scRNA analysis

Single Cell RNA Analysis on PBMCs of COVID-19

WANG, Han-Yi

Department of Biomedical Engineering, Chinese University of Hong Kong

Abstract

Motivation: COVID-19 has been a world crisis ever since its outbreak. Though vaccines and treatment strategies have been gradually developed, the cellular paths and specific immune responses remain unclear. This project applies the basics in singe cell RNA analysis to investigate the changes in major PBMC populations and aims to provide an insight to possible therapeutic cell targets.

Results: The distribution of major PBMC populations are different between healthy and COVID-19 individuals. A significant decrease is observed in T cell populations in COVID-19 patients while an increase in CD14+ monocyte is observed in one of the patients included.

Contact: 1155136569@link.cuhk.edu.hk

Supplementary information: The project codes are available at *scRNA_COVID_PBMC.zip file*.

1 Introduction

According to the World Health Organization (WHO), about 260 million confirmed cases and over 5 million deaths are caused by COVID-19. The severity of the dysfunction of immune response in patients depends on the sophisticated interaction between host, virus, and environment [1]. Lymphopenia, a condition of an abnormal decreased number of lymphocytes, is a common characteristic in severe COVID-19 patients. The T cell responses are impaired, and a significant decrease in CD4+ T cells and CD8+ T cells are found in past research [2, 3]. However, the underlying mechanisms remains unclear. In order to achieve a better understanding on the COVID-19's disease landscape, single cell RNA analysis on peripheral blood mononuclear cells have been conducted to provide a better resolution on the cell level responses.

PBMCs are the major immune cells in the human body. They are mainly composed of lymphocytes and monocytes, with a small proportion of dendritic cells. Due to their selective responses on different infections, PBMCs are widely used in research and toxicological applications. This project aims to provide an insight on the coordinated immune response of the pathogenesis of COVID-19 and thus aid in vaccine development and treatment strategies.

2 Methods

2.1 Dataset

The dataset in this project is adopted from this work [4]. Yu et al. performed sex-balanced sampling of peripheral blood mononuclear cells (PBMCs) from four categories of individuals: hospitalized COVID-19 patients, infected outpatients, exposed individuals (i.e., individuals who had close contact with COVID-19 patients), and healthy controls. A total of 48 PBMC samples were collected. The raw sequencing data was processed by 10x Genomics Cell Ranger software (v3.1.0). After demultiplexing the BCL files, FASTQ files were generated. The FASTQ files were aligned with STAR aligner to the human genome reference GRCh38 from Ensemble database. After feature barcoding and UMI counting, the gene expression matrix was stored. In this project, I extracted two healthy and two hospitalized patients for further downstream analysis.

Table 1. Dataset

Sample ID	Disease state	No. of cells	No of genes
22CBD6-00	Healthy	15406	36601
52BA23-00	Healthy	13801	36601
2DB5F9-00	Hospitalized	10560	36601
460F8D-00	Hospitalized	7460	36601

The dataset can be found on NCBI GEO series GSE171555.

2.2 scRNA Pipeline

The quality control criteria are based on this work [4] with several minor changes including additional droplet and specific gene removals to ensure a better quality control.

2.2.1 Quality Control

In basic filtering, cells with fewer than 500 genes and genes expressed in less than 10 cells are removed. Moreover, a high proportion of mitochondria gene might indicate cell apoptosis, due to the loss of cytoplasmic RNA leaving behind the relatively large size mitochondrial in the cells. Thus, cells with a percentage of mitochondrial genes over 10% are removed. Genes related to technical issues such as MALAT1 are removed. Other genes such as mitochondrial genes and genes related to hemoglobin (i.e., HBA, HBD, HBB etc.) are removed due to unwanted red blood cell contamination in the collection of PBMCs. Lastly, doublet removal is performed by Scrublet python doublet detector package. By simulating doublets through merging cell counts and predicting cells as doublets by the similarity of gene embeddings, Scrublet defines a doublet threshold and removes cells exceeding the threshold.

