

# Report: Uncovering the Unfolded Protein Response with Perturb-seq

Max Land  
Princeton University  
mland@princeton.edu

## Abstract

In the paper, "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response", researchers applied Perturb-seq to dissect the mammalian unfolded protein response (UPR) using single and combinatorial CRISPR perturbations (Adamson et al., 2016). Two genome-scale CRISPR interference (CRISPRi) screens identified genes whose repression perturbs ER homeostasis. In this report, we broadly survey the data generated by the CRISPRi screens via cluster analysis and principal component analysis. Next, we perform cluster analysis of gene expression data from the large scale mammalian UPR Perturb-seq data. Our results generated clusters that matched the clusters discovered in the original paper. In addition, we discovered and functionally annotated two new clusters: ER proteins and gene expression/mRNA pathway. These results substantiate the reliability of Perturb-seq as a powerful genome-wide profiling tool.

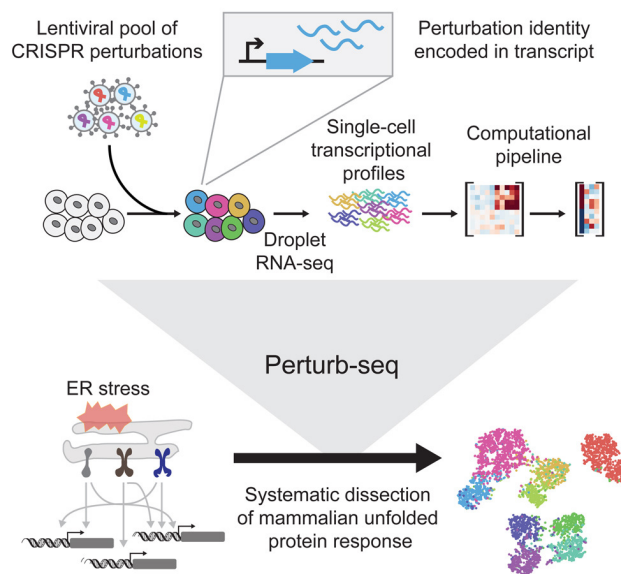


Figure 1: Graphical Abstract (Adamson 2016)

054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Background . . . . .	3
1.2	Perturb-seq . . . . .	3
1.3	Unfolded Protein Response . . . . .	3
1.4	UPR Experiments . . . . .	4
1.5	UPR Epistasis Experiment . . . . .	4
1.6	UPR Perturb-seq Experiment . . . . .	5
1.7	Our Own Analysis . . . . .	5
<b>2</b>	<b>Methods</b>	<b>5</b>
2.1	Data . . . . .	5
2.2	UPR Epistasis Experiment . . . . .	6
2.3	UPR Pertub-seq Experiment . . . . .	6
<b>3</b>	<b>Results</b>	<b>7</b>
3.1	UPR Epistasis Experiment: Cluster Analysis . . . . .	7
3.2	UPR Epistasis Experiment: PCA and t-SNE Projections . . . . .	9
3.3	UPR Perturb-seq Experiment . . . . .	9
<b>4</b>	<b>Discussion</b>	<b>13</b>
4.1	Results . . . . .	13
4.2	Pitfalls . . . . .	13
4.3	Next Steps . . . . .	13
<b>5</b>	<b>Author Contributions</b>	<b>14</b>
<b>6</b>	<b>References</b>	<b>14</b>
<b>7</b>	<b>Appendix</b>	<b>15</b>
7.1	Paper Figure S3 B . . . . .	15
7.2	Paper Figure 5A . . . . .	16

# 1 Introduction

## 1.1 Background

Functional genomics is a field in molecular biology that attempts to describe gene functions and interactions. The development of CRISPR has proven especially useful for probing genomic function in high-throughput. However, previous advances in pooled screenings have relied on simple phenotypic readouts that average properties of population (i.e. expression of a few exogenous reporters or cell viability). As a result, it is impossible to distinguish between mechanistically distinct perturbations that cause similar responses or when a bulk phenotype is driven by a sub-population.

## 1.2 Perturb-seq

To address these limitations, researchers built a highly parallel platform, termed "Perturb-seq", for single-cell functional genomics by pairing single-cell RNA sequencing (RNA-seq) with CRISPR-based transcriptional interference (CRISPRi), which mediates gene inactivation with high efficacy and specificity (Gilbert et al., 2013). They developed a robust cell barcoding strategy that encodes the identity of CRISPR-mediated perturbation in an expressed transcript, which is captured during single-cell RNA-seq analyses. This allows researchers to perform single-cell RNA-seq on each perturbed cell and see how knocking out a gene (or reducing expression) affects the transcriptome. Perturb-seq is readily implementable and scalable for parallel screening of large-scale single-cell RNA-seq and produces rich phenotypic output consisting of RNA-seq profiles for tens of thousands of single cells.

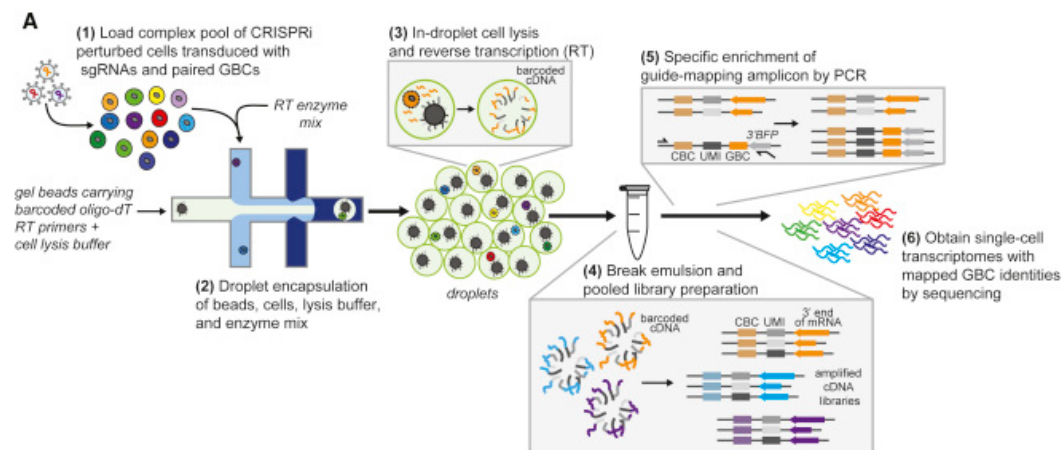


Figure 2: Schematic of Perturb-seq (Adamson 2016)

To parse these massive data sets, researchers also developed a novel analytic pipeline to decompose the noisy, high-dimensional single-cell data into more interpretable components. This enables the decoupling of the responses to a given perturbation within individual cells and from confounding effects, such as the cell cycle (Klein et al., 2015).

