PERSONAL NOTES

Single Cell Biology
- ML for PerturbSeq (Reva Chauhan)

Zebrafish embryogenesis model system

https://www.youtube.com/watch?v=4Um4LhWn_1E&list=PL1s-xR5JK306MetAfTHwPPMk_bVD97bZr&index=7
In the Zebrafish embryogenesis model system, scRNA-seq has been used to uncover the dynamics of gene expression during different stages of zebrafish development.

1. Dimensionality Reduction: To visualize and explore the high-dimensional data dimensionality reduction techniques such as Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE).
2. Clustering Analysis: Unsupervised clustering algorithms were applied to group cells with similar gene expression profiles into distinct clusters. Each cluster represented a unique cell type or developmental stage.
3. Pseudotime Analysis: Pseudotime analysis or trajectory inference was used to order the cells along developmental trajectories. This method estimates the developmental progression based on gene expression changes, revealing the temporal dynamics of cell fate decisions during embryogenesis.
4. Differential Gene Expression: The researchers identified genes that were differentially expressed between different developmental stages or along the pseudo time trajectory. These differentially expressed genes were associated with specific developmental events or cell fate decisions.
5. Gene Regulatory Network Inference: Using the scRNA-seq data, the researchers inferred gene regulatory networks that govern cell fate decisions during embryogenesis. They identified key transcription factors and signaling pathways involved in the regulation of gene expression.

The study uncovered the dynamic gene expression patterns and temporal progression of cell types during zebrafish embryogenesis.
It revealed distinct cell populations and transcriptional changes associated with various developmental stages.
lineage-specific gene expression programs and regulatory networks driving cell fate decisions were identified.
This allows us to characterize cell populations, identify key regulatory events, and gain a deeper understanding of the cellular processes that shape early development.

GENERAL NOTES

PerturbSeq is a method used to understand how changing genes or other parts of our genetic instructions affect how cells behave. It's like conducting experiments on a large group of cells to see what happens when we make specific genetic changes.

(Perturb-seq accurately identifies individual gene targets, gene signatures, and cell states affected by individual perturbations and their genetic interactions.)

EXPLAINED:

Let's imagine you have a recipe for baking cookies. Each ingredient in the recipe represents a gene or a regulatory element in our genetic code. Now, if you want to know how changing one ingredient affects the taste or texture of the cookies, you could systematically modify that ingredient while keeping everything else the same. For example, you might increase or decrease the amount of sugar in the recipe.

After making these changes, you would then bake different batches of cookies using the modified recipe. Once the cookies are ready, you can observe and measure how they turn out — maybe some are sweeter, while others are less sweet, and so on. These observations represent the phenotypic changes resulting from the genetic perturbations.

In a PerturbSeq experiment, instead of baking cookies, scientists work with cells. They modify specific genes or regulatory elements in the cells and then examine the resulting changes in how the cells function. This could involve studying various cellular characteristics like growth rate, response to stimuli, or production of specific molecules.

By performing these systematic genetic changes and measuring the resulting cellular behaviors, scientists can gain insights into how different genes or regulatory elements contribute to cell function and behavior. This knowledge can help us better understand diseases, identify potential therapeutic targets, or uncover new biological mechanisms.

Overall, PerturbSeq is a powerful tool that allows researchers to explore the impact of genetic modifications on cellular behavior, much like changing ingredients in a recipe to see how it affects the final product.

DETAILS:

Perturb-seq accurately identifies individual gene targets, gene signatures, and cell states affected by individual perturbations and their genetic interactions. We posit new functions for regulators of differentiation, the anti-viral response, and mitochondrial function during immune activation. By decomposing many high-content measurements into the effects of perturbations, their interactions, and diverse cell metadata, Perturb-seq dramatically increases the scope of pooled genomic assays.

Once the perturbations are made, scientists carefully observe and measure how the cells respond. They look at various characteristics of the cells, such as how they grow, how they respond to viruses, and how their mitochondria function.

By studying many different cells with various perturbations, scientists can identify which genes or genetic elements are responsible for specific changes in the cells. They can figure out which genes are important for cell differentiation (when cells become specialized for specific tasks), the body's response to viruses, and the functioning of mitochondria during immune activation (when the immune system is activated to fight off infections).

Perturb-seq also helps scientists understand how different genes and genetic elements work together. They can uncover genetic interactions, which means that the effects of one gene may depend on the presence or absence of another gene. It's like how certain ingredients in a recipe might interact with each other to create a particular flavor or texture.

To make sense of all this information, scientists break it down into smaller parts. They look at individual gene targets, which are specific genes affected by the perturbations. They also identify gene signatures, which are groups of genes that work together and show similar changes. Finally, they analyze cell states, which are different characteristics or behaviors of the cells.

By combining all these analyses and considering other information about the cells, like their characteristics or metadata, Perturb-seq allows scientists to study a wide range of genetic changes in a single experiment. It gives them a bigger picture of how genes and genetic elements affect cells, and it helps them discover new functions and understand complex interactions.

In simple terms, Perturb-seq is a technique that helps scientists understand how changes in genes affect cells. It allows them to study many cells at once and figure out which genes are important for different cell behaviors. It also helps them uncover how genes work together and discover new functions.

