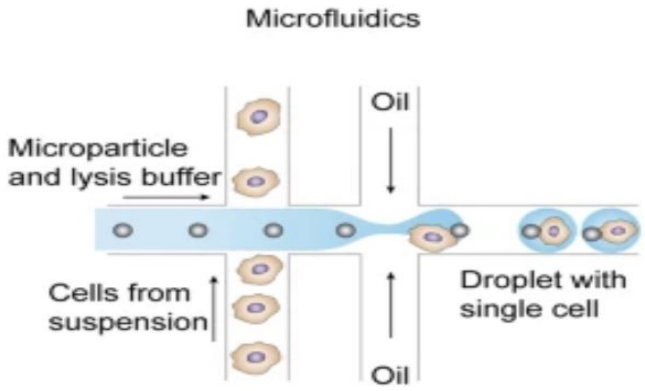
Uncovering Tumor Heterogeneity of Triple-negative Breast Cancer from Single-cell RNA Sequencing

Single-cell RNA sequencing

RNA sequencing (RNA-seq) is a genomic approach for quantitative analysis of messenger RNA molecules in a biological sample and is useful for studying cellular responses. RNA-seq has fueled much discovery and innovation in medicine over recent years. However, bulk RNA-seq often suffers from the complexity of tissues and fails to resolve the actual activity on the cell scale. Single-cell RNA-Seq is a more advanced technique used in biological research to quantify the transcriptome of a complex tissue by characterizing individual cells within that specific tissue. scRNA-seq can describe RNA molecules in individual cells with high resolution and on a genomic scale. In addition, scRNA-seq allows for the comparison of the transcriptomes among cells. Therefore, scRNA-seq has been to evaluate transcriptional differences within a population of cells, thus revealing unknown heterogeneity within a tissue. Similarly, assessments of transcriptional differences between individual cells have been used to identify rare cell populations that have been undetected in analyses of pooled cells, for example, malignant tumor cells within a tumor mass. Conclusively, heterogeneity analysis remains a core reason for embarking on scRNA-seq studies.

Procedures to perform scRNA-seq

Most established scRNA-seq experiments have adhered to a general methodological pipeline. The first and most important step for scRNA seq is single-cell isolation. The step is highly variable and flexible depending on the tissue type of interest. The common methods of single-cell isolation include mechanical, enzymatic, or combinatorial separation. Next, a micro reaction containing a single cell is necessary for following molecular biology reactions. Microfluidics is one of the most common devices to establish the environment because of its lost cost and low sample consumption. In the process of the droplet (microemulsion used to isolate individual cell reactions) formation, the single-cell suspension is combined with beads made with resin and flowed with oil to produce an emulsion.The droplet is created for a micro reaction that encapsulates a single resin bead and a single cell. Then detergent will destroy the cell, releasing its mRNA. The released mRNA will be captured by the polydT coated on the bead thanks to their polyA tails. Next, polydT-primed mRNA is converted to complementary DNA (cDNA) by reverse transcriptase. Depending on the scRNA-seq protocol, the reverse-transcription primers will also have other nucleotide sequences ligated to them, such as adaptor sequences for detection on sequencing platforms, unique molecular identifiers to mark unequivocally a single mRNA molecule, as well as cell barcodes to encode information on cellular origin. PCR then amplifies the cDNA. Enrichment of cDNA produces adequate copies of targeted DNA because next-generation sequencing requires a certain amount of material in order to generate accurate and reliable results. Next-generation sequencing is completed by loading samples into flow cells that match them into different base pairs and capture an image of the result. Base calling is done by identifying the different colors of clusters.



Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 50, 1–14 (2018). https://doi.org/10.1038/s12276-018-0071-8

scRNA-seq data analysis workflow

1. Alignment

Alignment is the process of mapping sequence reads to determine the location of each read on the whole genome to identify its gene name and location. Alignment is an important part of analyzing next-generation sequencing data. Its goal is to provide information about the genome location, discover new genes or transcripts, and provide information about which gene the mRNA comes from. The input of alignment is a short sequenced read produced by the sequencing platform, normally in form of a FASTQ file, which contains information about sequencing quality, read sequences, and machine information. The commonly used alignment algorithm is the Burrows-wheeler transform, which significantly reduces the complexity of computation. Raw sequencing data are processed and aligned to give count matrices.

