Analysis of single-cell RNA-Seq data from spinal cord

Alessia Valenti

Università degli Studi di Milano, Politecnico di Milano

The spinal cord is a tissue belonging to the nervous system, a complex network of different types of cells which through this structure is able to transmit information to the peripheral tissues (1). Due to the complexity of this network, single-cell RNA-seq analysis of cells originating from this tissue may provide precious information about the healthy and diseased tissue. Here we analyze data from healthy spinal cord of *Mus musculus* with the purpose of optimize the parameters for the identification of the principal types of cells in the tissue.

Single-cell RNA-Seq is a technique that allows the isolation of the trascriptome of each cell in a sample. Its potential it's related to the capacity to identify and distinguish between cellular types and sub-types and account for biological variability as well as detect rare, underrepresented cells in heterogeneous tissue. However, it still is a relatively new technique and the analysis pipelines differ very much depending on the workflow used for generating the data and even after this generation.(2). Due to its low cost, one of the most widely used methodology for this data generation is the droplet-based 10X Genomics protocol which isolate cells in single droplets where reagents for cell lysis, barcoding, reverse transcription and tagging through Unique Molecular Identifier (UMI) are enclosed. These steps are followed by PCR amplification and Illumina short-read sequencing.(2) Reads are then grouped by UMI, so to account for the amplification effect of the PCR and then counted by alignment to a reference genome. From here on, the count matrix obtained can be analyzed with several software tools such as Seurat, an R package for assessment of cell quality, analysis, and exploration of single-cell RNA-seq data.(3) Our investigation starts from the count matrix of the genes and include the recommended steps(4) for the analysis in Seurat of single-cell RNA-Seq data that leads to the clustering and identification of cell types in the spinal cord tissue from the species *Mus Musculus*.

Materials and Methods

Spinal cord samples were retrieved from 7 individuals of adolescent *Mus Musculus* with postnatal age P20-P23. The SRA for the experiment is SRA667466 with SRR6854181 as specific accession code for our data (5). Reads were produced processing the cells with the 10X Genomic Chromium Single Cell Kit Version 1 and sequencing with Illumina HiSeq 2500 (5, 6). Reads were aligned to the genome assembly GRCm38 and GENCODE v. 27(7) was used as genome annotation in order to generate the counts(8). The count matrix was retrieved from the Panglao database.(5) For the following analysis we used the R library Seurat v3.2.1(3) , integrated with the packages dplyr v1.0.2(9) for data manipulation, patchwork v1.0.1(10) and ggplot2 v3.3.2 (11) for data visualization.

We started loading the Rdata object downloaded, corresponding to a sparse count matrix, and creating a Seurat object. In doing so we used the two parameters 'min.cells = 3' and 'min.features = 200' to pre-filter genes and cells so that only genes expressed in at least 3 cells and cells that express at least 200 genes are kept. After retrieving the percentage of mitochondrial genes, we visualized its distribution together with the distributions of the number of total counts and the number of genes in each cell (fig. 1). We can observe two spikes in the distributions of number of detected genes, one at around 800 and one at around 1800-1900 genes. The counts number is variable as well and we can observe a higher spike at lower counts (around 1000 counts) and a second spike at around 5000 counts. The percentage of mitochondrial genes is low on average with a single spike at around 1%, indicating an overall a good quality of the cells, with only few of them having leaked the cytoplasm and retained only the mitochondrial genes. To better assess the quality of cells, since one of these properties alone won't be enough informative about the quality of the cells we used scatter plots to see the correlation between the number of counts in and the percentage of mitochondrial genes, and the number of counts and the number of detected genes. From the first graph (fig. 2a) we can visually assess that when the percentage of mitochondrial genes is high, the number of counts in the cell is low, probably due to the cell leaking the cytoplasm because of cell manipulation during the experiment. Another event that commonly happens, is the processing of two cells in one single droplet. This results in a high number of counts and a high number of features attributed to a single cell. This can be easily verified in fig. 2b: we can here observe that after around 3500 features cells with higher number of counts and features are less frequent, likely indicating this type of problem.

After this exploration the dataset was subset selecting cells which respected the following threshold:

- more than 300 genes detected but less than 3500, so to exclude cells that have leaked or problematic cells (with less than 300 genes) and doublets (with more than 3500 genes)
- a percentage of mitochondrial genes lower than 3
- a total number of counts higher than 500, so cells with a significant number of counts and that haven't leaked.

