Analysis of Single cell /nuclei data for Parkinson disease

Simranjit Kaur Virk

Department of Bioinformatics, IU School of Informatics and Computing at IUPUI, Indianapolis, IN, U.S.A. 46202

Abstract:

Motivation: Single cell sequencing has contributed a lot to our understanding of neurodegenerative diseases like Alzheimer diseases, but very limited progress has been made for Parkinson disease. Using computational and analytical tools we analyze and interpret plethora of rich information derived from Parkinson disease sc/snRNA-seq. We tried to identify conserved markers cell types within cell clusters to identify normal and diseased condition. The motive of this study was to analyze the single cell data for downstream analysis.

Results: We featured 33538 features across 48592 samples and found out that the top 10 gene expression in 27 differentiated cell clusters. Previously studied genetic marker were mapped to annotated cell types to identify specific cells. We were unable to get annotation of all the cell clusters because of limited availability of experiment data.

Contact: skvirk@iu.edu
Supplementary information: https://github.com/SKVirk27/Single-cell-nuclei-analysis.git

1 Introduction

Parkinson's disease (PD) is the common neurodegenerative disorder which was reported to affect over 1% of the population over the age of 60. As the overall incidence and occurrence of PD is increasing, there have been worldwide efforts to fight PD by understanding the disease at the single cell level using advanced technologies. PD has been characterized as a loss of dopaminergic neurons in the substantia nigra pars compacta. It results in motor symptoms such as stiffness, postural uncertainty, shiver at rest, and bradykinesia. Dopaminergic drugs and deep-brain stimulation have been used to alleviate these symptoms. It is important to understand the root cause of this disease to discover better treatments [1].

Next-generation RNA sequencing is a common technique for high-throughput transcriptome analysis. It had revolutionized our understanding of the molecular cause of human disease. Millions of cells or nuclei of a tissue can now be sequenced from a single experiment with the advent of DNA barcode and combinatorial indexing strategy [1].

The marker genes are known as the Cell-type-specific genes or cellular identity genes. It has been shown in experiments that they are highly expressed in one cell type, compared to other cell types. These genes are believed to be a key to the analysis of RNA transcriptional data. It is important to have knowledge of marker genes to fill the critical gaps in our understanding of cell biology and possibly the cellular origins of pathologies [2].

In this Study we have selected the publicly available raw data for Parkinson disease from Geo consortium for the controlled and diseased. We integrated the data and performed normalization, quality control and PCA analysis on the elected dataset. To remove technical noise, we performed batch effect analysis on integrated data. We tried to identify the cell type of the conserved markers in normal and diseased conditions. The availability of cell type specific transcriptome data has been a great challenge for many studies.

2 Methods

Dataset: A publicly available Parkinson dataset GEO (GSE184950) was selected. The selected tissue samples of the Subs Tania Nigra (Figure 1) were taken from the control and idiopathic PD postmortem brain tissues across various disease stages. The Single-nucleus RNA sequencing using the Chromium platform (10x Genomics, Pleasanton, CA) was performed with the Next GEM Single cell 3'GEX Reagent kit, and an input of ~10,000 nuclei from a debris-free suspension. Approximately 50mg of tissues were used for snRNA-sequencing. For our studies we have selected only 10 samples (3 control and 7 diseased) with an average age of 81 out of 32 available datasets.

