Analysis of single cell RNA-seq data from liver tissues from patients with nonalcoholic steatohepatitis

Complex Biological Systems

Paulina Duda

Abstract

The purpose of this work is to provide an workflow of the most common type of analysis for scRNA-seq (single cell RNA sequencing) data. I decided to conduct scRNA-seq analysis using data from article published in Natture in 24 March 2021 "NASH limits anti-tumour surveillance in immunotherapy-treated HCC". This paper include the part of the same analysis. In general I was guided by the choice due to the scope of my interests. Moreover my motivation was to improve my programming skills and to learn do scientific things. Workflow which was conduct in this paper in my opinion is worth to repeat particularly due to the novelty of research. Another value of my work is that we always do something a little differently so in this way we can validate the results obtained in the original work. Also We can prove that with the same data we come to the same conclusions. Furthermore it has significant value because we can detect some technical errors and shortcomings, sometimes more than we could expect. Summing up, my work has the main verification and educational value.

Introduction

Nonalcoholic steatohepatitis (NASH in short) is damage and inflammation of the liver caused by excess fat in the liver. NASH is similar to a type of liver disease caused by prolonged and heavy alcohol use, but only occurs in people who do not abuse it. Many people have a build-up of fat in their liver and in most people it doesn't cause any symptoms or problems, but in some people it causes inflammation, damages liver cells, and ultimately can lead to hepatocellular carcinoma (HCC in short). HCC is a common type of cancer that can have viral and non viral causes. I will focus on no viral couse that is NASH. Depending on the etiology of HCC, it became necessary to stratify patients on immunotherapy due to different immune response or lack of it at all. In humans, single-cell analysis has improved understanding not only releted to cancers but also, developmental processes and aging.

Overview of the workflow

Seurat is a suite of analysis tools including quality control and data mining for single cell RNA sequencing. The tool is used to identify and interpret heterogeneity from single cell transcriptomic measurements and to join different types of single cell data. Each scRNA-seq analysis is started with the control and improvement of the quality of the data we have at our disposal. In high-throughput methods, the basic step is to filter out the cell from barcodes that do not represent the cell. The simplest approach is to set a specific threshold for the minimum number of UMIs required to recognize a barcode as a cell. Another thing to filter out is percent of mRNA derived from the mitochondrial genome in cell, becouse those cells are often damaged or dying. Standard threshold is about 5%. In general quality control based on mtRNA amount, number of detected genes and total UMI's. Filtered data are then normalized. Normalization has to be performed becouse of different number of reads are varying between cells. This is significant step becounse quantity RNA per cell can be very different. By default, Seurat uses the "LogNormalize" normalization method, which normalizes the gene expression measurements for each cell by total gene expression. It then multiplies it by the scale factor (10,000 by default) and transforms the result logarithmically.

Feature selection works by calculating a subset of the features that show high variability between cells, meaning that they are strongly or weakly expressed in some cells. He discovers that the most variable genes are usually the most interesting for further analysis. Feature selection identifies genes with the strongest biological signal in relation to technical noise. By limiting further analysis to the most instructive genes. The effect is noise reduction which simplifies the analysis. Seurat (tools for single cell analysis) takes a non parametric approach to identify high up variable genes by empirically matching the relationship between variance and mean expression.

Another step of the general pipline analysis is dimensional reduction the are many methods avaible like linear principal component analysis (PCA) and non linear uniform manifold approximation and projection (UMAP) or t-distributed stochastic neighbor embedding (t-SNE). Most commonly used is PCA. This transformation retain on standard Eucidlean distance. After running, number of components depend on data that we have. The most common method is to plot the fraction of the variance explained by each element, then visually identify it the point where the curve makes a sharp turn. Another method is to plot the standard deviation of the major components to easily identify a break in the graph. If we operate on much more complex data it can be difficult to perform analysis using PCA. Then more appropriet method and current best would be UMAP. UMAP approximates the topology of the data using the cell to cell nearest neighborhood network and then estimates the low dimensional placed of the data which best maintains its structure. UMAP nowadays often replaces the aforementioned t-SNE.

