Application of scRNA-seq in Investigating Immune Cell Subtypes in PBMC

Leyao Li

Culver Academies, IN 46511, United States

Abstract. Bulk RNA sequencing produces population-average data, but unable to group the cells into different cell types and thus loses the heterogeneity. Hence, we applied single-cell RNA sequencing to the Peripheral Blood Mononuclear Cells (PBMC) dataset to investigate the underlying distinct cell types in the PBMC sample. To extract important information and reduce the dimensionality we have to investigate when facing complicated datasets, we applied the principal component analysis (PCA). We included the first 10 PCs, applied UMAP non-linear dimensionality reduction and clustered 9 cell subtypes and corresponding marker genes. We searched for the biomarkers for each cell type and assigned cell identities to these 9 immune cell types according to prior knowledge. Our results demonstrated the power of scRNA-seq in distinguishing cell subtypes in a complicated mix of samples, providing insights into important biological questions.

Keywords: Peripheral Blood Mononuclear Cells, principal component analysis, marker genes.

1. Introduction

Bulk RNA sequencing technologies [1, 2] have been widely used to study gene expression patterns at the population level in the past decade. With relatively homogeneous data, bulk RNA sequencing technique is very effective in analyzing data on a general level. However, certain problems lead to the insufficient resolution of bulk RNA sequencing. Bulk RNA sequencing is a technique that gives good data when testing a group of the same cell types or averaging the data of a group of cells, ignoring the cell heterogeneity, thus, leading to the importance of single cell RNA sequencing.

The single-cell RNA sequencing [3, 4] involves the droplet method, which groups one cell, and a magnetic bead with the same barcode (but different barcodes with different beads) within an oil droplet, compartmentalizing the lysis and amplification reactions in the buffer. This process is repeated for thousands of times, providing thousands or more reads of data for all the cells involved, providing more detailed data when tracing situations within a single cell instead of a group of cells [2]. As a result, we applied the single-cell RNA sequencing technique to fulfill the requirement of our experiment, which, instead of population-averaged data, will give clear information at the single-cell resolution and thus providing clear image of which transcripts comes from which cells. The detailed and specific data on single-cell conditions makes single-cell RNA sequencing effective in distinguishing cell types within a group of heterogenous cells, identifying cell state, locating disease within tissues such as cancer by comparing the result with other healthy cells, observing gene expression patterns, detecting cell to cell interaction, and determining cellular heterogeneity.

We analyzed the cells from the PBMC database with scRNA sequencing [5]. The high productivity and stability of peripheral blood mononuclear cells and enrichment of immune cells made them especially suitable for single-cell RNA sequencing application. Apart from that, we had obtained previous data from PBMC literature, so that the different cell types and corresponding biomarkers are ready for us. Therefore, we downloaded the PBMC dataset from 10x Genomics, which contains the immune cells from the human model, including B cells, T cells, Monocytes, and Platelets. This PBMC dataset included 2,700 single cells and 13,714 gene features were sequenced.

We applied scRNA-seq to this PBMC dataset, distinguished 9 cell clusters and matched them with prior knowledge of immune cell types. Our result demonstrated the potential of scRNA-seq for the investigation of cell clustering and heterogeneity. We discovered 4 main cell types: B cells, T cells, Monocytes, and Platelet. We further distinguished some cell subtypes of T cells and monocytes with

totally different biomarker genes. The marker genes for Naïve and Memory CD4 T cells are IL7R, and the marker gene for CD8 T cells is CD8A.

2. Results

To subset the good-quality data from the whole PBMC dataset, some of the quality control steps are necessary at the beginning. We utilized the violin plots to visualize quality of the whole dataset. The violin plots visualize the distribution of gene features (nFeature), transcripts (nCount), and mitochondria percentage (percent.mt) (Figure 1). Low-quality or empty cells exhibit low gene features, while cell doublets or multiplets exhibit high gene counts. As we observed here, the distribution of gene features is mainly located around 900 with a wide range from 200-2,000, indicating each cell has 200-2,000 different genes (Figure 1a). Therefore, we decided to apply the cutoff for single-cell quality control using "200 < nFeature < 2,500". Besides, the distribution of total numbers of transcripts (nCount) is mainly located around 2,000 with a wide range from 500-5,000, indicating each cell has 500-5,000 genes in total (Figure 1b). Low-quality or dying cells exhibit extensive mitochondria contamination within the cytoplasm, therefore the percentage of mitochondria present in the cytoplasm indicates the quality of the cells. The distribution of the percentage of released mitochondria in the cytoplasm is mainly located around 2%, with a range of up to 10% (Figure 1c). We applied percent.mt< 5% as the cutoff to ensure the high quality of an intact and healthy single cell. We applied these quality control steps to secure high-quality of the dataset for downstream analysis.