## 1.3 Unfolded Protein Response

Next, researchers used Perturb-seq to perform a systematic analysis of the mammalian unfolded protein response (UPR). UPR is triggered in the endoplasmic reticulum of cells following various disturbances within the cell such as misfolded proteins, infection, and membrane abnormalities. Three major distinct pathways are involved, controlled by the transmembrane sensor proteins ATF6, IRE $\alpha$ , and PERK. These sensors activate transcription factors (N-terminal cleavage product of ATF6, XBP1, and ATF6, respectively) to promote survival or trigger cell death (Walter and Ron, 2011).

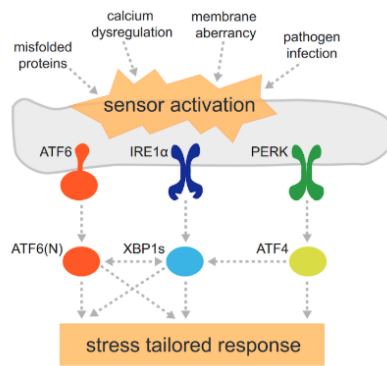


Figure 3: Mammalian Unfolded Protein Response (Adamson 2016)

Considering the diversity of potential inputs and the complexity of outcome, comprehensive characterization of the UPR in mammalian cells requires both unbiased profiling of the physiological stresses that activate the sensors and delineation of the complex transcriptional phenotypes for each input.

#### 1.4 UPR Experiments

To independently manipulate the three branches of UPR, researchers developed a method to simultaneously repress up to three genes with high efficacy. Then, they applied Perturb-seq with combinatorial repression of the UPR sensor genes to delineate the distinct transcriptional programs of the three branches. Next, they conducted a two-part experiment to dissect the mammalian UPR.

#### 1.5 UPR Epistasis Experiment

In the first experiment, researchers performed two genome-scale CRISPRi screens to identify genes important in maintaining ER homeostasis. sgRNAs were introduced in single cells to target each UPR sensor gene in all possible single, double and triple combinations. Cells were treated with chemical treatments that induce UPR: Thapsigargin and Tunicamycin, which inhibits ER calcium pump and N-linked glycosylation respectively. DMSO was used for control. After, the cells were sequenced.

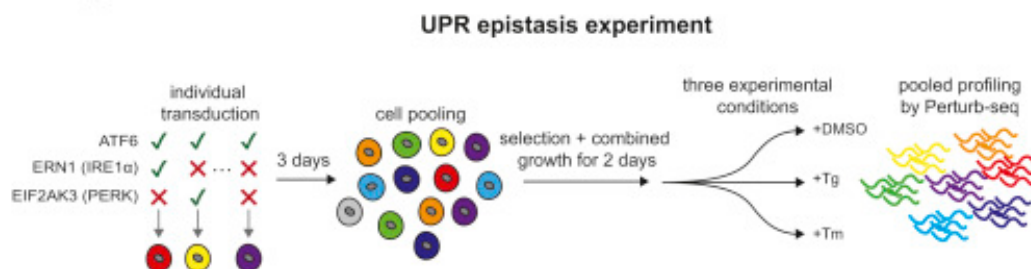


Figure 4: Schematic of UPR Epistasis Experiment (Adamson 2016)

From the ~ 15,000 cells sequenced, researchers selected UPR-inducing sgRNAs by looking at genes with the highest reporter phenotype activity in both screens. Using these genes, they next performed a Perturb-seq experiment.

## 1.6 UPR Perturb-seq Experiment

For this experiment, researchers applied Perturb-seq (as discussed previously) to a small CRISPRi library of 91 sgRNAs targeting 82 genes. These included many of the strongest hits identified in the previous experiment. They also included two negative controls.

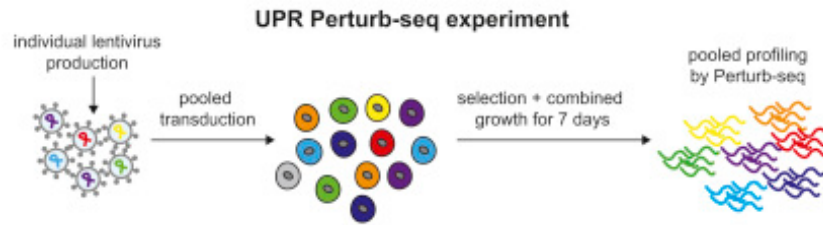


Figure 5: Schematic of UPR Perturb-seq Experiment (Adamson 2016)

Overall, this experiment generated  $\sim 65,000$  transcriptomes worth of gene expression data.

## 1.7 Our Own Analysis

Both the UPR Epistasis Experiment and UPR Perturb-seq Experiment generated large data sets. For the remainder of this report, we discuss the methods and results of our own analysis on the data, and compare with the results found by the original paper.