2.2.2 Dimension Reduction

The data is normalized, logarithmized, and scaled to unit variance before further processing. The top 2000 variable genes are selected to provide a suitable separation across cells. To reduce the dimensions of the data, principal component analysis (PCA) is performed to project the data to its top variances' axes. The top 25 PCs are extracted after analyzing the changes in the variance of PCs. For visualization, UMAP is calculated based on a neighborhood graph.

2.2.3 Clustering

Graph-based clustering includes the following two steps: computation of neighborhood graph and clustering of neighborhood graph. In the first step, the edges of the neighborhood graph are calculated by the K nearest neighbors according to a distance matrix expressing the similarity of genes. After deriving the neighborhood graph, the Louvain method for community detection is adopted. This modularity optimization algorithm defines clusters by maximizing the connections in a region compared to others. The resolution parameter in graph-based clustering defines the detailedness of the clusters (Figure 1).

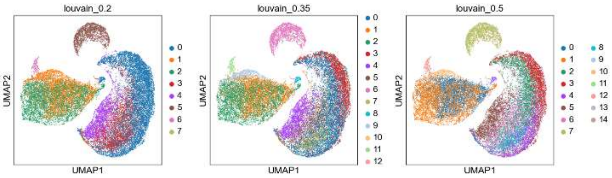

Fig. 1. Different resolutions of Louvain clusters. The number of clusters increase with a higher resolution.

2.2.4 Cell Type Prediction

Cell types are determined by differential gene analysis of each cluster. The ranks of the highly differential genes in each cluster are determined by the t-test. After identifying the marker genes in each cluster, the clusters are labelled with their corresponding cell type (Table 2).

Table 2. Gene markers and its corresponding cell types

Cell Type	Gene Markers
CD4 T cells	CD3E, CD3D, CD14
FCGR3A+ Monocytes	LYZ, CST3, FCGR3A
CD14+ Monocytes	LYZ, CST3, CD14
CD8 T cells	CD3E, CD8A
NK cells	NKG7, GNLY
B cells	CD79A, MS4A1
Conventional Dendritic cells	CST3, FCER1A
Plasmacytoid Dendritic cells	IL3RA, ITM2C

This major cell types and gene markers are identified by the 3kPBMC clustering tutorial of Scanpy

3 Results and Discussion

3.1 Results

The results section will be discussed in three parts: results of quality control, differential gene analysis, and cell type clustering.

3.1.1 Filtered gene results after quality control

After quality control, the total gene counts and the total number of unique genes in each cell has achieved a better linear relationship (figure 2A&2B). To show the specific changes in gene distribution after quality control, the boxplot (figure 2C&2D) of highly expressed gene fractions is plotted. Before gene filtering, the genes with the highest fraction of counts in each cell are mainly RBC contamination related genes (around 80% of the total counts) and the technical issue related MALAT1 gene (around 20% of the total counts). After gene filtering, these genes are removed, and a better distribution of informational gens are presented (figure 2D). For doublet removal, the cells predicted as doublets have approximately twice the number of total gene counts in other cells.

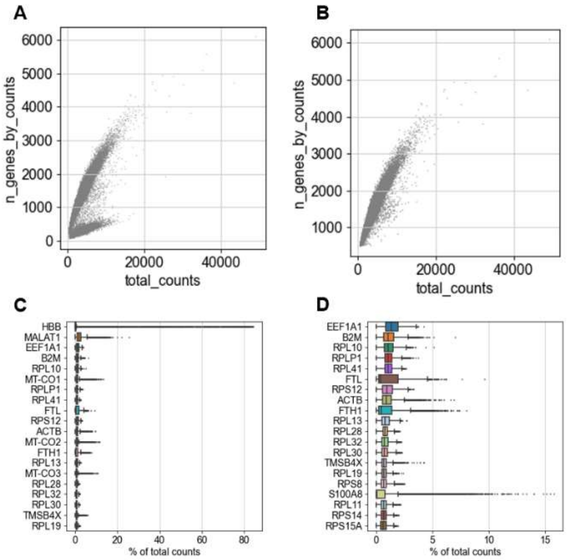

Fig. 2. (A) Scatter plot before quality control. (B) Scatter plot after quality control. (C) Highly expressed gene fraction before quality control. (D) Highly expressed gene fraction after quality control.