THE PROCESS:

1. Transcriptional effects: Cells have a process called transcription, where genes are used as templates to create molecules called RNA. These RNA molecules play important roles in the cell's functions. Perturb-seq helps scientists understand how changes in genes affect this transcription process.
Library preparation: This step involves converting the RNA of individual cells into DNA, amplifying it, and adding specific barcodes that uniquely identify each cell's RNA.

2. Genetic perturbations: Scientists use a powerful genetic tool called CRISPR to make specific changes to genes in cells. It's like editing the instructions of the cells' recipe book (genes) to see how it affects their behavior.
Perturbation: In this step, researchers introduce specific perturbations into the cells, such as genetic mutations, small molecule treatments, or gene knockdowns using techniques like CRISPR-Cas9.

Cell pooling: After the perturbation step, cells are pooled together, typically in a 96-well plate format.

3. Single-cell RNA-sequencing: Scientists then examine the RNA molecules in each cell individually. It's like studying the specific recipes each cell is using and how they're following those recipes.

4. Phenotypes: Phenotypes refer to the observable characteristics or behaviors of cells. Perturb-seq allows scientists to link the changes in genes (due to genetic perturbations) with the effects on cell behavior. They can see how changes in genes impact various complex cellular traits.
By examining the RNA profiles, scientists can see which genes are turned on or off in the cells with genetic changes. They can identify patterns and differences in gene expression that are associated with the differentiation process.

Additionally, Perturb-seq can be applied to study other cellular phenomena like aneuploidy (abnormal number of chromosomes), RNA processing (how RNA molecules are modified and regulated), and stress-related changes in the mitochondrial genome (the DNA inside the cell's energy-producing structures).

In simpler terms, Perturb-seq is a method that helps scientists understand how changes in genes affect the behavior of cells. They use a genetic tool called CRISPR to make specific changes in genes and then analyze the RNA molecules in each cell individually. By doing this, they can link the genetic changes to the observable traits or behaviors of the cells and gain insights into various complex processes in the cell, such as cell differentiation or how cells respond to stress.

TO DISCUSS

Doubts

Some questions I've come across:

1. How can data integration and normalization be effectively performed to account for technical variations and batch effects in perturb-seq datasets?
2. What is the optimal number of cells required to achieve sufficient statistical power for detecting subtle perturbation effects and ensuring result reproducibility through replication?
3. How can computational methods effectively account for cell type heterogeneity within perturb-seq datasets to focus on relevant cell subpopulations and avoid confounding effects?
4. How can the accuracy and reliability of computational predictions regarding gene functions and regulatory interactions be validated experimentally?
5. How can computational analysis capture and account for long-term or delayed effects of perturbations, and what methods are best suited for understanding the dynamic cellular responses observed in perturb-seq experiments?

PROJECT: Perturb-seq - Understanding Transcriptional Complexity in Dendritic Cells using Single-Cell Genomics and CRISPR-based Perturbations

Project Summary:
The project aims to investigate gene functions in mammalian cells, particularly focusing on dendritic cells' response to lipopolysaccharide (LPS). To achieve this, the study utilizes Perturb-seq which combines single-cell RNA sequencing (scRNA-seq) and CRISPR-based perturbations.

- So like immune cells & cells that activate them.

Aim of the Project:
The aim of the project is to uncover the functional roles of genes in mammalian cells, specifically in the context of dendritic cell responses to LPS, using the Perturb-seq approach. Through the integration of scRNA-seq and CRISPR data, the project seeks to provide a comprehensive understanding of immune activation processes and potential therapeutic targets.

Coding Aspect and Benefits:
The computational steps I will implement are crucial for uncovering functional insights from vast amounts of single-cell genomic data, enabling researchers to analyze and visualize gene expression landscapes, identify marker genes, and gain valuable knowledge about immune activation and cellular functions.

- We can collect the data, see how certain cells are disturbed, identify clusters and cell types, create visuals

