**Single Cell RNA-Seq Clustering Analysis: The Pooled Library and Its Subset**

**Guannan Shen**

**12-14-2018**

**Introduction**

Single cell RNA-seq (scRNA-Seq) is next-generation sequencing (NGS)-based technology for transcriptomics focused on the characterization of individual cells 1. This single-cell analysis will allow researchers to uncover new and potentially unexpected biological discoveries relative to traditional profiling methods that assess bulk populations, such as the discoveries of complex and rare cell populations 1. The scRNA-seq methods generate sparse gene expression profiles for thousands of single cells in one experiment. However, the sparsity and high complexity of the information often prevents full characterization of transcriptomes for individual cells 2. To extract more focused gene expression information from scRNA-seq libraries, the researchers have developed a strategy to physically recover the DNA molecules comprising transcriptome subsets 2. The cell-centric mode approach is to resample the transcriptome for rare cell types by another round of DNA sequencing 2. They resampled the transcriptomes of rare, single megakaryocytes (MK cell) from a complex mixture of lymphocytes from human health donor and analyzed them in a second round of DNA sequencing, yielding up to 20-fold greater sequencing depth per cell and increasing the number of genes detected per cell from a median of 1,313 to 2,002 2.

In my analysis, the principal components (PCs)-based clustering analysis was used to compare the patterns of the common pooled scRNA-seq library and the resampled MK cell-focused scRNA-seq library.

**Methods**

The data was deposited in GEO. The GEO sample (pooled data) GSM3374613 is the pooled scRNA-seq library, from the 10x genomics 3' end library. Peripheral blood mononuclear cells (PBMCs) were isolated from a health human donor. The GEO sample (resampled data) GSM3374614 is the resampled MK cell-focused scRNA-seq library, from the 10x genomics 3' end library.

Most single-cell RNA sequencing platforms incorporate a unique DNA sequence (unique molecular identifier, UMI) to mRNAs derived from a single cell. This sequence served as a molecular handle to resample targeted cell types. Particularly, the locked nucleic acid (LNA)-based hybridization was used to resample individual cell transcriptomes.

The clustering analysis using Seurat R package was conducted in Rstudio (Version 1.1.456) with R (version 3.5.1). In this study, the analysis was started with the count table for about 3,000 cells and 20,000 genes. The data was filtered by the sparsity of genes and cells. Genes expressed in at least 3 (~ 0.1% of the data) cells and all cells with at least 200 detected genes were kept for pooled data. For resampled data, these cutoffs were set as 2 and 100. In terms of QC (quality control) plots, I visualized gene and UMIs counts, plot their relationship, and exclude cells with a clear outlier number of genes detected as potential multiplets. The percentage of mitochondrial genes was also plotted. Thus, I also filtered cells based on the percentage of mitochondrial genes present. The high cutoffs for genes detected per cell and the percentage of mitochondrial genes were set as 2500 and 0.05, and 5000 and 0.05 for resampled data.

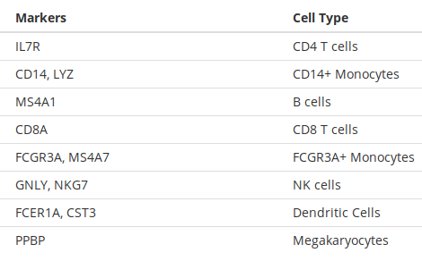
After removing unwanted cells from the dataset, the next step is to normalize the data. By default, a global-scaling normalization method “LogNormalize” was used that normalizes the gene expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default, 1,000 for resampled data), and log-transforms the result.

Then, the highly variable genes were found for downstream analysis. These were defined by following steps: the average expression and dispersion for each gene, places these genes into bins, and then calculates a z-score for dispersion within each bin. Once the highly variable genes were identified, a further normalization step was required to remove the uninteresting sources of variation. This could include not only technical noise, but batch effects, or even biological sources of variation (cell cycle stage). Linear models were constructed to predict gene expression based on user-defined variables. This is to regress out those noise. The scaled z-scored residuals of these models were stored and used for dimensionality reduction and clustering. In this analysis, I just regressed on the number of detected molecules per cell as well as the percentage mitochondrial gene content.

Next, the PCA, linear dimensional reduction, was performed on the scaled z-scored residuals. Several ways were used to visualize PCs. To overcome the extensive technical noise in any single gene for scRNA-seq data, the Seurat package clusters cells based on their PCA scores, with each PC essentially representing a ‘metagene’ that combines information across a correlated gene set. Determining how many PCs to include downstream is therefore an important step, which is essentially to define the true dimensionality of a dataset. An accurate selection of PCs will improve the downstream clustering a lot. Three approaches can be used. First, one can explore the genes in PCs and clusters to get more interpretable results. Second, the statistical significance of PCs can be considered, which is calculated by a resampling test inspired by the jackStraw procedure. Last but not the least, a more ad hoc method for determining which PCs to use is to look at a plot of the standard deviations of the principle components and draw your cutoff where there is a clear elbow in the graph.