Differential expression (DE in short) for an scRNA sequence is a comparison of non-single values for each gene (as in a population of cells - bulk RNA) but expression level distributions. Moreover, scRNA data are not grouped predetermined. Groups are defined based on expressions levels and this disrupt a main assumption in standard statistical test procedures. The non parametric Wilcoxon test was found to be the best solution in this case. For now, we expect further advances that would lead to more accurate statistical models for more complex datasets and experiments in general.

Materials and methods

I have decided to run the analysis on human samples, leaving out the mouse ones. Hence eliminate their comparison as well. Type of experiment was expression profiling by high throughput sequencing. I have chosen mRNA profiles of liver tissues from 2 patients (human research participants) with different pathological characteristics. Used scRNA-seq data described in article [1] are available at GEO under accession number: GSE159977 (PT-12, PT-21). For mentioned data analysis I used mainly Seurat 4.0.4 library. Another libraries that was used is: dplyr, patchwork, ggplot2, SingleR, celldex, RColorBrewer, SingleCellExperiment, data.table and Matrix.

Results

In the article "NASH limits anti-tumour surveillance in immunotherapy-treated HCC" they found increasing accumulation of non classical activated CD8+PD-1+ T-cells (cells with protein causes programmed cell death) in livers patients with NASH. PD-1 targeting immunotherapy (with NASH-induced HCC), administered at tumor initiation or after tumor diagnosis, has not proved beneficial, quite the contrary. This led to necrotizing hepatitis and an increase in the incidence of HCC. After immunotherapy, an increase in hepatic CD8 + PD-1 + T cells was confirmed in patients with NASH. The data shows that identify NASH, as potentially non-responsive to treatment in the context of HCC immunotherapy, which is a strong rationale for patient stratification. Analysis of scRNA-seq data of patients mentioned above is shown on plots below and in attached Rmarkdown file.