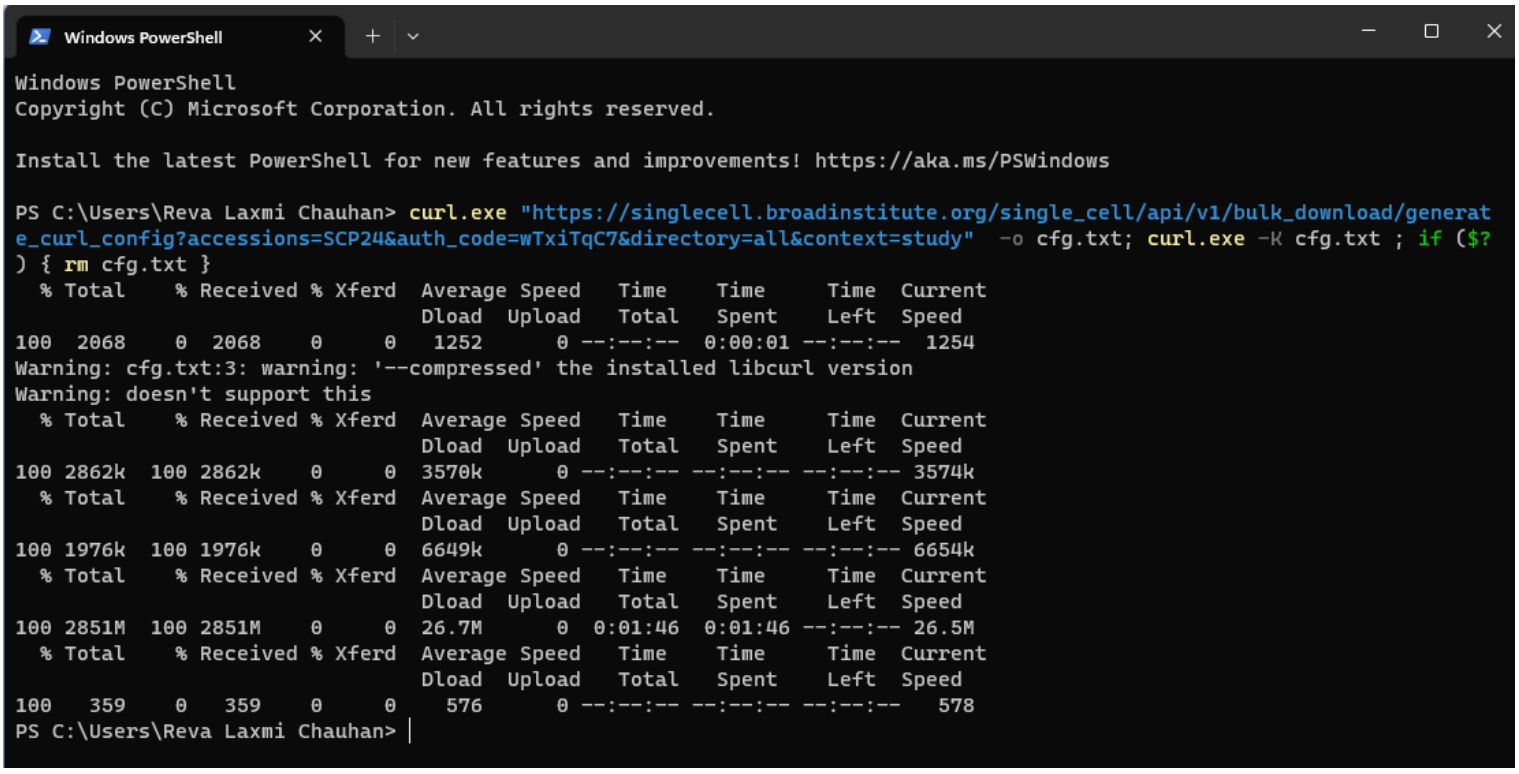
Conclusion:
By focusing on dendritic cells' response to LPS, the study offers novel insights into gene regulatory networks, differentiation, antiviral responses, and mitochondrial functions during immune activation.

Special Terms and Definitions:
1. Perturb-seq: A technique that combines single-cell RNA sequencing (scRNA-seq) with CRISPR-based perturbations to study gene function and cellular responses at a large scale.
2. Transcriptional Profiles: The collection of all the genes expressed and their respective expression levels in a given cell or tissue type, providing valuable insights into cellular functions.
3. CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats, a powerful gene editing tool used to introduce specific changes in the genome.
4. Dendritic Cells (DCs): Immune cells that play a crucial role in recognizing and presenting antigens to activate the immune response.
5. Lipopolysaccharide (LPS): A component found in the cell walls of certain bacteria, used in experiments to stimulate the immune response.
6. Gene Regulatory Networks: Complex interactions between genes and their products that control gene expression and cellular processes.
7. Dimensionality Reduction Techniques: Computational methods are used to reduce the number of variables in high-dimensional datasets, aiding in data visualization and analysis.
8. UMAP (Uniform Manifold Approximation and Projection): A dimensionality reduction algorithm used to visualize high-dimensional data in lower-dimensional spaces.

Project documentation

Downloaded sources from: https://singlecell.broadinstitute.org/single_cell/study/SCP24/perturb-seq

Downloaded it this way, but it is not that exactly.



‘‘Genetic screens help infer gene function in mammalian cells, but it has remained difficult to assay complex phenotypes—such as transcriptional profiles— at scale. Here, we develop Perturb-seq, combining single-cell RNA sequencing (RNA-seq) and clustered regularly interspaced short palindromic repeats (CRISPR)-based perturbations to perform many such assays in a pool. We demonstrate Perturb-seq by analyzing 200,000 cells in immune cells and cell lines, focusing on transcription factors regulating the response of dendritic cells to lipopolysaccharide (LPS). Perturb-seq accurately identifies individual gene targets, gene signatures, and cell states affected by individual perturbations and their genetic interactions. We posit new functions for regulators of differentiation, the anti-viral response, and mitochondrial function during immune activation. By decomposing many high content measurements into the effects of perturbations, their interactions, and diverse cell metadata, Perturb-seq dramatically increases the scope of pooled genomic assays.’’