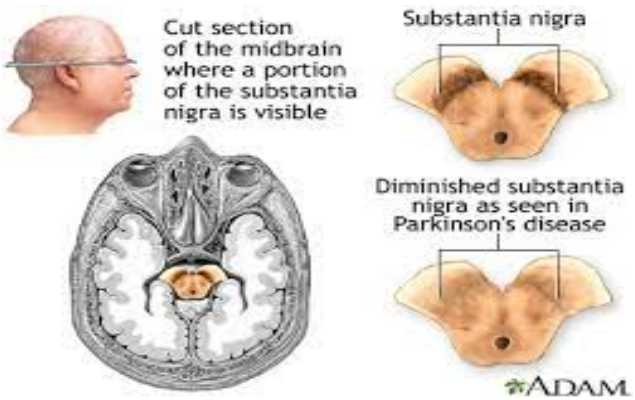

Figure 1 : Pictorial view of substantia nigra

snRNA-seq data preprocessing and pre-clustering analysis –

Starting from integrated the raw data using Seurat Package we get 33538 features across 136864 samples within 1 assay . We performed QC by removing low-quality cells with either too few genes (< 200) or an excessive number (> 800) of genes. Then we removed insufficiently detected cells by keeping 33538 features across 48592 samples within 1 assay. We calculated the Mitochondrial percentage and filtered cell nuclei containing greater than 10 percent mitochondria. We performed standard workflow steps to figure out if we see any batch effects. We performed integration to correct for batch effects(Figure 2). We scaled the data, ran PCA and UMAP analysis to visualize the integrated data. Pre-clustering analysis was also performed. Briefly, the UMIs data was first normalized by sequencing depth and then log-transformed using the Log Normalize method implemented in Seurat. About 2,000 most variable gene features were identified, scaled, and centered after regression out covariates. Next, dimensional reduction was performed using principal component analysis (PCA) based on the 2,000 most variable genes. The top 10 principal components as determined by an elbow approach were selected for integration of the snRNA-seq data across all sequencing libraries. Top 10 embeddings in the Harmony space were used for calculating 2D dimensional reductions by t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP). The same top 20 Harmony embeddings were also used to compute the nearest neighbor graph and the subsequent cell pre-clusters with the Louvain algorithm implemented in Seurat. This initial pre-clustering analysis resulted in 27 pre-clusters at resolution 0.2[3].

Cell annotation -We used SingleR Bioconductor R package to perform unbiased cell type recognition from single-cell RNA sequencing data, by referencing transcriptomic datasets of pure cell types to infer the cell of origin of each single cell independently. The automatic cell annotation was done using library cell dex. The cell dex package provides access to several reference datasets (mostly derived from bulk RNA-seq or microarray data) through dedicated retrieval functions. We have used the Human Primary Cell Atlas , represented as a Summarized Experiment object containing a matrix of log-expression values with sample-level labels [4]

Marker analysis –

We have first selected previously studied cell markers for Parkinson disease to know the cell type they expressed and then we manually found conserved markers of each cluster in our dataset to identify differentially expressed genes.

Cell cluster comparison –

We had compared cell clusters of control and diseased group to understand the difference between two samples to do future analysis for performing our future downstream analysis.

3 Results

The raw data was integrated and after quality control using Seurat object, we have received 33538 features across 48592 samples. The batch effect which occurs when non-biological factors in an experiment cause change in the data produced by the experiment were removed. To get accurate conclusions for further data analysis. In figure 2 it is clearly shown that after correcting batch effect the cluster are well segregated.

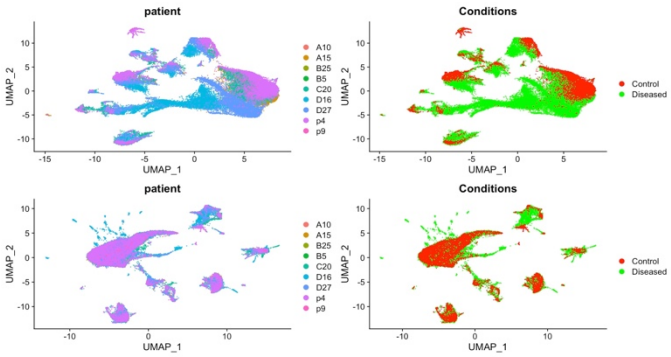

Figure 2 – Visualization of clusters before and after batch effect correction

3.1 **Gene expression -**

When we determine the genes that are most different in their expression between cells. These genes are used to determine which associated genes sets are responsible for the largest differences in expression between cells. The top 10 differentially expressed genes (Figure 3) between all the identified 27 cell clusters include N-acetylated alpha-linked acidic dipeptidase-like 2 (NAALADL2), Apolipoprotein E (APOE), carnosinase (CNDP1), potassium channel tetramerization domain containing 8 (KCTD8), Catenin Alpha 2 (CTNNA2), a Tubulin Isotype (TUBA1A), Protein Tyrosine-Phosphatase Receptor Z1(PTPRZ1), Proto spacer adjacent motif (PAM), RPL13(Ribosomal Protein L3) and Adenosylhomocysteinase Like 1 (AHCYL1). As per the finding from previous studies all the identified genes are associated with neurological disorders Parkinson disease except NAALADL2 which is more closely related to age.