## 2 Methods

### 2.1 Data

Sequencing Data from both experiments can be found online:

- GEO: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2406677> (Epistasis)
- GEO: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2406681> (Perturb-seq)

Only the outputs from cell ranger are necessary. There should be four files per experiment:

1. `barcodes.tsv` - List of guide cell barcodes.
2. `genes.tsv` - List of guide cell barcodes.
3. `raw_cell_identities.csv` - Cell identities spreadsheet: provides sequence, guid identity, read count, UMI count, coverage, etc.
4. `matrix.mtx` - Gene-barcode matrix: data entries are the UMI-collapsed counts of these reads, representing counts of molecules per cell as determined and filtered by cellranger. The rows are guide cell barcodes and the columns are gene identities. The matrix itself is in matrix market exchange format.

A note on barcodes and identifiers: In order to identify the sgRNAs, cells, and unique mRNA reads, researchers had to introduce several different markers or barcodes.

- Cell Barcode (CBC) - uniquely identifies each cell, links all sequencing reads from a given cell.
- Unique Molecular Identifier (UMI) - unique identifies each mRNA, enables counting of each captured mRNA molecule. This prevents counting of duplicates made PCR.
- Guide Barcode (GBC) - unique for each sgRNA introduced, marks specific cell perturbations and identifies a sgRNA.

The CBC and UMI are affixed to cDNA molecules during reverse transcription. They are read out by sequencing the mRNAs after the transcripts are pooled together.

## 2.2 UPR Epistasis Experiment

Note: There are useful pre-existing Perturb-seq library codes for processing single-cell expression matrix data.

Goal: Recreate Figure S3, B (See Appendix 7.1).

1. Process raw data, generate single-cell expression matrix from initial data files.
2. Remove multiplets - retain only singlets.
3. Filter out genes with zero expression.
4. Pull out control cells exposed to DMSO. Create new sub-population.
5. Filter remaining expression matrix for +thapsigargin samples.
6. Z-Normalize filtered matrix against DMSO control sub-population.
7. Average/Group cells by experimental condition (i.e. Thaps\_ATF6).
8. Filter for noisy genes (differentially expressed). Three different ways.
9. Perform cluster analysis.
10. Perform PCA and t-SNE projections to reduce dimensions of data.

Additional notes:

As mentioned in Introduction (Section 1.5), the UPR Epistasis Experiment consists of 8 genetic backgrounds (all single, double, and triple knockdowns of the 3 UPR branches) observed across 3 conditions (+DMSO, +thapsigargin, and +tunicamycin).

Step 5 and 7: we specifically filtered for

- DMSO\_control
- Thaps\_ATF6
- Thaps\_ATF6-IRE1
- Thaps\_ATF6-IRE1-PERK
- Thaps\_ATF6-PERK
- Thaps\_IRE1
- Thaps\_IRE1-PERK
- Thaps\_PERK
- Thaps\_control

Step 8: There are three ways we look for differential genes.

1. Deviations from the mean-CV relationship.
2. Kolmogorov-Smirnov (KS) test.
3. Random forest-based methods.

Step 9: For cluster analysis, we used euclidean distance metric and average linkage.

## 2.3 UPR Pertub-seq Experiment

Note: a demo of the following steps can be found in `uprExperimentAnalysisDemo.ipynb`.

Goal: Recreate Figure 5A (See Appendix 7.2).

1. Process raw data, generate single-cell expression matrix from initial data files.
2. Remove multiplets - retain only singlets.
3. Filter out genes with zero expression.
4. Group single-cells into sub-populations based on guide target.

5. Filter for noisy genes (differentially expressed).
6. Z-Normalize UMI counts.
7. Perform cluster analysis.

Additional notes:

Steps 1-5: The original matrix was ~ 65,000 cells by ~ 32,000 genes. Even after removing multi-plets and filtering out genes with zero expression, the matrix was still ~ 50,000 cells by ~ 22,000 genes. After grouping single-cells by guide target and selected for noisy genes, we were able to reduce the matrix to ~ 84 cells by ~ 478 genes.

Step 6: Next, we normalized expression distribution by Z-scoring. We were not able to normalize against the negative control. Instead, we first normalized the UMI counts within each cell to the me-dian UMI count within the population. The expression within the population is then Z-normalized. We recognize it makes more sense to normalize the expression data before selecting for noisy genes, but that matrix was too large for the computing power we had available at our disposal.

Step 7: Finally, we clustered the expression data and generated dendrogram heatmaps. We tested both Euclidean and Pearson distance metrics with average linkage clustering.

3 Results

3.1 UPR Epistasis Experiment: Cluster Analysis

The following clustering maps show differential expression of genes identified using different meth-ods. These are all cells treated with thapsigargin (thaps).