1. Genetic screens are used to understand gene function in mammalian cells.
2. It's difficult to study complex phenotypes, like transcriptional profiles, on a large scale.
3. The researchers propose a new method called "Perturb-seq."
4. Perturb-seq combines two techniques: single-cell RNA sequencing (RNA-seq) and CRISPR-based perturbations.
5. Perturb-seq allows researchers to analyze complex phenotypes in a large number of cells.
6. The researchers demonstrate Perturb-seq's effectiveness by studying around 200,000 cells from immune cells and cell lines.
7. The study focuses on transcription factors that regulate dendritic cells' response to lipopolysaccharide (LPS).
8. Perturb-seq accurately identifies individual gene targets, gene signatures, and cell states affected by specific perturbations and genetic interactions.
9. The study uncovers new functions for regulators of differentiation, the anti-viral response, and mitochondrial function during immune activation.

10. Perturb-seq breaks down various high-content measurements into the effects of perturbations, their interactions, and diverse cell metadata.
11. By using Perturb-seq, researchers significantly increase the scope of pooled genomic assays.

Documentation of the project.

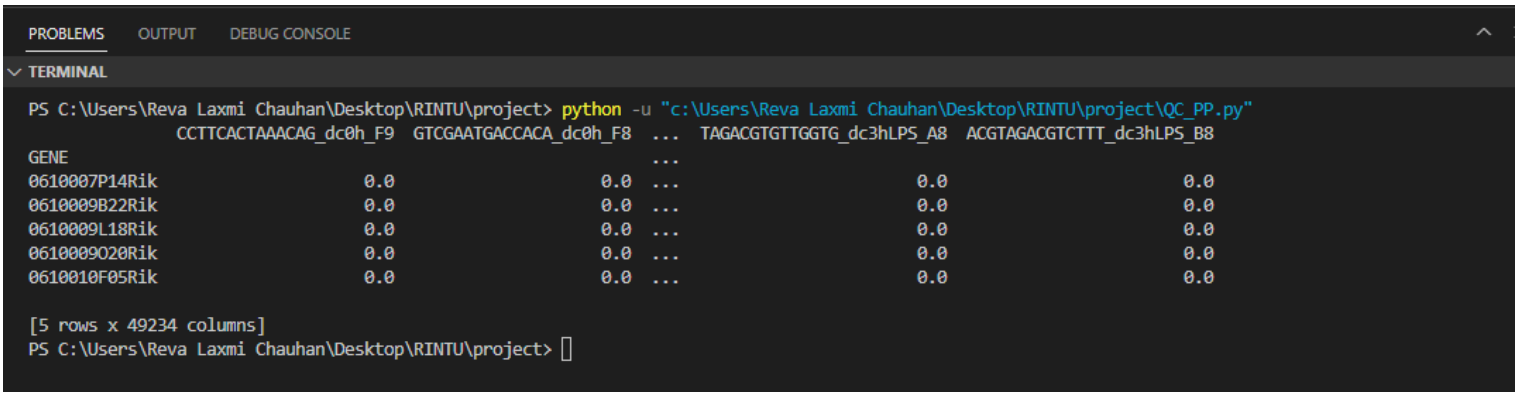
The files I downloaded
SCP24 which has
cluster -> pc_coords_fix.txt
expression -> dc_both_filt_fix_tp10k.txt
metadata -> cluster_assignment_bmdc.txt
file_supplemental_info.tsv

1. - Took some time to explore the contents of each file to understand their format and contents. Read accompanying documentation or data descriptions provided by the authors to get a clear idea of what each file contains.

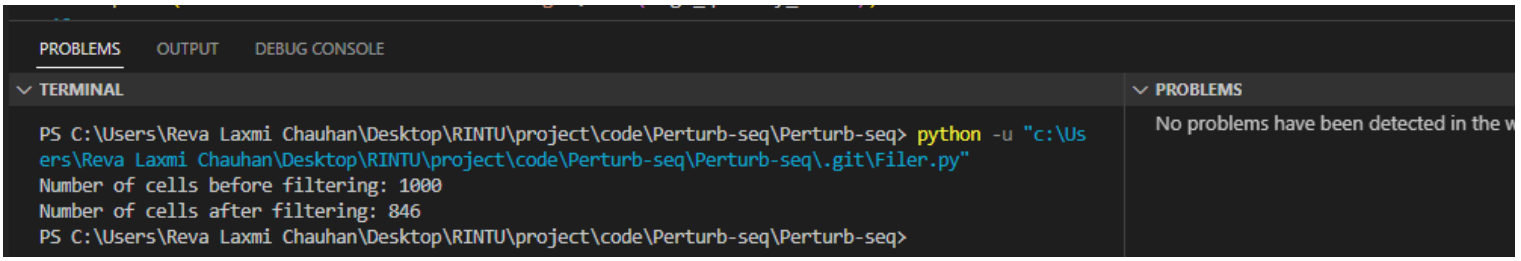
2. - Perform quality control and pre-processing on both the scRNA-seq and CRISPR data. This may involve filtering out low-quality cells, removing unwanted genes, and normalizing the data.

Performing quality control and pre-processing on scRNA-seq and CRISPR data is an essential step to ensure the reliability and accuracy of our downstream analysis. We have done the following according to a general outline of how to proceed with these tasks.