2. Quality control

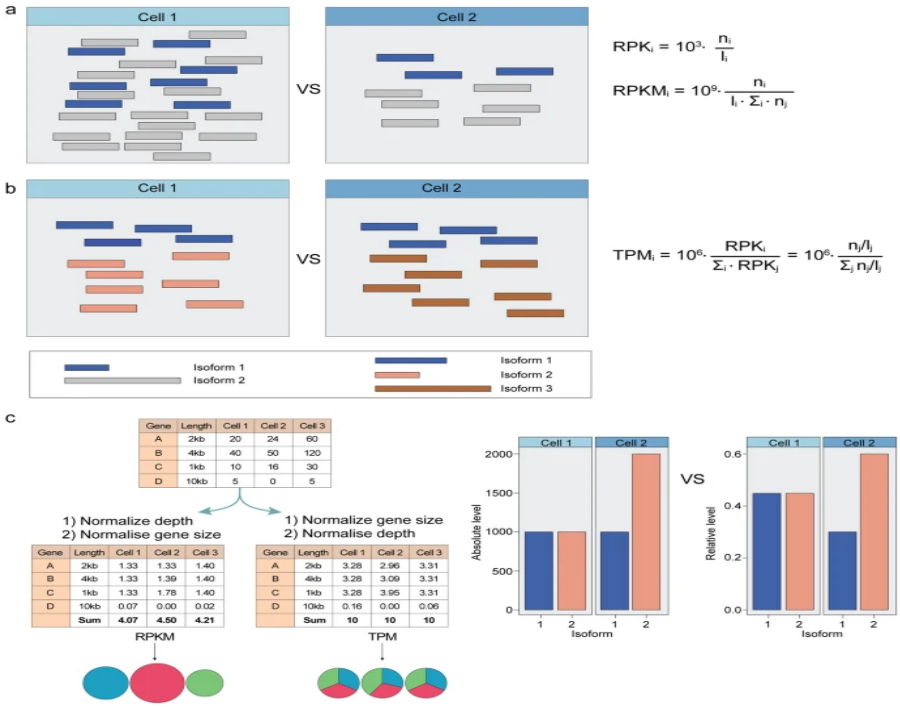
Before analyzing the single-cell gene expression data, I must ensure that all cellular barcode data are from viable cells. Cell QC is commonly performed based on three QC covariates: the number of transcripts per cell (n_Counts), the number of genes per cell (n_features), and the fraction of counts from mitochondrial genes (mt.percent). The distributions of these QC covariates are examined for outlier peaks that are filtered out by thresholding. These outlier barcodes can correspond to dying cells, cells with broken membranes, or doublets. For example, barcodes with a low count depth, few detected genes, and a high fraction of mitochondrial counts are indicative of cells whose cytoplasmic mRNA has leaked out through a broken membrane, and thus, only mRNA located in the mitochondria is still conserved. In contrast, cells with unexpectedly high counts and a large number of detected genes may represent doublets. Thus, high n_Count thresholds are commonly used to filter out potential doublets.

3. Normalization

Each count in a count matrix represents the successful capture, reverse transcription, and sequencing of a molecule of cellular mRNA. Count depths for identical cells can differ due to the variability inherent in each of these steps. Thus, when gene expression is compared between cells based on count data, any difference may have arisen solely due to sampling effects.

Normalization addresses this issue by scaling count data to obtain correct relative gene expression abundances between cells.

There are two common normalization methods. Transcripts per million or reads per kilobase per million mapped released. The difference between the two are operation orders. Transcripts per millions of firsts normalize to gene length and normalize to total transcripts. On the other hand, reads per kilobase per million mapped reads are the opposite. Normalization is an important part of sequencing because it removes the technical variability that includes technique variability or biological variability which causes inaccuracy or invalid data. Since sequenced RNA has differences due to technical problems. Normalization is applied to remove the difference in gene expression to produce correct data.



Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 50, 1−14 (2018). https://doi.org/10.1038/s12276-018-0071-8

4. Identify variable genes

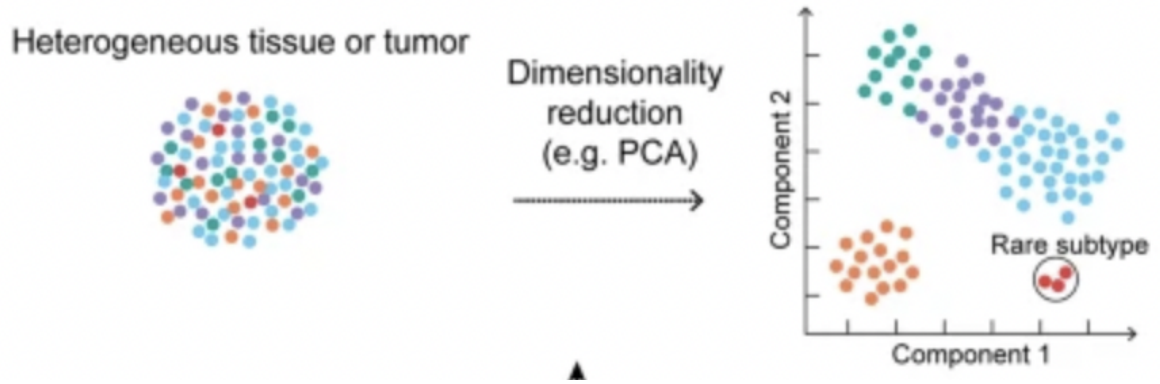
One single-cell RNA-seq dataset can contain expression values of up to 25,000 genes. Most of these genes are not informative for a given scRNA-seq dataset because they are consistently shared by all live cells, and many genes will essentially contain zero counts because they are silenced in specific cells. Even after filtering out these zero-count genes in the QC step, the feature space for a single-cell dataset can have over 15,000 dimensions. To ease the computational burden on downstream analysis tools, reduce the noise in the data, and visualize the data, one can use several approaches to reduce the dimensionality of the dataset.

The goal of identifying variable genes is to identify genes that show significant variation across different cell types, facilitating in distinguishing and clustering different cells. Mean variance is a common way to identify variable genes. The method is completed by plotting the variance of gene expression against the mean expression level among all the other samples or cells. Genes

that have a high variance value are considered more variable and are selected for downstream analysis.

5. Principal component analysis

After selecting variable genes, the dimensions of single-cell expression matrices can be further reduced by dedicated dimensionality reduction algorithms. These algorithms embed the expression matrix into a low-dimensional space, which is designed to capture the underlying structure in the data in as few dimensions as possible.
Principal component analysis (PCA) is a powerful and widely used technique for compressing high-dimensional datasets. In high-dimensional datasets, the number of features can be very large, making it extremely hard to visualize and analyze the data. PCA is a popular method that simplifies the data by finding a lower representation of the data but still preserving the important features and values of the dataset.



Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 50, 1–14 (2018). https://doi.org/10.1038/s12276-018-0071-8

6. Clustering analysis

After dimension reduction, clustering analysis is applied. Organizing cells into clusters is typically the first intermediate result of any single-cell analysis. Clusters allow us to infer the identity of cell populations. Clusters are obtained by grouping cells based on the similarity of their gene expression profiles. Expression profile similarity is determined via distance metrics, which often take dimensionality-reduced representations as input. A common example of similarity scoring is Euclidean distances that are calculated on the PC-reduced expression space.

