathways such as the integrated stress response (ISR) or the unfolded protein response (UPR), in hepatocytes in an intact liver (Fig 6F). Knockout of aa-tRNA synthetase genes such as Nars and Aars activated the ISR, as did the targeting of key ISR modulators such as Eif2α (Eif2s1) and Eif2b (Eif2b4); knockout of ER import and quality control genes such as Sec61a1, Dnajb9, and Sel1l activated the unfolded protein response. Knockout of the UPR effectors Xbp1 and Ire1 (Ern1) caused a decrease in UPR gene expression, suggesting some basal level of UPR activation in hepatocytes in *ad libitum* mice. Knockout of the ER calcium transport gene Atp2a2 activated both the UPR and the ISR.

Analysis of mean immunofluorescence intensity is simple, but also simplistic—a reduction of profoundly heterogeneous 18-color 3d fluorescent images into 18-dimensional vectors. We also explored deep learning-based approaches to grapple with the perturbation-imaging data. For example, we used our autoencoder model to generate a lower-dimensional embedding for each image and then performed unbiased clustering on the embeddings. We illustrated these embeddings and the clustering with UMAP, such as with the anti-CathB morphology embeddings (Fig 6G). We then identified perturbations that shifted the representation of cells between the different clusters. For example, knockout of the lysosomal cholesterol transport gene Npc1 caused significant enrichment of anti-CathB morphologies in clusters F and G, which have many distinct

anti-CathB punctae, and depletion from clusters A and E, which have very few anti-CathB punctae (Fig 6H, 6I). This is consistent with the accumulation of dysfunctional lysosomes that occurs when cholesterol export is blocked, such as in Niemann–Pick disease type C with NPC1 or NPC2 mutations. Acute knockout of the insulin receptor Insr also caused substantial accumulation in anti-CathB morphology clusters F and G.

We fasted some mice from the lenti-sgRNA cohort for 16 hours prior to perfusion and tissue preparation and conducted a parallel perturb-seq experiment from fasted tissue (Fig 6J). We used an energy distance metric to quantify the impact of each knockout on the distribution of global transcriptional states. The overall magnitude of transcriptional phenotype was similar between fasted and *ad libitum* mice (Fig 6K,L, r = 0.77). However, we noticed that many genes involved in autophagy and the lysosome (Atp6v0c, Atp6ap1, Lamtor2) and the broader endomembrane system (Dnm2, Arfrp1, Jtb, Zw10) had especially strong phenotypes in fasted animals in comparison to their phenotypes in *ad lib* animals. This may be due to a higher dependence on autophagy during fasting—the failure of autophagy during a crucial period may cause cellular chaos.

Zonated gene expression in hepatocytes contributes to the spatial division of liver function. This zonation is believed to be maintained by gradients of morphogens, oxygen, nutrients, and hormones, but the requirement for these different processes to maintain zonal gene expression in an individual hepatocyte in an adult mouse are unclear. We wondered whether our multimodal screening could shed light on mechanisms of hepatocyte zonal identity maintenance and dynamics.

Wnt signaling is an established driver of hepatocyte zonation. Wnt ligand secretion from central vein endothelial cells promotes pericentral gene expression in hepatocytes; low Wnt in the periportal region contributes to the periportal gene expression program. We scored all cells in the

Perturb-seq dataset according to zonated gene expression and looked at the distribution of zonal expression in cells with knockouts of central Wnt effectors and inhibitors (Fig 7A). Knockout of the Wnt transducer Ctnnb1 (β-catenin) caused a major relative decrease in the fraction of cells with pericentral gene expression, consistent with the role of Wnt signaling in pericentral expression. Conversely, knockout of the Wnt inhibitor Apc caused a relative decrease in the population of hepatocytes with periportal gene expression.

We looked across all knockouts in our experiment to identify the other drivers of zonal gene expression (Fig 7B,C). We also assessed periportal and pericentral expression as two separate scores, since the knockout of some core essential genes caused a general loss of zonal identity rather than a shift in one direction or the other. Apc knockout caused the strongest increase in pericentral expression, as well as a substantial decrease in periportal expression. Vhl knockout caused a major decrease in periportal expression and also increased pericentral expression, consistent with the idea that oxygen tension directly regulates zonal gene expression. Loss of Vhl activity in the oxygen-rich periportal region may cause the specific decrease in periportal expression. Knockout of Pten and the genes Zfp830 and Ckap4, which have not been studied in the context of hepatocyte zonation, also caused a shift toward a more pericentral-like expression state. Knockout of the Wnt receptor Lgr4 caused a strong increase in periportal gene expression, phenocopying Ctnnb1.

