**Single Cell RNA-Sequencing: An Application to 10x Genomics Mice Dataset**

IUPUI School of Informatics and Computing

Professor Juexin Wang: B573 PROGR FOR SCIENCE INFORMATICS

Semester Report: December 8th, 2022

Richard Mustaklem

**Abstract**

One of the largest problems with modern day science and more specifically, computer science, is the knowledge gap that is created between levels of education. From Sequence Search Services to Genome Sequencing, and even Data Collection, the educational demand for bioinformatics can become overwhelming. But within recent years, a new field has sprung to the frontier of bioinformatics. This paper will investigate different ways to interpret Single Cell RNA-sequencing (scRNA-seq) data and just how it can be understood within the context of mice data. One prime example is the Seurat Vignette1. The aim of this paper is to reduce the growing knowledge gap from the experts of scRNA-seq to a beginning scientist as well as investigate highly variable genes within the mice data.

**Introduction**

Within many fields of science, there is a growing knowledge gap between the expert and the novice. A field not too far from science, personalized medicine, has had this issue growing rapidly and globally (Haiech, 2012). Assumingly so, computational biology and data science will most likely follow a similar pattern. With the innovation of new fields of sciences occurring often, beginners like myself find it hard to keep up with the intricacies of bioinformatics and filtering the *junk* from the *gems*. However, many would agree that scRNA-seq is one of the biggest gems within the decade.

But scRNA-seq still has much room to grow before becoming the shining star we all want it to be. As of right now, single cell can “provide important information about fundamental characteristics of gene expression . . . scRNA-seq is also increasingly being used to trace lineage and developmental relationships between heterogeneous, yet related, cellular states in scenarios such as embryonal development, cancer, and myoblast” (Haque, 2017). What’s more, there are a multitude of datasets and databases being produced online that allow entry-level scientists to access and understand just how new pipelines or datasets can be configured within this new field. Contrary to these major advancements, a professor at IUPUI, Juexin Wang, explains that scRNA-seq has a “Big data challenge, sequencing sparsity,” and “Dropouts” (Wang, 2022).

This paper is aiming to investigate these problems and attempt to reduce the knowledge gap from an expert and novice. In doing so, this will improve the field of bioinformatics as well as strengthen the fundamentals of computer science for both sides of the playing field. Additionally, an investigation on the highly variable gene, Hbb-bs, will be investigated throughout the paper. The Hbb-bs gene “encodes a beta polypeptide chain found in adult hemoglobin, which consists of a tetramer of two alpha chains and two beta chains, and which functions in the transport of oxygen to various peripheral tissues” (reference 6). In other words, this gene is responsible for the production of beta-globin.

**Data Set**

The original code used to begin the project was provided by Rahul Satija, an Associate Professor of Biology at NYU, and the publisher of Seurat Vignette. The original code2 comes with a dataset provided by Satija however this project uses a 10xGenomics mice brain dataset .

The data is obtained from 10xGenomics3, a recent database that has arrived at the forefront of scRNA-seq technology. Login to the website, select resources, select datasets, and look for Single Cell Gene Expression on the left-hand side. Then select “Nuclei Isolation for Single Cell Gene Expression” and open the “5k Adult Mouse Brain Nuclei Isolated with Chromium Nuclei Isolation Kit.” Finally, download the Feature / cell matrix HDF5 (filtered) dataset and extract the contents to your computer.

The code produced by this project can be found at https://github.com/richardmustaklem/BIOINF-Semester-Final-Project-12-8.. It is a modified version4 of Satija’s code but for mice data instead of the dataset his code uses (human data).

**Methodology**

To begin this project, the steps of scRNA-seq analysis had to be defined (Luecken, 2019). After defining each step of the pipeline, changes had to be made for the new set of data. An example can be found on line 24 of the code. The “pattern” is important because each species has a different code that is unique for their set of data. More specifically, human data would have a pattern of “MT-” whereas mice data would use a pattern of “mt-”. On top of altering the code to fit to 10xGenomics mice data, there are examples that explore the before and after of computation behind certain sections of code, like normalization. The Seurat tutorial and many others fall short of these explanations. One example is looking at the normalization techniques used in scRNA-seq.

The workflow for this project is as follows. First, quality control and normalization will be used to identify the highly variable genes. Afterwards, the data needs to be scaled and dimensionally reduced. Once the data is reduced, the dimensionality will be determined for clustering. Once clustered, non-linear dimensional reduction (UMAP) will be used, leading to finding the differentially expressed features. Annotation will not be performed in this experiment, as that requires heavy computational power and time.

