Single-cell analysis of frontal cortex

Carla Casanova

Introduction

The present analysis is based on the dataset provided by Lake *et al.* (2018). They performed for their study both the *single-nucleus droplet-based sequencing (snDrop-seq)* and the *single-cell transposome hypersensitive site sequencing (scTHS-seq)*. Nevertheless, the selected series of data provided by them on the GEO database from the NCBI, was generated by isolating single DAPI+ nuclei (according to the Samples list) from the visual cortex (BA17), frontal cortex (BA6 or BA10), and cerebellar hemisphere from 6 different postmortem adult human brains and processed for snDrop-seq. So, for the current analysis the data selected is from the frontal cortex and only snDrop-seq was performed. Moreover, the technologies used for sequencing this data are Illumina MiSeq and Illumina HiSeq 2500.

In addition, the selected data set contains 24654 genes and 10319 cells as can be spotted in *Appendix 1*. However, the cell types found in the frontal cortex are:

- Ast: astrocytes
- End: endotelial cells
- Ex: excitatory neurons
- In: inhibitory neurons
- Mic: microglia
- Oli: oligodendrocytes
- Per: pericytes
- OPC: oligodendrocyte progenitor cell

Methods

A quality control (QC) step has been performed prior to the analysis. First, the percent of mitocondrial genes was computed and added to metadata in order to gather information about read counts of RNA (tnCount_RNA), the total number of genes expressed per cell (nFeature_RNA) and the mitocondrial genes (percent.mt) (*Appendix 2*). The different metrics have been represented with violin plots for each cell identity *Figures 1, 2 and 3*. As can be spotted, excitatory neurons present higher levels of gene expression and higher number of genes expressed than the rest of cell types, being Ex1 the cells with highest counts (>2500 genes). In contrast, End, In7 and Per cells showed fewer genes and expression levels compared to other cells. So, in order to avoid previous discrepancies of gene expression counts, as shown in *Appendix 3*, cells that have unique feature counts over 2,500 or less than 200 were filtered. Additionally, cells presenting more than 5% of mitocondrial counts were also filtered, nevertheless in this case no mitocondrial genes are found in the whole data set. As result, 24654 genes and 10164 cells passed the QC being removed only 155 cells. In addition, as shown in *Figure 4*, a Scatter plot among the counts of features and the RNA per cell type was computed.

Next step was data normalization by performing the global-scaling 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. So, the normalization step is done by doing:

$$ln[1 + (\frac{counts\ in\ a\ cell\ per\ one\ gene}{total\ number\ of\ counts\ in\ a\ cell} * 10000)]$$

Once data is normalized we are able to compare all the cells of the data set to find the more variable genes. It was done by directly modeling the mean-variance relationship inherent in single-cell data (which is implemented in the `FindVariableFeatures()` function). By default, 2000 features per dataset are returned and these will be used in downstream analysis, like PCA. Nevertheless, the top 10 more variable genes are represented in *Figure 5*. In addition, prior to the PCA analysis, a linear transformation ('scaling') was applied only to these variable genes in order to save time (since they are the input of the PCA analysis). This step is essential to give equal weight in downstream analyses, so that highly-expressed genes do not dominate. It is done by doing:

$$\frac{\textit{normalized data per one gene in a cell} \; - \; \textit{mean normalized data per one gene in the dataset}}{\textit{sd normalized data per one gene in the dataset}}$$

In addition, the results of the scaled data are stored in `pbmc[["RNA"]]@scale.data` and the normalized data is stored in `pbmc[["RNA"]]@data` while the raw data is stored in `pbmc[["RNA"]]@counts`.

Once data was filtered, normalized and scaled, the dimensionality reduction was performed by a PCA as shown in *Figures 6 and 7*. Moreover, Jackstraw and Elbow plots were also performed to asses the cutoff for dimension size as spotted in *Figure 8* and *Figure 9* respectively. The 20 dimensions found seem to be highly significant (small p-values) and the cutoff does not seem to be clear. The clustering was first performed with a resolution of 2 and different dimension. When using 20 dimensions, 26 clusters are found. In contrast, when using 15 dimensions, 24 clusters are found. Finally, when using 12 dimensions, 25 clusters are found. If the previous analyses are computed again with a resolution of 1, when using 15 dimension 18 clusters are found and when using 12 dimensions 17 clusters are found. As can be spotted dimensionality does not change significantly the results, but resolution. Since the current data set has more than 3K of cells, it is more likely that the most suited resolution is 2, hence more than 20 clusters can be distinguished. However, this result would be more close to the one obtained by the original study (which found 35 clusters).
Additionally, the non-linear dimensional reduction was performed by using the UMAP technique, in order to plot the clusters. Moreover, a differential gene expression analysis only for positive markers was also performed using the ROC test for all the clusters. Finally clusters were annotated based on their marker genes.

Results and Discussion

On the one hand, the current work has performed the clustering by selecting 12 dimensions and a resolution equal to 2 (since the data set is >3k of cells). As spotted in the Jackstraw and Elbow plots, seem that the first 12 dimensions offer the most part of the information which allows to compute the analysis. However, p-values are more similar until the 12^{th} dimension since from this point p-values go from $10^{-83}$to higher than 10^{-50}. Nevertheless, the most marked differences spotted among the p-values start from the 3^{th} dimension. As result, 25 clusters were found and the number of cells per cluster are shown in the *Appendix 4*. The clusters computed are also represented in the *Figure 10*. Additionally, a summary with counts and stats per each cluster has been computed as spotted in the *Appendix 5*. As spotted, the mean number of genes expressed per single-cell in each cluster ranges from ~422 to ~1709, so the QC step worked properly since cells with <200 or >2500 features are not found. In addition, the number of genes when grouping all the cells of a cluster ranges from 23875 to 730884.

On the other hand, a differential expression analysis was computed to find marker genes for each cluster from the human frontal cortex as spotted in *Appendix 6*. As spotted in *Appendix 7* and *Figure 11*, the top 20 markers were computed per cluster, even though most of the clusters presented less than 20 markers. Additionally, the cluster 0 presented only one marker which is represented in *Figure 12*. As can be spotted in the Heatmap (*Figure 12*) and the UMAP clustering (*Figure 10*), there are some cluster that share the same marker profile such as clusters 1 and 2, clusters 7-9 and clusters 14-16. *Panglao* and *Protein atlas* databases were consulted to associate the top markers of each cluster with different cell types, however just the markers with higher ROC values were considered.

In the case of clusters 1 and 2 the top marker is PLP1 which is highly related with *oligodendrocytes*. Even though two clusters of Oli were created, it is more likely that both populations are the same, since Lake *et al.* (2018) did not find subpopulations for this case. However, cluster 24 also was weakly associated with PLP1. Regarding to the cluster 3, the top gene marker is HS3ST4 which was reported by *Protein atlas* to be enriched in the *microglia*, however this marker was present only in the population 3.