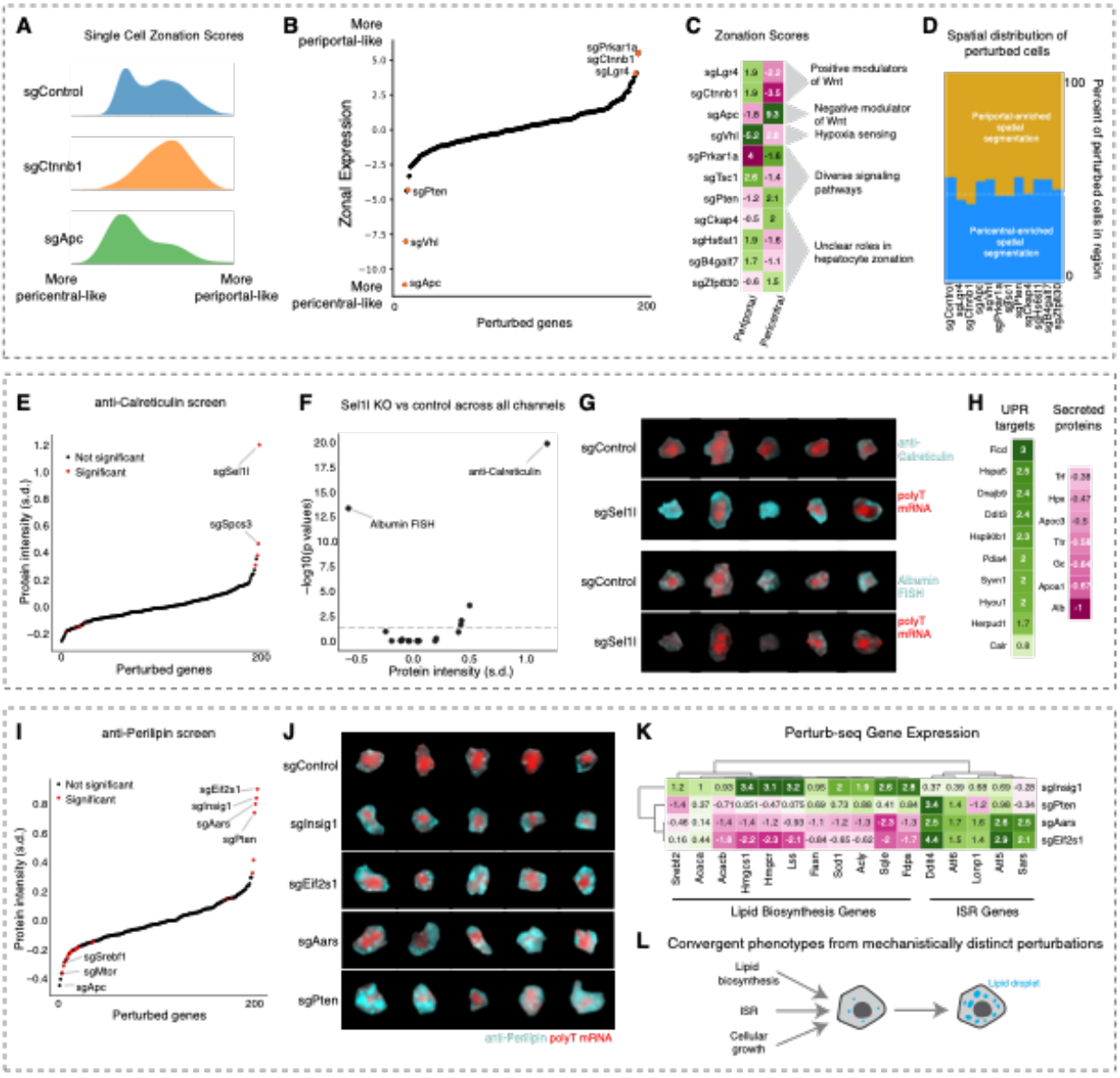

Figure 2.7: Exploring liver physiology with dual-mode screening

A) Kernel density estimate plots showing the distribution of zonation gene expression in single cells with control sgRNAs, sgRNAs targeting Ctnnb1, and sgRNAs targeting APC.
B) Ranking of perturbed genes by their impact on a set of lobule-zonation associated genes.
C) Heatmap summarizing the categories of genes whose perturbation has a large impact on zonated gene expression
D) Representation of the fraction of cells in periportal-enriched and pericentral-enriched spatially segmented zones, for the indicated knockouts. The white line represents the fraction of cells with control sgRNAs
E) Ranking of perturbed genes by their impact on anti-Calreticulin intensity in the imaging experiment. Red reflects FDR-corrected significance. (Figure legend continued on the next page.)

F) (Figure legend continued from the previous page.) Volcano plot showing the imaging channels whose intensity is significantly impacted by Sel1l knockout.
G) An unbiased sampling of cells with control sgRNAs and Sel1l sgRNAs, showing the anti-Calreticulin channel (top) and the Albumin FISH channel (bottom) alongside polyA FISH.
H) Heatmaps showing log2 fold change in the expression of UPR genes (left) and highly expressed secreted protein mRNAs (right), in the sequencing experiment.
I) Ranking of perturbed genes by their impact on anti-Perilipin intensity in the imaging experiment. Red reflects FDR-corrected significance.
J) An unbiased sampling of cells with control sgRNAs and sgRNAs targeting the indicated gene, showing the anti-Perilipin channel alongside polyA FISH.
K) A cluster-ordered heatmap showing log2 fold change in the expression of lipid biosynthesis genes and Integrated Stress Response genes for the indicated knockouts.
L) Diagram illustrating convergent mechanisms that all cause lipid droplet accumulation.

Intriguingly, knockout of protein kinase a subunit Prkar1a also caused a large increase in periportal gene expression, as did knockout of the Mtor inhibitor Tsc1. Knockout of the heparan sulfate synthesis gene Hs6st1 and the proteoglycan synthesis enzyme B4galt7 caused an increase in periportal expression and a decrease in pericentral expression, possibly due to a role for the extracellular matrix in conveying signaling molecules such as the Wnt ligand to their receptors.

