WWIEM Bioinformatics Service

97 Lisburn Road|
Belfast, UK| BT9 7BL
+44**********

Report for scRNA-seq Service

Customer Name,
Customer Address
City, State Zip
UK

QUEEN'S UNIVERSITY BELFAST
WELLCOME-WOLFSON INSTITUTE FOR EXPERIMENTAL MEDICINE

Comprehensive Data Analysis Report

This report contains the results of a completed scRNAseq experiment. Dataset are of Peripheral Blood Mononuclear Cells, containing 2700 single cells that were sequenced on the Illumina NextSeq 500.
Data analysis was performed using R version 4.1.1. Quality Control, analysis and data exploration were accomplished using Seurat package version 4.0.4.

Folder containing all data can be found here.

Table of Contents

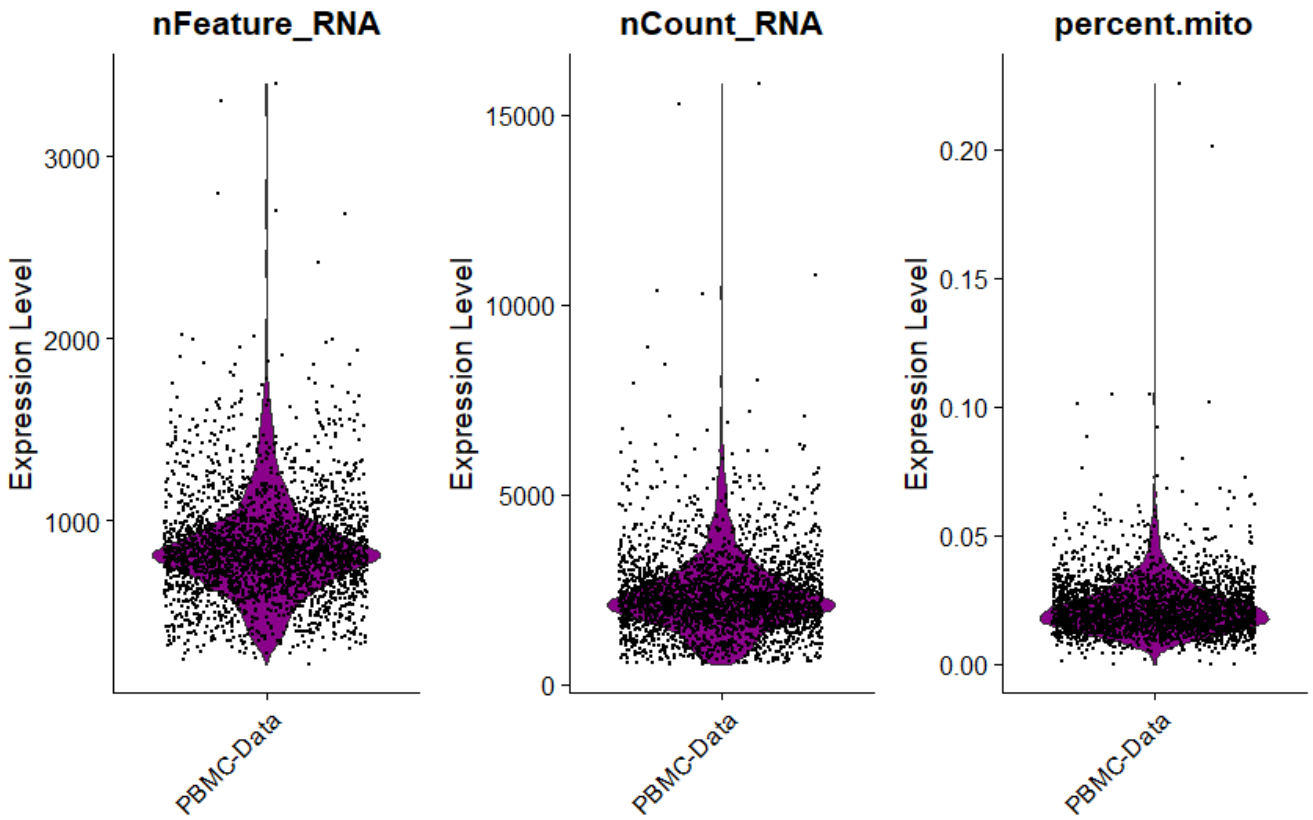
QC & Cell Selection

QC Metrics

Sequencing information:
There are 33694 genes and 4340 cells available for the upcoming analysis.