These thresholds, that were selected after the observation of the scatter plots and a trial and error process, revealed themselves to be the ones that better fit the data, cutting

[2] E-mail: alessia.valenti@studenti.unimi.it

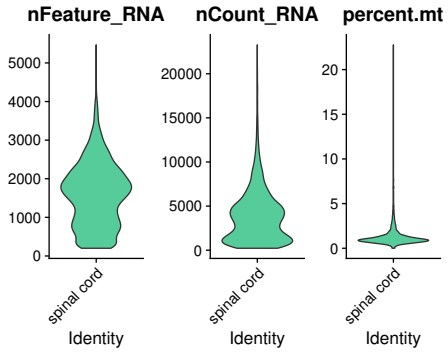

Fig. 1. Violin plot showing the distributions of the number of counts, number of genes and percentage of mitochondrial genes.

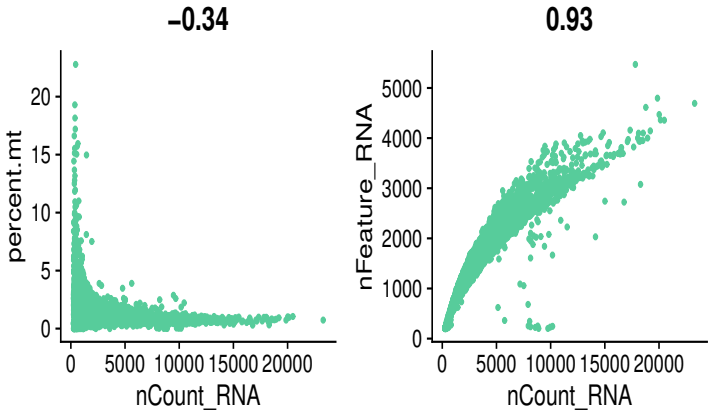

Fig. 2. (a) Scatter plot of the number of counts against the percentage of mitochondrial gene counts. At higher percentages of mitochondrial counts correspond lower counts. Pearson's correlation coefficient appeared to be -0.34 **(b)** Scatter plot of the number of counts against the number of genes detected in each count. The Pearson's correlation coefficient is very high: 0.94

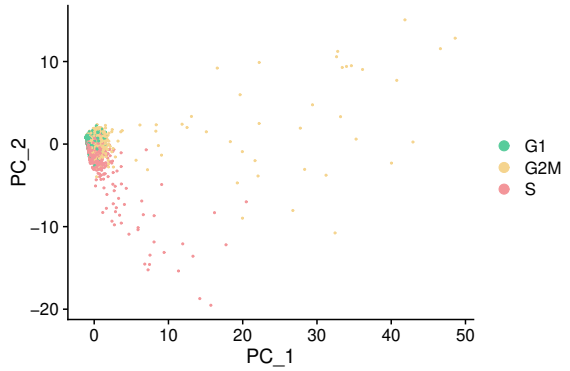

Fig. 3. Projection of the first two principal component of the PCA performed on cell cycle genes. Each cell was colored based on its phase of belonging, determined by its cell cycle score.

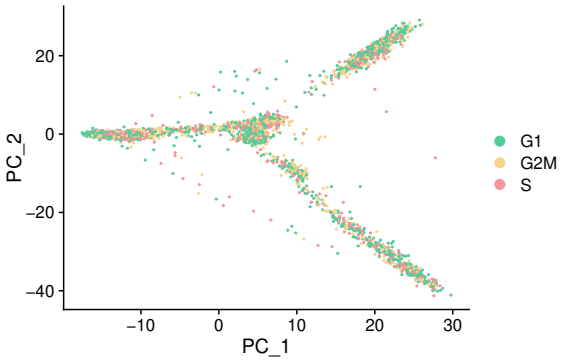

Fig. 4. Projection of the first two principal components of the PCA performed on all 4000 selected genes.

the noise from doublets and leaked cells without losing information. We reduced the number of cells from an initial number of 5129 to 4477, cutting out 652 cells.

Subsequently we normalized the data with the LogNormalization method where feature counts for each cell are divided by the total counts for that cell and multiplied by the scale factor, which is selected to be 10000. Then the results is natural-log transformed using the logarithm.(3) Since the total number of genes is very high, we select the top 4000 genes that are the most variable among all. Only after this step, we can scale the data so that the mean is 0 and the standard deviation is 1 for each gene among all cells.