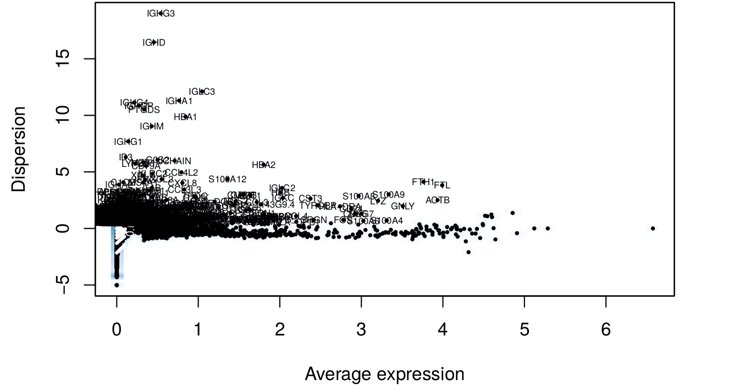
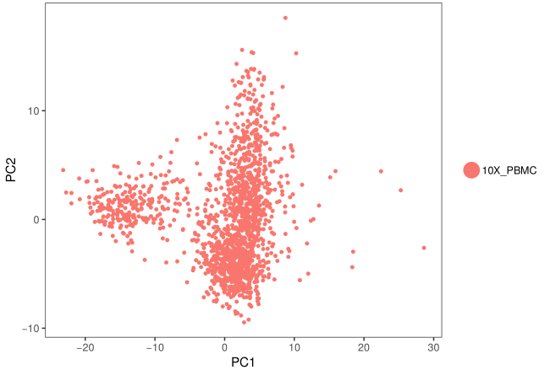
The Seurat package now uses a graph-based clustering approach. Briefly, the method embeds cells in a graph structure - for example a K-nearest neighbor (KNN) graph based on the Euclidean distance, with edges drawn between cells with similar gene expression patterns, and then attempt to partition this graph into highly interconnected quasi-cliques or communities. The resolution parameter for clustering is set between 0.6-1.2, typically, for single cell datasets of around 3,000 cells. Optimal resolution often increases for larger datasets.

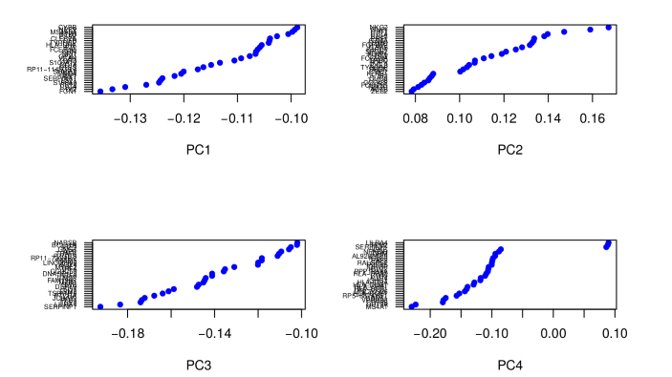
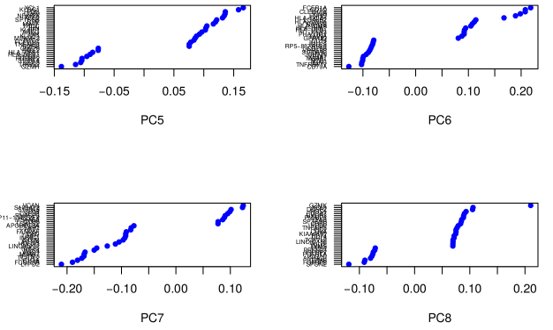
Finally, Seurat continues to use t-SNE as a powerful tool to visualize and explore these datasets. The cluster can then be labeled by top biomarkers defined via differential expression. The canonical biomarkers for cell populations in human PBMC were shown below.