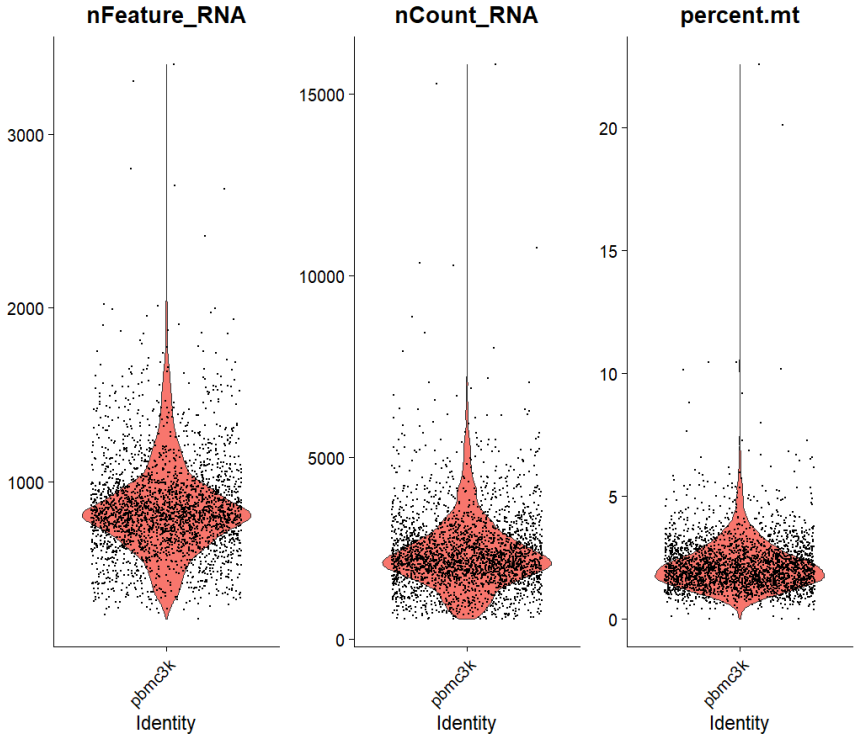

Figure 1. Violin plots of three different matrices, help to identify the distribution of single cells contents to ensure data quality. a) Distribution of the number of different genes in the cells in pbmc3k dataset. Y-axis indicates the number of different genes. B) Distribution of numbers of transcript in the cells in pbmc3k dataset. Y-axis indicates numbers of transcript. C) Distribution of the percentage of released mitochondria in the cytoplasm. Y-axis indicates the percentage of mitochondria RNA. Each dot indicates a single cell

We plotted the correlation of the parameters mentioned above to further double-check the sample quality (Figure 2). The total number of molecules (nCount_RNA) detected within a cell should correlate strongly with unique genes (nFeatures_RNA). The correlation between total number of

molecules and unique genes is 0.95, indicating a strong correlation (Figure 2a). We also compared the correlation between the total number of molecules and mitochondria percentage, which showed -0.13, indicating a weak correlation between the total number of molecules and mitochondria percentage (Figure 2b).

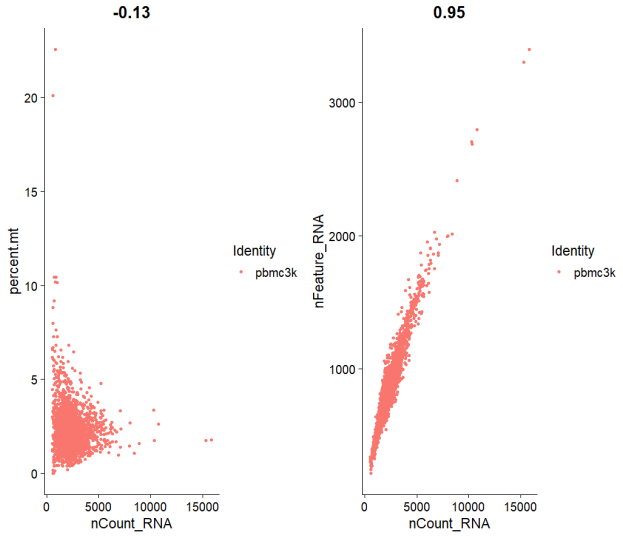

Figure 2. Correlation of different quality control. a) x-axis indicates the number of transcripts in the cell, y-axis indicates the percentage of mitochondria RNA. R = -0.13, which represents a weak correlation. b) x-axis indicates the number of transcripts, y-axis indicates the number of different genes in the cell. R = 0.95, which represents a strong correlation

We further applied volcano plot to find the highly variable features (i.e *they are highly expressed in some cells, and lowly expressed in others), which can help highlight the biological signals from different* cell *types.* The highly variable genes are the genes with distinct expression levels within different cells, indicating the effect of the marker genes (Figure 3). The more variable genes are better at identifying one single type of cell, and the less variable genes indicate the fact that they appear in many different types of cells but can't capture the differences between different cell types. We applied a standard to distinguish the variable genes and non-variable genes. The variable genes are marked in red, and the non-variable genes are marked in black and we labeled the top 10 variable genes.

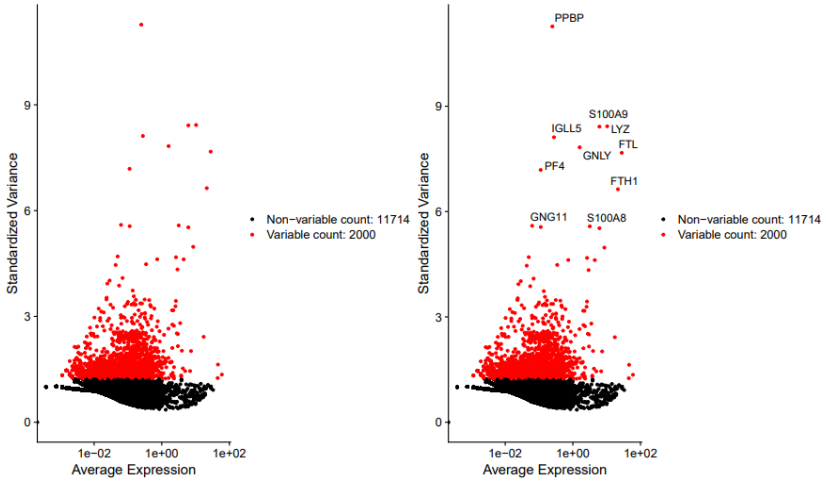

Figure 3. Volcano plot helps to identify highly variable genes. X-axis is the average expression of each gene; y-axis is the standardized variance of gene expression level distribution (among all cells). Each dot symbolizes a gene; the red dots indicate genes significantly variable to each other; the black dots indicates genes not variable to each other. The top 10 genes are labeled