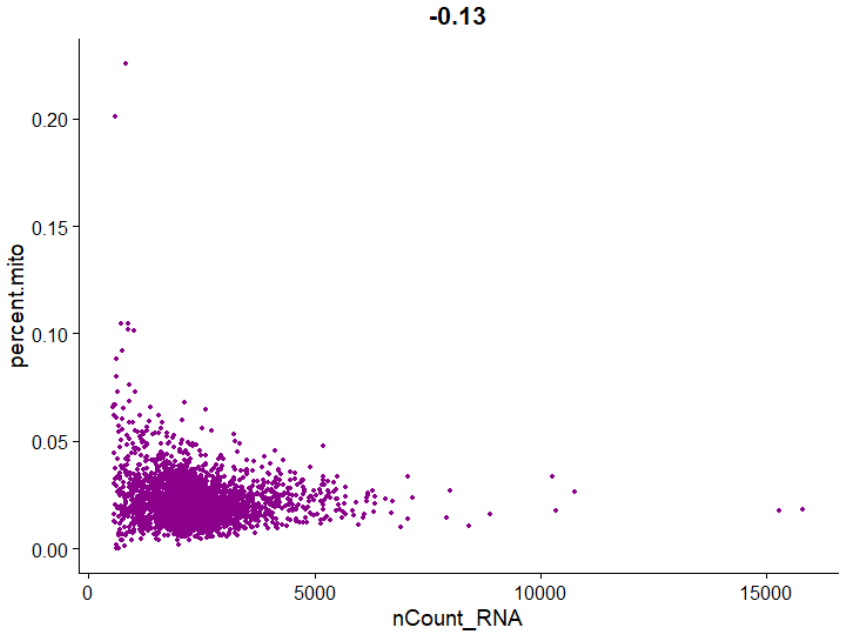
Violin Plot

This graph represents a violin plot showing the distribution of the expression level across different parameters.
The “n-Feature” plot shows the number of detected genes for every cell.
The “n-Count” plot shows the number of detected UMI for every cell.
The “percent.mito” plot shows the percentage of mitochondrial genes for every cell.
Black dots represent the values for individual cells, and the purple shapes show the distribution of the data.
For example, for “n-Feature” plot you can notice that the average number of genes per cell is about 900 and most of the cells have roughly around 700-1100 genes.

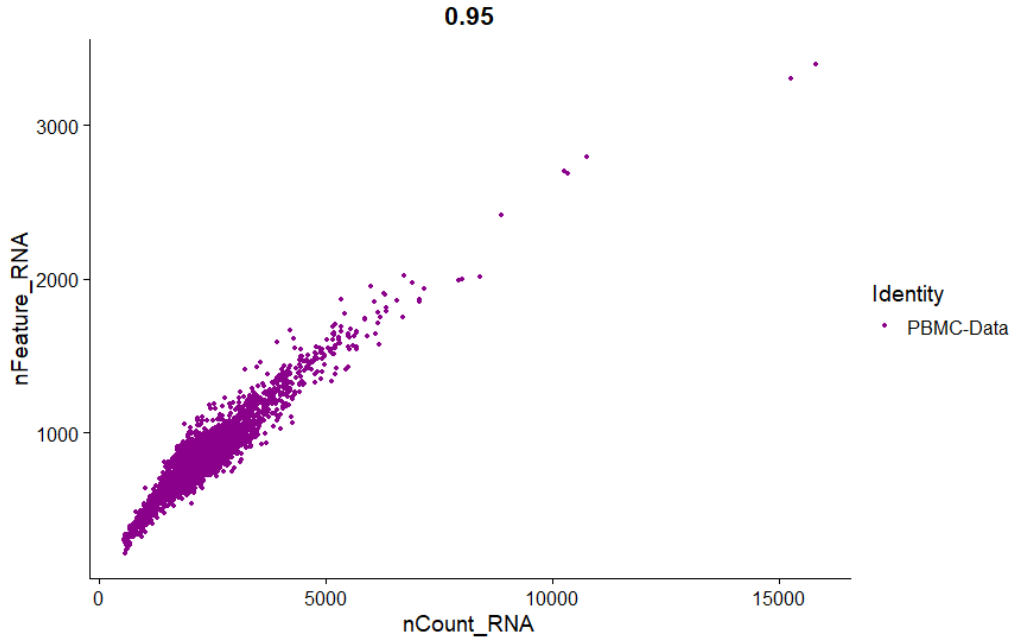


Feature Scatter plot

This graphs represent a feature scatter plot to visualize the relationship between “nCount-percent.mito” (a) and “nCount-nFeature” in (b).
One can notice a relatively low subset of cells with an aberrant level of high percentage of mitochondria (percent.mito) and a low level of UMI (nCount). These information will be crucial for filtration.



