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. 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. Finally, we perform cluster analysis of gene expression data from the large scale mammalian UPR Perturb-seq data.

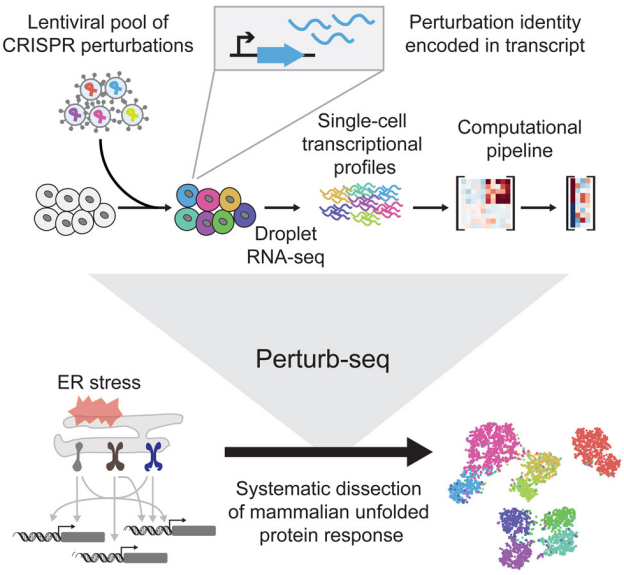

Figure 1: Graphical Abstract

Contents

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. 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 (Figure 1) 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.

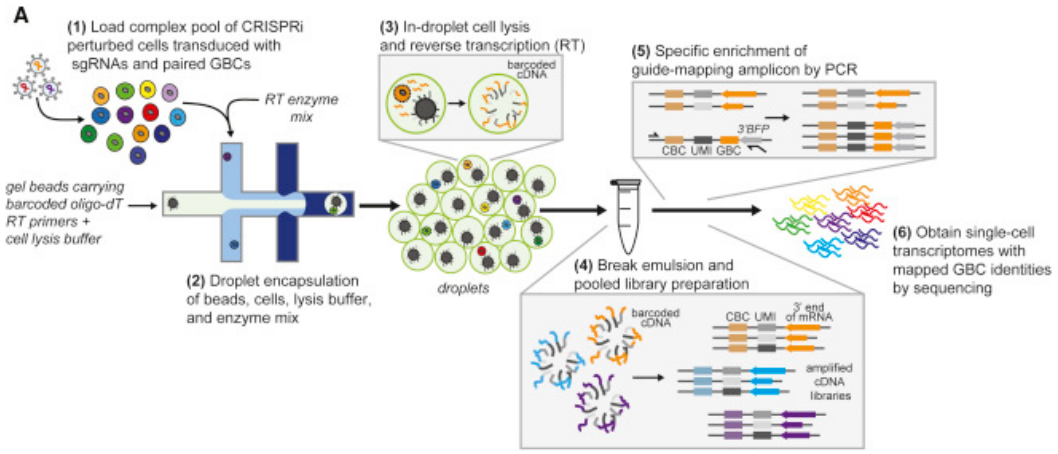

Figure 2: Schematic of Perturb-seq (Adamson 2016)

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.

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, IREα, and PERK. These sensors activate transcription factors (N-terminal cleavage product of ATF6, XBP1, and ATF6, respectively) to promote survival or trigger cell death.

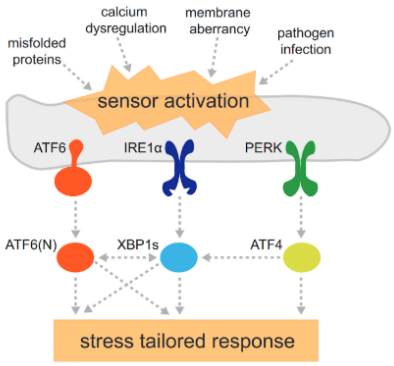

Figure 3: Mammalian Unfolded Protein Response

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.

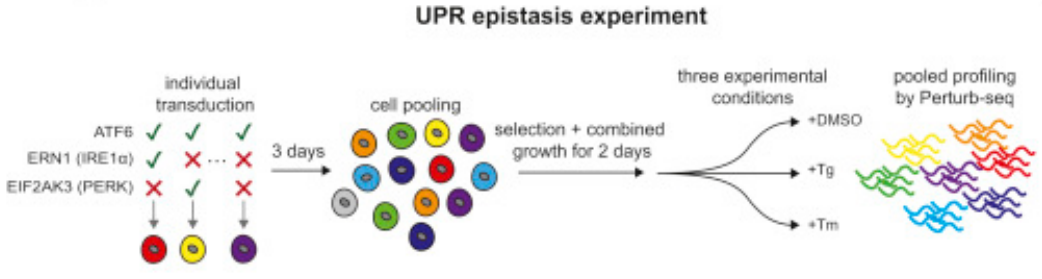

Figure 4: Schematic of UPR Epistasis Experiment

From the $\sim$ 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.