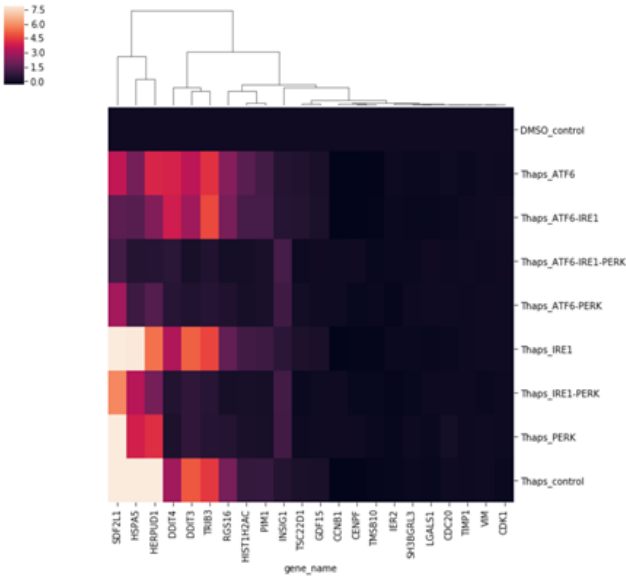


Figure 6: Mean CV

378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431

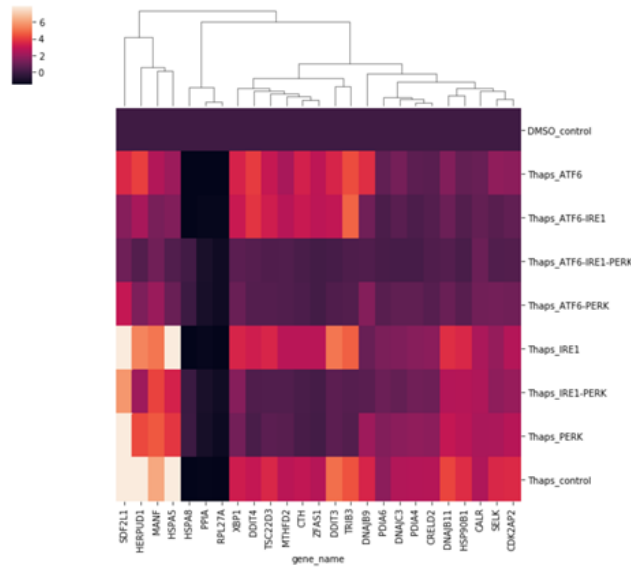


Figure 7: KS

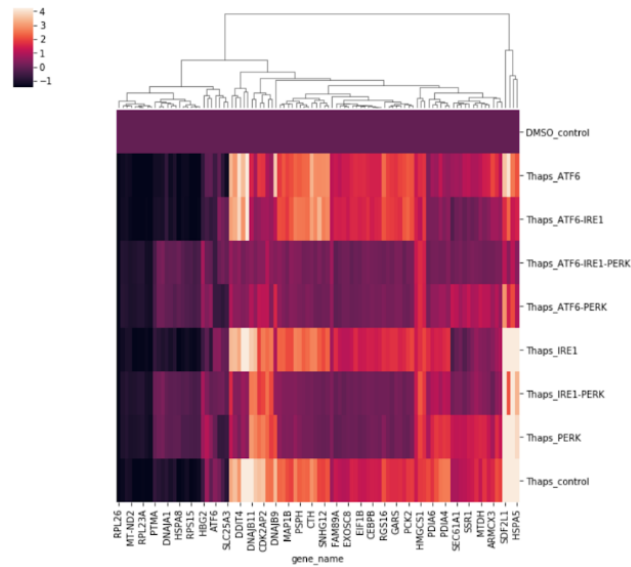


Figure 8: Random Forest

There isn't much to say about these figures. These were generated as a proof of concept type test to see if we could actually obtain results from cluster analysis. It is worth noting Random Forest classifier is better at discovering noisy genes with stronger expression than KS and Mean CV. In other words, Random Forest classifier seems to discover a desired number of highly variable genes.



### 3.2 UPR Epistasis Experiment: PCA and t-SNE Projections

Even though many genes may be differentially expressed, they often have some underlying relationship to one another. These correlations can be reduced to a certain number of factors. PCA identifies those components and t-SNE is a 2D projection that sorts the expression into clusters based on those components.

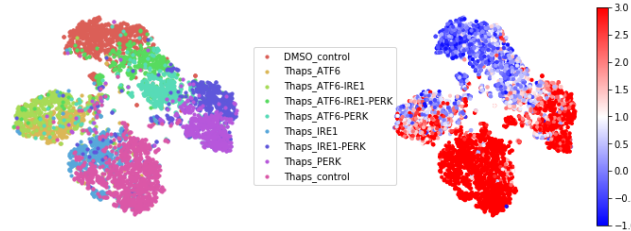


Figure 9: t-SNE projection of thapsigargin-treated cells (Left). Expression of HSPA5 (Right).

Comparing our t-SNE projection to Figure S3 B in the paper (See Appendix 7.1), we see that our figure does a pretty good job of clustering the thapsigargin-treated cells by experimental perturbations. It is worth noting our t-SNE plot has slightly different orientation and shape, but that is most likely due to a difference in components being identified by PCA.

Next, HSPA5 is a differentially expressed gene that encodes for a heat shock response protein located in the ER that is involved in proper folding/assembly of proteins. From the results above, we can deduce the activation HSPA5 is dependent on the ATF6 pathway. More specifically, it is active when the ATF6 pathway is not induced.

### 3.3 UPR Perturb-seq Experiment

For this section of our analysis, we perform hierarchical clustering of genes from the UPR Perturb-seq experiment. In the paper, their result is reported in Figure 5A (See Appendix 7.2). Their Figure 5A includes functional annotations of gene clusters with known ER pathways.

It is worth noting the researchers performed their cluster analysis using all of the genes, whereas we filtered for differentially expressed genes first before clustering. Next, while their cluster analysis uses the Pearson distance metric, we report three figures.

1. Euclidean Distance Metric - all differentially expressed genes.
2. Pearson Distance Metric - all differentially expressed genes.
3. Pearson Distance Metric - random subset 300/478 differentially expressed genes.

We box similar gene clusters between our results and the Figure 5A from the paper.