Principal component analysis (PCA) is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while

preserving the maximum amount of information, and enabling the visualization of multidimensional data [6]. To investigate the complicated dataset like scRNA-seq data, it's important to extract the most importance features (differentially expressed genes) while reducing the dimensionality. The are multiple ways to visualize the results of PCA analysis and help us determine the dimensionality of the dataset. We first applied the heatmap to identify the principal component (PC) threshold. Each horizontal line symbolizes a marker gene, and each vertical line symbolizes a cell. Yellow indicates a high expression level (or upregulation) of the differentially expressed genes in the PCs and purple indicates a low expression level (or downregulation) of PCs (Figure 4). From PC1 through PC8 (Figure 4a-h), the marker genes are decreasing in distinguishability, but are still distinguishable with high and low expression level sections featuring the strong separation of purple and yellow blocks. After PC8, the marker genes are almost completely undistinguishable, with high and low expression levels distributed across the map, and most of the cells are missing the data, indicating a mix of purple and yellow lines (Figure 4i-l). We tried to include the main PCs that can explain most of the data with fewer PCs by applying the elbow plot (Figure 5). The plot indicates a trend of decrease in the standard deviation of the principal components, and at PC 10 there is no distinguishing change in Standard Deviation, which indicates starting from PC 10 are mostly generally insignificant data.

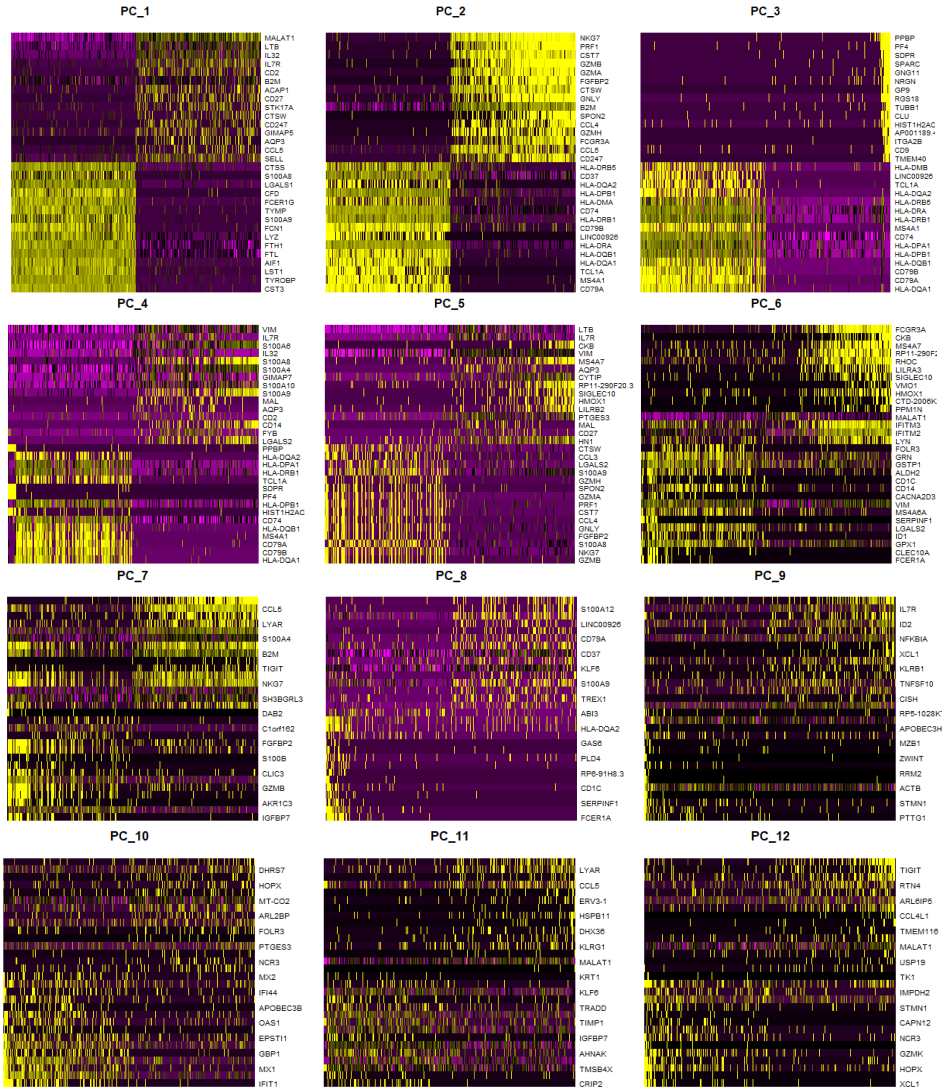

Figure 4. Heat map identifies principal component threshold. Each horizontal line symbolizes a marker gene, each vertical line symbolizes a cell. Yellow indicates high expression level of marker gene; purple indicates low expression level of marker gene. N = 10 is the cutoff for the downstream analysis

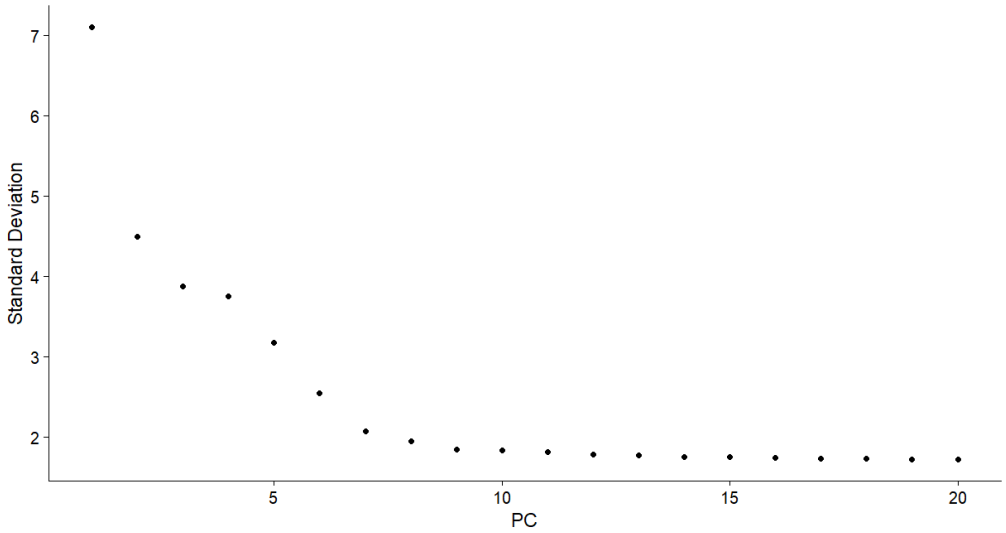

Figure 5. Elbow plot visualizes the cutoff of principal components. X-axis indicates the group number of principal components; y-axis indicates the standard deviation of data. PC 10 is chosen to be the cutoff