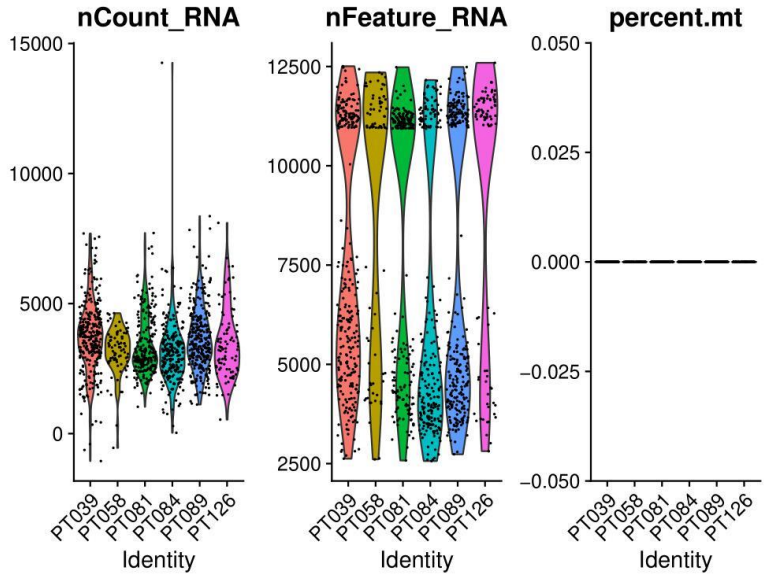
7. Annotation

On a gene level, clustered data are analyzed by finding the gene signatures of each cluster. These marker genes characterize the cluster and are used to annotate it with a meaningful biological label. This label represents the identity of cells within the cluster. The goal of annotation is to find characteristic markers for cells since each individual cell has its own unique marker.

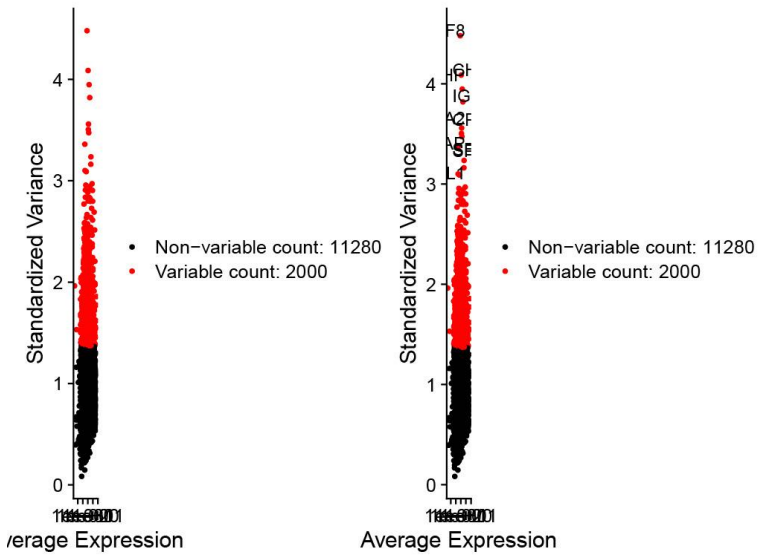
Analyzing scRNA-seq data from triple-negative breast cancer patients

I am applying the above-summarized workflow to analyze a dataset from triple-negative breast cancer that has been sequenced to reveal the composition of different cell types in the tumor. I installed the sequence completed RNA package from the National Center for Biotechnology and used dplyr, patchwork, and Seurat package in R studio to help us complete the workflow and correctly plot the result.