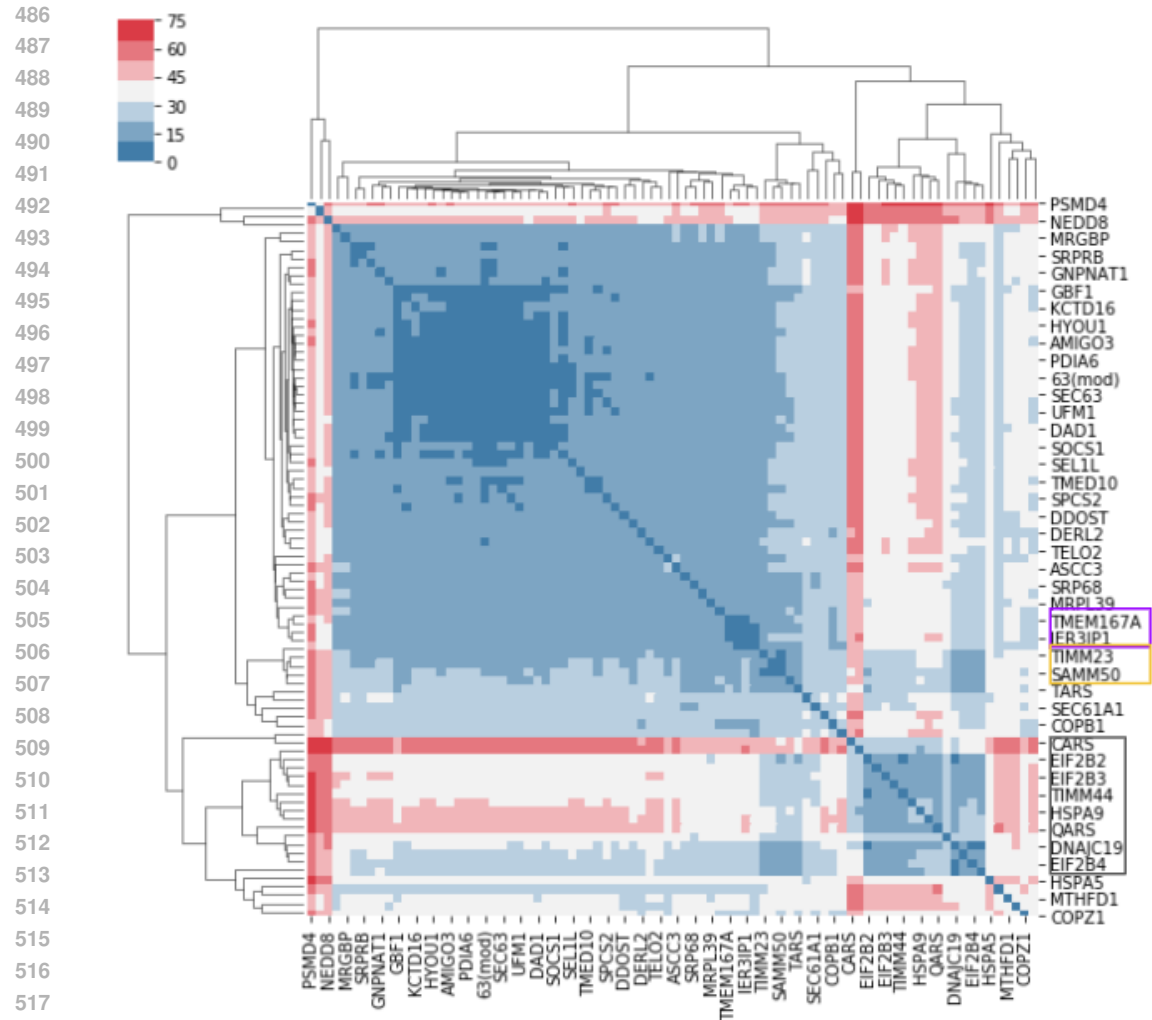


Figure 10: Euclidean Distance

Note, because this figure is generated using the Euclidean distance metric, more correlated genes are closer in distance together. Thus, the more blue, the more correlated two genes are (colors are inverted compared to paper's Figure 5A which uses Pearson). Keeping this in mind, we see there is a shocking similarity in the overall pattern in our figure and Figure 5A, despite having a different leaf orders.

Comparing specific gene clusters, we see there are five similar clusters that are functionally annotated in Figure 5A:

1. Purple: trafficking
2. Yellow: mitochondrial genes
3. Black: Aminoacyl tRNA synthetase, mitochondrial genes, translation initiation