Our initial lentiviral transduction appears to target cells in all zones of the lobule; cells with control sgRNAs exhibit a full range of zonated gene expression. In Perturb-seq, we measure the endpoint distribution of gene expression, but we do not know what happened during the time course of the perturbation or where each cell was in the tissue, prior to dissociation. What happens when a periportal hepatocyte gets a knockout such as sgApc that drives pericentral expression, or when a pericentral hepatocyte gets a pro-periportal sgRNA? The shift in endpoint zonal gene expression that we observe could be due to (1) *in situ* trans-"differentiation" without migration to an expression-appropriate zone, (2) trans-differentiation with migration, or (3) the death of cells in a knockout-inappropriate zone and/or the proliferation of cells in knockout-appropriate zone.

To discriminate between these hypotheses, we examined the perturbation imaging data. We automatically segmented the tissue we imaged into zones and looked to see whether knockouts that impacted gene expression impacted the position of cells within the zones (Fig 7D). We did not observe statistically significant changes to zonal location even with potent knockouts such as Apc and Ctnnb1, suggesting that, on the time scale of our experiment and with our statistical power, cells with zonation-impacting knockouts exhibit gene expression changes without migration and/or changes to survival or proliferation.

We noticed that knockout of the ERAD gene Sel1l caused a strong increase in anti-Calreticulin intensity (Fig 7E). Sel1l knockout also caused a significant decrease in Alb FISH

intensity, but had relatively small impacts on the other imaging channels (Fig 7F, 7G). Leveraging our paired data, we investigated the transcriptome-wide transcriptional changes associated with Sel1l knockout in hepatocytes. As expected for an ER chaperone, Sel1l knockout caused the upregulation of many ER chaperone mRNAs, such as Hspa5 and Dnajb9, as well as Calreticulin itself. Sel1l knockout decreased Alb mRNA, as expected from the imaging, but intriguingly it also caused the downregulation of other abundant mRNAs that encode for secreted plasma proteins, such as transferrin (Trf) and Apoa1 (Fig 7H). Altogether, the mRNAs downregulated by Sel1l knockout reflected an ~40% decrease in the total quantity of secreted protein mRNAs. This suggested that a major function of the unfolded protein response in hepatocytes *in vivo* may be to decrease the burden of secreted protein mRNAs, possibly directly through RIDD or indirectly through transcriptional downregulation.