- Load the scRNA-seq data from the "dc_both_filt_fix_tp10k.txt" file into our analysis environment. This data should contain gene expression values for each cell.



- Calculate metrics for each cell, such as the total number of expressed genes and the total count of unique molecular identifiers (UMIs). Remove cells with extremely low counts, as they may represent dead cells or technical artifacts.



```
PROBLEMS  OUTPUT  DEBUG CONSOLE
TERMINAL

PS C:\Users\Reva Laxmi Chauhan\Desktop\RINTU\project\code\Perturb-seq\Perturb-seq> python -u "c:\Users\Reva Laxmi Chauhan\Desktop\RINTU\project\code\Perturb-se
Original data:
             CCTTCACTAAACAG_dc0h_F9  ...  ACGTAGACGTCTTT_dc3hLPS_B8

[5 rows x 49234 columns]
Subsampled data with metrics after filtering:
             CCTTCACTAAACAG_dc0h_F9  GTCGAATGACCACA_dc0h_F8  ...  TotalExpressedGenes      TotalUMIs
GENE                                                        ...
RP23-81F19.2                0.000000                0.000000  ...                  765    2142.644392
Myl12a                      2.797507                2.210528  ...                39291  161282.239123
Tpk1                        0.000000                0.000000  ...                  361    1001.161411
Cst3                        5.198270                7.294834  ...                48492  311182.510767
Aga                         0.000000                2.210528  ...                 6693   17893.030483

[5 rows x 49236 columns]
PS C:\Users\Reva Laxmi Chauhan\Desktop\RINTU\project\code\Perturb-seq\Perturb-seq>
```

- Calculate the metrics anf filter step..

```
PS C:\Users\Reva Laxmi Chauhan\Desktop\RINTU\project\code\Perturb-seq\Perturb-seq> python -u "c:\Users\Reva Laxmi Chauhan\Desktop\RINTU\project\code\Perturb-seq\Perturb-seq\.git\CalculateMetrics_and_Filter
.py"
Original data:
             CCTTCACTAAACAG_dc0h_F9  GTCGAATGACCACA_dc0h_F8  CAGTTGGACAGATC_dc0h_F9  ...  GCTCAAGAAGAAGT_dc3hLPS_A8  TAGACGTGTTGGTG_dc3hLPS_A8  ACGTAGACGTCTTT_dc3hLPS_B8
GENE                                                                               ...
0610007P14Rik                      0.0                     0.0                     0.0  ...                        0.0                        0.0                        0.0
0610009B22Rik                      0.0                     0.0                     0.0  ...                        0.0                        0.0                        0.0
0610009L18Rik                      0.0                     0.0                     0.0  ...                        0.0                        0.0                        0.0
0610009O20Rik                      0.0                     0.0                     0.0  ...                        0.0                        0.0                        0.0
0610010F05Rik                      0.0                     0.0                     0.0  ...                        0.0                        0.0                        0.0

[5 rows x 49234 columns]
Number of cells before filtering: 11738
Number of cells after filtering: 11738
Data with metrics after filtering:
             CCTTCACTAAACAG_dc0h_F9  GTCGAATGACCACA_dc0h_F8  CAGTTGGACAGATC_dc0h_F9  GGACCCGATTCTCA_dc0h_H8  ...  TAGACGTGTTGGTG_dc3hLPS_A8  ACGTAGACGTCTTT_dc3hLPS_B8  TotalExpressedGenes      TotalUMIs
GENE                                                                                                       ...
0610007P14Rik                0.000000                0.000000                0.000000                0.000000  ...                   0.000000                   0.000000                 9751   27988.805421
0610009B22Rik                0.000000                0.000000                0.000000                0.000000  ...                   0.000000                   0.000000                 4086   11823.816344
0610009L18Rik                0.000000                0.000000                0.000000                0.000000  ...                   0.000000                   0.000000                  905    2693.373403
0610009O20Rik                0.000000                0.000000                0.000000                0.000000  ...                   0.000000                   0.000000                  525    1530.063480
0610010F05Rik                0.000000                0.000000                0.000000                0.000000  ...                   0.000000                   0.000000                 3057    8397.315256
...                               ...                     ...                     ...                     ...  ...                        ...                        ...                  ...          ...
mt-Nd1                       6.054766                4.258687                3.913100                0.000000  ...                   3.133174                   4.614814                46063  223831.130031
mt-Nd2                       2.797507                3.045599                2.175101                3.881299  ...                   2.038946                   2.297767                36675  135471.454923
mt-Nd4                       5.919431                4.722418                3.529636                2.976021  ...                   4.048521                   2.297767                44945  209366.083985
mt-Nd5                       2.797507                0.000000                0.000000                0.000000  ...                   0.000000                   3.143054                14648   41792.904813
mt-Nd6                       0.000000                0.000000                2.175101                0.000000  ...                   0.000000                   0.000000                 4101   10858.634031
```