(a)



(b)

Variable Gene Detection

Variable Genes Detection across the single cells

In this step, highly variable genes are calculated and targeted for downstream analysis. This is achieved by calculating the average expression and dispersion for each gene, placing these genes in bins, and then calculating a z-score for the dispersion in each bin.

Below an overview of the genes detected after the computation.

Detected Genes	Mean	Variance	Standardized Variance
AL627309.1	0.003411676	0.003401325	0.9330441
AP006222.2	0.001137225	0.001136363	0.9924937
RP11-206L10.2	0.001895375	0.001892500	0.9627290
RP11-206L10.9	0.001137225	0.001136363	0.9924937
LINC00115	0.006823351	0.006779363	0.9062135
NOC2L	0.107278241	0.159514698	0.7849309

Data Scaling

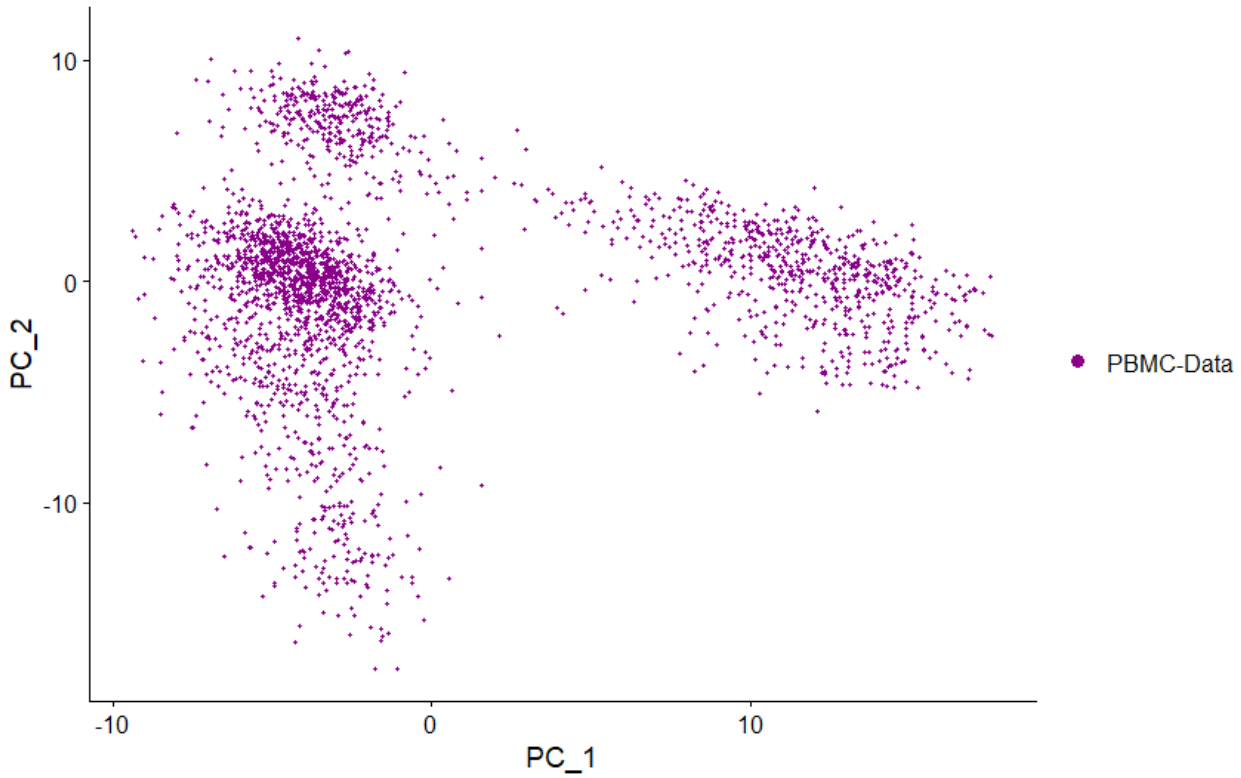
Technical noise, but also batch effects, or even biological sources of variation are all sources of variation. These unwanted variations must be removed to improve downstream analysis and clustering. To smooth out the effect of these signals, we built linear models to predict gene expression based on user-defined variables. The scaled z-scored residuals of these models will be used for dimensionality reduction and clustering.

Dimensional Reduction

We used the highly variable genes for centering and scaling. Then, we performed a principal component analysis on them using *RunPCA* function from *Seurat* package. We removed the signal-to-noise ration by selecting the significant principal components, according to p-values, using the *JackStraw* function that uses permutation tests.
Cells were then clustered and embedded into a graph structure in PCA space.
The clustering results is shown and explored below through different graphs.