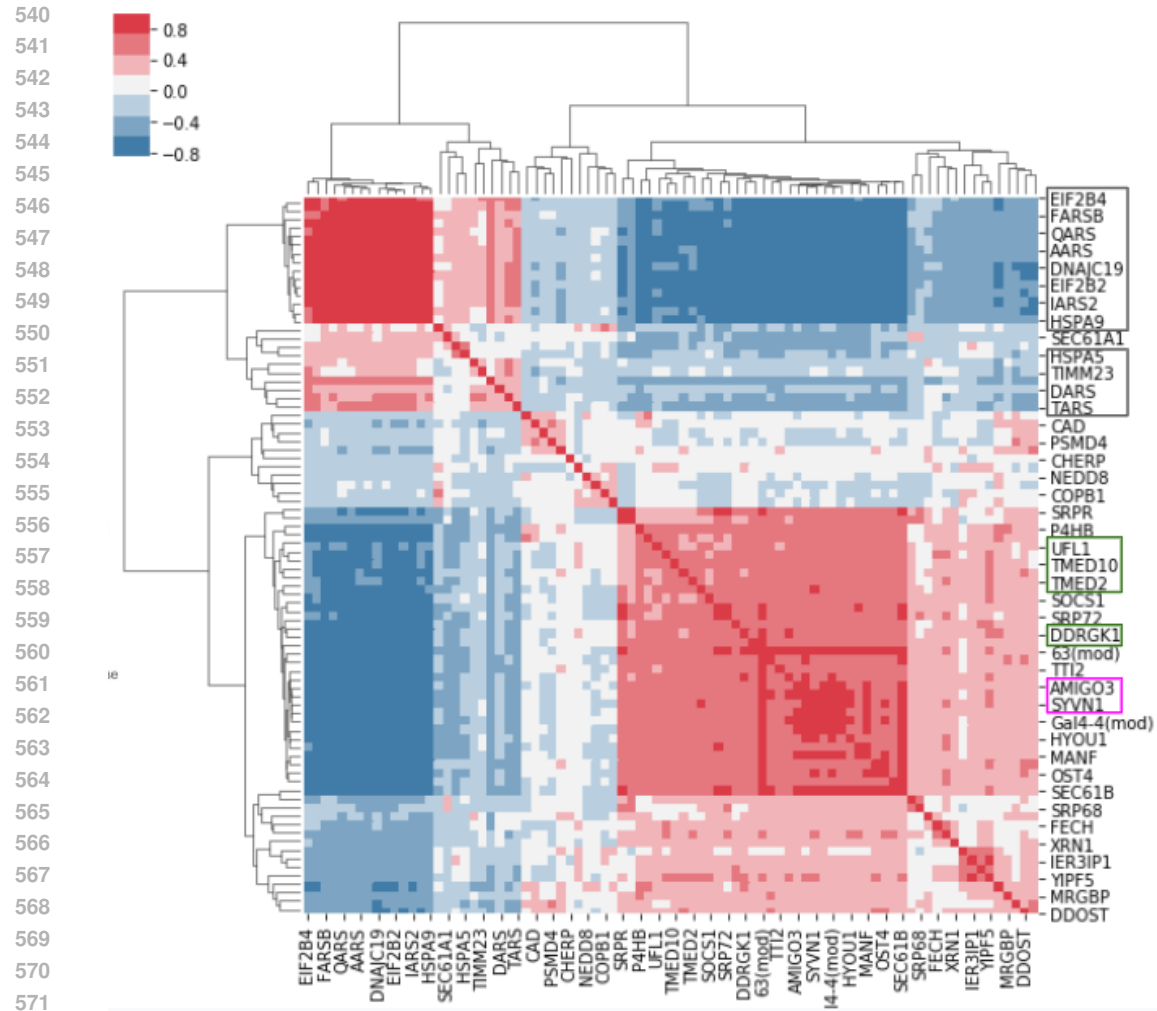


Figure 11: Pearson Distance (All Genes)

It is interesting to see that our cluster heatmap generated using the Pearson distance metric looks more different from the paper's Figure 5A than our Euclidean distance figure.

Nonetheless, we see there are six similar clusters that are functionally annotated in Figure 5A:

1. Black: Aminoacyl tRNA synthetase, mitochondrial genes, translation initiation
2. Green: trafficking complex, UFMylation
3. Pink: ERAD E3 Ligase

In case you were curious, UFM stands for ubiquitin fold modifier. UFMylation refers to the process of adding a ubiquitin tag to a substrate for protein degradation.

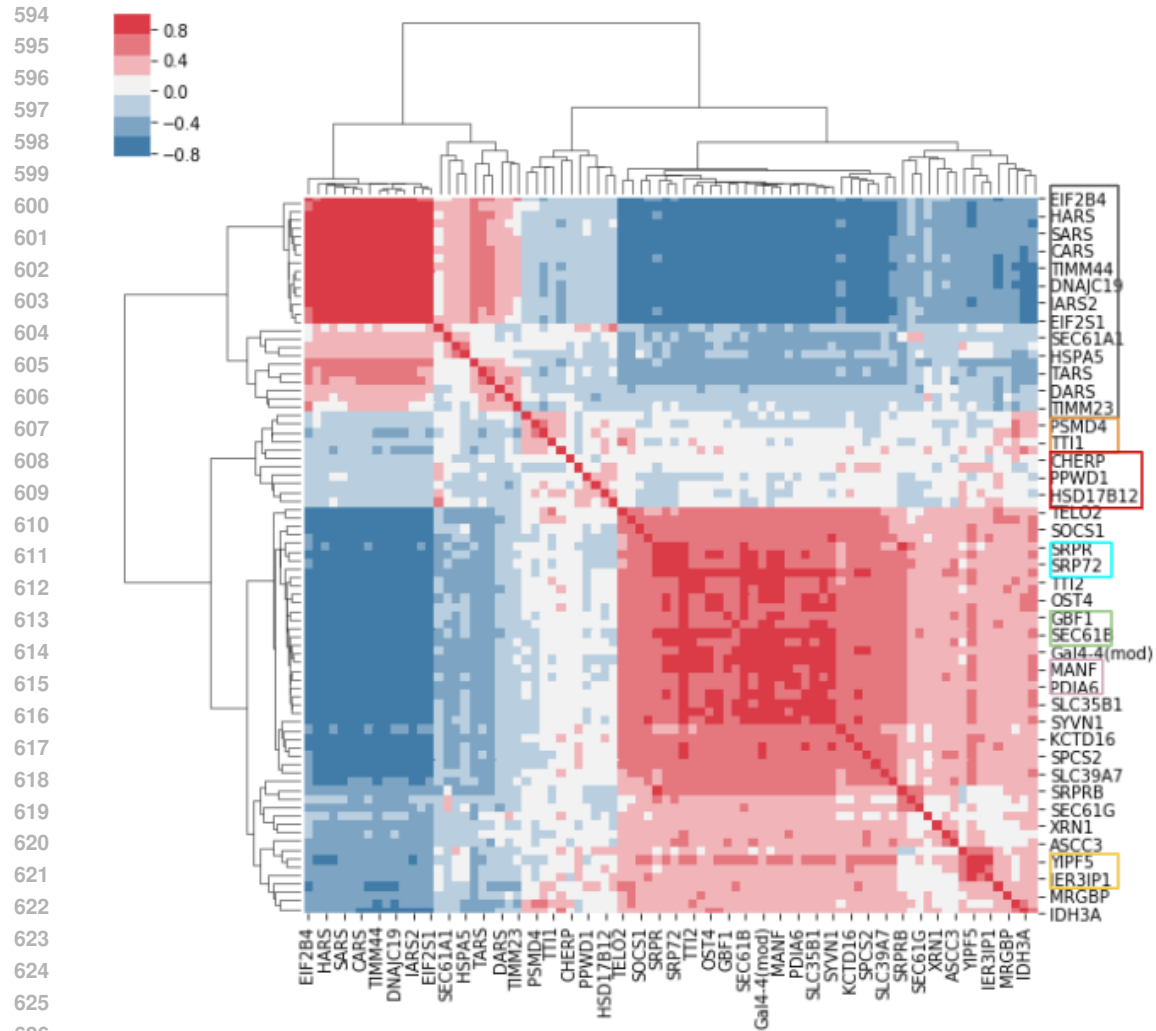


Figure 12: Pearson Distance (Random Subset of Genes)

Taking a random subset of differentially expressed genes, it makes sense the generated cluster heatmap looks very similar to the previous figure using Pearson distance metric with all genes. However, the similar clusters discovered are quite different.