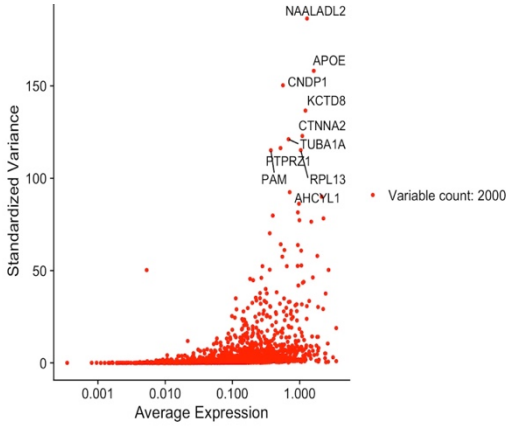

Figure 3 – The Top 10 genes which are differentially expressed in our dataset.

3.2 Cluster Analysis and Annotations –

Analysis of clusters of normal and diseased using Seurat clustering tool we identified many cell clusters which are present in diseased but absent in control samples. (Figure 4)

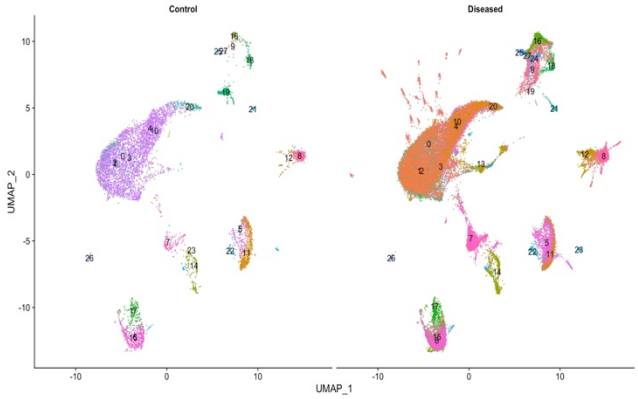

Figure 4 – Cell cluster comparison between control and diseased

We have analyzed cluster 13 which was present in diseased and absent in control condition. The top expressing gene identified with high P-value was **DIRAS family GTPase 3 (**DIRAS3). This gene is usually linked to ovarian and breast cancer. As per GWAS central [4] this gene was identified as a marker in one of the Parkinson study. We have tried to find out conserved markers for each cluster. For example, for cluster 11 between diseased and normal we identified Phosphatidylinositol-5-phosphate 4-kinase type-2 alpha (PIP4K2A), which is related to bipolar disorder. The gene was highly expressed in diseased compared to normal. Due to time limitations we were unable to capture all the gene markers in every cluster.

Marker identification without cell association is meaningless so we tried to annotate our cell using SinglR inbuilt cell-dex library. The cells were automatically annotated. In figure 5 cell cluster identification is described. Some clusters were not annotated because of a limitation in data availability regarding brain tissue.

There are ways to annotate each cell manually however it is extremely time-intensive work. For this project we avoided to annotate cell manually.

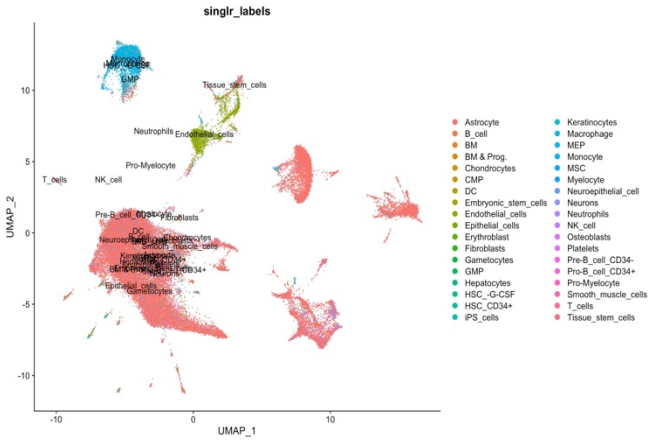

Figure 5 – Cell annotations using SingleR package for Parkinson datasets.