To visualize the cell clusters, we utilized the dimensionality reduction method UMAP. A total number of nine clusters are presented in this UMAP (with PC = 10), each representing a subtype of cell (Figure 6). Cluster 0-4 have more cell numbers compared to cluster 5-8, with cluster 8 being the most separated and having the fewest cell numbers. The nine clusters can be classified into four main clusters based on the relative distance between cells in the two UMAP dimensions. Clusters 3 and 8 are two separated clusters, while clusters 0, 2, 4, and 6 and clusters 1, 5, and 7 are classified into two big clusters. The relatively close distance on the UMAP dimensions between these two big clusters indicates the similarity of cells within the clusters. Therefore, clusters 0 and 2 are most similar, while clusters 4 and 6 (in cluster 0/2/4/6) are slightly further, but these four clusters share more similarities compared to the other cell clusters.

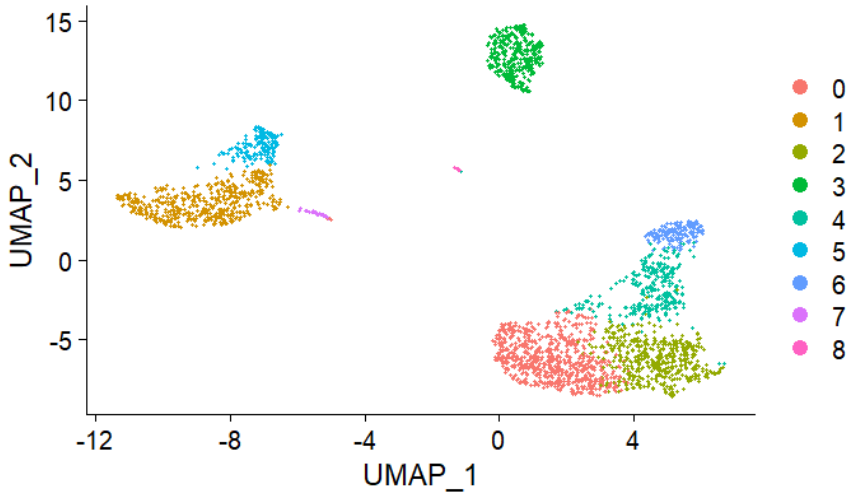

Figure 6. UMAP cell clustering identifies nine different cell clusters in pbmc3k dataset (samples/disease conditions). UMAP_1 and UMAP_2 indicate two different dimensions. Each color represents a cell cluster

To visualize the expression level of marker genes in cell clusters in the UMAP, we applied the feature plots (Figure 7). *CCR7* is highly expressed in clusters 0, 2, and 3, and less expressed in clusters 1, and 5. It is not effective because it is expressed in many clusters without distinction (Figure 7a). *S100A9* is highly expressed in cluster 1, and less expressed in clusters 5, and 7, it is effective because it is distinctively expressed in cluster 1 (Figure 7b). *LTB* is highly expressed in clusters 0, 2, 3, and 4; and less expressed in clusters 5, and 7. it is not efficient because it is expressed in many clusters

without distinction (Figure 7c). *TCL1A* is highly expressed in cluster 3, it is efficient because it is distinctively expressed in cluster 3 (figure 7d). *CCL5* is highly expressed in clusters 4, 6, and 8, it is not efficient because it is expressed in many clusters without distinction (figure 7e). *FCGR3A* is highly expressed in clusters 5 and 6, it is not efficient because it is expressed in many clusters without specificity (figure 7f). *GNLY* is highly expressed in cluster 6, it is efficient because it is distinctively expressed in cluster 6 (figure 7g). *FCER1A* is highly expressed in cluster 7, it is efficient because it is distinctively expressed in cluster 7 (figure 7h). *PF4* is highly expressed in cluster 8, it is efficient because it is distinctively expressed in cluster 8 (figure 7i).