We see there are seven similar clusters that are functionally annotated in Figure 5A:

1. Black: Aminoacyl tRNA synthetase, mitochondrial genes, translation initiation
2. Orange: Triple T complex
3. Neon blue: SRP-mediated protein targeting
4. Green: Nonessential translocon
5. Yellow: trafficking

Furthermore, we discovered two similar cluster that were not annotated in the paper. We went ahead and annotated them ourselves:

1. Red: Gene expression/mRNA pathway
2. Light Pink: ER proteins

## 4 Discussion

In this report, we chose to recreate results from the paper, "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response" (Adamson et al., 2016). In the paper, researchers developed Perturb-seq, which combines droplet-based single-cell RNA-seq with a strategy for barcoding CRISPR-mediated perturbations, allowing many perturbations to be profiled in pooled format. Perturb-seq was then used to investigate mammalian unfolded protein response.

### 4.1 Results

The paper consisted of a two part experiment. For the UPR Epistasis Experiment, we successfully recreated the t-SNE projection plot of thapsigargin-treated cells (Figure S3 B in the paper). Furthermore, we found random forest identifies the genes that are best at classifying cells into different populations. A foray into genes that have been assigned high importance revealed many to be sub-population specific. An example is HSPA5, which we were able to deduce as linked to the ATF6 pathway.

For the second part of the experiment, the UPR Perturb-seq Experiment, we successfully processed and clustered many GBs of single-cell RNA-seq data. We were able to use both Euclidean and Pearson distance metrics and get similar cluster results to the paper (Figure 5A) despite filtering first for differentially expressed genes. This makes intuitive sense, since the gene expression matrix is large and sparse and most data entries are zero so all the relevant information is contained in the differentially expressed genes anyways. Also, Euclidean distance behaves similar to Pearson if we Z-normalize the expression matrix. This is because Pearson is not sensitive to linear transformations of the data. Overall, the gene clusters found are consistent with known interactions, such as those involved in protein trafficking, SRP-mediated protein targeting, and ubiquitylation (UFMylation and ERAD E3 Ligase). We were also able to annotate two new clusters, discovered when we took a random subset of differentially expressed genes. It is also worth noting that our random subset clustering analysis produced the most similar number of gene clusters to the paper's Figure 5A. It is worth looking more into the validity of the results of random sampling cluster analysis. If they can be shown to produce similar and/or better results compared to cluster analysis that takes into account all genes, it would prove very useful when dealing with large gene expression matrices.

Ultimately, the results from our cluster analysis of Perturb-seq data align with that of the paper—both reach conclusions that are consistent with existing knowledge. This substantiates the reliability of Perturb-seq and makes a strong case for its utilization in future research.

### 4.2 Pitfalls

As is common with algorithms applied to novel experimental data in computational biology, it was hard to objectively evaluate quality of the results of our analysis. For the most part, we were only able to cross-validate our clusters with those already identified and functionally annotated by researchers in the original paper (Adamson et al., 2016). While we were able to functionally annotate two new clusters, that took a lot of additional research into known protein functionality. As seen in Figures 10-12, there are many clusters generated by our analysis that remain unannotated.

On another note, another pitfall we encountered was the fact the raw single-cell expression data was extremely large and we didn't have access to the proper computing power. While the computing clusters did provide us with memory, our run-time was still too slow. Ultimately, we had to significantly reduce the size of our original expression matrix (averaging by guide targets and filtering for differentially expressed genes). The original analysis performed in the paper did not need to reduce the size of the expression matrix. While we were still able to retain most of the relevant information by filtering for differentially expressed genes, there must still be some loss of information in our analysis compared to that of the original paper.

### 4.3 Next Steps

For the UPR Epistasis Experiments, we can reconstruct the t-SNE plots with Tunicamycin instead of Thapsigargin. Furthermore, we can try determine the branch specificity (ATF6, XBP1, ATF4) and sensitivity of each gene.

For the UPR Perturb-seq Experiment, we should rerun our cluster analysis, filtering for differentially expressed genes using random forest. The current method used mean CV to identify differentially

expressed genes, but our results from the UPR Epistasis Experiments suggest random forest is better at discovering highly variable genes.

On a broader scale, the biggest challenge of Perturb-seq is on the analytical side. Current techniques (Zheng et al., 2016) have allowed researchers to generate RNA data from ~50,000 cells in ~10 min. As technologies improves, sequencing costs will decline. However, the data generated by Perturb-seq is intrinsically noisy. Novel analytical approaches in addition to current large-scale analyses (Dixit et al., 2016) of single-cell data is necessary to gain a systematic understanding of the complex regulatory programs at work within cells.

## 5 Author Contributions

The work on the project was pretty collaborative overall. After agreeing on the paper, Katie and I both decided to recreate the figures from the analysis of both the UPR Epistasis and Perturb-seq Experiments. Katie took the lead on the generating the figures for the UPR Epistasis Experiment and I took the lead on the analysis of the UPR Perturb-seq Experiment. We both learned SLURM in order to use Princeton's computing clusters. After generating the figures, we analyzed the results together and made the presentation together.

Ultimately, we replicated only a portion of the genomics data analysis performed in the paper. We made some progress decoupling the branches of the UPR and uncovered subtle subpopulations within cells of the same time. Nonetheless, we successfully demonstrated the potential of Perturb-seq as a powerful tool in future research.