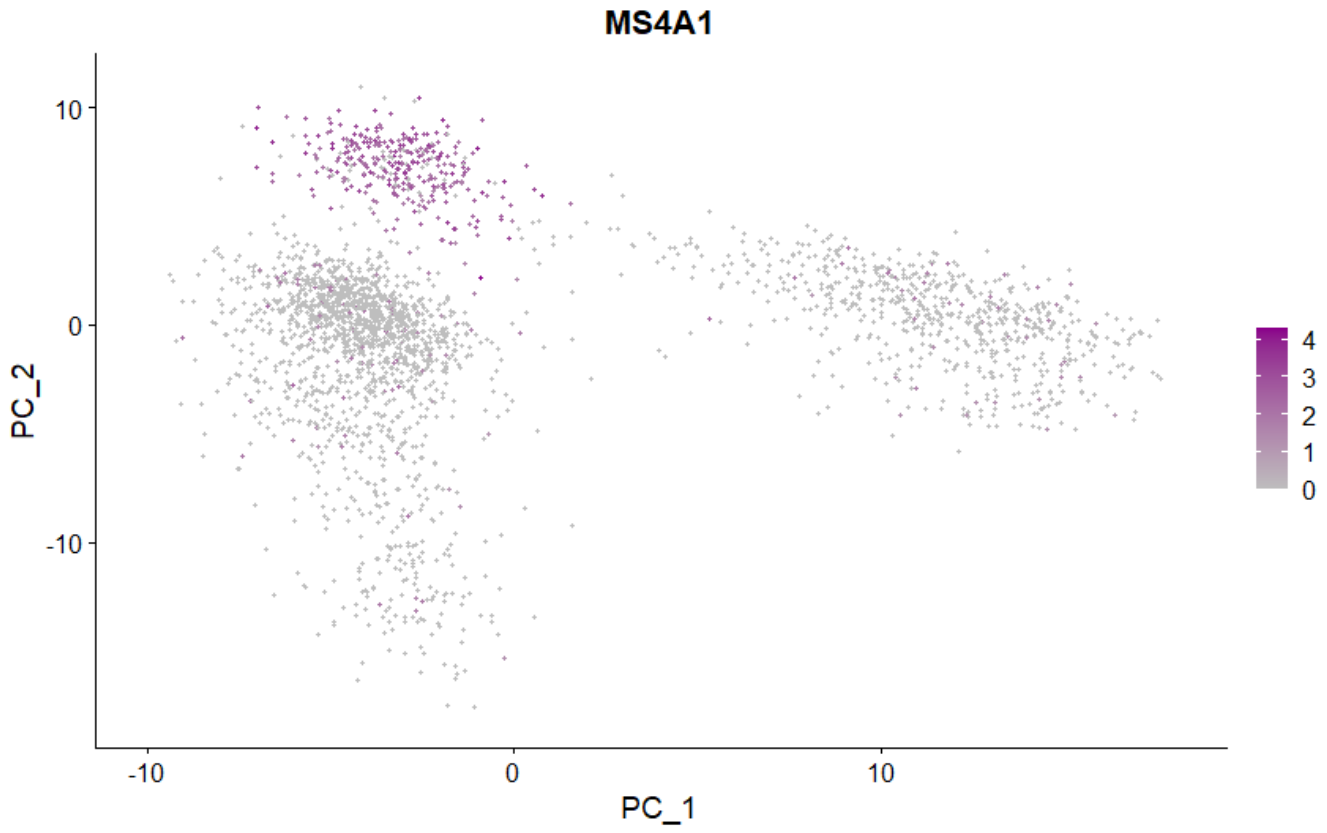
DimPlot

The graph represents a 2D scatter plot. Each point shows a cell and its position is determined by the reduction performed.



FeaturePlot

The graph represents a dimensional reduction plot where we coloured cells by a quantitative feature, here the feature was “MS4A1”.
We can notice in which cluster the feature is highly present.