Before proceeding with the Principal Component Analysis (PCA), we checked whether the cells cluster based on the cell cycle phases. In order to do so, the cell cycle score was computed and, based on it, to each cell was assigned a cell cycle phase and a PCA on cell cycle genes was performed. In fig. 3 we can observe the results by plotting the first two principal components: even if colors are more localized in certain regions, no clear cluster or distinction between cell cycle phases is detected and therefore we decided not to regress it out. The fact that no cell cycle effect is present in our cells is confirmed by the plots of successive principal components of this PCA (not shown) and by the plots of the principal components produced with PCA on all the 4000 selected genes (fig.4), which does not show any particular pattern. Still, the purpose of PCA is originally dimensionality reduction which reduce both noise and the gene expression space. In fact, Principal Component Analysis purpose is to linearly combine vectors into new vectors, called principal components, which variance is the largest between the possible principal components that are independent from the previous principal components. Thus, each principal component will have standard deviation that is smaller than the principal component that came before it.

This concept is clearly showed by the elbow plot (fig. 5): while first principal components have a high standard deviation, this diminish until the standard deviation almost do not change anymore. Through the elbow plot, we can choose the number of significant principal components to use for the clustering, in this case 22 which is also the number of the last principal component which change of percentage of variation

is more than 0.05%.

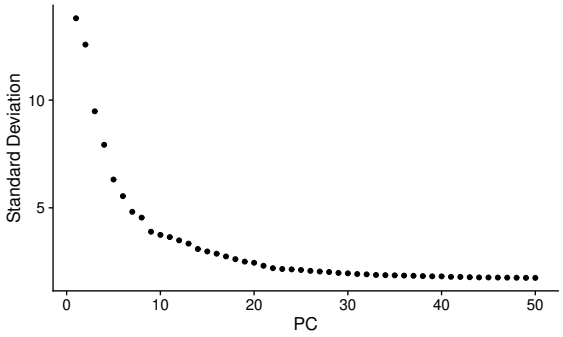

Fig. 5. Elbow plot. The standard deviation of each principal component is plotted together with the number of the PC. We can spot a flattening of the curve around PC20.

Next step would be clustering, that in Seurat is usually covered by two steps: the construction of a shared nearest neighbor (SNN) graph which exploit the k-nearest neighbor algorithm and then calculate the neighborhood overlap, and the identification of clusters based on the SNN graph. For the first step we selected the first 22 principal components and kept the default k = 20 for the k-nearest neighbor algorithm while, for the second step, we set the resolution parameter to 1.0. For visualization, we used the Uniform Manifold Approximation and Projection, a dimensional reduction technique particularly useful for this purpose. Early results are shown in fig. 6. We identified 22 clusters and for each one, the top 3 marker genes were selected based on the average logarithmic fold change of their expression in that cluster. We used them as input gene set in the Panglao search function, an online tool that helps find cell types where a certain set of genes are expressed (12). After the identification of cell types, these were assigned to the clusters, which overlapped, meaning that more clusters corresponded to the same cell type. Results are shown and discussed in the next section.

Results

With the SNN algorithm 22 clusters were identified, which converged to 8 putative cell types and 1 cluster of unknown type. In figure 7 the previous 22 clusters are tagged with the corresponding cell types. The distinction between clusters is coherent with the identified classes and, with the exception of few cells, no mixed clusters are present. All cells are of types expected to be in a tissue such as the spinal cord. The expression of some of the most meaningful marker genes are shown in 9 and 10. Especially in this last graph, aside from few exceptions, each gene displayed is specifically and highly expressed only in the clusters of the type of cell it is marker of, confirming the accuracy of our analysis.

Oligodendrocytes are the most numerous cells, despite the removal, performed in the original experiment and therefore inherited by our data, of more than 200,000 of these from the hindbrain (not analyzed here) and the spinal cord to better balance the number of oligodedrocytes in these regions. Their high presence, that comes to include 71% of all cells in the spinal cord, its due to their function of support in the long-range neurotrasmission (6).