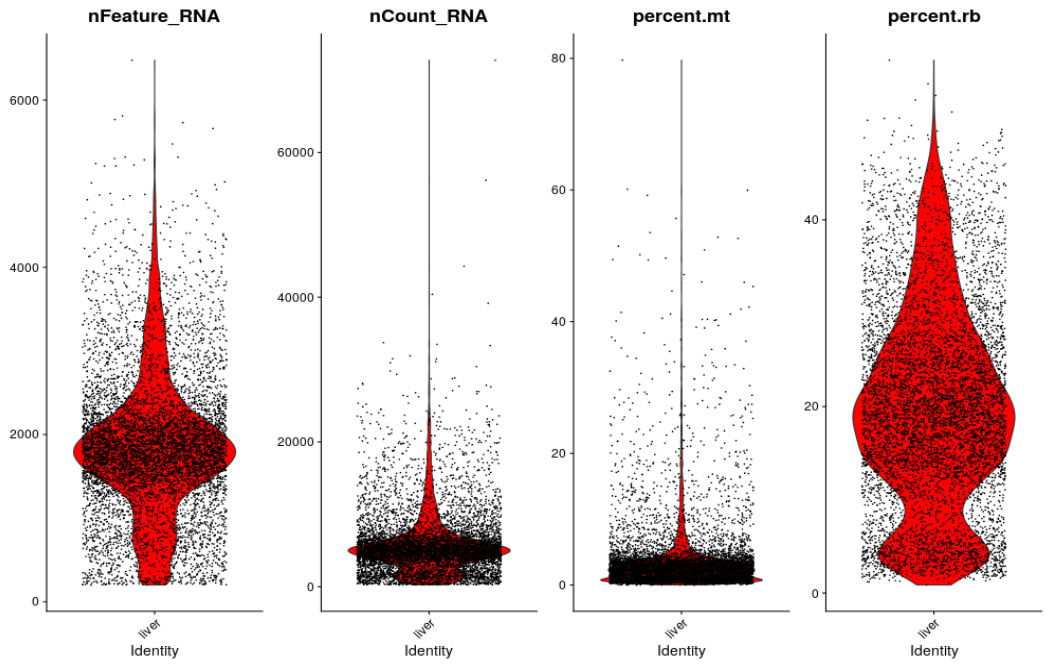

Figure 1: Visualize QC metrics

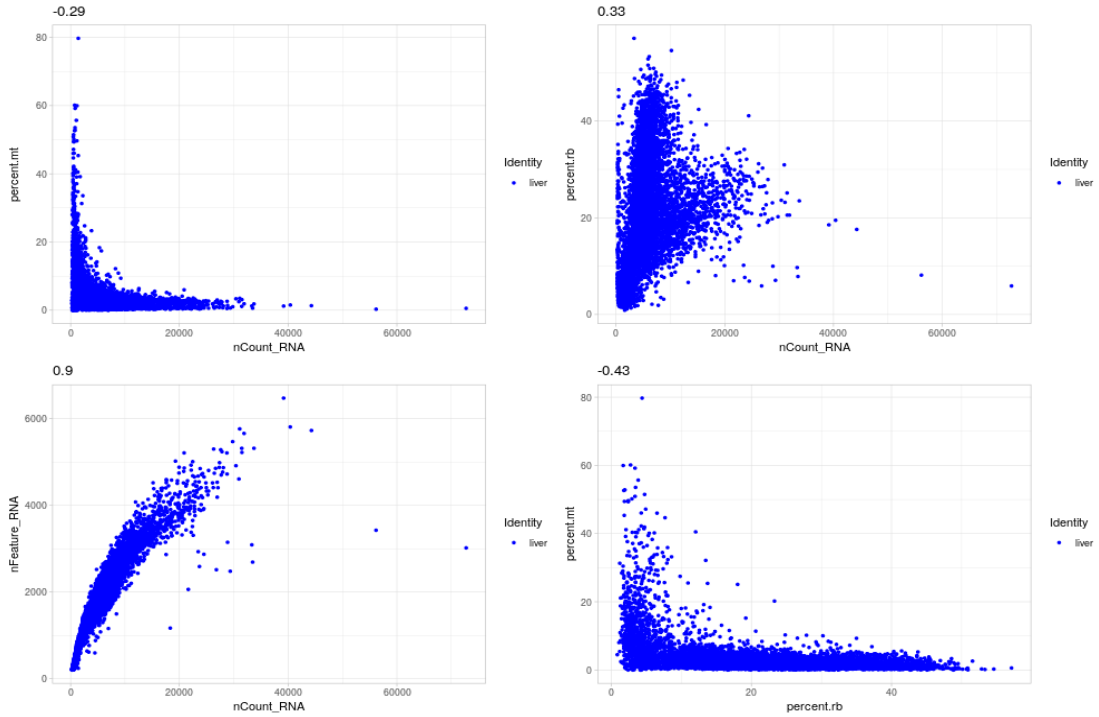

Figure 2: Plot metadata features against each other and see how they correlate (Pearson).