When looking at inhibitory populations, cluster 6 present inhibitory markers such as ERBB4 and ZNF385D which are reported on the *Protein atlas* database. However, ZNF385D was exclusively associated to inhibitory neurons and it is also found as marker in clusters 11 and 18. Another marker found to be exclusively related with inhibitory neurons is SYNPR (reported by the *Protein atlas*) which is also expressed in clusters 5 and 11. Additionally, the ADARB2 marker is also associated with inhibitory neurons and OPC, which was present in clusters 5, 18 and 19. However, these populations are more likely to be inhibitory if considering the whole gene marker profile. Other markers found in inhibitory populations were ROBO2 (highly associated with 11 and 21) and GRIK1 (clusters 11, 18, 19 and 21). In contrast, when analyzing excitatory populations, Lake *et al.* (2018) reported that excitatory neurons expressed CBLN2 gene, which was found as marker in cluster 7 and 16. Moreover, Darmanis *et al.* (2015) also reported SLC17A7 as another marker found in excitatory populations, which is found to be expressed in clusters 12 and 14. However, since the marker profile of cluster 14 is similar to cluster 15 it is more likely that they are also excitatory neurons. Another gene marker IQCJ-SCHIP1 was reported by the *Protein atlas* database to be associated with excitatory neurons, which is present in the cluster 9. Additionally, MAP1B was weakly associated with the cluster 22 and it is also associated to excitatory populations among others.

Regarding to cluster 10, gene markers associated with OPC were found such as LHFPL3 and TNR as spotted on *Protein atlas* and *Panglao* respectively. Finally, when analyzing clusters 13, 17 and 23, the main gene marker found is SLC1A2 which is highly associated with *astrocytes* and it is reported in the both databases previously seen, moreover it was also found by Lake *et al.* (2018). Moreover, GPC5 marker is also related with astrocytes and it was strongly associated with the previous clusters. In addition, despite clusters 20 and 25 are not strongly defined, they presented PLXDC2 and ATP1A2 respectively, which are related with astrocytes among other according to *Protein atlas* and *Panglao*. So, they are more likely to be astrocytes since they are also closer to clusters 13, 17 and 23.

The gene expression among clusters of the main markers are represented in *Figure 13*. Additionally, the annotated clusters are also represented in *Figure 14*.

Conclusions

Several subpopulations of excitatory and inhibitory neurons were found according to the results obtained by Lake *et al.* (2018). Nevertheless, they were able to diferenciate more subpopulation within these communities. However, the current analysis did not found endotelial cells and pericytes which were supposed to be present in the data set. Since some populations were not clearly defined, such as cluster 20 and 25, it is more likely that the resulted annotation was wrong for this type of populations due to weak association with gene markers. Moreover, the marker genes were strongly defined for some populations (such as 1, 2, 3 and 10 among others), but very weakly to others, such as population 4 (it only presented two marker genes with ROC values ~0.4). Since the cutoff for dimensionality size used in downstream analyses was not defined clearly, it can be assumed that the resolution of the current work could be improved in order to find stronger associations with gene markers. Maybe, the 15-20 dimensions are needed instead of 12.

References

1. Lake, B., Chen, S., Sos, B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol 36, 70–80 (2018). https://doi.org/10.1038/nbt.4038

2. Spyros Darmanis, Steven A. Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M. Shuer, Melanie G. Hayden Gephart, Ben A. Barres, Stephen R. Quake. A survey of human brain transcriptome diversity at the single cell level. Proceedings of the National Academy of Sciences of the United States of America., 112(23), 7285-7290 (2015). https://doi.org/info:doi/

3. Protein Atlas (consulted on Febrary 2022). https://www.proteinatlas.org

4. PanglaoDB (consulted on Febrary 2022). https://panglaodb.se/index.html

Appendix

Appendix 1. Exploring Seurat object and cell types in the selected data set.

```
# Visualize Seura object (count matrix)
pbmc
```

```
## An object of class Seurat
## 24654 features across 10319 samples within 1 assay
## Active assay: RNA (24654 features, 0 variable features)
```

```
# Explore cell types on the dataset
table(pbmc$orig.ident)
```

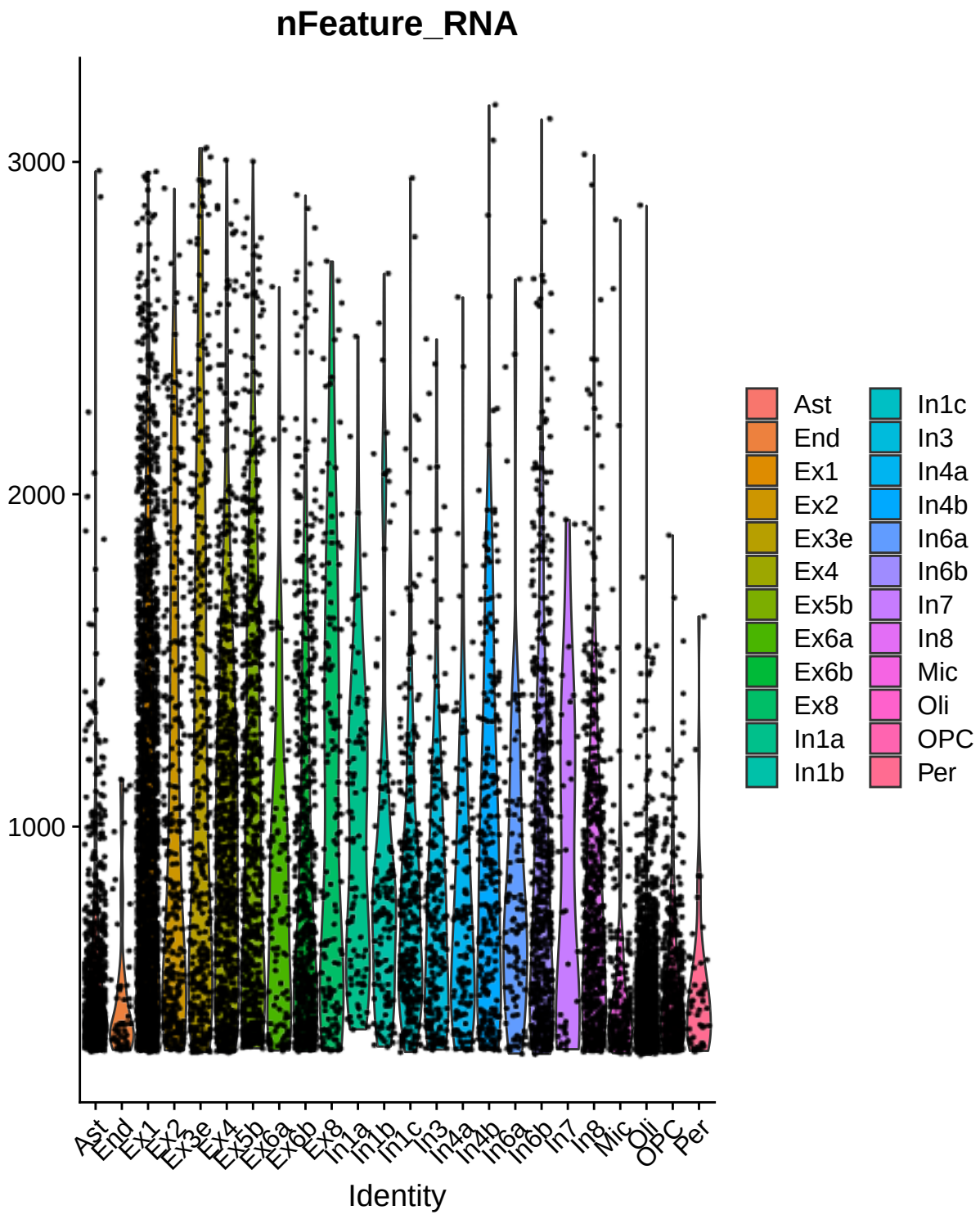
```
##
##  Ast  End  Ex1  Ex2 Ex3e  Ex4 Ex5b Ex6a Ex6b  Ex8 In1a In1b In1c  In3 In4a In4b
##  737   51 1917  310  372  705  640  109  576  142   87  120  280  209  124  239
## In6a In6b  In7  In8  Mic  Oli  OPC  Per
##  101  556   33  588  187 1758  433   45
```