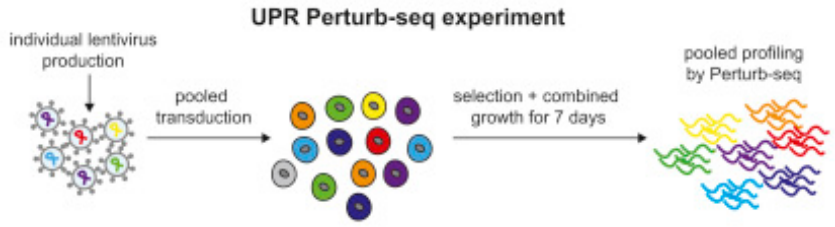

Figure 5: Schematic of UPR Perturb-seq Experiment

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 6.1).

1. Process raw data, generate single-cell expression matrix.
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. KS test.
3. Random forest-based methods.

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 6.2).

1. Process raw data, generate single-cell expression matrix.
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 $\sim$ 65,000 cells by $\sim$ 32,000 genes. Even after removing multiplets and filtering out genes with zero expression, the matrix was still $\sim$ 50,000 cells by $\sim$ 22,000 genes. After grouping single-cells by guide target and selected for noisy genes, we were able to reduce the matrix to $\sim$ 84 cells by $\sim$ 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 median 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 methods. These are all cells treated with thapsigargin (thaps).