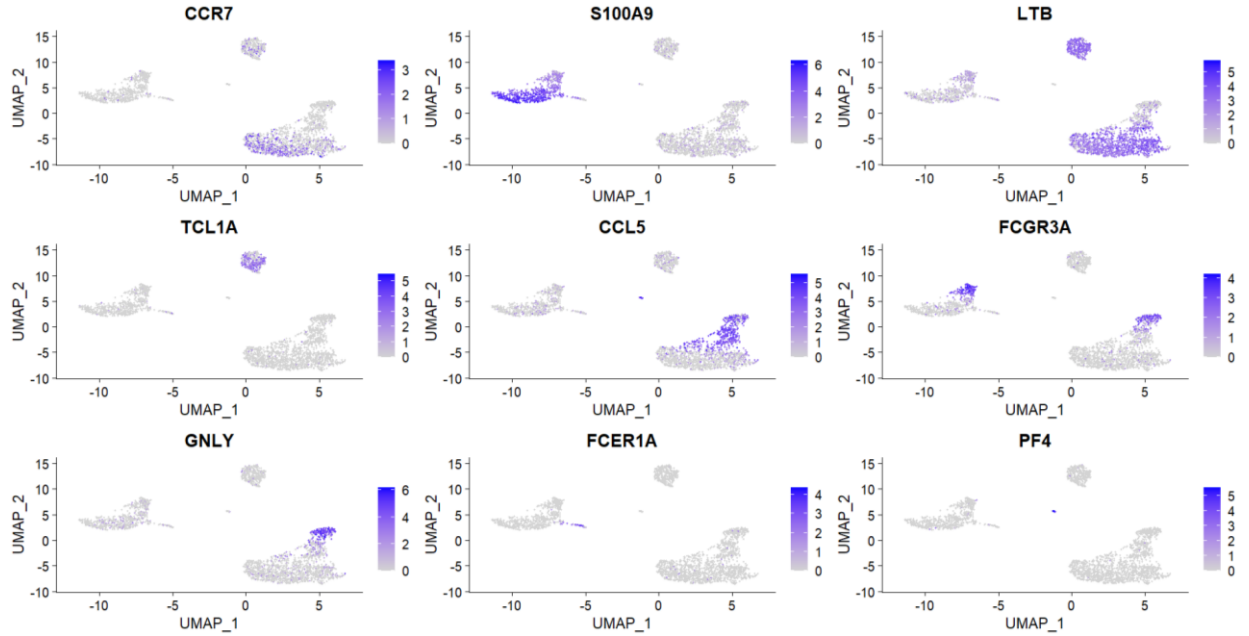

Figure 7. Feature plots of embedded marker gene expression level in the UMAP. Purple dots symbolize the distribution of marker genes. The color bars indicate different levels of marker gene expression. Darker color represents higher expression, lighter color represents lower expression. a-i represents marker gene for all nine clusters, cluster 0: CCR7; cluster 1: S100A9; cluster 2: LTB; cluster 3: TCL1A; cluster 4: CCL5; cluster 5: FCGR3A; cluster 6: GNLY; cluster 7: FCER1A; cluster 8: PF4

To investigate the effect of the marker genes, we applied the violin plots, which presented the expression level of certain marker genes in different clusters (Figure 8). The *CCR7* marker gene is not efficient because it is similarly expressed in clusters 0 and 2, making this marker gene undistinguishable (Figure 8a). *S100A9* is generally expressed in all clusters, but with the highest expression level in cluster 1, it is distinctive and thus effective (Figure 8b). *LTB* is ineffective because it is generally expressed in all clusters with high expression levels in many clusters, making it undistinguishable (Figure 8c). *TCL1A* is effective because it has a high expression level in cluster 3, making it resolvable (Figure 8d). *CCL5* is undistinguishable because it has high expression levels in many clusters (Figure 8e). *FCGR3A* is not distinctive because it has high expression levels in clusters 5 and 6, making these two clusters unresolvable from each other (Figure 8f). *GNLY* is good because it is generally expressed in all clusters but with an especially high expression level in cluster 6 (Figure 8g). *FCER1A* has a high expression level in cluster 7, making it distinguishable and thus practical (Figure 8h). *PF4* is ideal because it has a high expression level in cluster 8, making it differentiable (Figure 8i).

