Overview of Analysis Methods in scRNA-seq Data and Applications in Identifying Ciliopathy Disease Progression in Neurodevelopmental Disorders

Ariba Huda

University of North Carolina at Chapel Hill

BIOS540H Independent Study Literature Review

Abstract

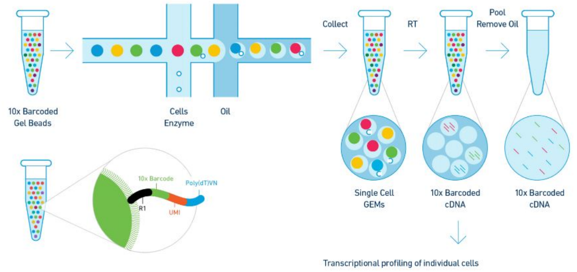
This overview will discuss advances in single cell RNA sequencing (scRNA-seq) technology and in statistical analysis methods used to process data and determine cell type, developmental stage, and gene ontology. ScRNA-seq technology has become the modern approach for understanding the heterogeneity and complexity of RNA transcripts within individual cells, as well as revealing the composition of different cell types and functions within highly organized tissue samples. As more and more scRNA-seq data analysis methods and technologies are being developed, there is an increasing need to determine the most optimal workflow to process highly variable and sparse scRNA-seq data. This review will summarize the workflow of scRNAseq data analysis methods, starting from quality control and preprocessing, determining differential gene expression, dimensionality reduction and clustering analysis, to applications in pseudotime trajectory analysis and determining gene ontology. Lastly, this workflow will be assessed in the lens of ciliopathy disease progression in neurodevelopmental disorders by applying scRNA-seq analysis methods to ciliary gene databases. A ciliopathy is a genetic disorder that affects cilia cellular structures or cilia function. Neurological defects are commonly found in ciliopathies, further highlighting a necessity for primary cilia function in neurodevelopment. Using scRNAseq technology, ciliary genes can be correlated to different brain cell types, brain layer location, and developmental stages. This information is useful in better understanding the cellular mechanisms behind ciliopathy disease progression in different neuron types and ages.

1 Introduction

ScRNA-seq technology measures RNA from individual cells without the need for selective cell purification. These techniques can be summarized by three characteristics: scope (number of cells), granularity (number of genes or epigenetic features), and spatial resolution [3]. scRNAseq has a special application in neuroscience as it helps characterize cellular types and markers of neural circuits. Methods for the physical separation of individual cells can be done using plate-based or droplet-based sorting methods. Droplet-based scRNA-seq is more efficient due to the small reaction volume and ability to rapidly process thousands of cells in microfluidic devices. Droplet-based scRNA-seq methods count the 5' or 3' ends of mRNA molecules, noting unique molecular identifiers (UMIs) to avoid duplicates [6]. In neurodevelopmental research, scRNA-seq has limits as it cannot be applied to frozen post-mortem brain tissues due to the rupturing of cell membranes in frozen temperatures. Thus, an RNA from single nuclei can be sequenced instead (snRNA-seq). Single-cell technology has been able to recognize significantly more morphologically and functionally distinct neuron types compared to traditional microscopy methods. A very useful application of scRNA-seq is lineage tracing through artificial labeling, specifically through pseudotime analysis. scRNA-seq has become a pivotal technology for neuroscience research and how applications in lineage tracing can help determine the disease progression across cell populations.

Due to technical limitations and biological factors, scRNA-seq data are more complex than bulk RNA-seq data, thus

Figure 1: ScRNAseq droplet-based technology workflow from 10x genomics website.