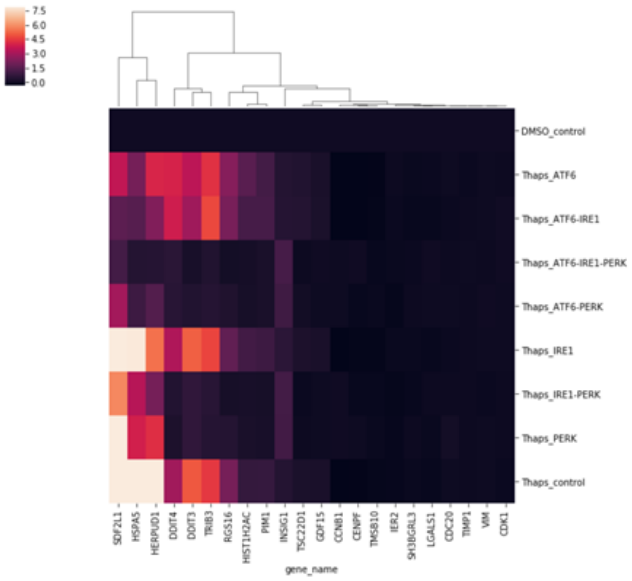

Figure 6: Mean CV

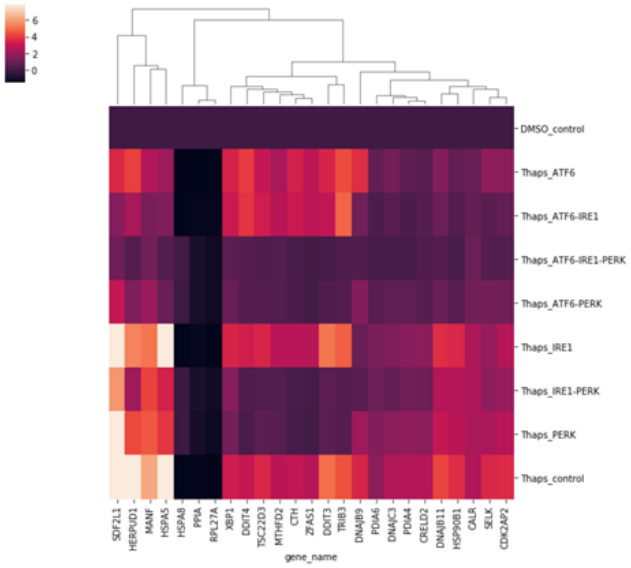

Figure 7: KS

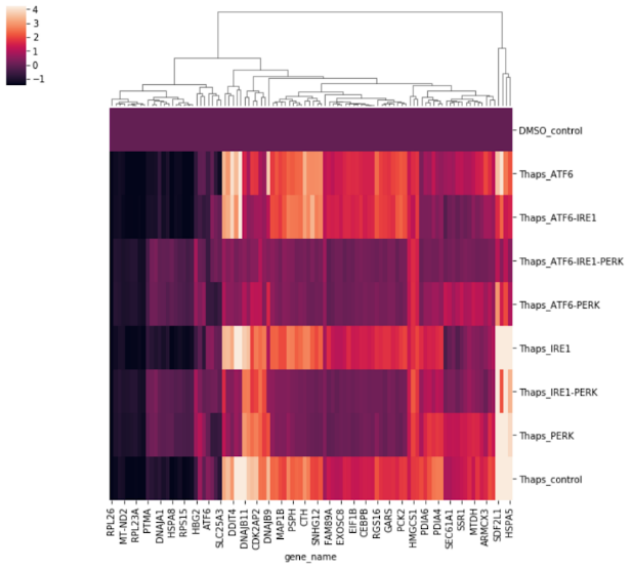

Figure 8: Random Forest

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.

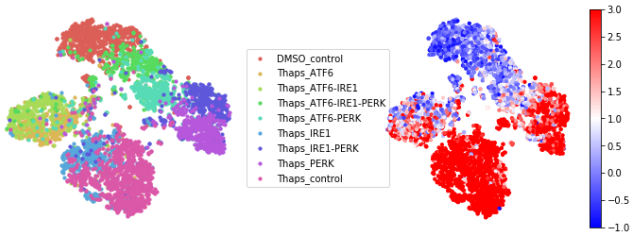

Figure 9: t-SNE projection of thapsigargin-treated cells (Left). Expression of HSPA5 (Right).

Comparing our t-SNE projection to Figure S3 B in the paper (See Appendix 6.1), 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 6.2). 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.