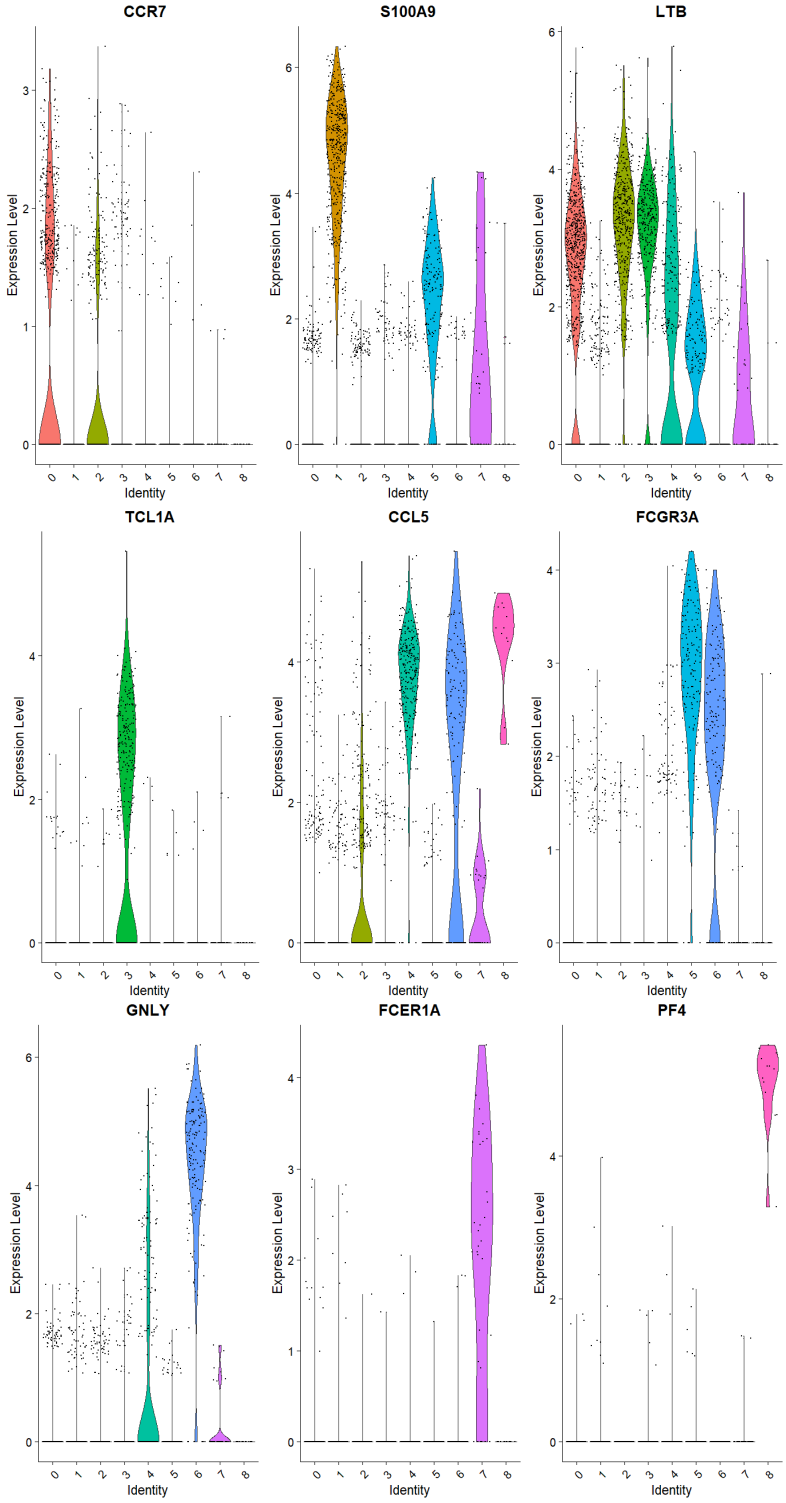

Figure 8. Violin plot displays the specificity of marker gene in each cluster. X-axis indicates nine clusters; y-axis indicates expression level of marker genes. A) CCR7 marker gene is expressed in cluster 0, 1. B) S100A9 marker gene is expressed in cluster 1, 5, 7. C) LTB marker gene is expressed in cluster 0, 2, 3, 4, 5, 7. D) TCL1A marker gene is expressed in cluster 3. E) CCL5 marker gene is expressed in cluster 2, 4, 6, 7, 8. F) FCGR3A marker gene is expressed in cluster 5, 6. G) GNLY marker gene is expressed in cluster 4, 6, 7. H) FCER1A marker gene is expressed in cluster 7. I) PF4 marker gene is expressed in cluster 8

We further visualized the specificity of different marker genes within their corresponding clusters by utilizing a dot plot (Figure 9). *CCR7, LTB, CCL5,* and *FCGR3A* are not effective. *CCR7* has a similar percentage of cells expressing *CCR7* in three clusters while a minority of cells in cluster X

are expressing *CCR7*, making it undistinctive. *LTB* has similar expression levels and percentages in many clusters, making it undistinctive. *CCL5* has similar expression levels and percentages in many clusters, making it not distinctive. *FCGR3A* has a high expression level and similar expression percentage in FCGR3A + Mono cluster and NK cluster, making these clusters undistinctive. *S100A9, TCL1A, GNLY, FCER1A, and PF4 are effective. S100A9* has a high expression level and percentage in the CD14 + Mono cluster, making it distinctive. *TCL1A* is only expressed in the B cell cluster, making it very distinctive. *GNLY* has a high expression level and percentage in the NK cluster, making it distinctive. *FCER1A* is expressed in the DC cluster with many cells expressing *FCER1A*, making it distinctive. *PF4* has a high expression level and percentage in the Platelet cluster, making it distinctive.

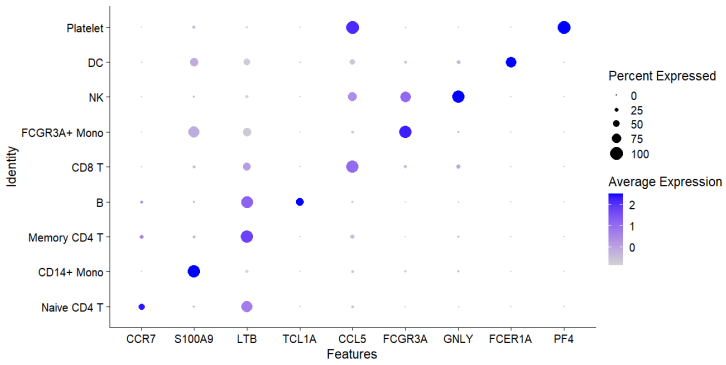

Figure 9. Dot plot distinguishes the specificity of marker genes in nine clusters. The x-axis symbolizes the marker genes. The y-axis symbolizes the nine cell types. the dot size indicates the percentage of the cells that express certain marker genes. The color bar indicates the average expression level of marker genes in each cluster

To display the marker gene specificity with all the clusters as a whole, we applied the heatmap to determine the heterogeneity of the cluster genes. The width of each group indicates the number of cells in the cluster. The ladder shape indicates that the marker gene is effective as it is expressed in a unique cluster (Figure 10). From cluster 0 to cluster 8, a clear stair-like pattern is shown from the top left corner to the bottom right corner, indicating each cluster has a certain amount of marker genes that are uniquely expressed in a single cell cluster. This tests for the effective marker genes in each cell cluster in the database.