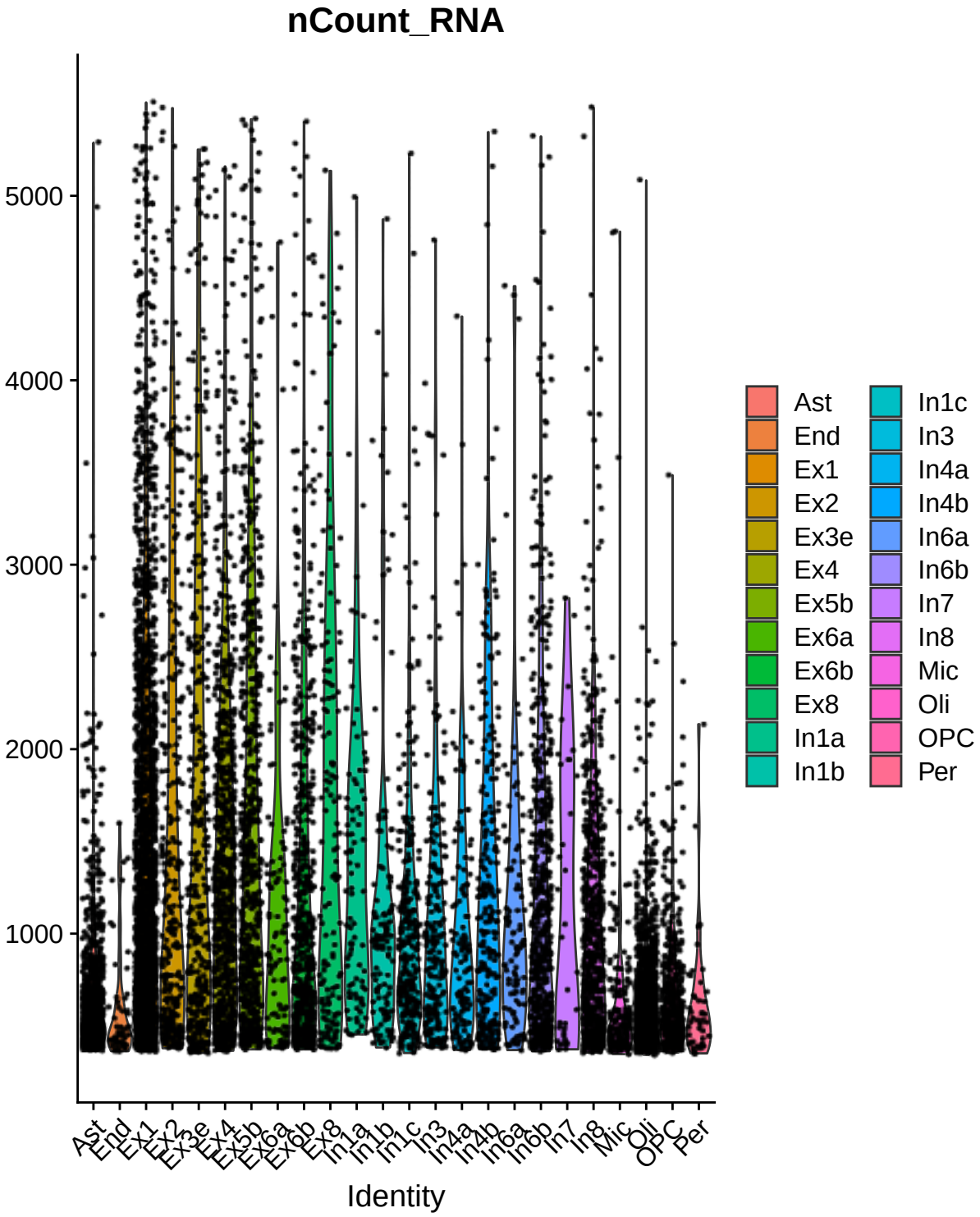
Appendix 2. Computing percent of mitochondrial genes and exploring the computed metrics for QC control.

```
# The [[ operator can add columns to object metadata. Let's add the proportion
# of mitochondrial genes
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")

# Show QC metrics for the first 5 cells
head(pbmc@meta.data, 5)
```

```
##                       orig.ident nCount_RNA nFeature_RNA percent.mt
## Ex1_fcx8_GGACGCCTTTAA        Ex1        412          359          0
## Ex1_fcx1_GTTCCAGCACGA        Ex1       4976         2964          0
## Ex1_fcx1_CGAGGCTAAAGG        Ex1       3927         2150          0
## Ex1_fcx1_CTTCAGCACGTA        Ex1       3808         2192          0
## Ex1_fcx1_GACCTTTCAGGC        Ex1       3838         2350          0
```

Figures 1, 2 and 3. Violin plots for the different metrics: number of genes, number of reads and mitocondrial genes respectively.


nFeature_RNA
3000
2000
1000
Ast
End
Ex1
Ex2
Ex3e
Ex4
Ex5b
Ex6a
Ex6b
Ex8
In1a
In1b
In1c
In3
In4a
In4b
In6a
In6b
In7
In8
Mic
Oli
OPC
Per
Identity
Ast
End
Ex1
Ex2
Ex3e
Ex4
Ex5b
Ex6a
Ex6b
Ex8
In1a
In1b
In1c
In3
In4a
In4b
In6a
In6b
In7
In8
Mic
Oli
OPC
Per


nCount_RNA
5000
4000
3000
2000
1000
Ast
End
Ex1
Ex2
Ex3e
Ex4
Ex5b
Ex6a
Ex6b
Ex8
In1a
In1b
In1c
In3
In4a
In4b
In6a
In6b
In7
In8
Mic
Oli
OPC
Per
Identity
Ast
End
Ex1
Ex2
Ex3e
Ex4
Ex5b
Ex6a
Ex6b
Ex8
In1a
In1b
In1c
In3
In4a
In4b
In6a
In6b
In7
In8
Mic
Oli
OPC
Per