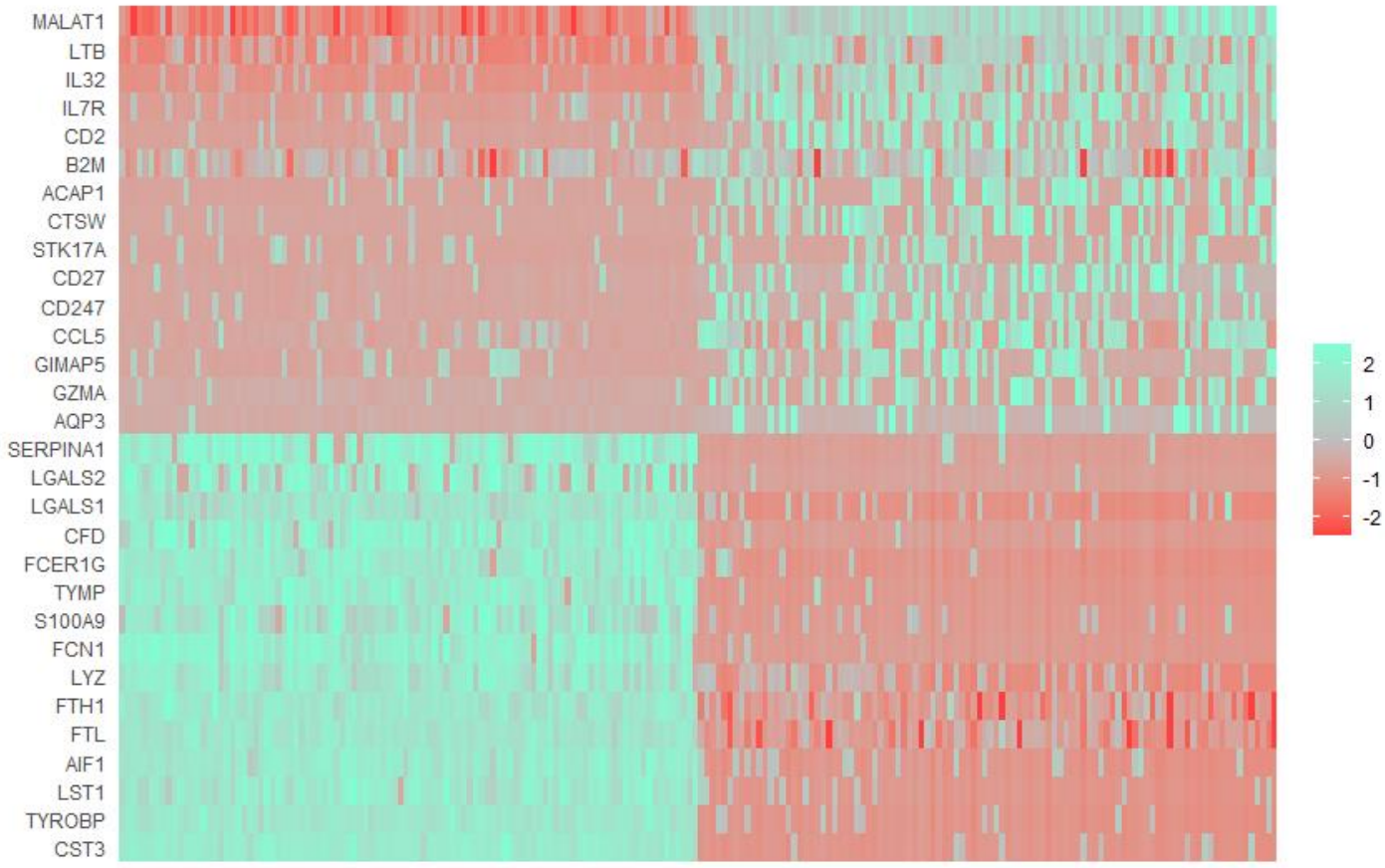


Variable FeaturePlot

This graph represents a variable feature plot which shows the variable features in the data. The purple points represent the variable count (2000) and the grey points show the non-variable counts (11714). Here we labelled the 10 most highly variable genes.

Dimheatmap

The graph below represents a heatmap focusing on the principal component PC-1.
Here, the genes and cells are ordered by the scores of the PCA.