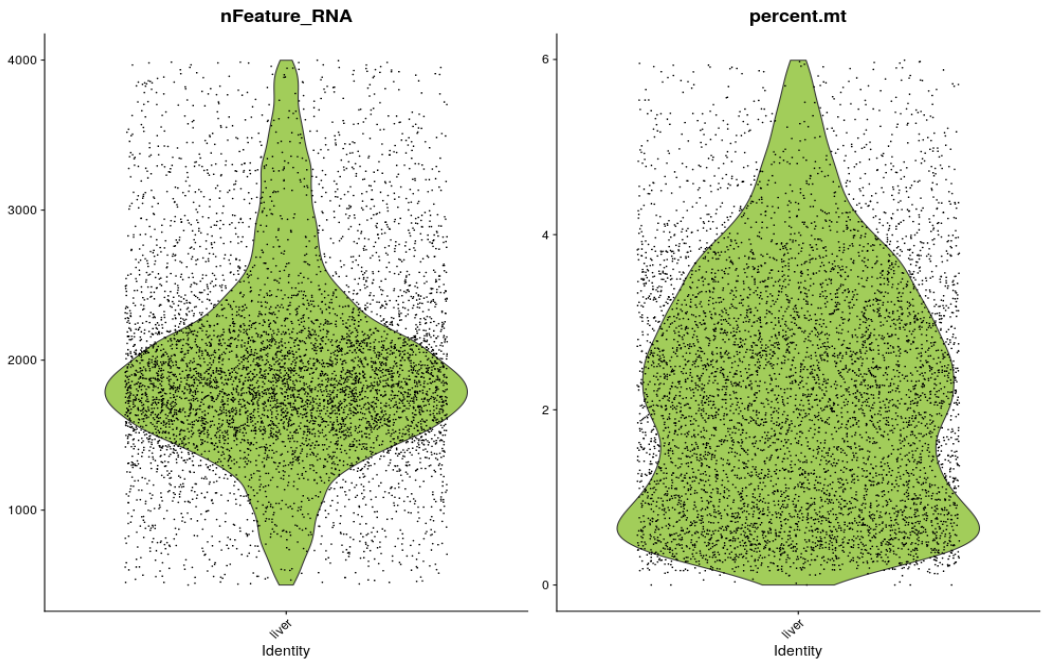

Figure 3: Visualize QC metrics after filtering

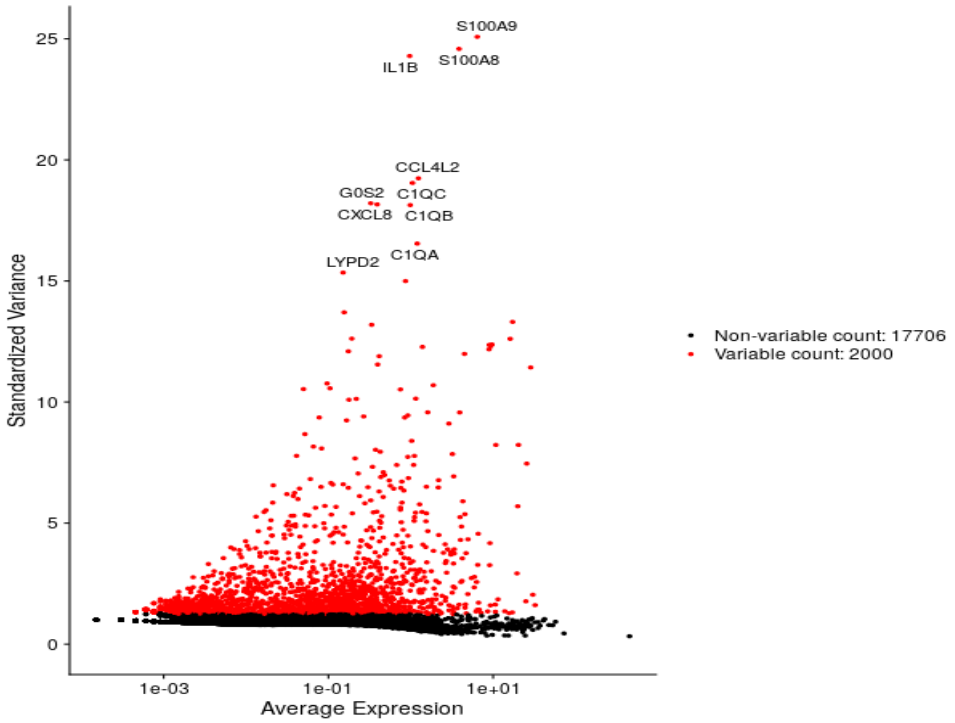