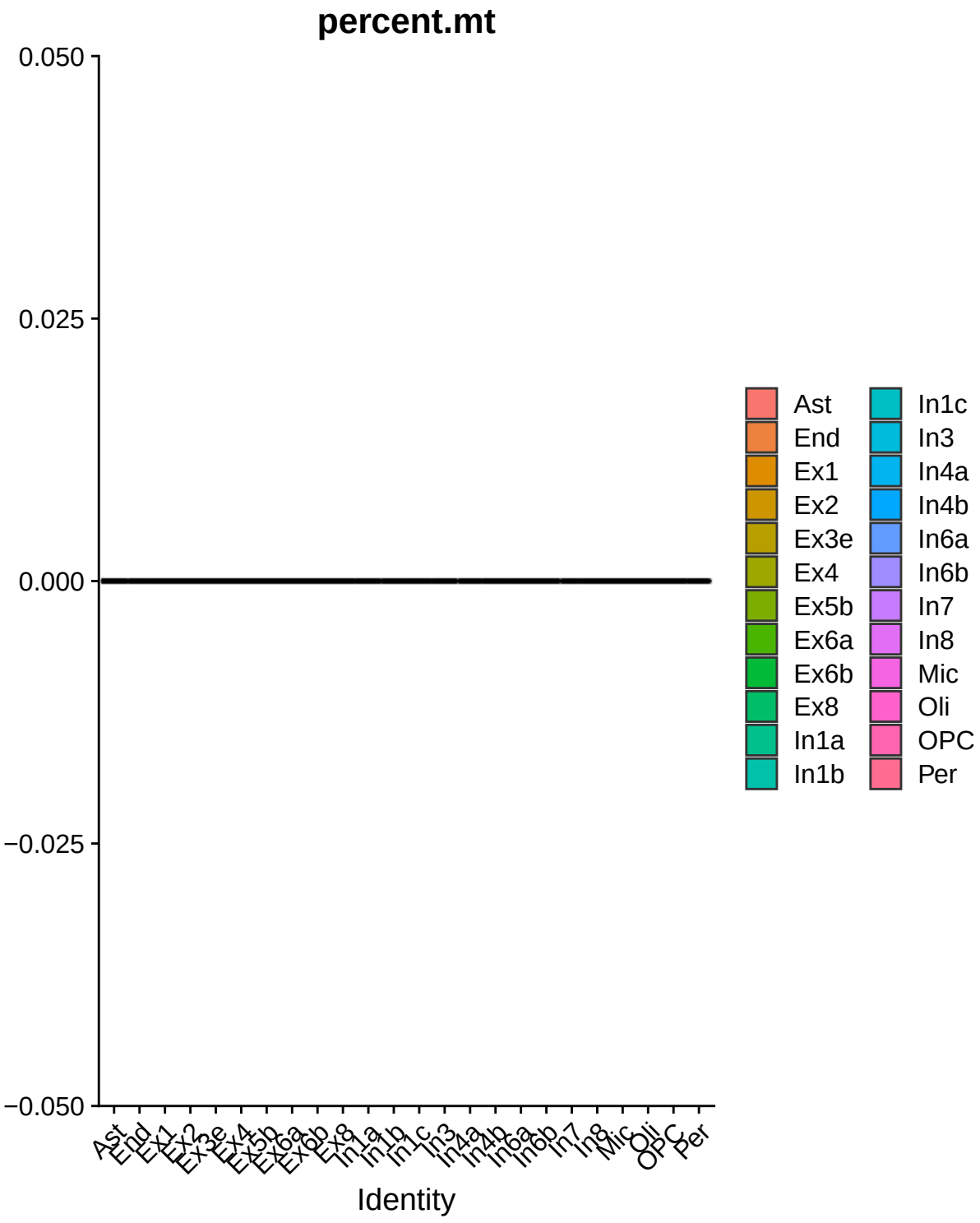
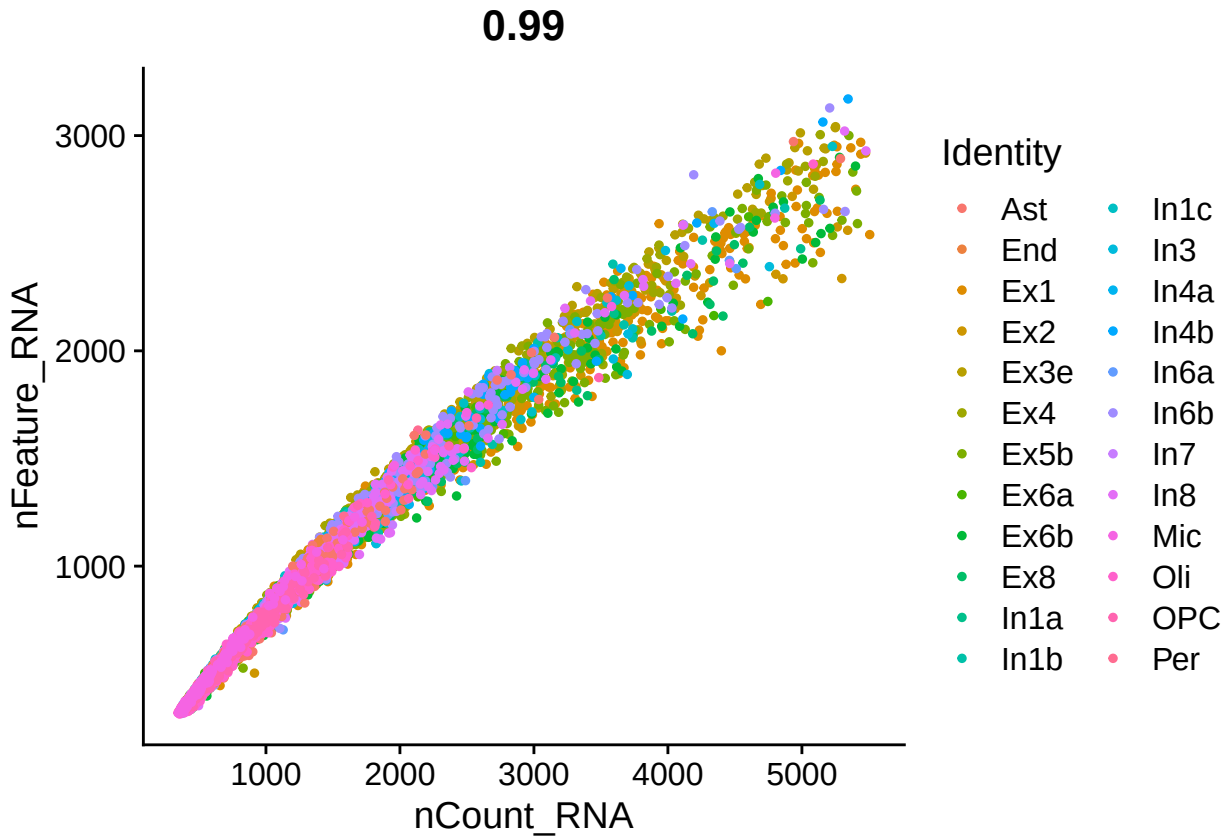


Figure 4. Scatter plot of total counts of features versus total counts of RNA.