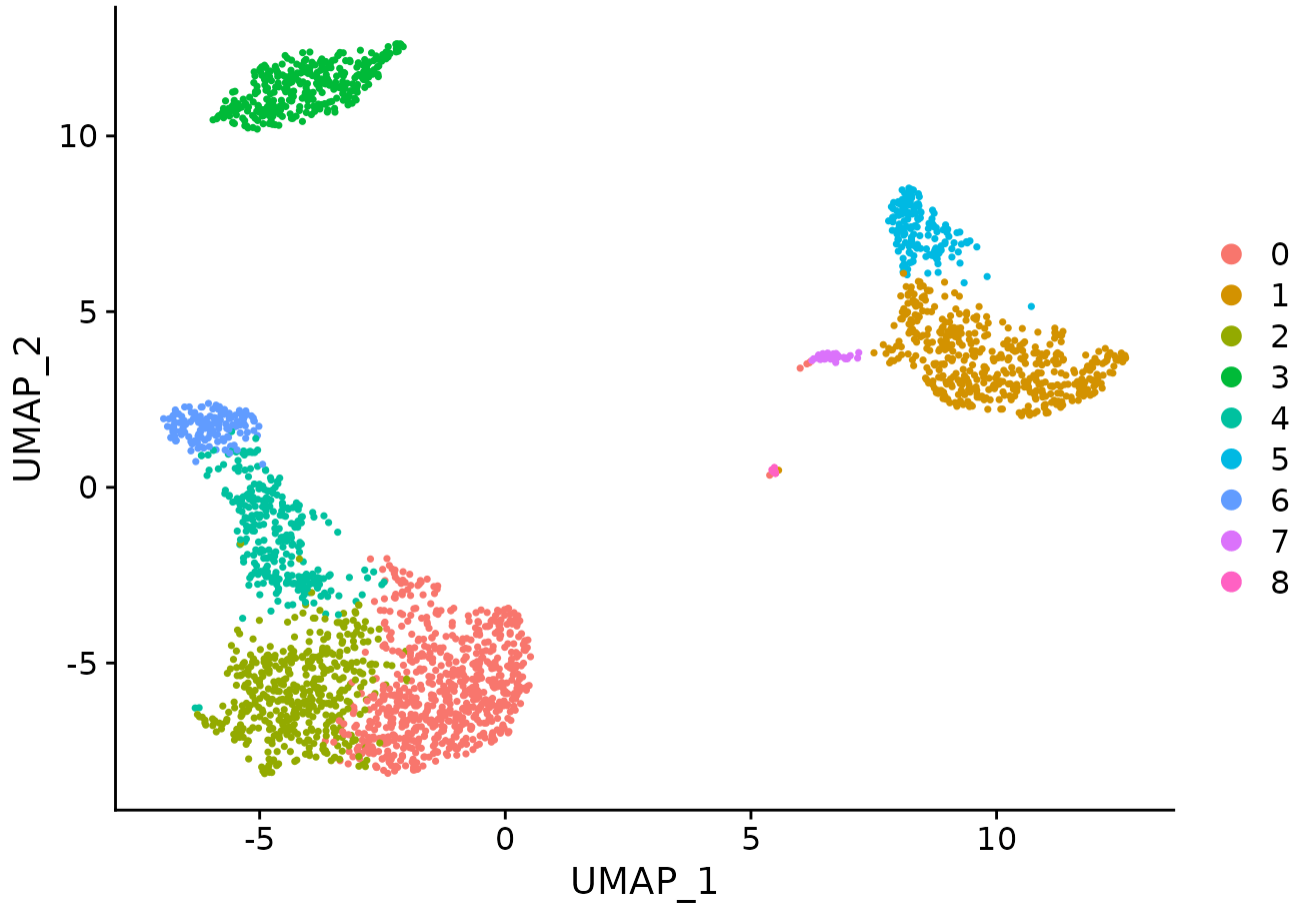
Cell Clustering

Graph-based clustering

Seurat is used for clustering which apply a graph-based approach through embedding cells in a graph structure. Mainly, it draws edges between cells with similar gene expression profiles, then attempts to split this graph into highly interconnected communities.
Based on this, a KNN graph is first built according to the Euclidean distance in the PCA space, and the weights of the edges between two cells are refined according to the shared overlap in their local neighbourhoods (Jaccard similarity). We use modularity optimization techniques, here Leuven algorithm, to merge the cells iteratively and optimize the standard modularity function.
We use *FindNeighbors* function to compute the nearest neighbor graph and compute SNN, and *FindClusters* function to compute nodes, edges, run Louvain algorithm and extract the number of communities.

Non-linear Dimensional Reduction (UMAP)

Here, we can visualize clustered cells.



DGEA

Differential gene expression is performed by *Seurat* through defining positive and negative markers of a single cluster compared to all other cells. A gene must be detected at a minimum of 25% in either of the two group of cells.
We can either perform this process for all clusters, or test groups of clusters vs. each other, or against all cells.

An overview of all markers of the first cluster

Marker	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
RPS27	1.665366e-115	0.6903363	0.998	0.992	2.283883e-111
RPS12	1.641196e-114	0.6865994	1.000	0.992	2.250736e-110
RPL32	5.489165e-114	0.6048309	0.998	0.995	7.527841e-110
RPS6	8.769487e-111	0.6376655	1.000	0.995	1.202647e-106
S100A4	1.051721e-107	-2.0927252	0.637	0.861	1.442331e-103

Complete list can be found here.