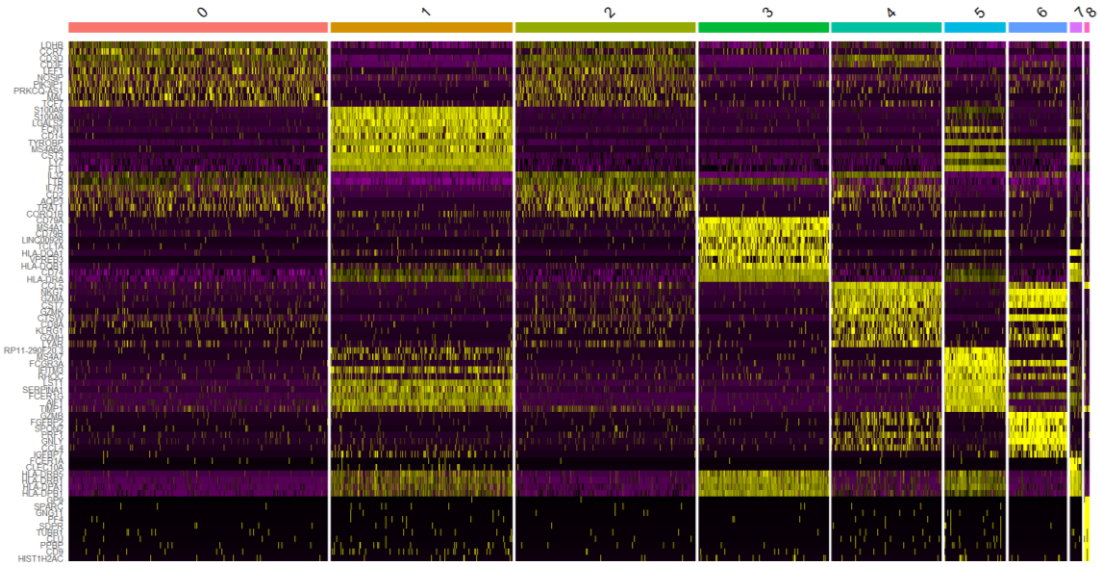

Figure 10. Heat map distinguishes the specificity of marker genes for all nine clusters. Each horizontal line symbolizes a marker gene, each vertical line symbolizes a cell. Yellow indicates high expression level of marker gene; purple indicates low expression level of marker gene. The map is separated into nine groups, each representing a cluster of cells

We searched for the top marker genes for each cluster and tried to match them with prior knowledge. We identified cluster 0/5/6/7/8 with Top2 marker genes (Table 1). *CCR7* is the marker gene for Naïve CD 4 T immune cells. *FCGR3A* is the marker gene for FCGR3A+ Monocyte immune cells. *GNLY* is the marker gene for NK immune cells. *FCER1A* is the marker gene for DC immune cells. *PPBP* is the marker gene for Platelet immune cells (Table 2). *LYZ* is the marker gene for CD14+ Monocyte immune cells. *IL7R* is the marker gene for Memory CD4+ T immune cells. *MS4A1* is the marker gene for B immune cells. *CCL5* is the marker gene for CD8+ T immune cells.

Table 1. Top2 marker genes for each cluster ranked by fold change

	Gene	Cluster	Avg_log2FC	p.val	Pct.1	Pct.2
1	CCR7	0	1.36	9.57e- 88	0.447	0.108
2	LDHB	0	1.09	3.75e-112	0.912	0.912
3	S100A9	1	5.57	0	0.996	0.215
4	S100A8	1	5.48	0	0.975	0.121
5	LTB	2	1.27	1.06e- 86	0.981	0.643
6	AQP3	2	1.23	2.97e- 58	0.42	0.111
7	CD79A	3	4.31	0	0.936	0.041
8	TCL1A	3	3.59	9.48e-271	0.622	0.022
9	CCL5	4	3.10	5.61e-202	0.983	0.234
10	GZMK	4	3.00	7.25e-165	0.577	0.055
11	FCGR3A	5	3.31	3.51e-184	0.975	0.134
12	LST1	5	3.09	2.03e-125	1	0.315
13	GNLY	6	5.32	3.13e-191	0.961	0.131
14	GZMB	6	4.83	7.95e-269	0.961	0.068
15	FCER1A	7	3.87	1.48e-220	0.812	0.011
16	HLA-DPB1	7	2.87	1.67e- 21	1	0.513
17	PPBP	8	8.59	1.92e-102	1	0.024
18	PF4	8	7.29	9.25e-186	1	0.011

Table 2. Assignment of each cell cluster with marker genes

Cluster ID	Marker Genes	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	CD14, LYZ	CD14+ Mono
2	IL7R, S100A4	Memory CD4+ T
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet

We further labeled the UMAP cell clusters with these immune cell types and subtypes concerning their relative distance between each cluster (Figure 11) based on the prior knowledge from the previous literatures. The closer clusters indicate they are more similar, and the more separated clusters indicate they are less similar (in terms of gene expression pattern). We noticed that there are four main clusters, which are the T cell (T cell-like) cluster, Monocyte cluster, B cell, and Platelet. In T cell clusters, Naïve CD4 T and Memory CD4 T have proximity to each other because they share the same CD4 receptor marker. CD8 T cell is a little bit further from these two CD4 T cell subtypes, but they still share many T cell common features. And NK cells are located the furthest in these big clusters, share the least similarity within the T cell clusters. In Monocyte clusters, CD14+ Monocyte and FCGR3A+ Monocyte are close to each other since they share similar Monocyte cell properties, and DC cells are more separated from them, indicating they share the least amount of properties with

the Monocytes within this big cluster. B cells are located far away from T cells and Monocytes, indicating that it is very different from these two cell types. Platelet is also separated, showing its low similarity with all the other cell types.

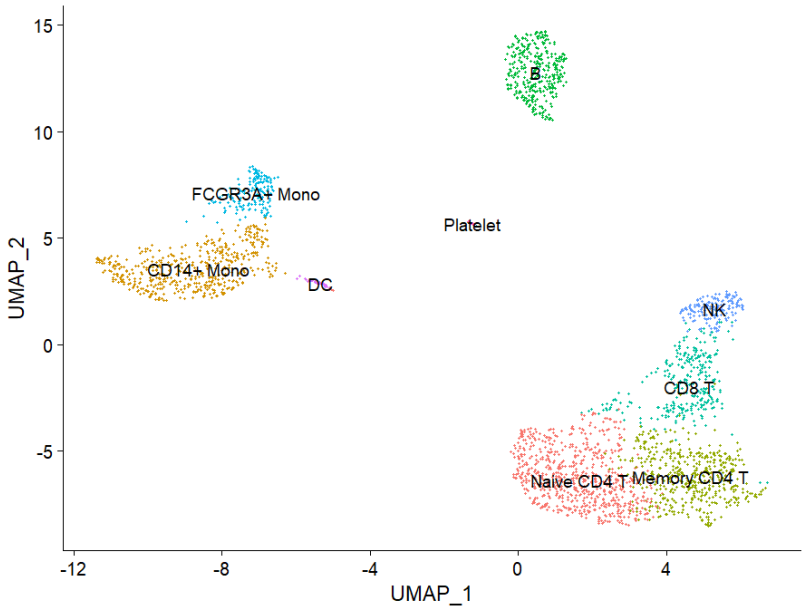

Figure 11. UMAP cell clustering identifies nine different cell clusters in pbmc3k dataset (samples/disease conditions). UMAP_1 and UMAP_2 indicate two different dimensions. Each color represents a cell cluster. The name labeled on the cluster indicates the cell type

With our results listed above, we presented the effectiveness of scRNA-seq in identifying different cell subtypes within a large sample size, indicating the future application of scRNA-seq to significant biological questions with great heterogeneity that bulk RNA-seq is unable to reach.