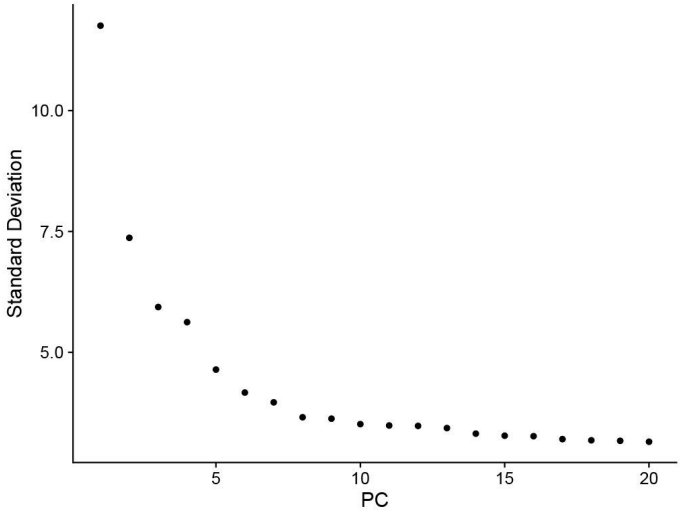
The first step of single-cell RNA sequencing is quality control, where I used a violin plot to show the nCount_RNA, nFeature_RNA, and mitochondria percentage of the sequenced cells. From the violin plot, I can tell that the data set has been normalized from a uniform distribution of nCount_RNA and nFeature_RNA, and controlled mitochondria percentage. The number of genes detected in nFeature RNA is relatively high due to doublets caused by two cells in one single droplet during single-cell analysis.



Next, I find the variable of these genes by applying variance stabilizing transformation to the genes. In this step, the dataset was filtered to keep only genes that are “informative” of the variability in the data. Thus, highly variable genes are often used. Variance stabilizing transformation (VST) is a statistical technique that is used to transform data in a way that the variance of the transformed data is approximately constant across the range of the data. This is often done to prepare the data for further analysis or modeling, as many statistical methods assume that the variance of the data is constant. In this process, I compute a score for each gene and choose the top two thousand most variable genes for the next step, PCA.

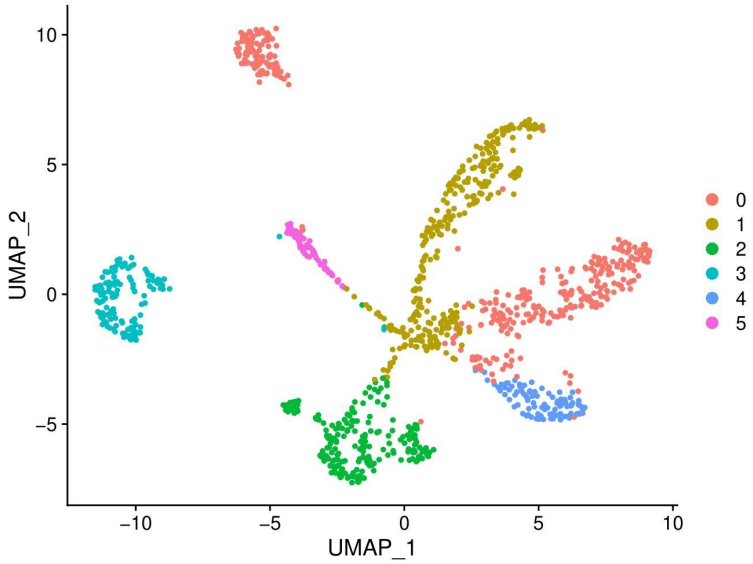


Now I perform principal component analysis (PCA) on the genes that are only highly variable in the dataset. Reduced dimensions are generated through linear combinations of feature space dimensions. PCA visualizes the graph of the first two principal components from the dataset which generates a scatterplot where the X-axis represents a gene and Y-axis represents the loadings of the genes. Using an elbow plot, it identified the different inclusions of variance in each component. Usually, the first principal component contains the most variance within the dataset. The variance decreases in the following components and the first 8 components cover almost all the variance. Hence, I only need the first 8 components for dimension reduction.