Hepatic steatosis, characterized by the accumulation of lipids within hepatocytes, is linked to metabolic syndrome and can progress to serious diseases such as NASH and fibrosis. We used our imaging screen to identify *in vivo* drivers of steatosis. Acute knockout of the lipid biosynthesis inhibitor Insig1, the growth inhibitor Pten, the integrated stress response inhibitor Eif2α (Eif2s1), or the alanine-tRNA synthetase Aars caused steatosis, as measured by accumulation of anti-Perilipin (Fig 7I,J).

We examined the Perturb-seq data from the same mice to identify the transcriptome-wide changes associated with each perturbation (Fig 7K). Consistent with its function as an inhibitor of lipid biosynthesis gene transcription, Insig1 knockout caused the accumulation of mRNAs such as cholesterol biosynthesis genes such as Hmgcr and FDPS and fatty acid biosynthesis genes such as Acaca and Fasn. Knockout of Aars and Eif2s1 (Eif2α) drove upregulation of ISR target genes such as Ddit4 and ATF5. Intriguingly, Aars and Eif2s1 knockout actually caused a decrease in most

lipid biosynthesis mRNAs. Knockout of Pten caused a distinct transcriptional response that could not be summarized either as the ISR or lipid biosynthesis.

These four knockouts apparently drove three separate transcriptional responses but a convergent steatotic phenotype. Had we only conducted Perturb-seq, we would not have identified any common functions of Insig1, Pten, and Eif2s1/Aars; if we only had an immunofluorescence imaging screen, we would have grouped them as four knockouts with the same function. Analyzing the two modalities together, we propose that the knockouts cause steatosis through three separate mechanisms—(1) Insig1 through the transcriptional activation of lipid biosynthesis genes; (2) Eif2s1/Aars through the apparent sequestration of free lipids into lipid droplets, possibly as a component of the ISR that may serve to protect the cytoplasm and other organelles from stressed cells from biophysically disruptive lipid molecules, and (3) a distinct Pten-associated mechanism, possibly including uptake from plasma, some synthesis increase, sequestration, and/or the repurposing of organellar stores (Fig 7L).

Discussion

The ability to conduct massively parallel genetic screens with complex readouts in native tissue represents an emerging tool to generate rich *in vivo* genotype-phenotype maps that capture diverse regulators of organismal and cellular physiology with unprecedented depth and precision. Here, we develop a general method for multimodal mosaic screens with Perturb-seq- and multiplexed imaging-based phenotyping. We conduct screens in hepatocytes in mouse liver and demonstrate how the resulting data enables causal dissection of multifarious aspects of liver biology, from zonation to steatosis to the signaling of nutritional status.

Two paired goals of *in vivo* rich functional screening are: (1) to identify the phenotypic consequences of each genetic perturbation, thereby shedding light on the function of the targeted gene in the cell and tissue context where it was perturbed, and (2) to map the full variety of cellular and multicellular phenotypes that can manifest in tissue, which can reveal the relationship between multiple aspects of phenotype such as morphology and transcriptome. Our results demonstrate the value, for both goals, of multimodal phenotyping combining both multiplexed imaging and transcriptome-wide mRNA measurement. For example, we find that knockouts targeting Insig1, Eif2s1, and Pten all cause a convergent morphological phenotype with the dramatic accumulation of lipid droplets, but entirely distinct transcriptional responses. From a gene function perspective, our paired results indicate that the function of these three genes impinges on lipid homeostasis, but that they either impact lipid droplets through distinct mechanisms or have additional functions unrelated to lipid droplets. From a phenotype perspective, our results highlight a decoupling between lipid droplet morphology and the cellular transcriptome and suggest that both must be measured to capture the full state of a cell.

Our results illustrate the importance of *in vivo* studies of physiology. For example, the regulation of hepatocyte zonation is challenging to study *in vitro*. Hepatocytes rapidly lose their zonal identity when isolated and introduced into cell culture; even complex organ-on-a-chip and organoid models are imperfect recapitulations of tissue architecture, and new findings in those contexts may often require *in vivo* confirmation. Here, we avoid these issues by directly probing the regulation of zonal gene expression in the liver of a mouse. We identify novel candidate regulators of zonation such as Hs6st1 and B4galt7 and investigate the consequences of expression-neighborhood mismatch. We anticipate future experiments with genome-scale perturbation

libraries and multiplex perturbation vectors to comprehensively classify zonation regulators and to identify epistatic relationships between the associated signaling pathways.