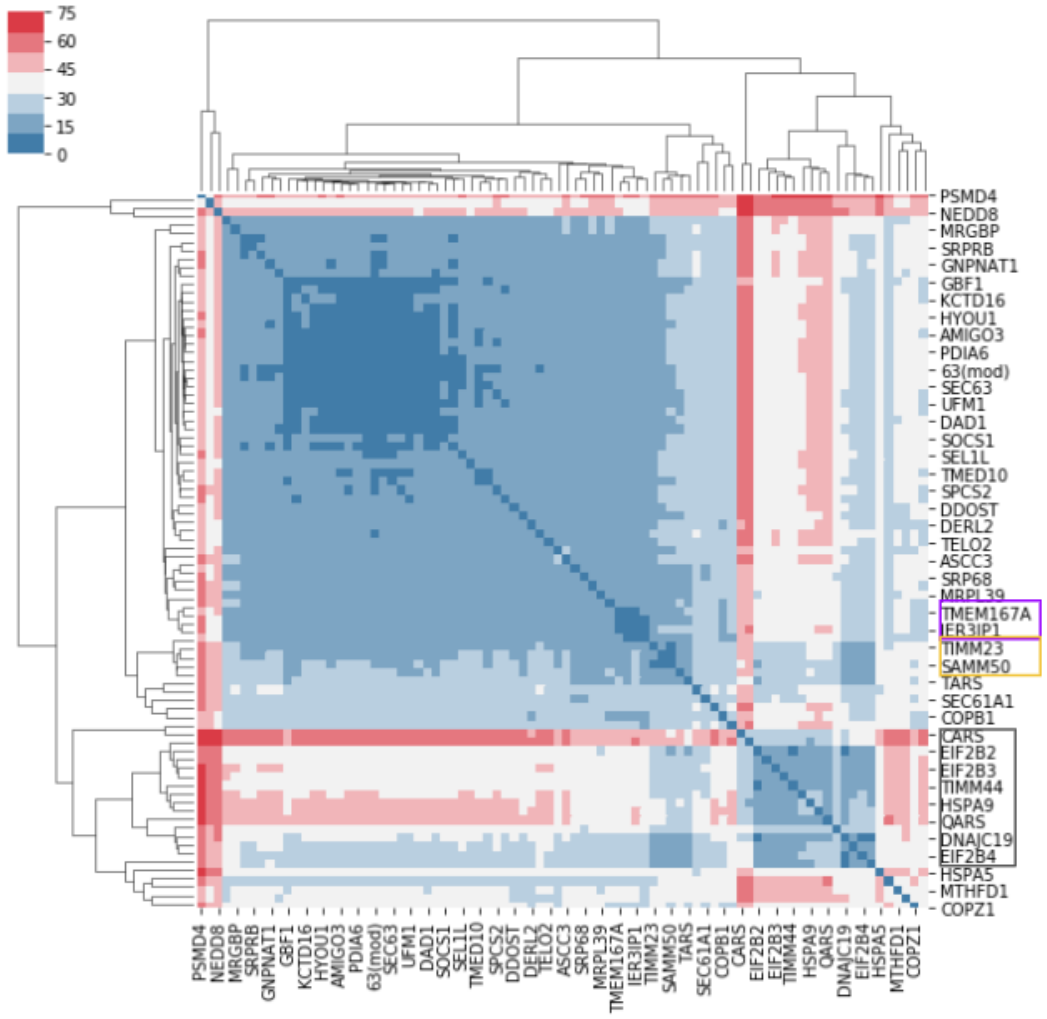

Figure 10: Euclidean Distance

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 similiarity 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 sythetase, mitochondrial genes, translation initiation

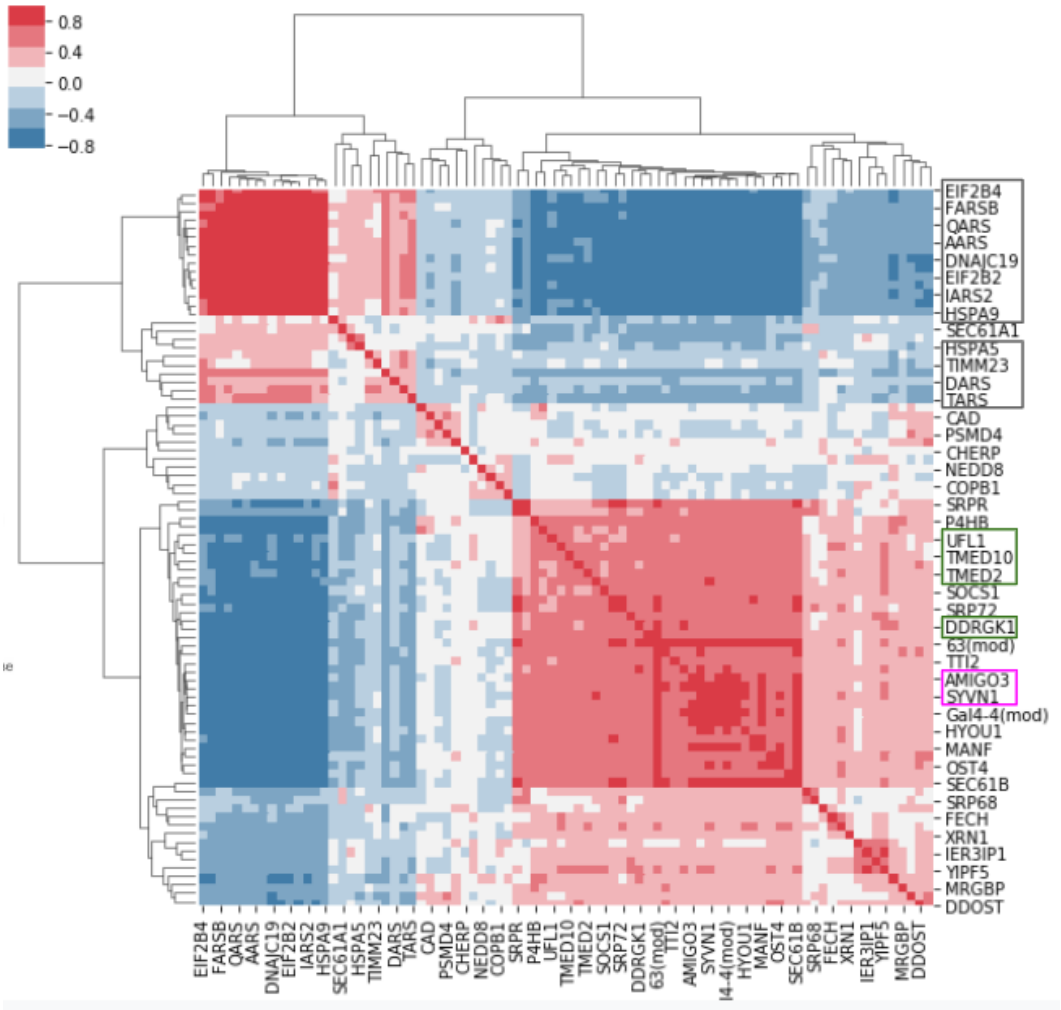

Figure 11: Pearson Distance (All Genes)

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 sythetase, 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.