**Results**

The first set of results stems from the Quality Control section, seen in Figures 1 and 2. From Figure one, we can see this mice data has an abnormal amount of noise, given the percent mitochondria column having a portion of the data hovering 40%. MBD is short for Mice Brain Data. The clean data, shown in Figure 2, is sub-setting only up to 15% of the mitochondria column.

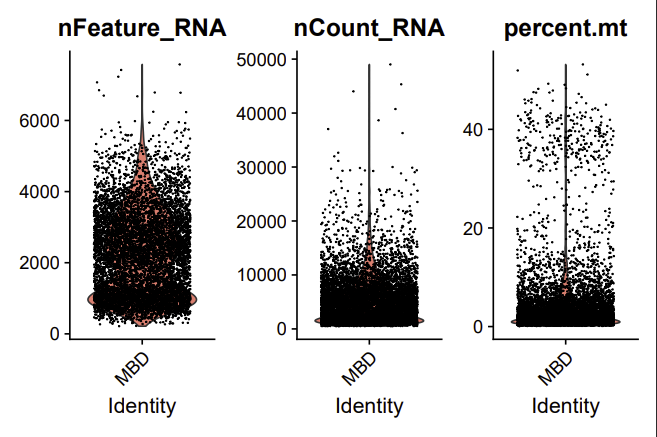
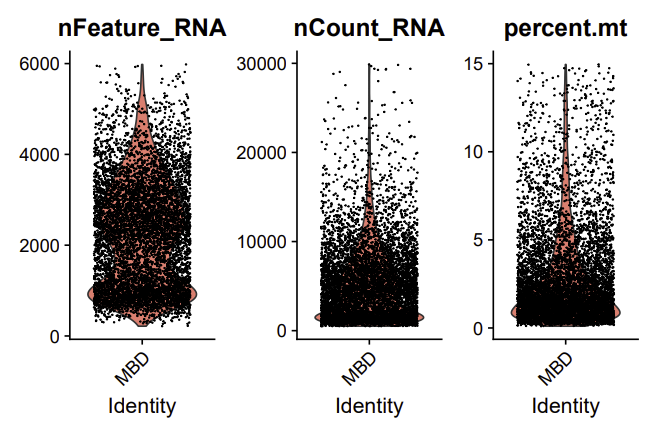
****

Figure 2: Post-Quality Control mice data. The percent.mt column has less noise and the data is more concise and reliable

Figure : Pre-Quality Control mice data. The percent.mt column has plenty of noise, it will need to be cleaned

To visualize normalization, a significant gene was picked from Figure 3 that can show the before and after of normalization in Figures 4 and 5. The gene, Hbb-bs, will be tracked throughout the project to help visualize the processes and steps taken within scRNA-seq and because it is one of the most significant genes in this study, it will be beneficial to the conclusion as well.

**Chart, scatter chart

Description automatically generated**A picture containing chart

Description automatically generated

Figure 3: A plot showing the expression of a gene versus the variability. Typically, scientists will want to investigate highly expressed and highly variable genes

Figure 4: Highly variable gene that will be investigated. Pre-normalization shown

**Diagram, scatter chart

Description automatically generated**

Chart

Description automatically generatedAfter normalization, the data was scaled and had linear dimensional reduction applied to the dataset. To determine the dimensionality of the dataset, an elbow plot was used that is seen in Figure 6. Once the plot starts to level off, the dimensionality can be determined. This dataset has a dimensionality that ranges from 1 to 25 principal components.

Figure 6: An elbow plot showing the dimensionality of the dataset. Can test the first 25 PC for analysis.

Figure 5: Highly variable gene that will be investigated. Post-normalization shown. More information can be gathered from the above plot.

Afterwards, a clustering of the cells was done to see where Hbb-bs groups amongst other expressed genes. This can be seen from Figure 7. Figure 8 pairs nicely, as it will display which Principal Component Hbb-bs is apart of as well.