We conducted our screens in the liver, which has a relatively small number of constituent cell types and limited spatial variation in comparison to other organs such as the brain, and is a target for lentiviral perturbation vectors similar to those established for *in vitro* genetic screens. Our approach of neonatal lenti-sgRNA delivery to transgenic Cas9 mice produced effective mosaics and potent knockout phenotypes. Future multimodal mosaic screens will build on advances in technology for the delivery of perturbation reagents, such as engineered AAVs that target the various organs with high efficiency. AAV delivery of sgRNAs has recently been used for Perturb-seq and we anticipate a relatively simple adaptation for paired imaging and fixed-cell Perturb-seq. However, AAV delivery can suffer from problems of high viral multiplicity-of-infection, which may make pooled screens more challenging to interpret.

An important area to explore further is identifying molecular regulators of various forms of cell-cell interactions, which often play a vital role in regulating tissue homeostasis and disease. The core logic of our analyses involves identifying the phenotypic differences that separate cells and/or cell neighborhoods that have received a genetic perturbation from those that have received a control perturbation. This logic can be applied both to perturbed cells as well as their neighbors to identify non-cell autonomous effects. Although in our initial study we did not find significant phenotypes from analyses of non-cell autonomous interactions with the set of perturbations chosen for this study, we envision future studies looking at diverse physiological processes – in the liver, including pathologies such as inflammation, immune cell recruitment, and fibrosis – will benefit from a spatial approach.

The analysis of multiplexed, high-resolution single-cell imaging data is still an emerging area of research, with much potential for further progress. We found that individual cells even in wildtype animals exhibit daunting morphological complexity across multiple imaging channels as well as diversity across cells and spatial zones. This complexity presents analytical challenges as well as opportunities for fundamental advances in understanding *in vivo* regulators of subcellular structure, as well as the relationship of that structure to cellular function. Although our VQ-VAE model effectively represents heterogeneity in individual protein channels across a variety of cell types, physiological states, and perturbations, it does not explicitly model the correlated features across different channels within the same cell. Future developments in deep learning methods, such as the use of visual transformers or multimodal models that can explicitly represent paired imaging and transcriptomic data, may substantially increase our ability to resolve subtle cellular phenotypes. Further development in transfer learning methods that can bridge between different datasets and imaging modalities may allow one to relate the disrupted cellular organization seen under different genetic perturbations with changes in cells in diseased human or mouse tissue measured using traditional immunofluorescence or hematoxylin and eosin staining. Such an approach could enable the 'reverse engineering' of disease phenotypes, to predict the most likely genetic perturbation that would result in the image of a pathological cell.

Although our experiments demonstrate the power of multimodal functional interrogation of diverse genes *in vivo* at the scale of hundreds of genes, future experiments will increase the scale of both phenotype and genotype measurements. In terms of phenotype, the number of genes measured spatially using RCA-MERFISH, as well as the number of proteins assayed through multiplexed immunofluorescence, could easily be increased in future experiments to obtain a more unbiased picture of cellular state through imaging. In terms of genotype, a clear future direction is

to increase the number of perturbations to the genome-wide scale, in order to enable truly systematic maps of gene function. This scale will be powerful both in terms of identifying the effects of perturbing individual genes on cellular features such as zonation identity or pS6RP staining intensity, but perhaps more importantly in terms of identifying the relationships between perturbations. When grouping different perturbations into pathways or complexes, the number of correlations between perturbations increases quadratically with the number of genes perturbed. As such, increasing to the genome-wide scale – a 100X increase in the number of genes perturbed – would give 10,000X as much information about the relationship between different genes. While achieving such a scale through sequencing is currently primarily limited by cost, reaching genome-wide scale through imaging will require substantial improvements in microscopy hardware to improve imaging throughput and biochemical methods to increase the efficiency of perturbation barcode readout. Analysis methods that aim to deconvolve phenotypes from higher multiplicity of infection using ideas from compressed sensing could further dramatically increase throughput.