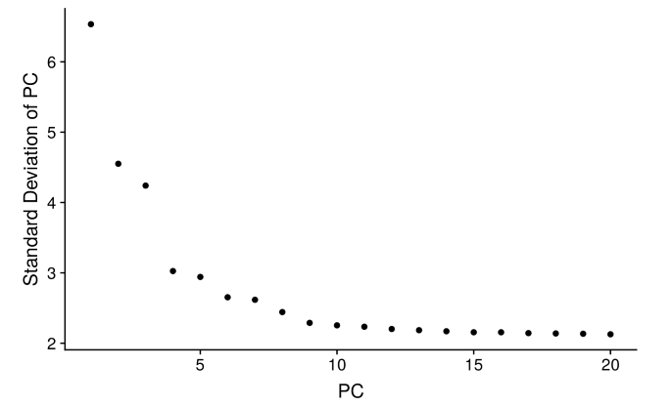
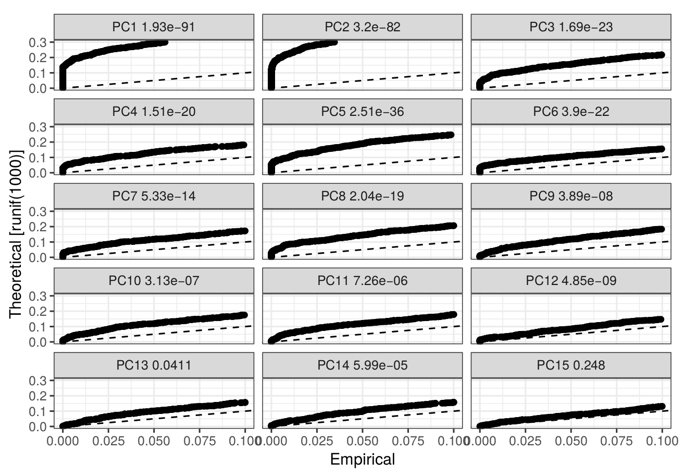
**Results**

Both pooled data and resampled data showed good qualities after filtering the count table. As shown below, 2800 highly variable genes were identified for pooled data and 301 highly variable genes were identified for resampled data (figure not shown here).

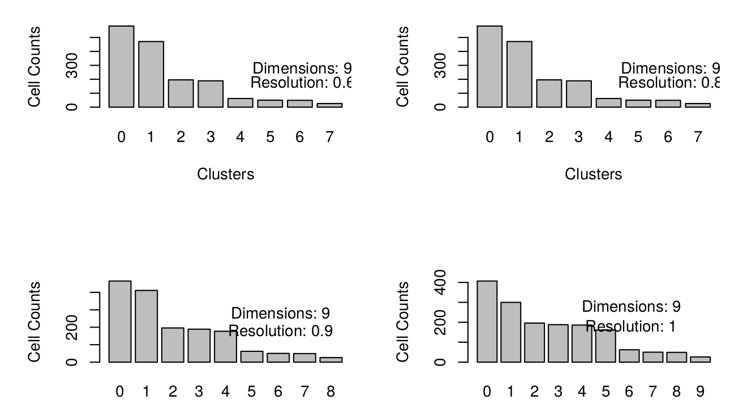
 

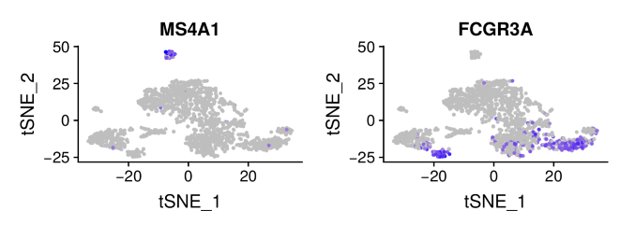
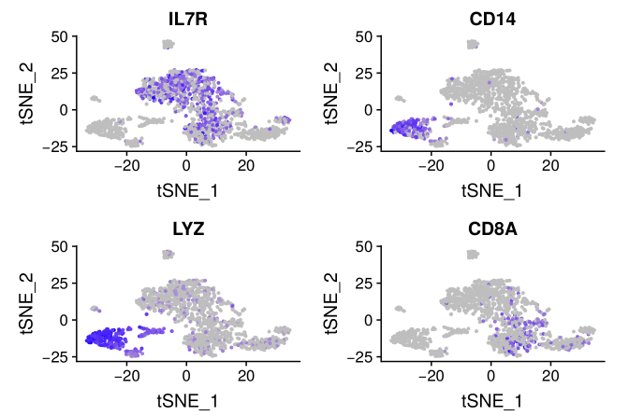
 

The clustering of pooled data based on the first two PCs was shown above. Typically, for scRNA-seq of human PMBC sample, the first two PCs are not adequate to separate about 10 different major cell types. The genes that define the first eight PCs were also shown above. This can help to define the cutoff for number of PCs to use, but not heavily applied in this analysis.