An overview of all markers of cluster 5 from clusters 0 and 3

Marker	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
TYROBP	2.785763e-256	5.086003	0.994	0.120	3.820396e-252
S100A8	1.233247e-253	5.812617	0.979	0.097	1.691257e-249
S100A9	6.106459e-246	6.257452	0.996	0.169	8.374398e-242
LGALS2	1.316367e-243	4.085707	0.906	0.032	1.805266e-239
FCN1	1.768538e-240	4.192268	0.951	0.096	2.425373e-236

Complete list can be found here.

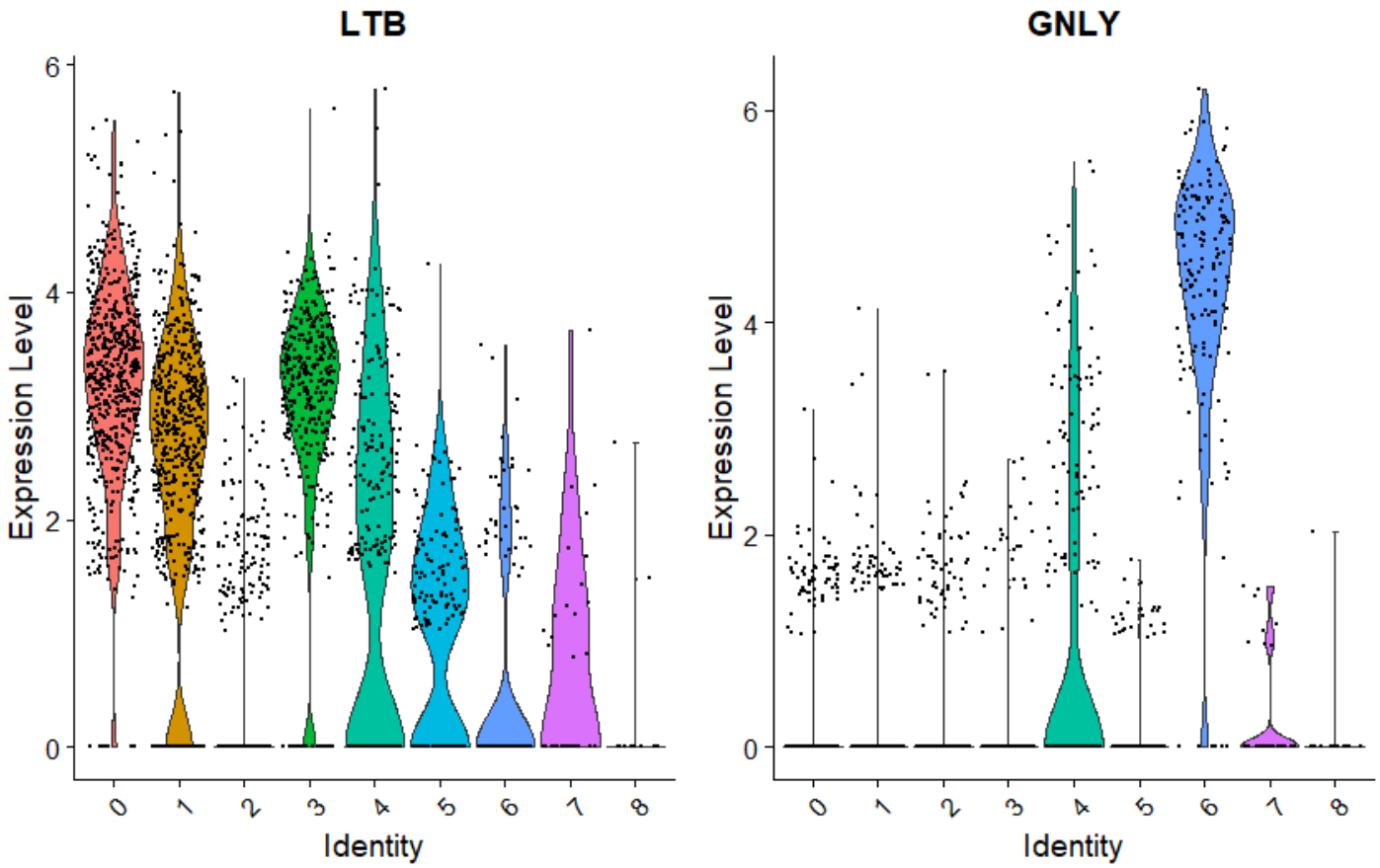
Markers for every cluster compared to all cells

List of all markers can be found here.