Beyond scaling these experiments along the phenotype and genotype dimensions, we anticipate that our approach can be expanded to perturb multiple cell types in multiple different tissues, potentially in species other than mice, as well as different types of perturbations than CRISPR cutting. Embryonic or neonatal lentiviral delivery can be used to target other organs beyond the liver, including skin, heart, and brain. AAV-based delivery will further expand the number of accessible organs and cell types, potentially even spanning the majority of tissues in an animal. AAV-based approaches can also be applied in non-murine species, including non-human primates. Expanding beyond the hepatocytes may require novel analysis approaches or much larger numbers of cells per perturbation than the approximate order-of-magnitude $10^{2.5}$ we recover in this study. Future screens may also incorporate alternate CRISPR perturbation modalities such

as CRISPRi or CRISPRa or methods such as ORF overexpression, which would involve simple modifications of fixed-cell Perturb-seq and RCA-MERFISH and could potentially work in wild-type animals of a variety of species.

In total, our study presents a new, scalable approach for the detailed large-scale construction of *in vivo* genotype-phenotype maps with both transcriptomic and imaging readout. We anticipate that the large-scale construction of multimodal *in vivo* genotype-phenotype maps across a variety of organs, species, and perturbation modalities will provide a crucial functional complement to existing cell type atlasing efforts. These large-scale, multimodal maps will enable both basic discovery biology as well as enable the training of sophisticated AI models that can capture the rich structure of genetic and cell function.

Some aspects of our study limited the scope of our results. (1) We targeted only ~200 genes that we had reason to suspect were likely to have phenotypes in the liver, limiting the number of pairwise comparisons we could make between knockouts and the chance for serendipitous discoveries of novel gene functions, (2) we delivered perturbations only to hepatocytes and sampled a limited number of cells per perturbation, limiting our ability to understand how the phenotypic impacts of knockouts vary across cell types in the liver with high statistical power, and (3) our immunofluorescence panel covers only a subset of cellular structures and cell signaling pathways.

REFERENCES

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

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

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

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

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

Bock, C., Datlinger, P., Chardon, F., Coelho, M.A., Dong, M.B., Lawson, K.A., Lu, T., Maroc, L., Norman, T.M., Song, B., et al. (2022). High-content CRISPR screening. Nat Rev Methods Primers *2*, 8.

Boeing, S., Williamson, L., Encheva, V., Gori, I., Saunders, R.E., Instrell, R., Aygün, O., Rodriguez-Martinez, M., Weems, J.C., Kelly, G.P., et al. (2016). Multiomic Analysis of the UV-Induced DNA Damage Response. Cell Reports *15*, 1597–1610.

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

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

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

Cross, B.C.S., Lawo, S., Archer, C.R., Hunt, J.R., Yarker, J.L., Riccombeni, A., Little, A.S., McCarthy, N.J., and Moore, J.D. (2016). Increasing the performance of pooled CRISPR–Cas9 drop-out screening. Sci Rep-Uk *6*, 31782.

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

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

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

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

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

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

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

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

Frangieh, C.J., Melms, J.C., Thakore, P.I., Geiger-Schuller, K.R., Ho, P., Luoma, A.M., Cleary, B., Jerby-Arnon, L., Malu, S., Cuoco, M.S., et al. (2021). Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. Nat Genet *53*, 332–341.

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

Funk, L., Su, K.-C., Feldman, D., Singh, A., Moodie, B., Blainey, P.C., and Cheeseman, I.M. (2021). The phenotypic landscape of essential human genes. Biorxiv 2021.11.28.470116.

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

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

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

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

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

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

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

Itzhak, D.N., Tyanova, S., Cox, J., and Borner, G.H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. Elife *5*, e16950.

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

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

Jost, M., Chen, Y., Gilbert, L.A., Horlbeck, M.A., Krenning, L., Menchon, G., Rai, A., Cho, M.Y., Stern, J.J., Prota, A.E., et al. (2017). Combined CRISPRi/a-Based Chemical Genetic Screens Reveal that Rigosertib Is a Microtubule-Destabilizing Agent. Mol Cell *68*, 210-223.e6.

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

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

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

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

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

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

Ma, P., Mahoney, M.W., and Yu, B. (2013). A Statistical Perspective on Algorithmic Leveraging. Arxiv.