Figure 4: Plot most variable genes

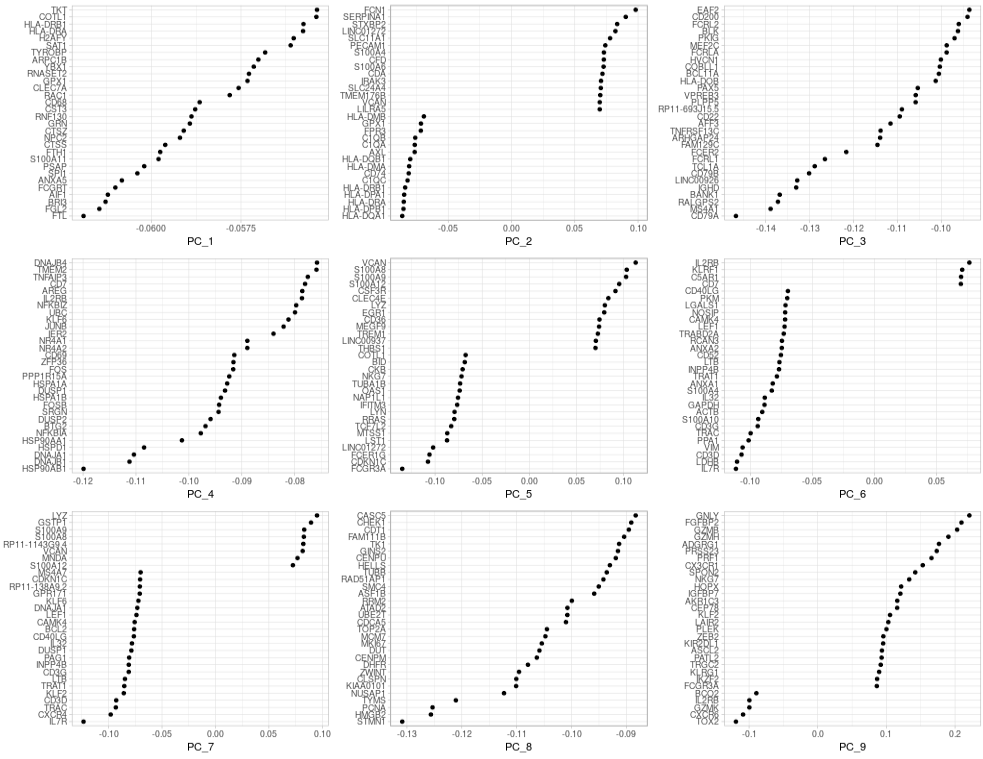

Figure 5: Plot PCA loadings

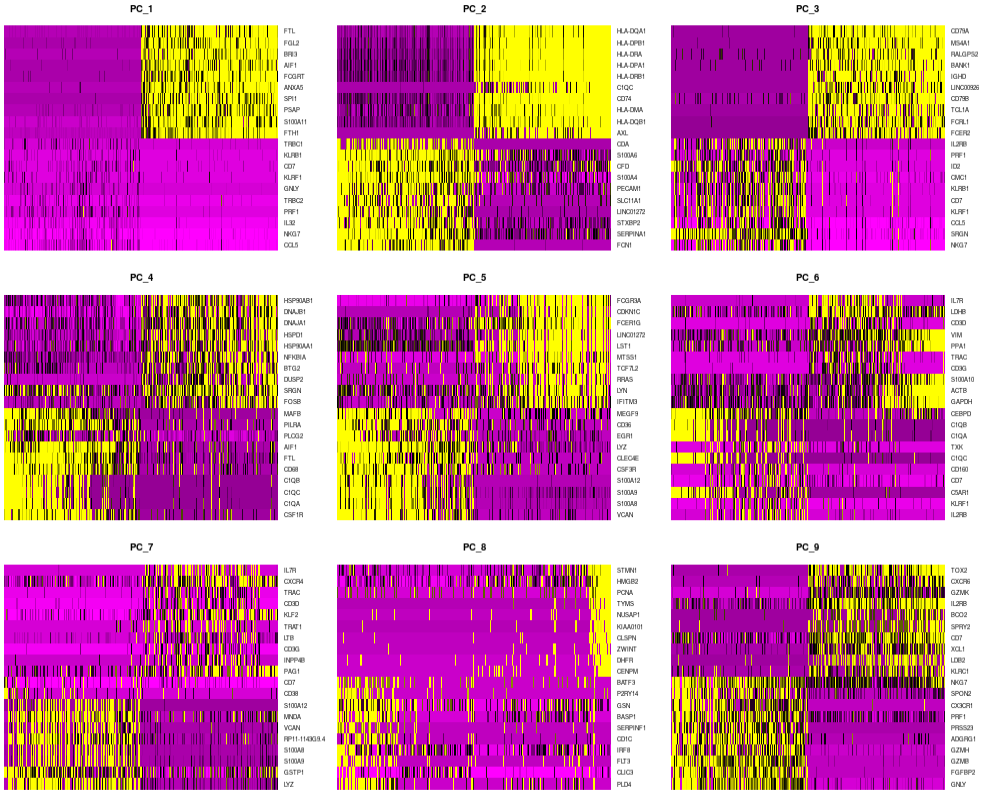