Visualisation of Marker Expression

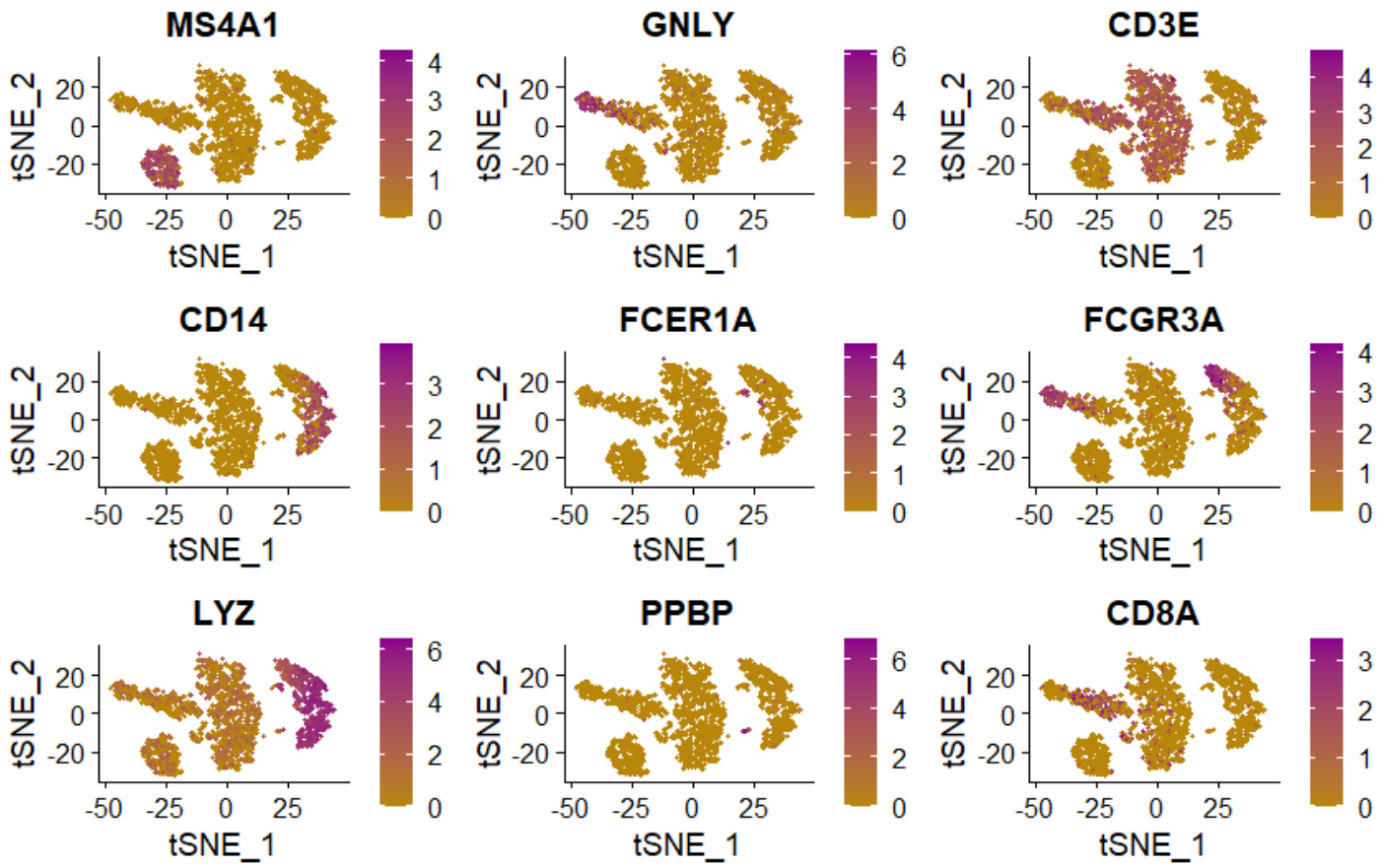
VlnPlot

It shows expression probability distributions across the clusters. Here an example of "LTB" and "GNLY".



FeaturePlot

It shows gene expression on a tSNE plot. Example of 9 different genes.



Heatmap

It shows an expression heatmap for all cells and the top 20 markers for the eight clusters.

Cell type identity assignment

In this step, we use canonical markers to match our clusters to known cell types. Mainly, after identifying the differentially expressed genes on each cluster, we annotate cell clusters according to curated known cell markers.
Here we used Panglaodb for retrieving our canonical markers.