3. Methods

3.1. Data Acquisition and Quality Control

To ensure that we capture a single cell in each droplet, we filtered the cells by the following parameters: 1. feature >200 to remove empty/low-quality droplets; 2. Feature <2500 to remove duplets or multiplets; 3. Mitochondria percentage < 5% to remove dying cells. We further, validate our data by comparing the correlation between RNA_Count and RNA_Feature (r=.95). We applied a global scaling normalization method (LogNormalize) to normalize feature expression to each cell by the total expression.

3.2. Linear Dimensional Reduction, Using Principal Components Analysis (PCA)

We applied (FindVariableFatures) to identify the differential expressed genes. We identified the top ten features. To determine the dimensionality of this dataset, we applied RunPCA to obtain the principal component groups. We visualized the results by DimHeatMap and ElbowPlot. (main body shape). First, we visualized the DimHeatMap to display the gene expression patterns in 500 cells. (main body-pattern). Second, we applied the ElbowPlot to obtain the elbow position. Both methods identified PC10 as the cutoff for the downstream analysis.

3.3. Non Liner Dimensional Reduction (UMAP)

To further reduce the dimensionality in this dataset, we applied a non-linear dimensional reduction called the UMAP [7] and helped visualize the cell clusters in 2D. we obtained 9 distinct clusters using PC=10, and these 9 clusters form into four groups. The cell clusters in the same group show some similarities in the gene expression type.

3.4. Cluster Marker Gene Identification

We applied the "FindAllMarkers" function to obtain the top two biomarkers for each cluster. We validated/visualized the marker gene identification for each cluster using four different visualization methods. a) we applied "FeaturePlot" to embed gene expression of these marker genes onto different clusters using UMAP settings; b) we further compared the marker genes among nine clusters to visualize the distribution of expression levels using "ViolinPlot"; c) we visualized top two marker genes for all nine clusters as a whole to validate the specificity of marker gene identification with "DotPlot"; d) finally, we applied HeatMap to display top 10 marker genes for each cluster.

3.5. Assign Cell Type According to Prior Knowledge

According to prior knowledge, we know the marker gene for different cell types. From the maker genes in our dataset, we can match them with the cell types using the shared common marker genes. We first compared the top two marker genes and found the following marker genes: PPBP for cluster 8 is the marker gene for platelet; FCER1A for cluster 7 is the marker gene for dendrite cell; GNLY for cluster 6 is the marker gene for natural killer cells; FCGR3A for cluster 5 is the marker gene for FCGR3A+ monocytes; CCR7 for cluster 0 is the marker gene for Naïve CD4+ T cells. The current top 2 marker genes couldn't support assigning cell type identities to all clusters, therefore we extend to the top 5 marker genes. We further validated NKG7 for cluster 6 is the marker gene for natural killer cells; MS4A7 for cluster 5 is the marker gene for FCGR3A+ monocytes; MS4A1 for cluster 3 is the marker gene for B cells; LYZ for cluster 1 is the marker gene of CD14+ monocytes; IL7R is the marker gene for CD4+ T cells (both naïve and memory CD4+ T cells). We assigned cluster 2 to memory CD4+ T cells by IL7R because we excluded cluster 0 to naïve CD4+ T cells by CCR7. To further accurately assign cluster 4, we extended to the top 10 marker genes and assigned cluster 4 to CD8+ T cells by the expression of CD8A.

References

[1] Wang, Zhong, et al. "RNA-Seq: A Revolutionary Tool for Transcriptomics." Nature Reviews Genetics, vol. 10, no. 1, Jan. 2009, pp. 57–63, doi: 10.1038/nrg2484.

[2] Conesa, Ana, et al. "A Survey of Best Practices for RNA-Seq Data Analysis - Genome Biology." BioMed Central, 26 Jan. 2016, genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8. Accessed 12 Aug. 2024.

[3] Jovic, Dragomirka, et al. "Single-cell RNA Sequencing Technologies and Applications: A Brief Overview." Clinical and Translational Medicine, vol. 12, no. 3, Mar. 2022, doi: 10.1002/ctm2.694.

[4] Kashima, Yukie, et al. "Single-Cell Sequencing Techniques from Individual to Multiomics Analyses." Nature News, Nature Publishing Group, 15 Sept. 2020, www.nature.com/articles/s12276-020-00499-2. Accessed 12 Aug. 2024.

[5] Oelen, Roy, et al. "Single-Cell RNA-Sequencing of Peripheral Blood Mononuclear Cells Reveals Widespread, Context-Specific Gene Expression Regulation upon Pathogenic Exposure." Nature News, Nature Publishing Group, 7 June 2022, www.nature.com/articles/s41467-022-30893-5. Accessed 12 Aug. 2024.

[6] Wold, Svante, et al. "Principal Component Analysis." Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1–3, Aug. 1987, pp. 37–52, doi: 10.1016/0169-7439(87)80084-9.

[7] McInnes, Leland, et al. "UMAP: Uniform Manifold Approximation and Projection." Journal of Open Source Software, 2 Sept. 2018, joss.theoj.org/papers/10.21105/joss.00861. Accessed 12 Aug. 2024.