3.3 Genetic marker association with cell type –

We tried to link the already identified Parkinson gene markers in previous studies with our cell clusters. The reason is to understand gene expression changes of these genes at a cellular level (Figure 6). The Conserved markers selected were glucosylceramidase beta 1 (GBA), Leucine rich repeat Kinase 2 (LRRK2), parkin RBR E3 ubiquitin protein ligase (PRKN), Synuclein alpha (SNCA), Parkinsonism associated deglycase (PARK7), PTEN induced kinase1 (PINK1), Leucine rich repeat kinase 2 (LRRK2), and Vacuolar protein sorting (VPS35) were visualized using feature plots.

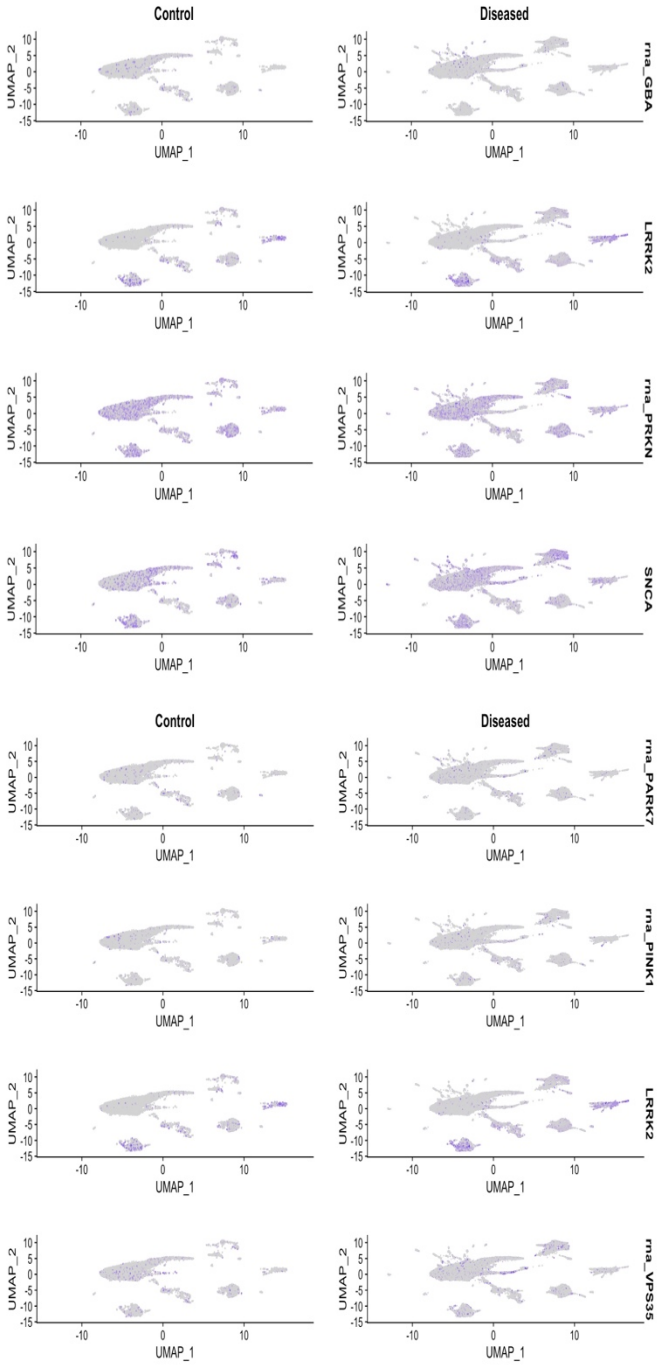

Figure 6 – Feature plots of Genetic markers genes of Parkinson disease.

The Feature plots of gene markers clearly showing in figure 6 the gene expression difference between different cluster in two conditions. For example, gene expression of PRKN is more in diseased condition than control in different cell clusters.

Last we did some visualization to view the gene expression change in different cell types.