Figure 6: Heatmap of each PCs

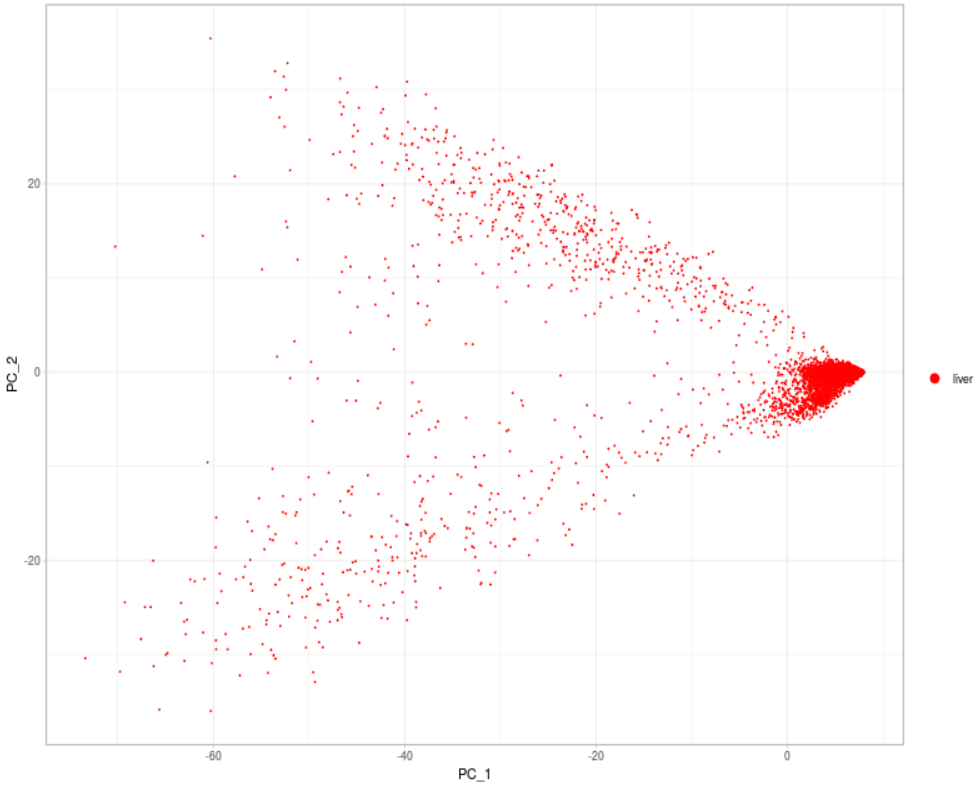

Figure 7: Visualize reduced representations (used PCA)