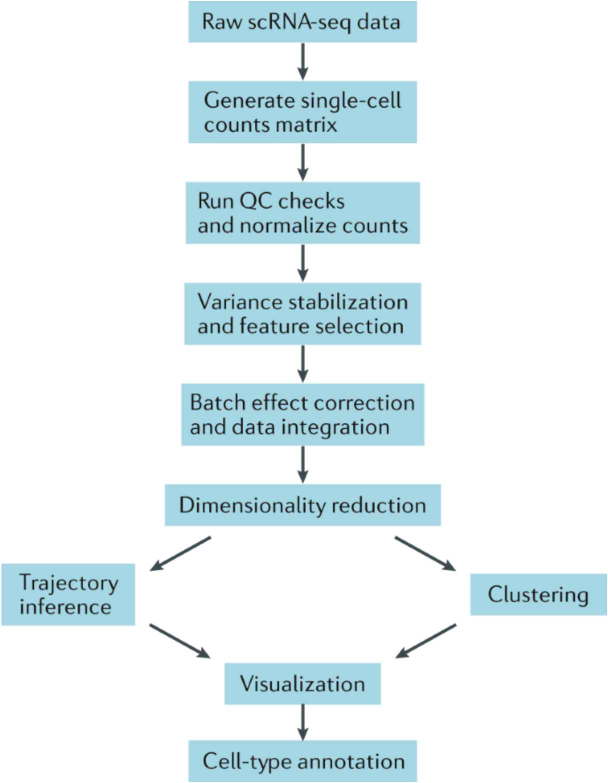
the large variability of scRNA-seq data brings challenges in data analysis. Many different data analysis methods are being created but they must ensure accuracy and reproducibility of results. Attributed to the low amount of starting material, limitations of scRNA-seq include low capture efficiency and high dropout rates. ScRNA-seq produces nosier and more variable data compared to bulk RNA-seq. A variety of tools have been designed to conduct diverse bulk RNA-seq data analyses, but many of those methods cannot be directly applied to scRNA-seq data. After the experimental isolation of single cells, the quality of the data is most importantly indicated by the quantification of read alignment and expression [8]. With technologies such as 10X genomics, read alignment and expression is already quantified and formatted. Following this, a detailed analysis workflow is needed.

2 scRNAseq Data Analysis Methods

2.1 Quality Control + Preprocessing

Quality control is a vital step in scRNA-seq data analysis to ensure the validity of experimental results. In order to make detailed conclusions about cellular heterogeneity related to disease progression, it must be ensured that only high quality cells are included in the analysis. Sequence data is read into a count matrix which is then used to perform quality control on the dataset. The most commonly used metrics to filter out poor quality cells include cells with low numbers of UMIs and detected genes, empty doublets or multiplets (when a droplet either has no single cell or multiple cells), high presence of ambient RNA, and high mitochondrial gene percentage [10]. Following this, the data is normalized so that we can assume all cells have the same number of mRNAs and scaled such that the gene expression across the cell population follows a normal distribution. This ensures that highly-expressed genes do not skew the analysis. Then, principal component analysis (PCA) is done to reduce dimensions and data complexity while preserving as much important information as possible.

Figure 2: Overall outline of scRNAseq data analysis workflow.


Raw scRNA-seq data
Generate single-cell counts matrix
Run QC checks and normalize counts
Variance stabilization and feature selection
Batch effect correction and data integration
Dimensionality reduction
Trajectory inference
Clustering
Visualization
Cell-type annotation

2.2 Determining Differentially Expressed Genes (DEGs)

Various novel statistical and machine learning methods are used to optimize precision in identifying cellular markers using differential expression (DE) analysis of transcriptomic data. These methods are typically done via re-ranking and prioritizing genes post-DE analysis with supervised feature selection methods for selecting optimal genes with the most relevance in predicting target variables while also achieving minimal redundancy [1]. Then machine learning classification was used to assess the discriminatory power of selected genes. DE analysis was conducted by calculating the foldchange with respect to non-survival followed by a Benjamini-Hochberg correction method. Then various machine learning classification techniques can be done using Random Forest, eXtreme Gradient boosting, and/or Logistic Regression. Feature selection uses Random Forest Feature Importance and Minimum Redundancy and Maximum Relevance. Overall marker gene discovery was assessed by measuring Accuracy (ACC), Sensitivity (Sn). Specificity (Sp), Matthews correlation coefficient (MCC) and Area under ROC curve (AUC) [1]. scRNAseq analysis requires a comprehensive overall statistical workflow used to assess the most relevant genes in a sample to determine the most appropriate cellular markers.

2.3 Dimensionality Reduction + Clustering Analysis

In scRNAseq analysis, there is a preference for uniform manifold approximation and projection (UMAP) for non-linear dimensionality reduction due to its high efficiency, reproducibility, and meaningful organization of cell clusters [4]. Nonlinear dimensionality reduction methods are becoming more meaningful for cell type clustering analysis as it is able to avoid overcrowding of representation, in comparison to typical linear methods (PCA). Becht, et al., compares both UMAP and another non-linear method, tSNE, and concludes that UMAP is able to better represent typical multi-branches continuous trajectories of cellular phenotypes and development [4]. UMAP in combination with some sort of community detection algorithm, usually by default the Louvain algorithm, is able to iteratively group similar cells together. Then, the most differentially expressed genes in each cluster can be matched to known cell markers to annotate cell type. This step is crucial in knowing what cell populations are mostly represented in our scRNA samples.