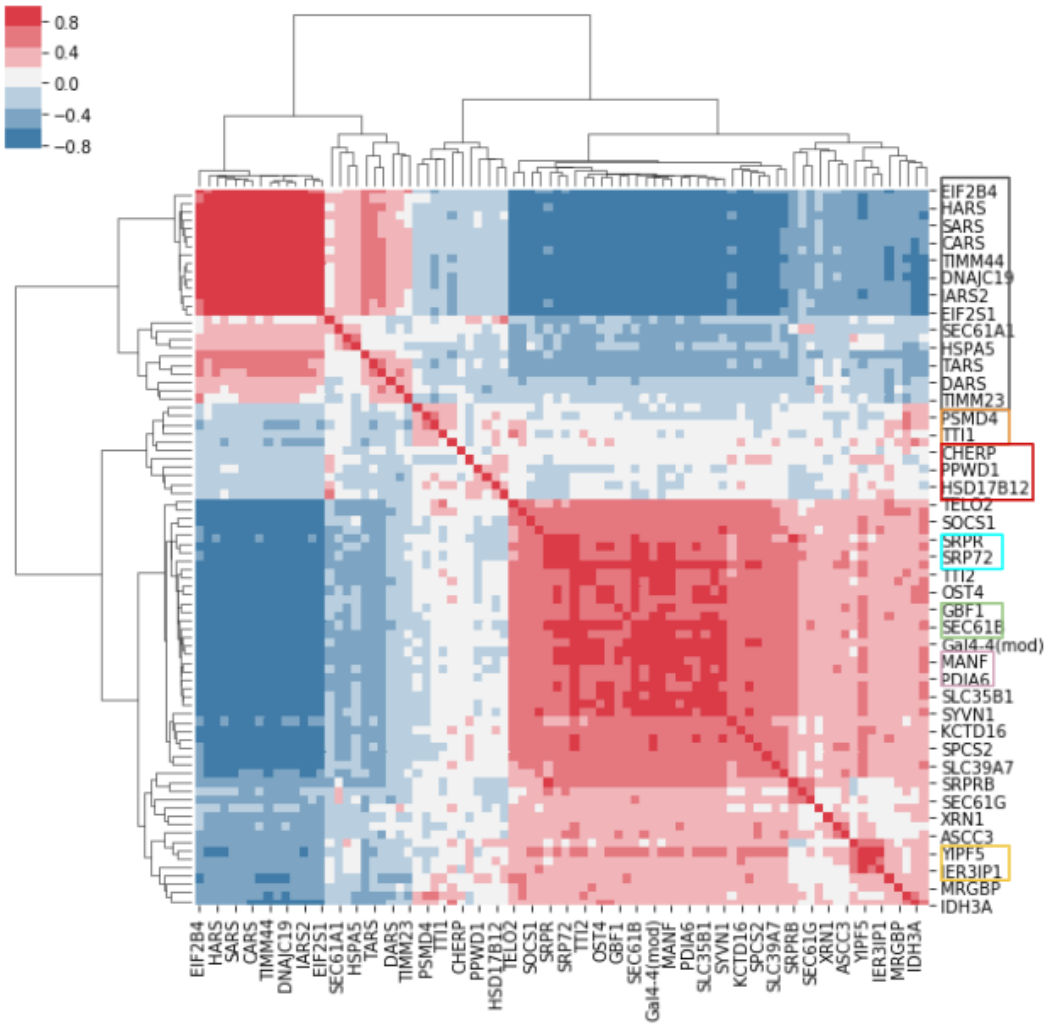

Figure 12: Pearson Distance (Random Subset of Genes)

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 sythetase, 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 Conclusion

4.1 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". 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.

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 subpopulation 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 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.

5 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.