3.1.2 Differential Gene Analysis on Louvain Clusters

The top 25 genes are ranked by t-test scores in all clusters which are shown in Figure 3. The marker genes are identified, and labelled clusters are shown in the dot plot in Figure 4.

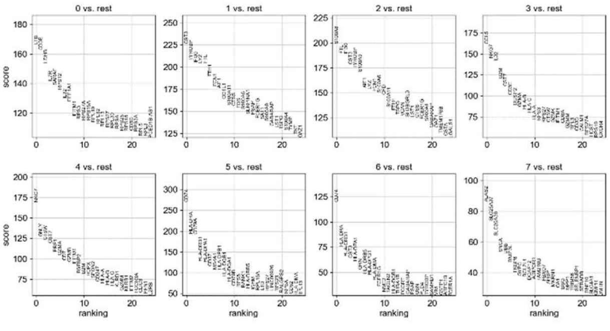

Fig. 3. Ranking results of t-test.

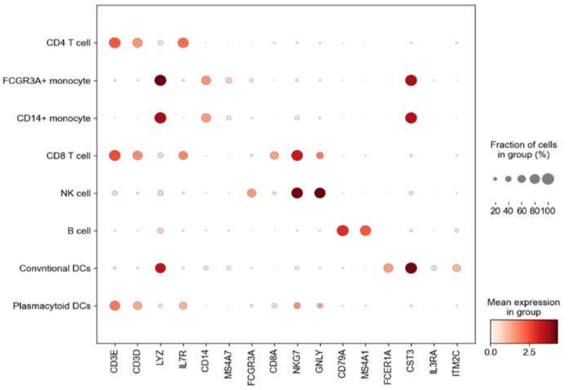

Fig. 4. Dot plot of marker genes and major PBMC cell types. The cell type of each cluster is identified based on the highest fraction of marker genes.

3.1.3 Visualization of Clusters of Major Cell Types of PBMCs

The clustered results are shown in UMAP with fixed random seed to ensure reproducibility. Different resolutions of the Louvain algorithm were tested and fixed at 0.2 at last which generates eight clusters corresponding to the eight main cell types of PBMCs. Cells of healthy individuals and COVID-19 patients have different distribution. It is obvious that the patients have a significant decrease in T cells, including CD4 T cell and CD8 T cell.

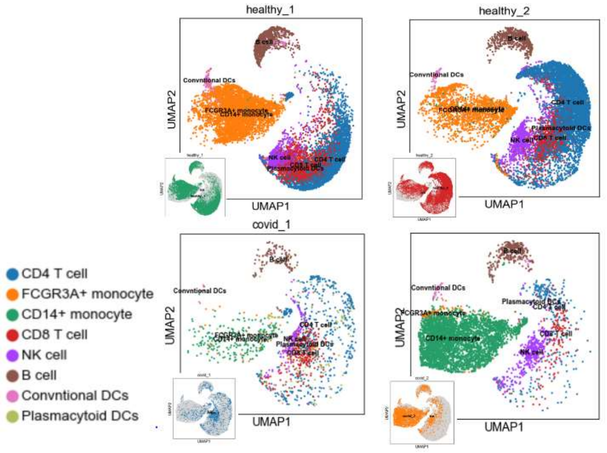

Fig. 5. Louvain clusters of major PBMCs cell type.

3.2 Evaluation

Since the adopted dataset was not labeled, I utilized Ingest python package to automatically transfer the labels of a labelled PBMCs reference data to the unlabeled dataset in this project. Ingest fits the unlabeled dataset by projection onto the reference dataset by PCA and map the labels by the KNN classifier.

The overall manual labeling performance is acceptable since seven out of eight clusters are labelled correctly. The only incorrect cluster is the “megakaryocytes”, which was labelled as “plasmacytoid dendritic cells” in the manual labeling. After investigating the dot plot (figure 4) again, the plasmacytoid dendritic cell cluster was shown to have similar gene expression as CD4 T cell and CD8 T cell, leading to a difficulty in distinguishing and thus an unaligned classification with the Ingest results.