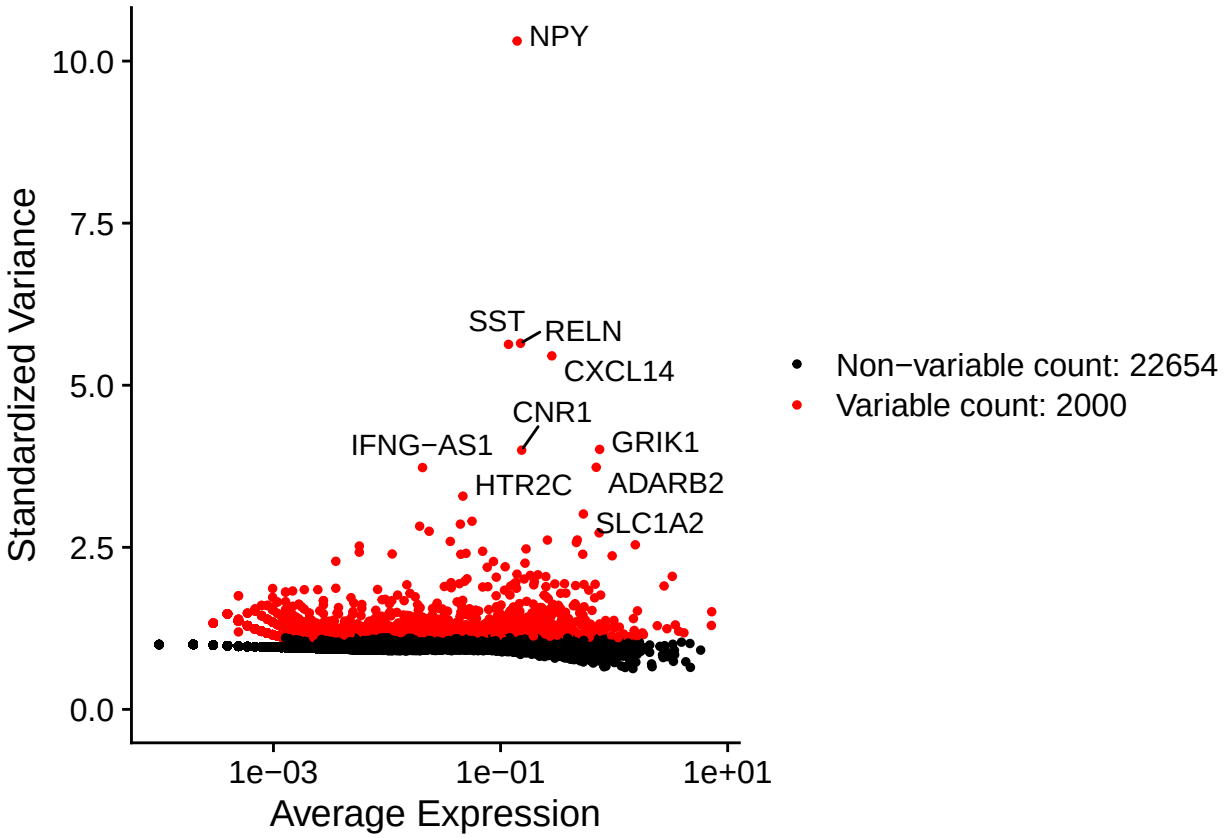


Appendix 3. QC filter for cells with counts between 200-2500 and less than 5% of mitocondrial counts.

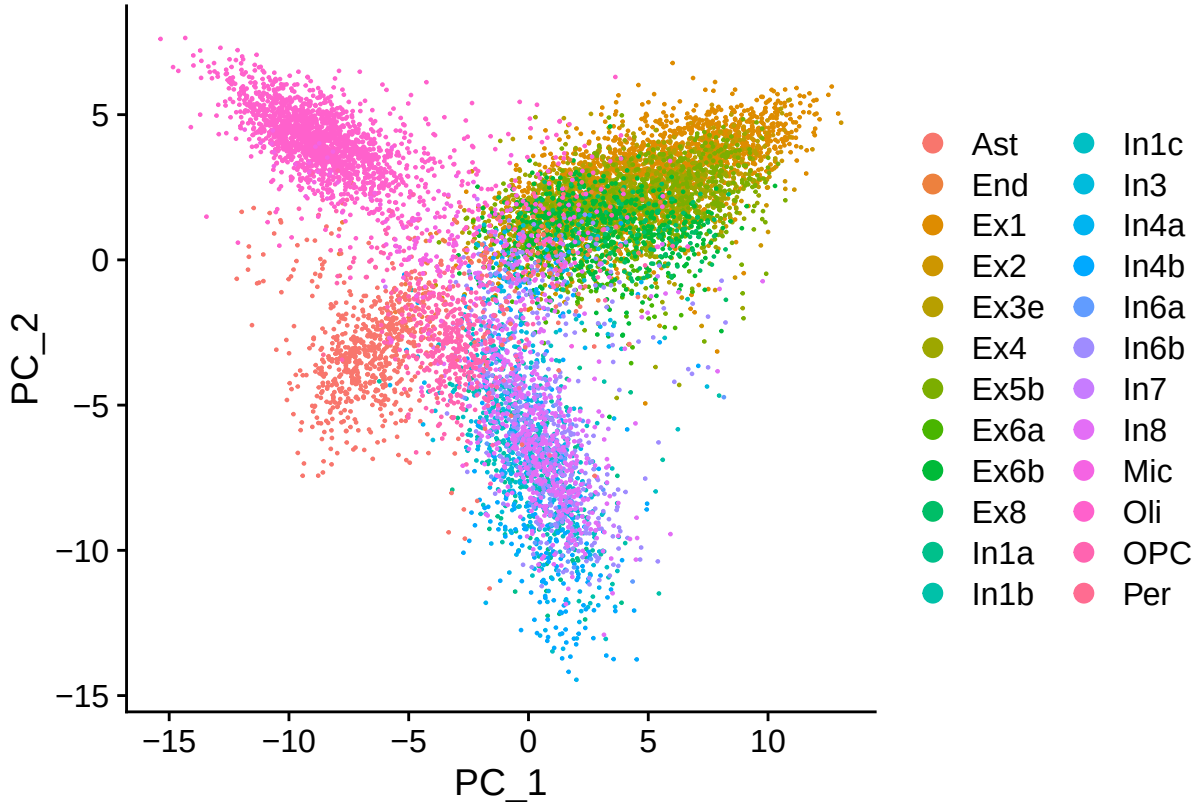
```
# Filter bad quality cells and genes
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt <
    5)

# Visualize the final number of cells and genes
pbmc
```

```
## An object of class Seurat
## 24654 features across 10164 samples within 1 assay
## Active assay: RNA (24654 features, 0 variable features)
```

Figure 5. Most variable genes among different cell types from human frontal cortex.