Chart, scatter chart

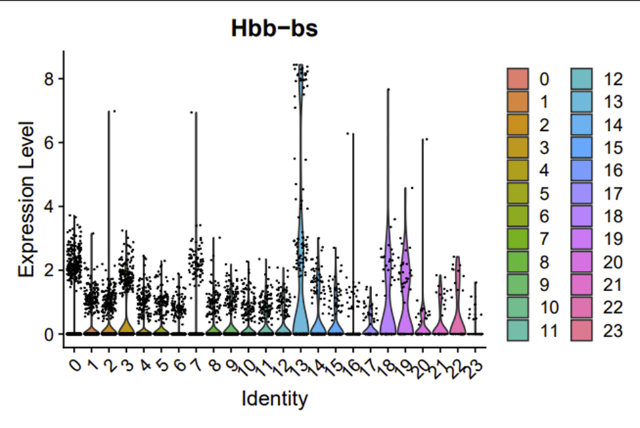
Description automatically generated****

Figure 7: UMAP used to show where Hbb-bs gene is clustered. Is shown in the middle

Chart, scatter chart

Description automatically generated

Figure 8: Hbb-bs expression level shown within each PC. 13 seems to have the most expression compared to other PC, although 18-19 are close

Figure 9 is showing the overall clustering with the dataset and how each PC is related to one another. As we can tell from the figure, Hbb-bs is not related to many other PC, given that it is isolated in the middle and not near other groups. This could mean that Hbb-bs plays a large role in gene expression since it is highly expressed and so variable, yet it is isolated from the other groups of cells telling us that it is an entirely different group of cells that aren’t similar to the others in this dataset.

Figure 9: UMAP showing the relationship between different PC clusters. Hbb-bs in the middle, isolated from the rest of the data set

**Discussion**

Once the results showed the isolation of Hbb-bs gene expression, this could provide insight to what is occurring within the dataset. Based off Figures 7, 8, and 9, Hbb-bs is not heavily related to the rest of the dataset. This is determined due to the clustering of Hbb-bs being isolated from many other clustered groups. Hbb-bs is commonly expressed in red blood cells (Hbb, 2022). Knowing this information, we can determine that PC 13 is likely comprised of red blood cell related gene expression within this 10xGenomics dataset. To further aid annotation, there are computational (machine learning) and manual methods available that will generate annotation for clustering. However, with lack of time or lack of computational power, annotation was skipped for this project. In the future, annotation can be used to verify Hbb-bs gene expression and to strengthen the confidence on the results. Additionally, there are other methods that can be run to achieve similar results in future projects; one being tSNE. Although UMAP is popular for scRNA-seq, many scientists still use tSNE, as both have promising results. The second goal of this project was to decrease the knowledge gap required to run and understand scRNA-seq. This measurement was hard to quantify because the focus of the project was to produce understandable code for both a novice and expert. To achieve this goal in the future, using a survey and asking participants to read, test, and run the code would allow more power and results for this goal.

**Footnotes**

1. <https://satijalab.org/seurat/index.html>
2. <https://satijalab.org/seurat/articles/pbmc3k_tutorial.html>
3. <https://www.10xgenomics.com/resources/datasets/5k-adult-mouse-brain-nuclei-isolated-with-chromium-nuclei-isolation-kit-3-1-standard3>
4. <https://github.com/richardmustaklem/BIOINF-Semester-Final-Project-12-8>

**References**

1. Haiech, Jacques, and Marie-Claude Kilhoffer. “Personalized Medicine and Education: The Challenge.” *Croatian Medical Journal*, U.S. National Library of Medicine, Aug. 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3428815/.
2. Haque, Ashraful, et al. “A Practical Guide to Single-Cell RNA-Sequencing for Biomedical Research and Clinical Applications - Genome Medicine.” *BioMed Central*, BioMed Central, 18 Aug. 2017, genomemedicine.biomedcentral.com/articles/10.1186/s13073- 017-0467
3. Wang, “Graph Neural Network frameworks for single-cell data analyses”, Lecture 3, B627: Advanced Seminar in Biohealth Informatics, IUPUI, Indianapolis, September 22, 2022)
4. “ScRNA-Seq Analysis (Dimensionality Reduction, Clustering, Identifying DE Genes).” *| Notebook.community*, Greene Laboratory, notebook.community/greenelab/GCB535/37- scRNAseq-II/scRNAseq\_inclass\_2.
5. Luecken, Malte. *Current Best Practices in Single‐Cell RNA‐SEQ Analysis: A Tutorial*. 19 June 2019, www.embopress.org/doi/full/10.15252/msb.20188746.
6. “Hbb-BS Hemoglobin, Beta Adult s Chain [Mus Musculus (House Mouse)] - Gene - NCBI.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, 26 Sept. 2022, www.ncbi.nlm.nih.gov/gene/100503605.