Identify cell types/subtypes of microglia that are dysregulated by induced the antiretroviral drugs

Mustafa AbuElqumsan, Xin Liu, and Prof. Shau-Jun TANG

2024-04-29

Contents


```
knitr::opts_chunk$set(error = FALSE, include=FALSE,  eval=FALSE, echo=TRUE, warning=FALSE)
```

The structuring of data

we had generated our data from 10x genomics, we had 2463920 single cell that were sequenced on the lllumina NexSeq 500, the returning unique molecular identified (UMI) is it count matrix from cell. the values in this matrix represent the number of molecules for each feature that means it is a gene in row that are detected in each cell.

Creating the object to maintain the data

The object has been created using Seurat package, to save all data required count matrix and the resulting analysis as like scaled data, PCA, and clustering results

Merge individual Seurat Objects into Single Seurat Object

Due to our goal is to matching shared cell types across data, even though the integration could result in a loss of biological resolution. But we had preferred here to done the intergration for our data to check up the effect of integration into the efeciancy of clustering of Seurat package.

Calculate the percentage of counts originating from a set of features "Visualize feature-feature relationships" and then Visualize QC metrics that are used to fillter cells.

By this step we had filtered cells based on our criteria :- * The number of unique genes detected in each cell * Similarly, the total number of molecules detected within a cell. * Percentage of reads that map to the mitochondrial genome

Normailization of the data

we removed unwanted cells from dataset, by such step we trying to employ a global-scalling normalization method "LogNormalize"that normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (10.000 by default) and log-transforms the result

Identification of highly vaiable features(Feature selction)

we can show a subset of features that exhibit high cell-to-cell variation in dataset.

Scaling the data

we applied a linear transformation as the standard pre-processing step before to dimensional reduction done it, to shifts the expression of each gene and by that the mean expression across cells is Zero, and Scale the expression of each gene by that the variance across cells is 1. We could regress out heterogeneity associated with cell cycle stage as well

we can visualizing both cells and features that define the PCA

we can explore of the primary source of heterogeneity in our data we trying to decide which PCs to include for further downstream analysis, by which both cells and features are ordered according to their PCA scores. this for exploring correlated feature set

Elbow Plot

Is a ranking of principle component based on the precentage of variance explained by each one

clustering this dataset which would return predominantly batch-specific clusters

The principle of clustering here is distance by k-nearest neighbour (KNN)graph with edges drawn between cells with similar feature expression patterns, and then attempt to partition this graph into highly interconnected communities, that can happened by euclidean distance in PCA sapce, and refine the edge weights between any two cells based on the shared overlap in their local neighborhoods (jaccard similarity). and then clustering process could been done by apply modularity optimization techniques such as the Louvain algorithm or SLM to iteratively group cells together, with the goal of optimizing standard modularity function, (UMAP/tSNE) are non-linear dimensional reduction techniques, to visualize and explore the dataset, but the goal of these algorithms is to learn underlying sturcture in dataset, in order to place similar cells together in low-dimensional space. therefore, cells that are grouped together within graph-based clusters.these methods aim to preserve local distance in dataset(ensuring that cells with very similar gene expression profiles co-localize), but often do not preserve more global relationships. But we should take in our considerations to avoid biological conclusions solely on the basis of visualization techniques

Identifying differentially epressed features

we looking to identifying clustering via differential expression (DE), by identifies positive and negative markers of a single cluster, compared to all other cells. we can use several tests for differential expression. e.g. the ROC test returns the classification power for individual marker ranging from 0-random, to 1-perfect. and then we show expression probability distributions across clusters, visualize feature expression.

we find the epression heatmap for given cells and featurs

We are plotting the top 20 markers for each cluster

Assigning cell type identity to clusters

We had used canonical markers to easily match the unbiased clustering to known cell types