Oligodendrocytes progenitor cells (OPC) are found near the oligodendrocyte cluster, even if mostly well separated from it, presumably due to the relationship the two classes of cells have. Neurons are divided in more clusters of cells, probably because each one contains neurons of a different type, such as GABAergic and glutamineric neurons (6). However, aim of this analysis was not to identify cell sub-types, so we grouped all neuronal clusters under the general category "Neurons". Of the other cell types identified, we paid particular attention to the two small clusters of ependymal cells and pericytes, which only contains 16 and 17 cells respectively. Their presence was not detected when using a low number of PC such as in fig. 11. For the creation of this plot, only 10 PCs where used and ependymal cells were not detected probably because absorbed by the bigger cluster of oligodendrocytes. Instead pericytes are not recognized when using a lower resolution such as 0.5 (fig. 12) because merged with the endothelial cells. Moreover, in both cases the OPC included the cluster that was previously tagged as unknown and the neuronal clusters are more compact, especially in fig. 11, probably because these configurations lack the information that separates different neurons.

Concerning the cluster tagged as "Unknown", one of its recurrent marker gene was *HMGB2* which is a marker gene for the cell cycle phase G2M, so it is possible that this cluster represent a group of cells, possibly OPC, that are proliferating. Increasing the resolution or the number of PC has neither improved nor worsened the final results, even when more initial clusters were found since anyhow they converged to the same 8 cell types.

For what concerns the number of variable genes selected, we notice that selecting a lower number of variable genes, for example 2000 instead of 4000, does not change the final number of cluster found. However, what changes is the definition of these cluster and in particular we observe that the ependymal cells are closer to the oligodendrocytes and the pericytes are closer to the endothelial cells (fig 13), which is consistent with the fact that they merge with the closer bigger cluster when the number of PC or the resolution are low. When increasing the number of variable genes used for PCA (in fig. 14), the endothelial cells split in two clusters. This may be due to the presence of sub-types of endothelial cells, however for the purpose of this study we preferred to show the graph obtained with the selection of the 4000 most variable genes. Anyhow, we obtained overall good results even when changing the number of genes used as input for PCA and changes were minimal. This was true also when considering changes in the threshold: wider or stricter thresholds did not entail particular differences in the number and type of clusters, plausibly because of an overall medium-to-good quality of the unfiltered cells.

Comparing our results to the ones published in the Panglao repository (see fig. 8) (5), we can notice a few differences. First of all, the number of cells that were retained for the analysis is slightly lower than the number we kept, the former being 4383 and the latter 4477. The second most visible difference is the fact that pericytes are present only in our results. This was surprising to us since in most of the configurations we used these cells were present and forming a well defined cluster. Moreover, it has been well established the function of pericytes in the central nervous system where they play an active role as cellular constituent of the blood-brain barrier and form the

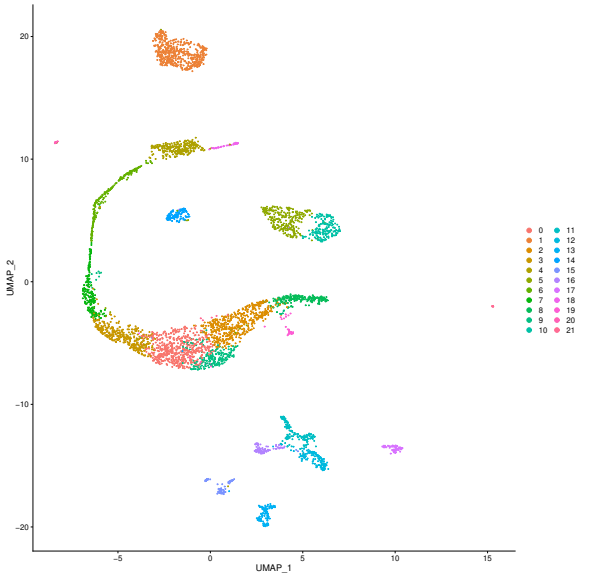

Fig. 6. Projection of the first two UMAP components highlighting the presence of the 22 clusters found with the 22 first PCs and a resolution = 1.0.

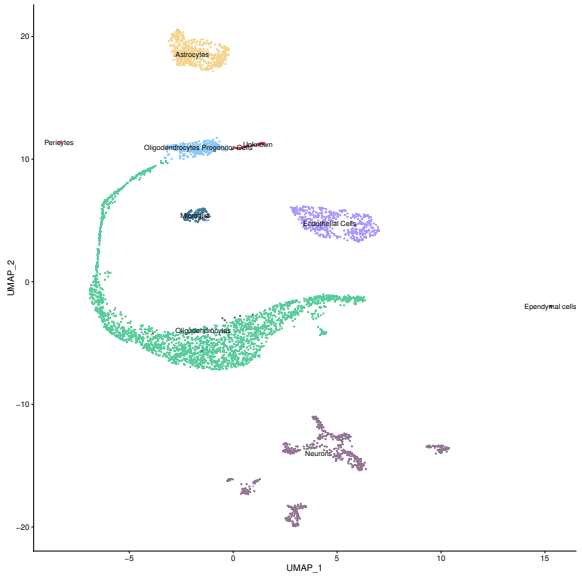

Fig. 7. Projection of the first two UMAP components with the label corresponding to the cell types identified through their marker genes.