Maes, T., Mascaró, C., Tirapu, I., Estiarte, A., Ciceri, F., Lunardi, S., Guibourt, N., Perdones, A., Lufino, M.M.P., Somervaille, T.C.P., et al. (2018). ORY-1001, a Potent and Selective Covalent KDM1A Inhibitor, for the Treatment of Acute Leukemia. Cancer Cell *33*, 495-511.e12.

Malovannaya, A., Li, Y., Bulynko, Y., Jung, S.Y., Wang, Y., Lanz, R.B., O'Malley, B.W., and Qin, J. (2010). Streamlined analysis schema for high-throughput identification of endogenous protein complexes. Proc National Acad Sci *107*, 2431–2436.

Mamińska, A., Bartosik, A., Banach-Orłowska, M., Pilecka, I., Jastrzębski, K., Zdżalik-Bielecka, D., Castanon, I., Poulain, M., Neyen, C., Wolińska-Nizioł, L., et al. (2016). ESCRT proteins restrict constitutive NF-κB signaling by trafficking cytokine receptors. Sci Signal *9*, ra8–ra8.

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

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

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

Mimitou, E.P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., et al. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nat Methods *16*, 409–412.

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

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

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

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

Pan, J., Kwon, J.J., Talamas, J.A., Borah, A.A., Vazquez, F., Boehm, J.S., Tsherniak, A., Zitnik, M., McFarland, J.M., and Hahn, W.C. (2022). Sparse dictionary learning recovers pleiotropy from human cell fitness screens. Cell Syst.

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

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

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

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

Quirós, P.M., Mottis, A., and Auwerx, J. (2016). Mitonuclear communication in homeostasis and stress. Nat Rev Mol Cell Biology *17*, 213–226.

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

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

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

Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry, J.M., Riordan, D.P., Srinivas, N., et al. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat Biotechnol *38*, 954–961.

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

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

Rubin, A.J., Parker, K.R., Satpathy, A.T., Qi, Y., Wu, B., Ong, A.J., Mumbach, M.R., Ji, A.L., Kim, D.S., Cho, S.W., et al. (2019). Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. Cell *176*, 361-376.e17.

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

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

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

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

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

Sharma, B., Xie, L., Yang, F., Wang, W., Zhou, Q., Xiang, M., Zhou, S., Lv, W., Jia, Y., Pokhrel, L., et al. (2020). Recent advance on PTP1B inhibitors and their biomedical applications. Eur J Med Chem *199*, 112376.

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

Smits, A.H., Ziebell, F., Joberty, G., Zinn, N., Mueller, W.F., Clauder-Münster, S., Eberhard, D., Savitski, M.F., Grandi, P., Jakob, P., et al. (2019). Biological plasticity rescues target activity in CRISPR knock outs. Nat Methods *16*, 1087–1093.

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

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

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

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a Cancer Dependency Map. Cell *170*, 564-576.e16.

Wainberg, M., Kamber, R.A., Balsubramani, A., Meyers, R.M., Sinnott-Armstrong, N., Hornburg, D., Jiang, L., Chan, J., Jian, R., Gu, M., et al. (2021). A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nat Genet *53*, 638–649.

Wang, E., Zhou, H., Nadorp, B., Cayanan, G., Chen, X., Yeaton, A.H., Nomikou, S., Witkowski, M.T., Narang, S., Kloetgen, A., et al. (2021). Surface antigen-guided CRISPR screens identify regulators of myeloid leukemia differentiation. Cell Stem Cell *28*, 718-731.e6.

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

Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J.B., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowicz, M.L., Kropiwnicki, E., Jagodnik, K.M., et al. (2021). Gene Set Knowledge Discovery with Enrichr. Curr Protoc *1*, e90.

Yu, L., Myers, G., Ku, C.-J., Schneider, E., Wang, Y., Singh, S.A., Jearawiriyapaisarn, N., White, A., Moriguchi, T., Khoriaty, R., et al. (2021). An erythroid-to-myeloid cell fate conversion is elicited by LSD1 inactivation. Blood *138*, 1691–1704.

Zheng, H., Qi, Y., Hu, S., Cao, X., Xu, C., Yin, Z., Chen, X., Li, Y., Liu, W., Li, J., et al. (2020). Identification of Integrator-PP2A complex (INTAC), an RNA polymerase II phosphatase. Sci New York N Y *370*.

Publishing Agreement

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

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

DocuSigned by:
Reuben Saunders
8B6CB1E17AF14BB...
Author Signature

8/21/2024
Date