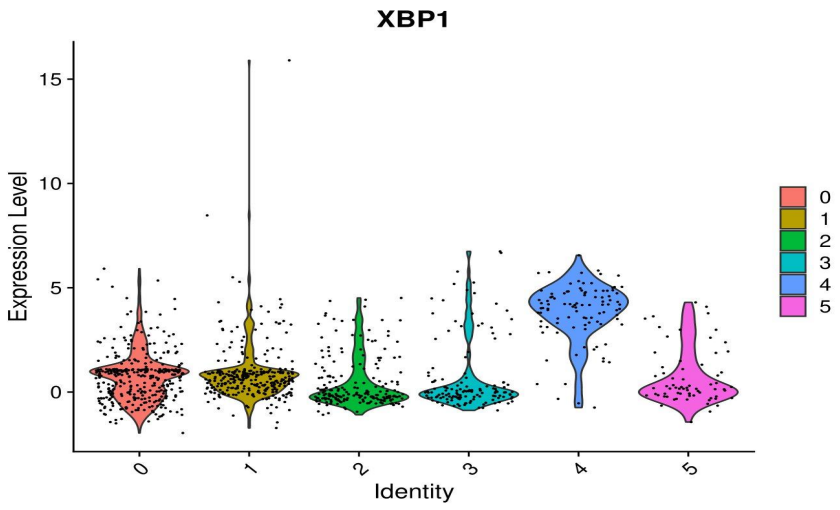
After PCA I apply dimension reduction. I only use the first 8 components of the data because the components after that do not provide useful information for this process. This means that cells that are close to each other in the high-dimensional space defined by the first eight principal components will be considered neighbors. I also use the function “find clusters” to organize different cell types into different clusters, which calculated the Euclidean distance among points and clustered them according to the distances. After this, I graphed them using

the uniform manifold approximation and projection (UMAP) algorithm, a technique for dimension reduction that organizes cells that have similar identities to a single cluster. From the graph, I can tell that there are 6 different clusters that represented 6 different types of cells.

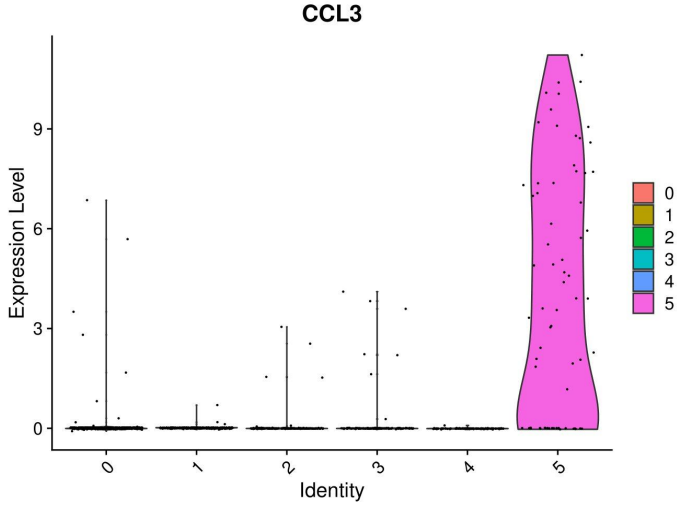


Since I know most cell types that are included in the tumor, I can create a violin plot of a specific gene in the dataset to ascertain the expression level across different clusters. For the unknown cell types, I can apply the “find marker” function to find the most variable gene to annotate the unknown cell types. From the violin graph the higher the expression level the more I can confirm the cell identity of each cluster. To figure out the single cell type expression cluster that corresponds with the correct gene, I use The Human Protein Atlas website to find more information about the genes.