neurovascular unit together with endothelial cells, astrocytes and neurons (13). The hypotheses we can propose is that the resolutions used to obtain the results is not high enough for the correct distinction of the pericytes from the endothelial cells or that some genes that are expressed in both cell types, such as *VTN* which is a marker for both clusters, may have led to the allegedly wrong conclusion that the pericytes cluster was of endothelial cells. A further hypothesis could be that since the pericytes cluster was very small and we used more cells than the ones used in the published data, maybe the pericytes were left out from these. However in our analysis even with stricter threshold for cell filtering (nFeature > 300, nFeature < 2500, percentage.mt < 3 and nCount > 500) we were able to identify pericytes, thus we tend to exclude this last speculation. Other divergences that we can spot are the two clusters of unknown cells that were present in the published data, one near the neurons and the other constituting a well defined cluster on his own. Near the OPC, no "unknown" cluster was detected, unlike in our outcome, maybe because it was joint to the OPC cluster. All the other cells and clusters resembled the ones we computed.

Bibliography

1. O Bican, A Minagar, AA Pruitt, The spinal cord: A review of functional neuroanatomy. *Neurol. Clin.* **31**, 1 – 18 (2013) Spinal Cord Diseases.
2. F S., T L., L I., N M., B M., Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data. *F1000Research* **7**, 1297 (2018).
3. T Stuart, et al., Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
4. S Lab, Seurat - guided clustering tutorial, https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html (2020).
5. PanglaoDB, https://panglaodb.se/view_data.php?sra=SRA667466&srs=SRS3059941 (2020).
6. Z A., et al., Molecular architecture of the mouse nervous system. cell. *Cell* **174(4)**, 999–1014.e22 (2018).
7. H J., et al., Gencode: the reference human genome annotation for the encode project. *Genome research* **22(9)**, 1760–1774 (2012).
8. F O, G LM, B JLM, Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database (Oxford)* (2019).
9. H Wickham, R François, L Henry, K Müller, *dplyr: A Grammar of Data Manipulation*, (2020) R package version 1.0.2.
10. TL Pedersen, *patchwork: The Composer of Plots*, (2020) R package version 1.0.1.
11. H Wickham, *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York), (2016).
12. PanglaoDB, https://panglaodb.se/search.html (2020).
13. P Dore-Duffy, K Cleary, *Morphology and Properties of Pericytes*. (Humana Press, Totowa, NJ), pp. 49–68 (2011).

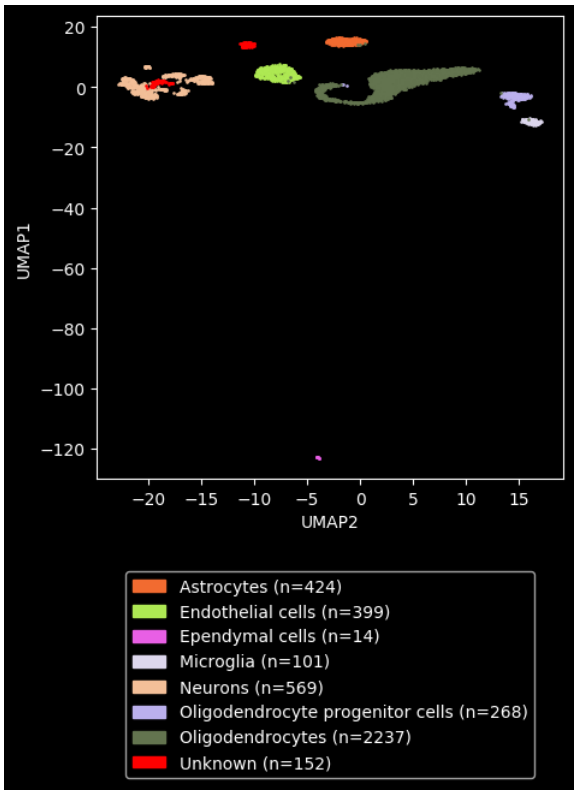

Fig. 8. UMAP projection of the results published in the Panglao repository(5)