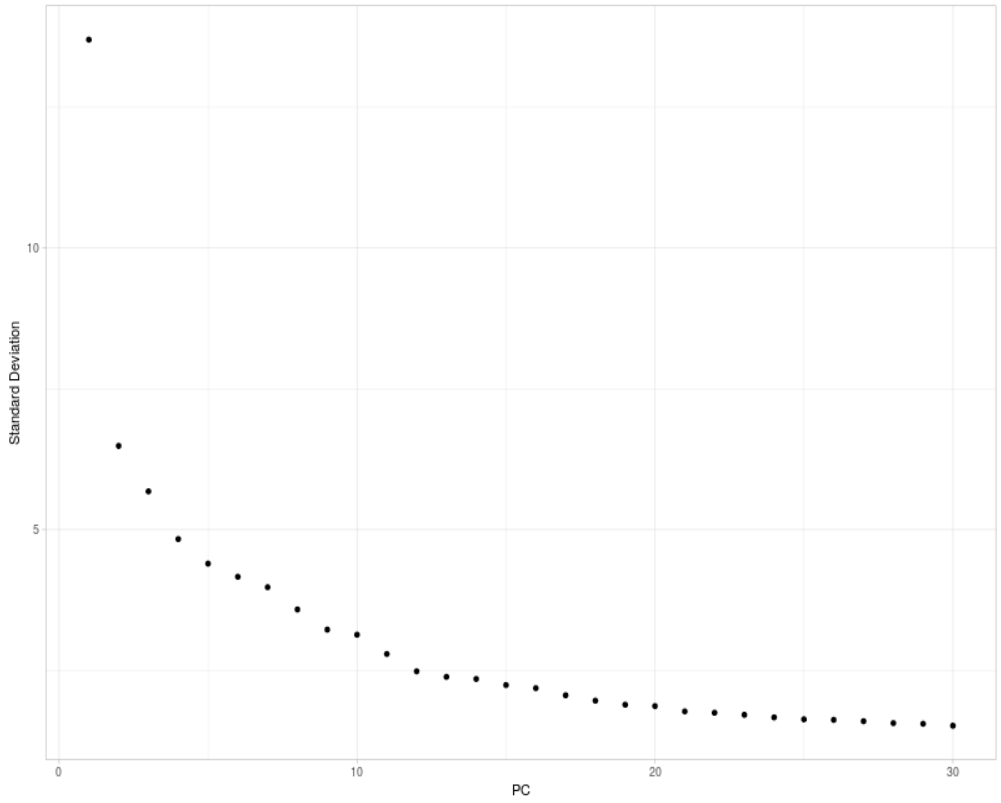

Figure 8: Check how many PCs can be used without information loss

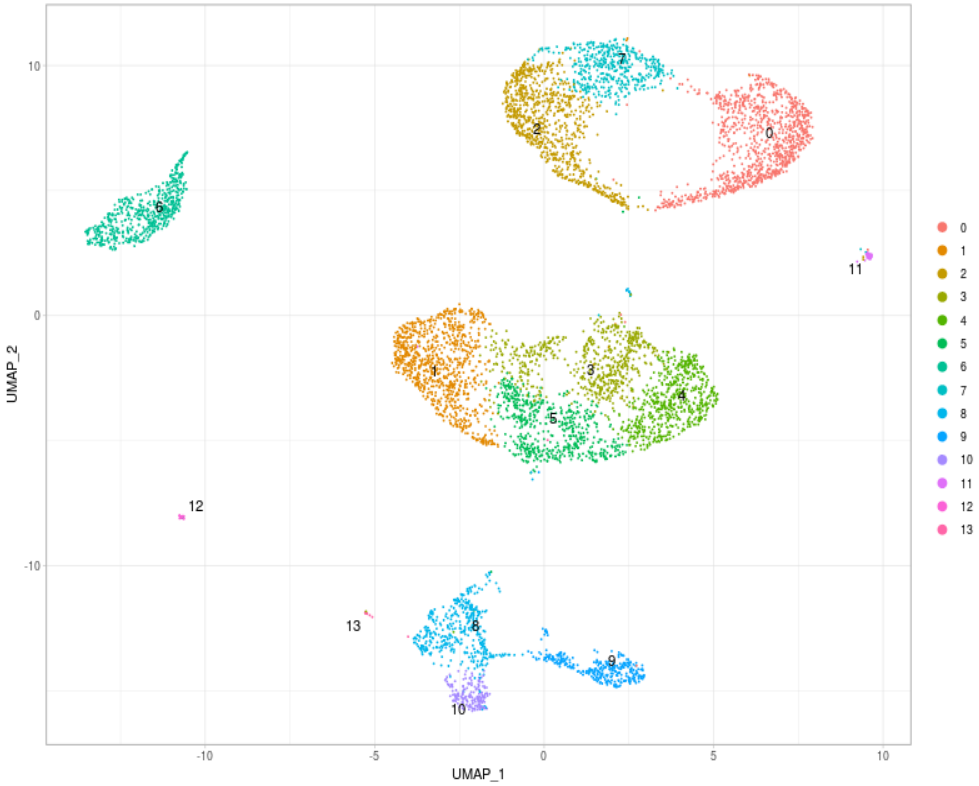

Figure 9: Visualize reduced representations (used UMAP)

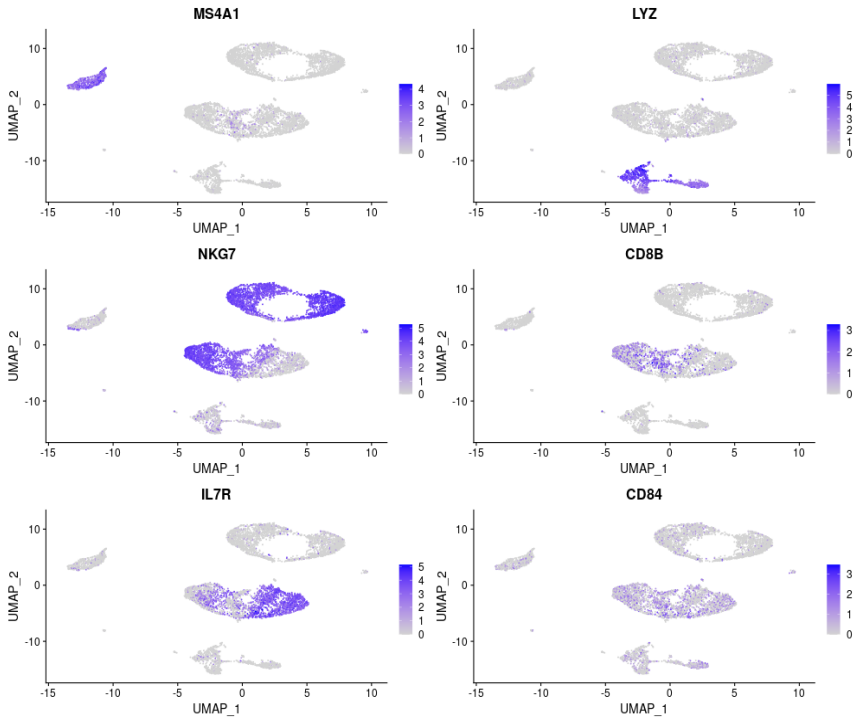

Figure 10: Visualizes feature expression (B cells, monocytes, natural killers CD8 T, CD4 T, CD84, respectively)

Figure 11: Visualize the annotations

References

1. Pfister, D., Núñez, N.G., Pinyol, R. et al. NASH limits anti-tumour surveillance in immunotherapy-treated HCC. Nature 592, 450–456 (2021). https://doi.org/10.1038/s41586-021-03362-0

2. Andrews, T.S., Kiselev, V.Y., McCarthy, D. et al. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc 16, 1–9 (2021). https://doi.org/10.1038/s41596-020-00409-w