Here we have

1. OG Data: This part displays the first few rows of the original scRNA-seq data loaded from the file.
2. Number of Cells Before Filtering: This line of output tells us the total number of cells present in the original dataset before any filtering is applied.
3. Number of Cells After Filtering: This line of output tells us the total number of cells remaining in the dataset after applying the filtering step to remove cells with extremely low counts. This number represents the count of "high-quality" cells that meet the criteria defined by the filtering thresholds.
4. Data with Metrics After Filtering: This part displays the subsampled data (or the entire data if not subsampled) after applying the filtering step. It shows the data with two additional columns: 'TotalExpressedGenes' and 'TotalUMIs'. These columns represent the calculated metrics for each cell in the dataset.
   - `TotalExpressedGenes`: For each cell, it indicates the total number of genes expressed (genes with non-zero expression values). A higher value suggests a more active cell with a diverse gene expression profile.
   - `TotalUMIs`: For each cell, it indicates the total count of Unique Molecular Identifiers (UMIs) across all genes. UMIs are used to distinguish individual RNA molecules. A higher value suggests a higher complexity of RNA molecules, which can reflect the transcriptional activity of the cell.
5. The "Data with Metrics After Filtering" section shows the filtered dataset, where only cells meeting the specified thresholds for 'TotalExpressedGenes' and 'TotalUMIs' are retained. These are considered "high-quality" cells without extremely low counts, and they will be used for further analysis and downstream applications.
6. The filtering step ensures that cells with very low gene expression and UMI counts, which may represent dead cells or technical artifacts, are removed from the analysis to improve the quality of downstream analyses. The remaining "high-quality" cells are more likely to represent viable and biologically relevant cells in the scRNA-seq dataset.

3. Identifying Cell Clusters:

- Use the "cluster_assignment_bmdc.txt" file to identify the cell clusters from the scRNA-seq data. This file likely contains information about the cluster assignment for each cell.

```
Original scRNA-seq data:
             CCTTCACTAAACAG_dc0h_F9  ...  ACGTAGACGTCTTT_dc3hLPS_B8
GENE                                 ...
0610007P14Rik                   0.0  ...                        0.0
0610009B22Rik                   0.0  ...                        0.0
0610009L18Rik                   0.0  ...                        0.0
0610009O20Rik                   0.0  ...                        0.0
0610010F05Rik                   0.0  ...                        0.0

[5 rows x 49234 columns]