## 6 References

1. Britt Adamson, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, Luke A. Gilbert, Max A. Horlbeck, Marco Y. Hein, Ryan A. Pak, Andrew N. Gray, Carol A. Gross, Atray Dixit, Oren Parnas, Aviv Regev, Jonathan S. Weissman (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response, *Cell*, Volume 167, Issue 7, Pages 1867-1882.e21, ISSN 0092-8674, <https://doi.org/10.1016/j.cell.2016.11.048>.
2. A. Dixit, O. Parnas, B. Li, J. Chen, C.P. Fulco, L. Jerby-Arnon, N.D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens *Cell*, 167 (2016), pp. 1853-1866 this issue
3. L.A. Gilbert, M.H. Larson, L. Morsut, Z. Liu, G.A. Brar, S.E. Torres, N. Stern-Ginossar, O. Brandman, E.H. Whitehead, J.A. Doudna, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes *Cell*, 154 (2013), pp. 442-451
4. A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D.A. Weitz, M.W. Kirschner Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells *Cell*, 161 (2015), pp. 1187-1201
5. P. Walter, D. Ron. The unfolded protein response: from stress pathway to homeostatic regulation. *Science*, 334 (2011), pp. 1081-1086
6. G.X.Y. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S.B. Ziraldo, T.D. Wheeler, G.P. McDermott, J. Zhu, et al. Massively parallel digital transcriptional profiling of single cells *bioRxiv* (2016), 10.1101/065912

## 7 Appendix

### 7.1 Paper Figure S3 B

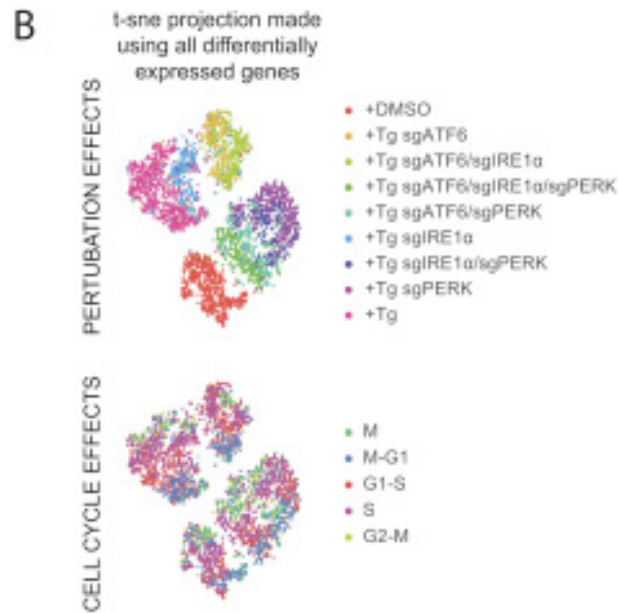


Figure 13: Analysis of thapsigargin-treated cells (t-SNE projections) (Adamson 2016)

## 7.2 Paper Figure 5A

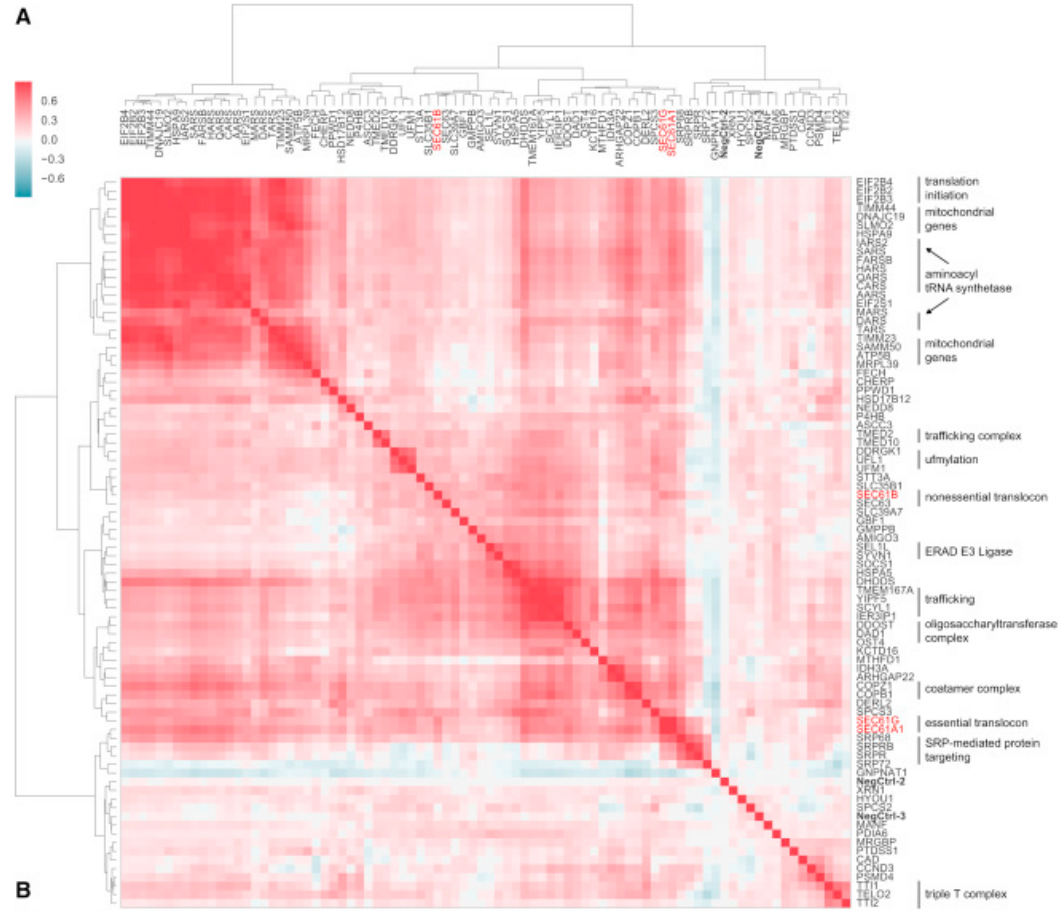


Figure 14: Clustering of genes from UPR Perturb-seq experiment. Heatmap displays correlations between hierarchically clustered average expression profiles from all cells bearing sgRNAs targeting the same gene (identified by GBCs). Functional annotations are indicated. (Adamson 2016)