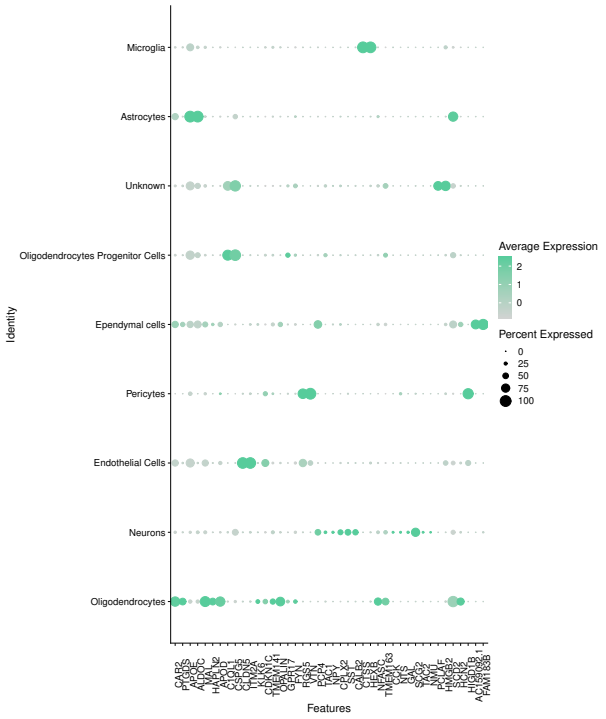

Fig. 9. Dotplot showing the expression of selected marker genes in different types of cells

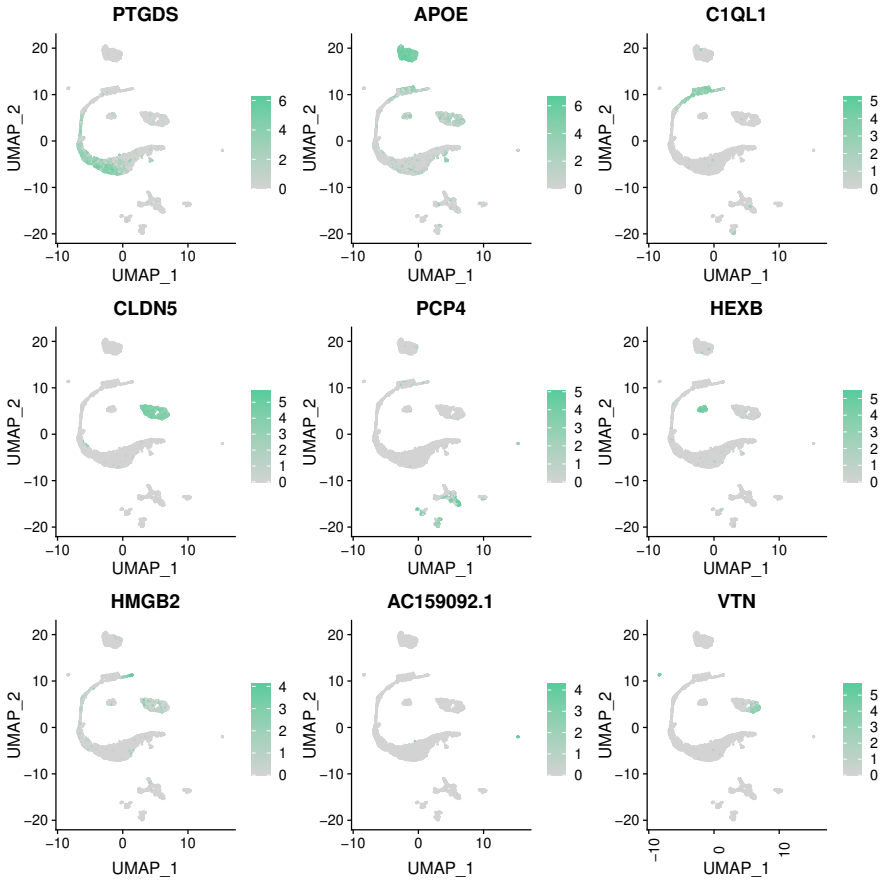

Fig. 10. Heatmap of the expression of selected marker genes over the UMAP plot of clusters. We can notice that almost all genes are specific for a cluster of a cell type