Cluster assignment data:
                    NAME CLUSTER    SUB-CLUSTER
0                   TYPE   group          group
1  CCTTCACTAAACAG_dc0h_F9     0hr  0hr_m_Nfkb1_3
2  GTCGAATGACCACA_dc0h_F8     0hr      0hr_multi
3  CAGTTGGACAGATC_dc0h_F9     0hr    0hr_m_Rel_2
4  GGACCCGATTCTCA_dc0h_H8     0hr      0hr_multi
```

```
Merged data:
Empty DataFrame
Columns: [CCTTCACTAAACAG_dc0h_F9, GTCGAATGACCACA_dc0h_F8, CAGTTGGACAGATC_dc0h_F9, GGACCCGATTCTCA_dc0h_H8, GTTGAGTGCATACG_dc0h_F9, TATCCAACTCTTCA_dc0h_F8, AGAAGATGCCCACT_dc0h_E8, CCCTACGAATGCCA_dc0h_F8, TAGGTTCTCTCTCG_dc0h_G8, CTCTAATGGAATAG_dc0h_G8, GTTAACCTAACCGT_dc0h_F9, AACTCGGACGGGAA_dc0h_E8, TCGCCATGATCGAC_dc0h_F8, GTTATAGAGAATCC_dc0h_H9, CCAATTTGTTCCAT_dc0h_G9, CTATAAGAGCCATA_dc0h_F8, GCACCACTCCCTTG_dc0h_G9, CGGTACCTCCCGTT_dc0h_G9, GGACGAGACATGGT_dc0h_F9, TCGACGCTCTGATG_dc0h_H8, GAGTACTGTTCTCA_dc0h_F9, CCCTGAACTTGTGG_dc0h_G8, GCGCATCTAAGATG_dc0h_F8, GCCATCACAAACGA_dc0h_G8, CTCCGAACGCATCA_dc0h_G9, TCAGCGCTAGGCGA_dc0h_F9, AACGTTCTGGAAAT_dc0h_H8, CCGGTACTGCAGAG_dc0h_H9, TACAATGAACGACT_dc0h_F8, TTCTTACTTTCCAT_dc0h_H8, GTGTGATGCGAATC_dc0h_F9, ATGCGATGAGCCAT_dc0h_H8, ACGTTTACCATTTC_dc0h_G8, GGAAGGTGTAAGCC_dc0h_H8, TGGCACCTCATTTC_dc0h_E8, GGCGCATGTCGTGA_dc0h_F9, ACGAACACAGAAGT_dc0h_F9, ATGCTTTGCTTGAG_dc0h_E8, GTAATAACCTAGCA_dc0h_E8, TATGCGGAAGAGTA_dc0h_F9, ACAATAACATGCCA_dc0h_H8, TCAGGATGAGCAAA_dc0h_E8, GACTGTGACTTGGA_dc0h_G8, ATCTTTCTCTTCCG_dc0h_G8, TAACACCTATCGAC_dc0h_G9, TCAGGATGATACCG_dc0h_G9, CGAGCGTGGCGGAA_dc0h_G8, CAGCTCACTAGTCG_dc0h_E8, GGATGTACTTACTC_dc0h_H8, GACGATTGTACAGC_dc0h_F9, GGAGGATGAAGGCG_dc0h_F9, AAGTGCACTGTGCA_dc0h_F8, TGTAATGATCGATG_dc0h_G9, GCAAGACTACCCAA_dc0h_H9, GTATCTACTGTGCA_dc0h_H9, ACGCTGCTGGACAG_dc0h_F9, CATCAGGAATGTCG_dc0h_F9, TGGACTGATGGATC_dc0h_H8, AGAAGATGTCGTAG_dc0h_G9, TCAATAGACTGTTT_dc0h_H8, TAAGTAACTCCTCG_dc0h_E8, AAAGCCTGGGACAG_dc0h_E8, AGCCAATGCCACAA_dc0h_F8, GACTACGACACTAG_dc0h_F9, AAGTCCGATGGTAC_dc0h_G8, TGGTTACTTCTTAC_dc0h_H8, CGCAGGACGCGAGA_dc0h_F9, AATCCTTGGTCATG_dc0h_H9, ACACCAGATTGGCA_dc0h_F9, TACGATCTCCGTTC_dc0h_F8, CATGGATGCACACA_dc0h_H9, ATAATCGATGCCCT_dc0h_H9, TACCGGCTGTGTAC_dc0h_H9, TAGCGATGGCTTCC_dc0h_E8, ATTACCTGGGACAG_dc0h_F8, CAAGAAGACGTTGA_dc0h_H9, ATCGCGCTCTGAGT_dc0h_G8, AACGCAACTACTGG_dc0h_H8, CTAATGCTGACGTT_dc0h_H9, CGCTAAGAGCCTTC_dc0h_H9, GTCAATCTTTGTCT_dc0h_G9, CAGATCGATGGTCA_dc0h_E8, AACTGTCTGCAGAG_dc0h_H9, GGGTAACTTGTCGA_dc0h_G9, AAGTCCGACGAATC_dc0h_H8, GCGCATCTATCTTC_dc0h_G9, ATTACCACAGACTC_dc0h_F9, CCCGAACTTGTTCT_dc0h_F9, TATTTCCTGAGGTG_dc0h_E8, AGCAAGCTGACACT_dc0h_E8, AACCCAGATCCTCG_dc0h_H9, GTAATATGACTCAG_dc0h_G8, GTACCCTGGATAGA_dc0h_G9, GTTGAGTGCACTAG_dc0h_H9, ACAATAACACCACA_dc0h_F8, GTGACAACACCACA_dc0h_F8, AGACCTGATTCGCC_dc0h_E8, CCGATAGATCTGGA_dc0h_H9, AGCAAGCTCGTAAC_dc0h_G9, GAAGTCTGAAGGGC_dc0h_F8, ...]
Index: []
```