Figures 6 and 7. 2D scatter plot of PCA analysis and Heatmap of 15 dimension from human frontal cortex



dataset.

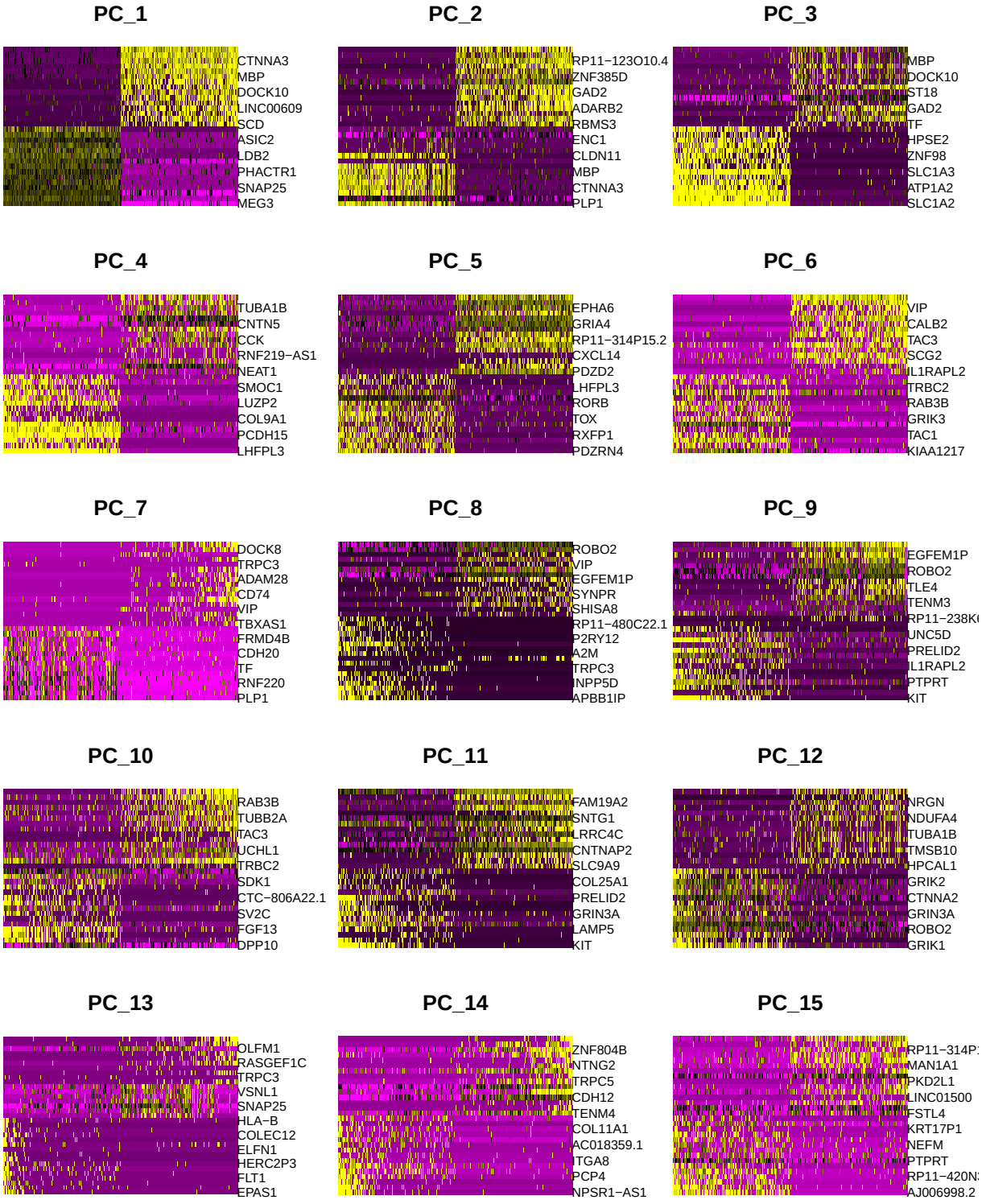


Figure 8. JackStraw plot of 20 dimentions from human frontal cortex data set.