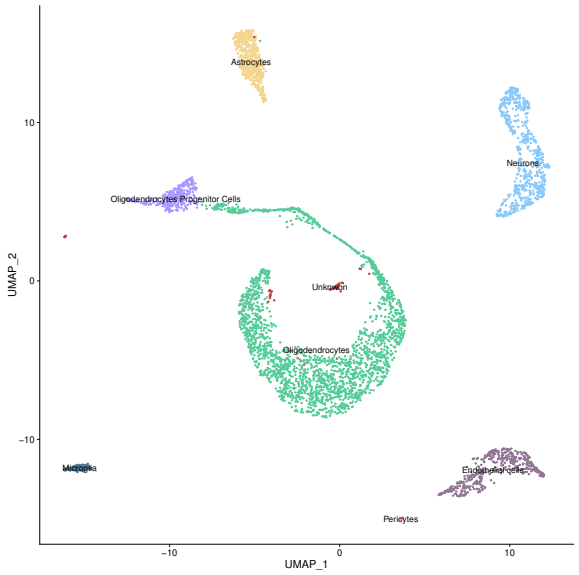

Fig. 11. Projection of the first two UMAP components obtained with the 4000 most variable genes, the use of 10 PC and a resolution for the clustering of 1.0.

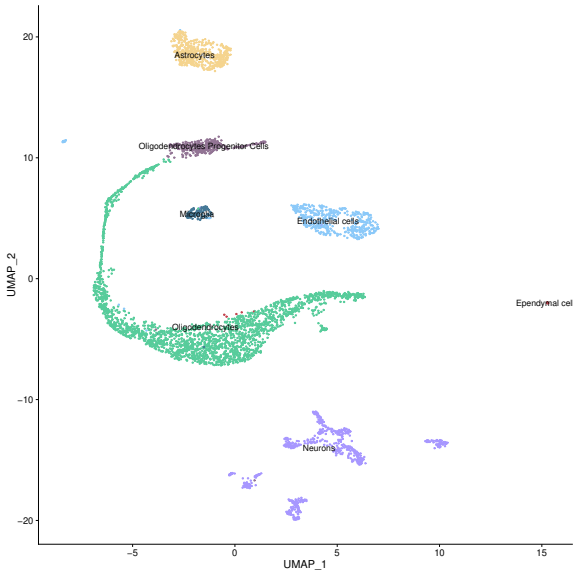

Fig. 12. Projection of the first two UMAP components obtained with the 4000 most variable genes, the use of 22 PC and a resolution for the clustering of 0.5.

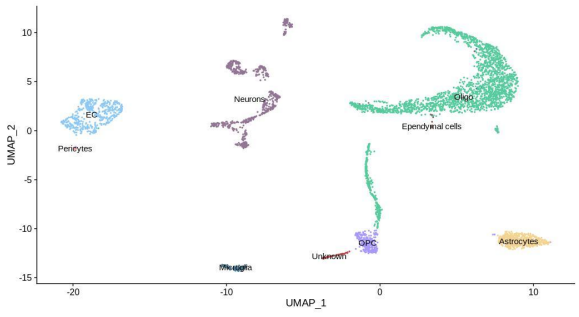

Fig. 13. Projection of the first two UMAP components obtained with the 2000 most variable genes, the use of 22 PC and a resolution for the clustering of 1.0 (EC = endothelial cells, OPC = Oligodendrocytes Progenitor Cells, Oligo = Oligodendrocytes).

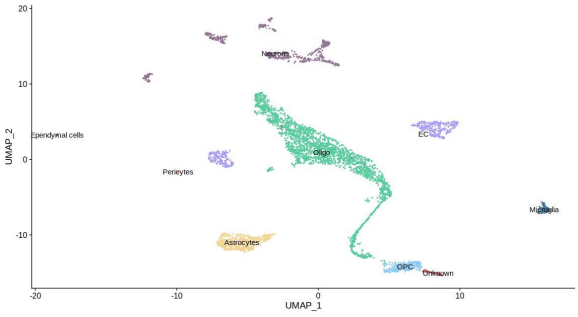

Fig. 14. Projection of the first two UMAP components obtained with the 5000 most variable genes, the use of 22 PC and a resolution for the clustering of 1.0 (EC = endothelial cells, OPC = Oligodendrocytes Progenitor Cells, Oligo = Oligodendrocytes).