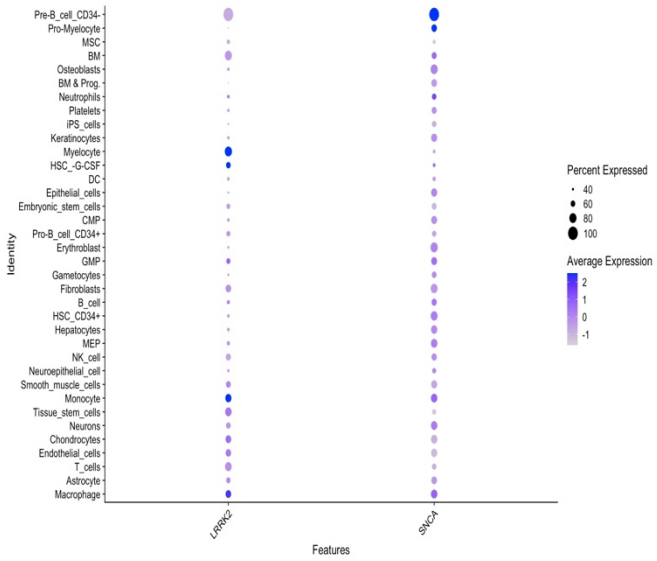

Figure 6: Visualization of Gene expression changes across different cell types.

In Dot plot (Figure 6) the LRRK2 gene showing high expression in Myelocyte and SNCA showing high expression in CD34 cell. This shows that how single cell sequencing can provide a meaningful information on basis of different cell types.

4 Discussion:

The Developing sc/snRNA-seq approaches had become a vital tool for separating heterogeneity and composition of different cell types. They provided us the significant insights into molecular mechanisms of neurological disease such as Alzheimer and other complex human diseases such as different cancer types. Here, we have analysed the publicly available Parkinson dataset for post-mortem Humana brain tissue Substania Nigra on 10 samples including 3 control and 7 diseased. This data set had shown the "clear distinction from that of the previous mouse midbrain DA neurons, which are largely based on the TFs determining the formation and differentiation of DA neurons during mouse brain development."[[3]. The motive of this project was to understand the basic workflow of analysing of the single cell/nuclei RNA dataset. In our studies we were able to identify the top 10 genes which had high expression as compared to other features in the dataset. All the identified genes as per the previous studies had proven the link with Parkinson disease. Cell clustering analysis further showed how different cell cluster present in one condition and absent in other condition. The feature plots helped us to visualize the different gene expression in different cell clusters. By performing cell annotation on our dataset, we were able to capture the different expression of same gene in different cell type. We had tried to perform a lot more downstream analysis but because of limited knowledge and technical difficulties we were unable to achieve it. During this project we have learnt a lot and figure out many tools and technique to analyse the single cell data more deeply. The only limitation I found of analysing the single cell data is that is consider each cell cluster as individual sample however as per my knowledge every cell functionality is dependent on each other. Limitation of this study is prior to quality control we didn't perform doublet cell detection and cell were not manually annotated which in return had provided many false results of cell type.

Future Goal –

After developing good skills in analysing single cell data now my main goal is to integrate it with GWAS (Genome wide association studies) to understand how variants can influence gene expression. My area of focus is neurological diseases.

Conflict of Interest: none declared.

5 References

1. L. S. Ma SX, "Single-Cell RNA Sequencing in Parkinson's Disease. Biomedicines.," PMID: 33916045, 2021 .

2. W. J. L. K. R. Yixuan Qiu, "Identification of cell-type-specific marker genes from co-expression patterns in tissue samples," *Bioinformatics,* vol. Volume 37, no. 19, 1 October 2021, p. Pages 3228–3234, 2021

3. Qian Wang, "Single-cell transcriptomic atlas of the human substantia nigra in Parkinson's disease," bioRxiv, Department of Neurology, Evelyn F. McKnight Brain Institute, Brain Endowment Bank, University of Miami Miller School of Medicine, FL 33136, USA, 2022.

4. J. M. A. F. D. a. D. B. Aaron Lun*, "Bioconductor," 14 June 2020. [Online]. Available: https://bioconductor.org/packages/devel/bioc/vignettes/SingleR/inst/doc/SingleR.html.

5. GWAS CENTRAL Available : https://www.gwascentral.org/generegion/phenotypes?q=DIRAS3&t=ZERO&m=all&page=1&format=html.