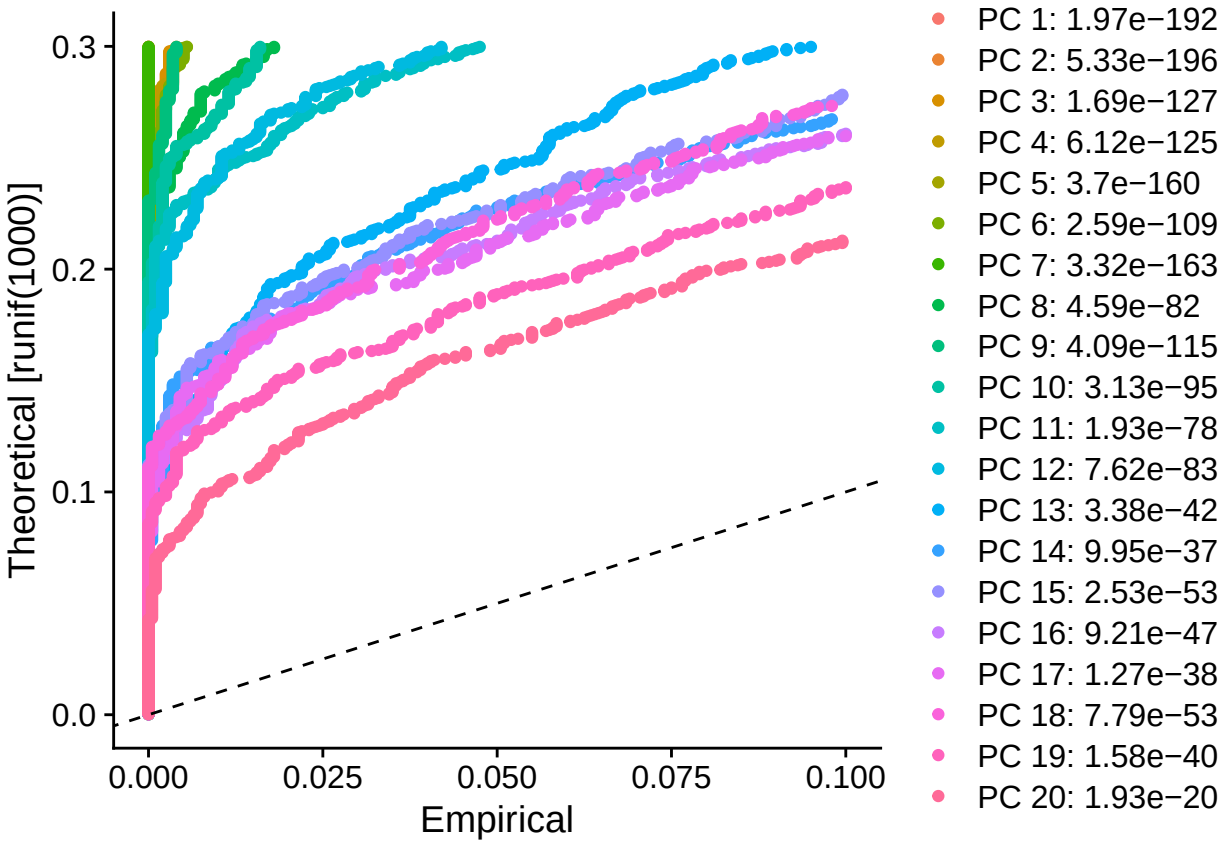
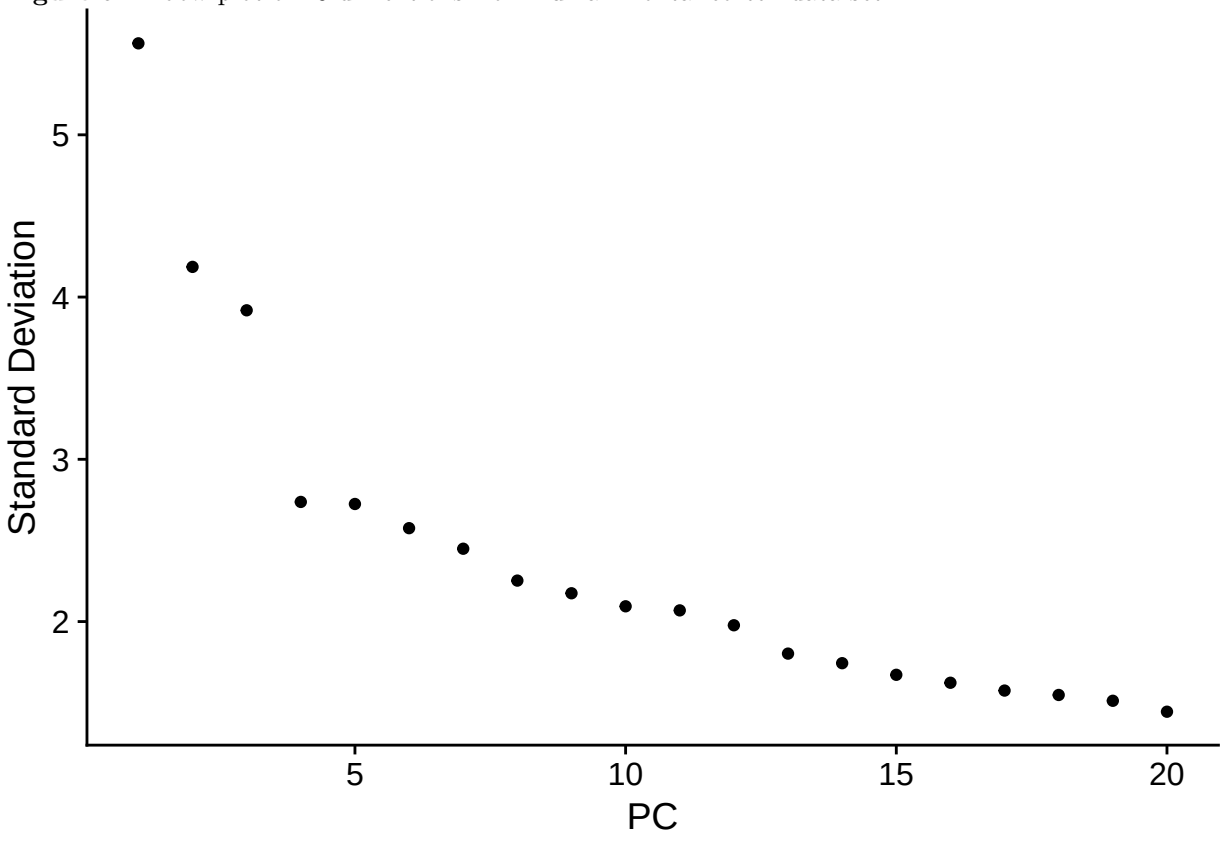


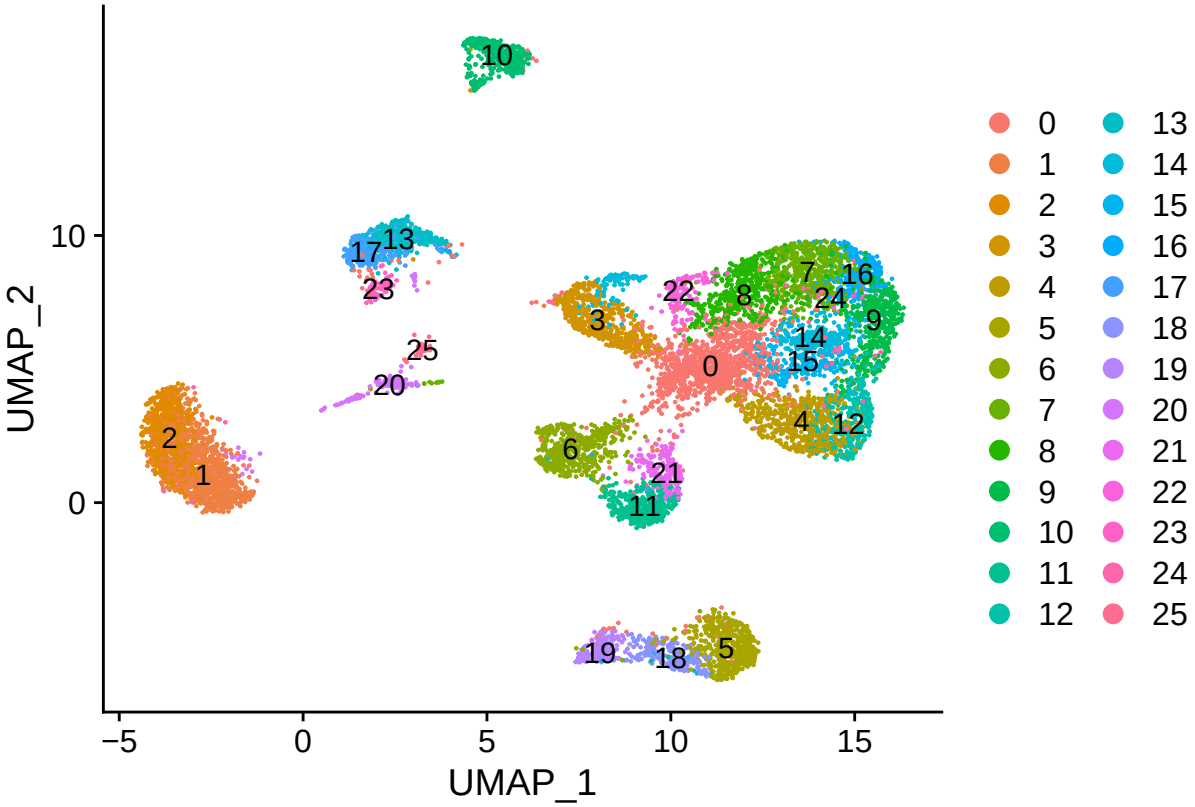
Figure 9. Elbow plot of 20 dimentions from human frontal cortex data set.