- The cluster assignments are correctly merged with the scRNA-seq data based on the cell IDs. By doing so, the cluster assignment data is now properly associated with each cell in the scRNA-seq dataset, allowing for downstream analyses and exploration of different cell populations and their characteristics.

4. Assigning Perturbations to Cells:

- Link the CRISPR perturbations to the corresponding cells in the scRNA-seq data. This step is crucial for associating perturbation effects with gene expression changes.

```
PS C:\Users\Reva Laxmi Chauhan\Desktop\RINTU\project\code\Perturb-seq\Perturb-seq> python -u "c:\Users\Reva Laxmi Chauhan\Desktop\RINTU\project\code\Perturb-seq\Perturb-seq\.git\Assign_pertubations_to_cells.py"
scRNA-seq data:
             CCTTCACTAAACAG_dc0h_F9  ...  ACGTAGACGTCTTT_dc3hLPS_B8
GENE                                 ...
0610007P14Rik                   0.0  ...                        0.0
0610009B22Rik                   0.0  ...                        0.0
0610009L18Rik                   0.0  ...                        0.0
0610009O20Rik                   0.0  ...                        0.0
0610010F05Rik                   0.0  ...                        0.0

[5 rows x 49234 columns]

CRISPR perturbation data:
                    NAME CLUSTER    SUB-CLUSTER
0                   TYPE   group          group
1  CCTTCACTAAACAG_dc0h_F9     0hr  0hr_m_Nfkb1_3
2  GTCGAATGACCACA_dc0h_F8     0hr      0hr_multi
_dc0h_G8, ATCTTTCTCTTCCG_dc0h_G8, TAACACCTATCGAC_dc0h_G9, TCAGGATGATACCG_dc0h_G9, CGAGCGTGGCGGAA_dc0h_G8, CAGCTCACTAGTCG_dc0h_E8, GGATGTACTTACTC_dc0h_H8, GACGATTGTACAGC_dc0h_F9, GGAGGATGAAGGCG_dc0h_F9, AAGTGCACTGTGCA_dc0h_F8, TGTAATGATCGATG_dc0h_G9, GCAAGACTACCCAA_dc0h_H9, GTATCTACTGTGCA_dc0h_H9, ACGCTGCTGGACAG_dc0h_F9, CATCAGGAATGTCG_dc0h_F9, TGGACTGATGGATC_dc0h_H8, AGAAGATGTCGTAG_dc0h_G9, TCAATAGACTGTTT_dc0h_H8, TAAGTAACTCCTCG_dc0h_E8, AAAGCCTGGGACAG_dc0h_E8, AGCCAATGCCACAA_dc0h_F8, GACTACGACACTAG_dc0h_F9, AAGTCCGATGGTAC_dc0h_G8, TGGTTACTTCTTAC_dc0h_H8, CGCAGGACGCGAGA_dc0h_F9, AATCCTTGGTCATG_dc0h_H9, ACACCAGATTGGCA_dc0h_F9, TACGATCTCCGTTC_dc0h_F8, CATGGATGCACACA_dc0h_H9, ATAATCGATGCCCT_dc0h_H9, TACCGGCTGTGTAC_dc0h_H9, TAGCGATGGCTTCC_dc0h_E8, ATTACCTGGGACAG_dc0h_F8, CAAGAAGACGTTGA_dc0h_H9, ATCGCGCTCTGAGT_dc0h_G8, AACGCAACTACTGG_dc0h_H8, CTAATGCTGACGTT_dc0h_H9, CGCTAAGAGCCTTC_dc0h_H9, GTCAATCTTTGTCT_dc0h_G9, CAGATCGATGGTCA_dc0h_E8, AACTGTCTGCAGAG_dc0h_H9, GGGTAACTTGTCGA_dc0h_G9, AAGTCCGACGAATC_dc0h_H8, GCGCATCTATCTTC_dc0h_G9, ATTACCACAGACTC_dc0h_F9, CCCGAACTTGTTCT_dc0h_F9, TATTTCCTGAGGTG_dc0h_E8, AGCAAGCTGACACT_dc0h_E8, AACCCAGATCCTCG_dc0h_H9, GTAATATGACTCAG_dc0h_G8, GTACCCTGGATAGA_dc0h_G9, GTTGAGTGCACTAG_dc0h_H9, ACAATAACACCACA_dc0h_F8, GTGACAACACCACA_dc0h_F8, AGACCTGATTCGCC_dc0h_E8, CCGATAGATCTGGA_dc0h_H9, AGCAAGCTCGTAAC_dc0h_G9, GAAGTCTGAAGGGC_dc0h_F8, ...]
Index: []

[0 rows x 49237 columns]
PS C:\Users\Reva Laxmi Chauhan\Desktop\RINTU\project\code\Perturb-seq\Perturb-seq>
```

What we are doing here is:

- We are using the `pd.merge()` function to merge these two datasets based on a common identifier column 'NAME'. The merge is performed to link the cluster assignment information to the corresponding cells in the scRNA-seq data.

- The resulting merged_data DataFrame contains both the scRNA-seq information and the cluster assignment information linked together.
- This code specifically focuses on linking the cluster assignment data to the scRNA-seq data, which is crucial for downstream analyses, such as exploring the clusters, calculating cluster-specific statistics, visualizing clusters, or identifying marker genes for each cluster.

INCOMPLETE:

5. Gene Expression Data:
   - The "dc_both_filt_fix_tp10k.txt" file likely contains the expression levels of genes in different cells from dendritic cells (DCs). We will use this as our scRNA-seq input.

6. CRISPR Perturbation Data:
   - The "file_supplemental_info.tsv" or other relevant files may contain information about the CRISPR perturbations. This data will be used to associate specific perturbations with the gene expression changes that we should have previously observed in step 4.

7. Integration of scRNA-seq and CRISPR Data:
   - This step involves associating the gene expression changes observed in scRNA-seq with the specific CRISPR perturbations performed. We may need to perform statistical analysis or correlation tests to identify the effects of each perturbation on gene expression.

8. Visualization and Analysis:
   - Visualize the integrated data using appropriate tools like scatter plots, heatmaps, or dimensionality reduction techniques (e.g., t-SNE or UMAP). This will help us understand how the CRISPR perturbations impact the gene expression landscape in different cell clusters.

9. Interpretation and Conclusion:
   - Based on the integrated analysis, we will interpret the results and draw conclusions about the effects of CRISPR perturbations on gene expression and cellular functions in the context of dendritic cell responses to lipopolysaccharide (LPS).

RESOURCES

https://pubmed.ncbi.nlm.nih.gov/30979687/
https://www.sciencedirect.com/science/article/pii/S1074761319301268
https://singlecell.broadinstitute.org/single_cell/study/SCP739/single-cell-transcriptomics-of-human-and-mouse-lung-cancers-reveals-conserved-myeloid-populations-across-individuals-and-species
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5181115/#:~:text=Perturb%2Dseq%20accurately%20identifies%20individual,mitochondrial%20function%20during%20immune%20activation.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8263256/
http://projects.sanderlab.org/scperturb/

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1926-6
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1218-y
https://www.nature.com/articles/s41576-023-00586-w

How the data is, how its generated (understand the chemistry in order to understand the ML)
https://github.com/asncd/MIMOSCA

This is what they have used for their analysis
https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

---

For the future:
Pertrub seq- understand how the data set works.