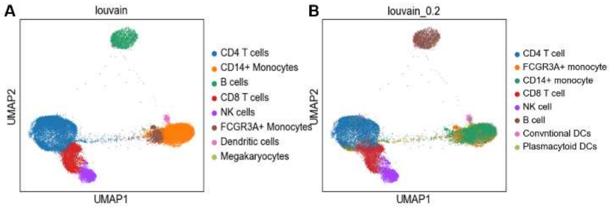

Fig. 6. (A) Transferred annotation from labeled PBMC reference dataset. (B) Manual annotation from differential gene analysis.

3.3 Discussion

3.3.1 Decrease in T cells in COVID-19 Individuals

As most immune cells showed a decrease, the decrease in T cells is of the most significant. This result aligns with past studies which mostly indicates a decrease in the number of total T cells, especially in CD4+ and CD8+ subset [5]. Moreover, various research has shown that patients with a larger decrease in T cells usually have more severe disease status [5]. Since virus clearance depends highly on T cells, it is essential to increase the number of T cells in patient’s body to facilitate recovery.

3.3.2 Increase in CD14+ Monocytes in Patient covid_2

CD14+ monocytes are responsible for the phagocytosis of foreign substances and killing of infected host cells. Past research indicates a positive correlation between expansion of CD14+ monocytes and disease severity [6]. Thus, it is reasonable to infer that patient covid_2 might be in a more sever disease state than patient covid_1. However, very limited studies focused on the changes in monocytes in COVID-19 individuals. A comprehensive single cell analysis into the dynamic changes in monocytes from the time of infection to convalescence should be further conducted.

4 Conclusion

In conclusion, this project highlights the population changes in major cells of PBMCs and provides and insight to the pathogenesis of COVID-19. Future improvements include a more detailed quality control such as sex filtering, since past studies have shown a higher susceptibility to COVID-19 in men than in women [7]. More detailed quality control techniques including removal of batch effect and dropouts can also be performed. Moreover, a higher resolution in the Louvain clustering can lead to more clusters which helps in the analysis of subpopulations of major PBMCs such as anti-virus-specific T cell clones. With a deeper understanding of the immune cells, possible cellular components for targeted therapy or vaccine development can be developed to relieve the global burden of COVID-19.

References

[1] S. Samadizadeh, M. Masoudi, M. Rastegar, V. Salimi, M. B. Shahbaz, and A. Tahamtan, "COVID-19: Why does disease severity vary among individuals?," *Respir Med,* vol. 180, p. 106356, Apr-May 2021, doi: 10.1016/j.rmed.2021.106356.

[2] R. Zhou *et al.*, "Acute SARS-CoV-2 Infection Impairs Dendritic Cell and T Cell Responses," *Immunity,* vol. 53, no. 4, pp. 864-877 e5, Oct 13 2020, doi: 10.1016/j.immuni.2020.07.026.

[3] Z. Chen and E. John Wherry, "T cell responses in patients with COVID-19," *Nat Rev Immunol,* vol. 20, no. 9, pp. 529-536, Sep 2020, doi: 10.1038/s41577-020-0402-6.

[4] C. Yu *et al.*, "Mucosal-associated invariant T cell responses differ by sex in COVID-19," *Med (N Y),* vol. 2, no. 6, pp. 755-772 e5, Jun 11 2021, doi: 10.1016/j.medj.2021.04.008.

[5] B. Diao *et al.*, "Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19)," *Front Immunol,* vol. 11, p. 827, 2020, doi: 10.3389/fimmu.2020.00827.

[6] A. Rajamanickam *et al.*, "Dynamic alterations in monocyte numbers, subset frequencies and activation markers in acute and convalescent COVID-19 individuals," (in English), *Sci Rep-Uk,* vol. 11, no. 1, Oct 12 2021, doi: ARTN 20254
10.1038/s41598-021-99705-y.

[7] P. Brodin, "Immune determinants of COVID-19 disease presentation and severity," *Nat Med,* vol. 27, no. 1, pp. 28-33, Jan 2021, doi: 10.1038/s41591-020-01202-8.