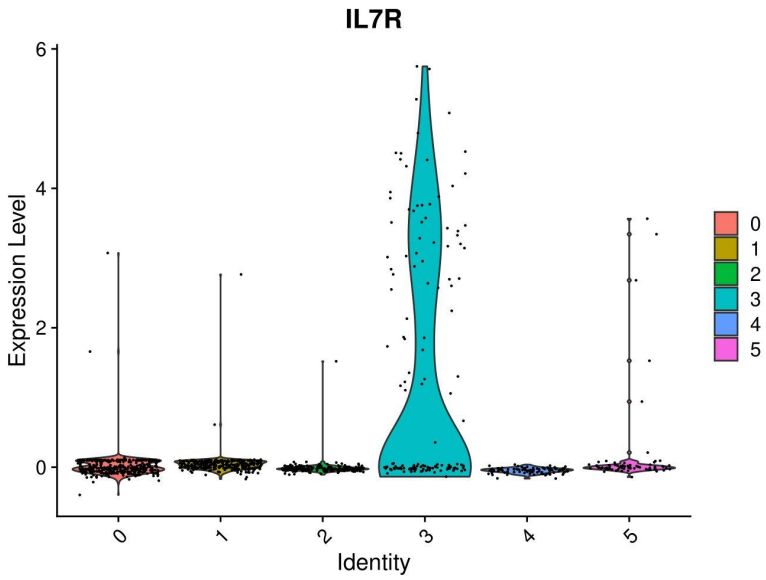
From the violin graph of gene XBP1, I found it is uniquely expressed in cluster 4. XBP1 is a gene that is expressed by B cells to differentiate into plasma cells. Therefore I can confirm that cluster 4 is a B-cell cluster.



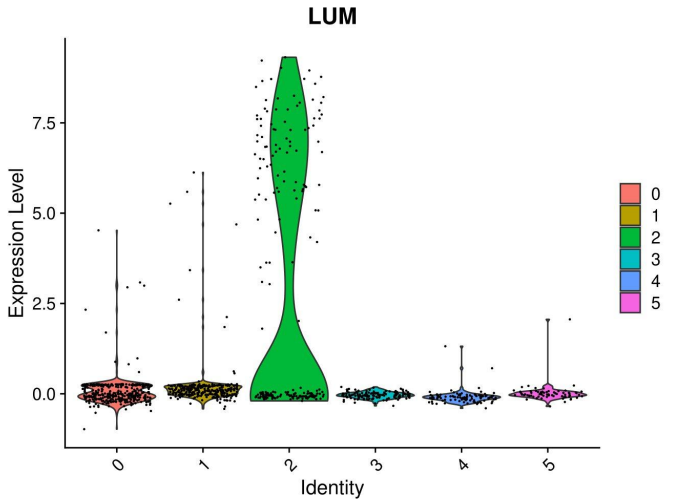
Similarly, based on gene expression of CCL3, the cytokine chemokine ligand 3, I concluded that cluster 5 is myeloid cells.



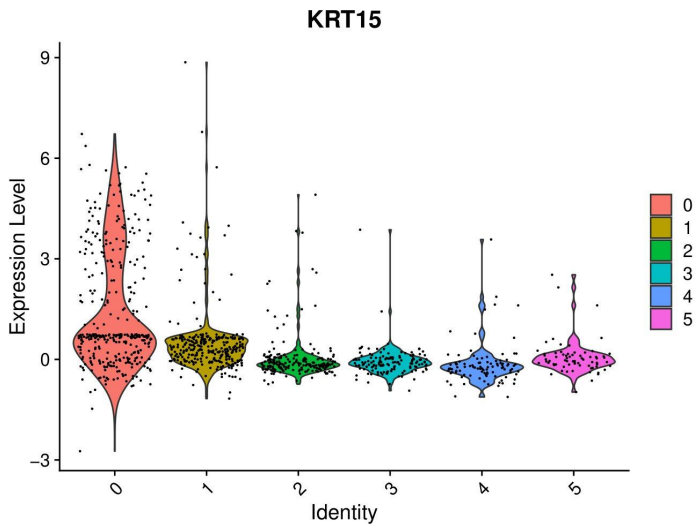
The graph of gene IL7R is the most dominant in cluster 3 and so cluster 3 is T cells.



The graph of gene LUM, which encodes lumican for extracellular matrix, indicates the highest expression level in cluster 2 and is identified as connective tissue cells.



The gene KRT15 has the highest expression level in cluster 0 and is identified as a Cancer cell. However, cancer cell clearly diverged into two major populations in UMAP, which suggested that the cancer cell shared different characteristics and likely formed heterogeneous population.



Finally, I annotate different clusters by assigning each cluster a cell name based on its own expression and labeling its name next to its cluster. Looking As a result, we revealed different types of cells with different clusters.