6 Appendix

6.1 Paper Figure S3 B

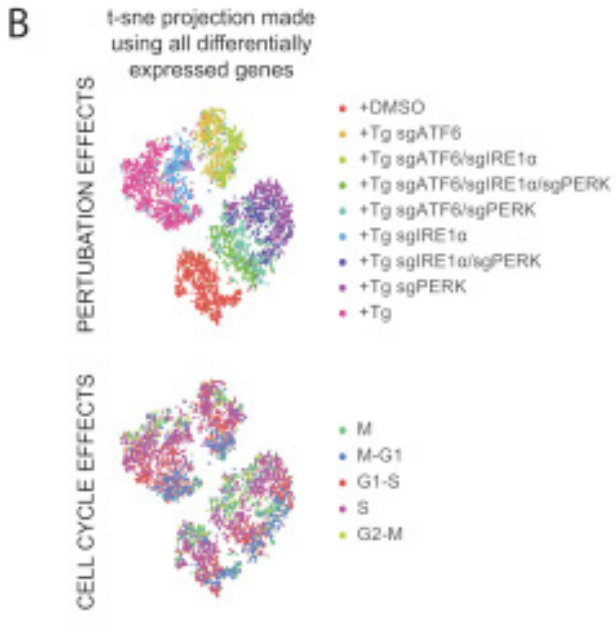

Figure 13: Analysis of thapsigargin-treated cells (t-SNE projections)

6.2 Paper Figure 5A

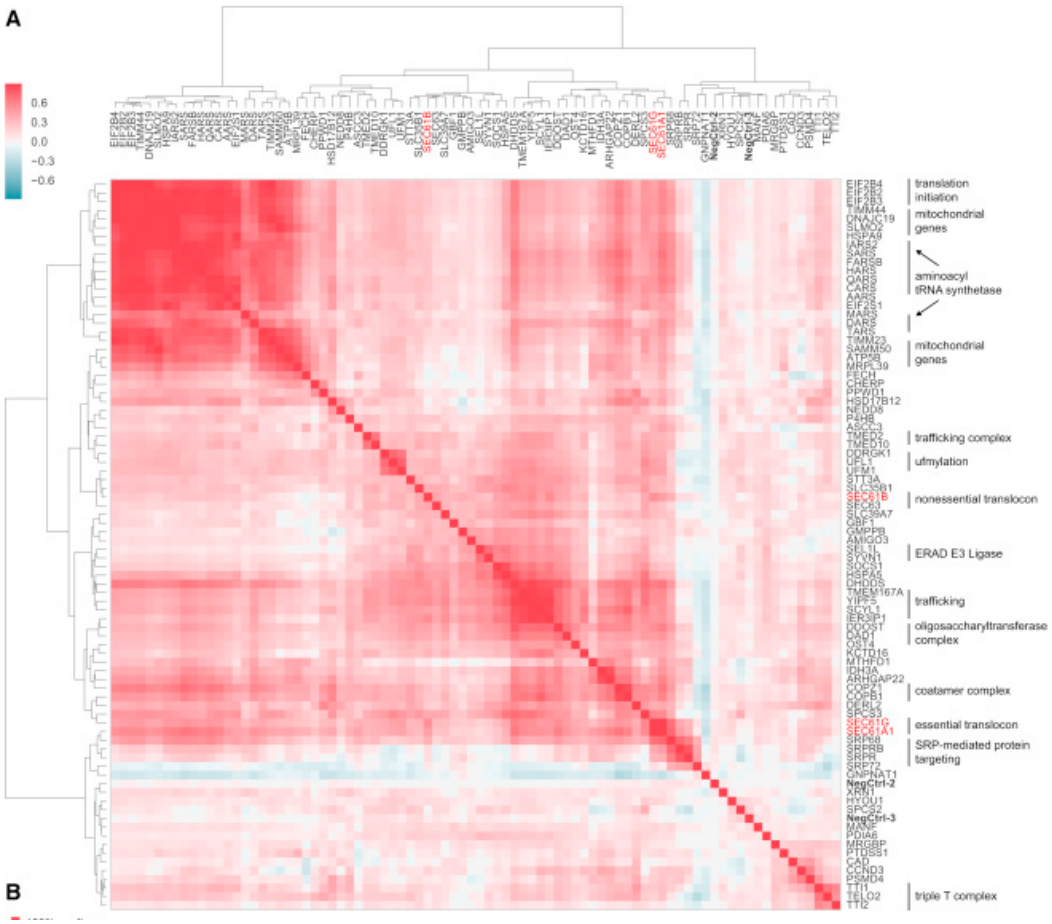

Figure 14: Clustering of genes from UPR Perturb-seq experiment. Heatmap displays correlations between hierarchically clustered average expression profiles from all cells bearing sgRNAs targeting the same gene (identified by GBCs). Functional annotations are indicated.