2.4 Trajectory Analysis

One exciting application of scRNA-sequencing in the ability to quantify cell differentiation through trajectory inference (TI). TTI allows researchers to study cellular dynamics, specifically development patterns, from scRNA-seq data. These methods infer a graph-like structure that maps cells to compare properties over pseudotime, an abstract unit of progress during cellular dynamic processes. TI analysis assumes that the biological process of interest is dynamic, data is sampled to sufficient depth (such that even very transient states to be presented) and changes in gene expression are gradual during the developmental process [7]. Various computation approaches can be used, including dimensionality reduction, clustering, graph traversal, probabilistic methods, and RNA velocity to assist in downstream analyses including trajectory visualization, differential expression, alignment, and gene regulatory network (GRN) inference [7]. Due to the high variability in gene expression with a cell, there is a large amount of uncertainty in the preciseness of ordering of cells by differentiation processes. Due to this, Campbell and Yau suggest the implementation of probabilistic modeling techniques to quantify this uncertainty and to include it in the trajectory inference analysis [5].

3 Application to Ciliopathies Disease Mechanisms in Brain

To utilize scRNAseq data in better understanding the cellular mechanisms behind ciliopathies, a list of ciliary gene-specific cellular markers is useful. A database and study created by Arlotta, et al., has done scRNAseq data analysis on diseased mice to offer a comprehensive look at mouse neocortex development and cellular diversity [2]. The study took single-cell RNA samples everyday during embryonic corticogenesis and at early postnatal stages as well as a spatial transcriptomics time course. Using differential gene expression analysis, a diffusion pseudotime-based approach was used to inference cell differentiation trajectories after doing initial scRNA-seq pre-processing, analysis, and clustering. As a result, the reconstructed developmental trajectories assisted in the inference of spatial organization and the gene regulatory programs that influence lineage decisions. This is significantly useful because the developmental map can also pinpoint origins of lineage-specific abnormalities that are linked to diseased mice. Using this database as a source of gene lists and expression levels across development to crosscheck with human homolog ciliary genes will help to determine if any cilia specific gene mutations are more prevalent in certain developmental states or tissue layers. To cross check with ciliary genes, a paper by Inglis, et al, provides a compiled list of identified ciliary proteins from bioinformatic, genomic, and proteomic studies and discusses how cilia organelle functions and related disease mechanisms [9]. This compiled list is very useful for cross checking results of DEG analysis post scRNA-seq analysis and determining if the predictive machine learning methods used provide consistent results. In combination with a similar list from Reiter and Leroux, which offers more gene markers and associated molecular pathways that contribute to ciliopathies [11]. This compiled list will allow for attributing the markers determined significant after DEG analysis and classification to certain ciliopathy mechanisms. The list includes ciliopathies across the human body, so the gene list used in my analysis will include human genes and mouse homologs with the highest percent similarity to human genes. To begin the next steps of the project, data will be extracted from these sources to pursue the most common ciliary genes and what molecular disease pathways they are associated with, and then match the gene list with the specific cell type, brain tissue layer, and developmental stage data to determine if we see any correlations between any variables.

References

[1] Abbas, M., and El-manzalawy, Y. Machine learning based refined differential gene expression analysis of pediatric sepsis. *BMC Medical Genomics 12*, 122 (August 2020).

[2] Arlotta, P., and Aviv Regev, title = Molecular logic of cellular diversification in the mouse cerebral cortex, J. . B. y. . .

[3] Armand, E. J., Li, J., Xie, F., and Luo, C. Single-cell sequencing of brain cell transcriptomes and epigenomes. *Neuron 109*, 1 (Jan 2021), 11–26.

[4] Becht, E., McInnes, L., and Healy, J. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology 37* (December 2019).

[5] Campbell, K., and Yau, C. Order under uncertainty: Robust differential expression analysis using probabilistic models for pseudotime inference. *PLOS Computational Biology* (November 2016).

[6] Chen, G., Ning, B., and Shi, T. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in Genetics 10*, 317 (April 2019).

[7] Deconinck, L., Connodt, R., and Salens, W. Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology* (September 2021).

[8] Gagnon, J., Pi, L., Ryals, M., and Wan, G. Recommendations of scrna-seq differential gene expression analysis based on comprehensive benchmarking. *National Center for Biotechnology Information 12*, 6 (June 2022).

[9] Inglis, P., Boroevich, K., and Leroux, M. Piecing together a ciliome. *Trends in Genetics 22*, 9 (2006).

[10] Jiang, R., Sun, T., Song, D., and Li, J. Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome Biology 23*, 31 (January 2022).

[11] Reiter, J., and Leroux, M. Genes and molecular pathways underpinning ciliopathies. *Nature 18* (2017).