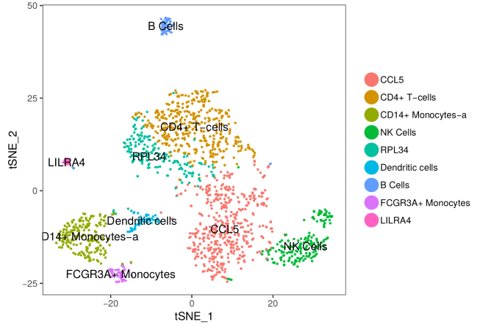
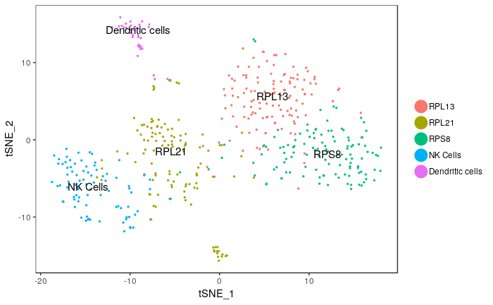


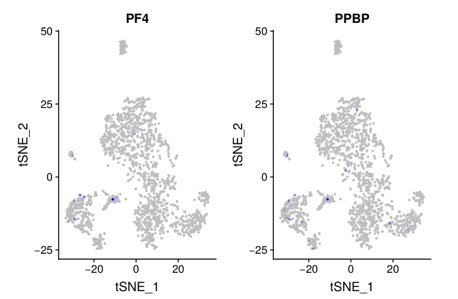
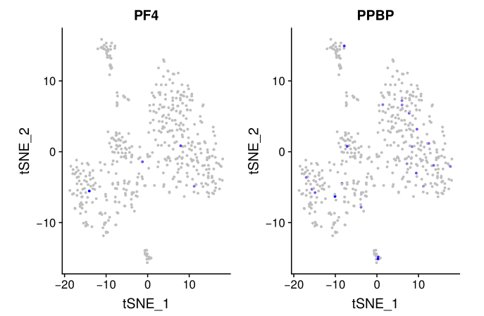
The statistical significance of PCs and the elbow plot were shown above. The first 11 PCs showed very clear deviations between the empirical and theoretical distributions, which suggests these PCs might be included for downstream clustering. The PC elbow plot on the right showed a cutoff ranging from 8 to 10. Eventually, first 9 PCs were chosen for both pooled data and resampled data. The resolution parameter was set as 0.9 and 1.0 for these datasets, respectively. These settings returned 9 clusters for pooled data as shown below.



The distribution of biomarkers, such as MS4A1 of B cells, FCGR3A of FCGR3A+ monocytes, IL-7R of CD4+ T cells, CD14 and LYZ for CD14+ monocytes and CD8A for CD8+ T cells were shown above. For example, CD8+ T cells did not have a good clustering in this analysis, whereas, the CD14+ monocytes were shown to contain sub-clusters.

Eventually, the overall clustering of pooled data (upper left) and resampled data (upper right), and the distributions of PF4 marker and PPBP marker of pooled data (lower left) and resampled data (lower right) were shown above. 4 MK cells were resampled based on the PF4 marker in the resampled data. As the clustering shown above, the idea of resampling in a second round of sequencing is to increase the library depth for rare cell types, such as the MK cell shown here. The resampled data lost cells and genes information besides the targeted cell. Thus, the right way to design the study and analyze the data based on the resampling technique should be as: the first round of sequencing (pooled data) serves as the data to define the clustering and the rare cell population; the second round of sequencing (resampled data) then applied to target this cell population by more genes detected in the subset of this population for a better characterization.

**Summary**

My analysis did not reproduce the results in the paper, completely. This could largely due to the two key steps in the clustering analysis, removing unwanted sources of variation and selection of the number of PCs. There is a lack of information for me to control the technical noise, such as batch effects and cell-cycle variation, and I did not explore genes and cells in PCs thoroughly to determine relevant sources of heterogeneity.

In addition, for resampled data, the common clustering approach for scRNA-seq is not appropriate. Since this data contains more focused information on 4 MK cells, approaches used for traditional RNA-seq data analysis might be helpful.

**References**

1. Single-cell RNA sequencing technologies and bioinformatics pipelines, Byungjin Hwang, *Experimental & Molecular Medicine,* volume 50, Article number: 96 (2018).
2. Recovery and analysis of transcriptome subsets from pooled single-cell RNA-seq libraries, Kent A. And Jay R. Hesselberth, bioRxiv preprint first posted online Sep. 5, 2018.
3. SATIJA Lab, Seurat R package, <https://satijalab.org/seurat/pbmc3k_tutorial.html>.