Appendix 4. Number of cells per cluster when using 12 dimensions from human frontal cortex data set.

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 1195  918  722  532  530  519  518  517  491  450  418  392  373  366  318  302
##   16   17   18   19   20   21   22   23   24   25
##  265  254  204  203  185  182  152   59   53   46
```

Figure 10. Clusters generated by the single-cell analysis from human frontal cortex.



Appendix 5. Table with counts grouped by cluster of RNA, genes, cells and the mean of features per single-cell.

```
# Visualize the mean number of genes in a single-cell by cluster
table_clusters <- pbmc@meta.data %>%
    group_by(pbmc@meta.data$seurat_cluster) %>%
    summarize(nCount_RNA = sum(nCount_RNA), nFeature_RNA = sum(nFeature_RNA), cells = n()) %>%
    mutate(mean_nFeatures = nFeature_RNA/cells)
```

Appendix 6. Marker genes from human frontal cortex grouped by cluster and ordered by log2FC.

```
# Group the results by cluster and order by log2FC
pbmc.markers %>%
    group_by(cluster) %>%
    slice_max(n = 2, order_by = avg_log2FC)
```

```
## # A tibble: 50 x 8
## # Groups:   cluster [26]
```

```
##    myAUC avg_diff power avg_log2FC pct.1 pct.2 cluster gene
##    <dbl>    <dbl> <dbl>      <dbl> <dbl> <dbl> <fct>   <chr>
##  1 0.728    0.569 0.456      0.821 0.978 0.8   0       KCNIP4
##  2 0.861    1.95  0.722      2.81  0.834 0.192 1       PLP1
##  3 0.769    1.69  0.538      2.44  0.656 0.186 1       MBP
##  4 0.866    2.05  0.732      2.96  0.819 0.128 2       MOBP
##  5 0.92     2.02  0.84       2.92  0.947 0.197 2       PLP1
##  6 0.865    1.96  0.73       2.82  0.814 0.159 3       HS3ST4
##  7 0.753    1.64  0.506      2.37  0.577 0.08  3       EGFEM1P
##  8 0.702    0.734 0.404      1.06  0.755 0.433 4       PCLO
##  9 0.71     0.492 0.42       0.710 0.972 0.75  4       SYT1
## 10 0.913    2.01  0.826      2.89  0.925 0.181 5       ADARB2
## # ... with 40 more rows
```

Appendix 7. Computing top 20 positive markers per cluster from human frontal cotex.

```
pbmc.markers %>%
    group_by(cluster) %>%
    top_n(n = 20, wt = avg_log2FC) -> top20
```

Figure 11. Heatmap of top 20 gene markers per cluster.

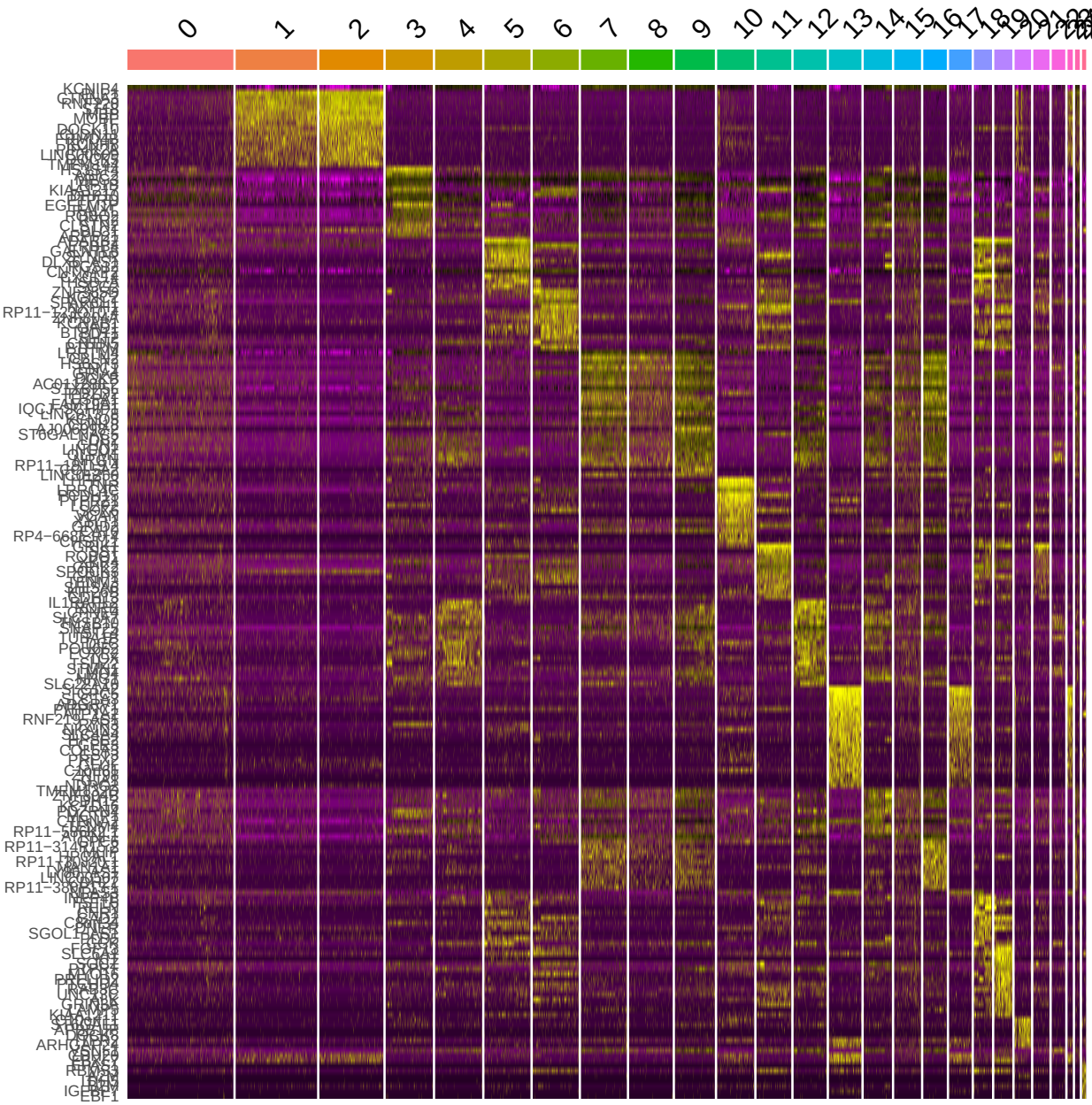


Figure 12. Violin plot for the only gene marker of cluster 0 from human frontal cortex.

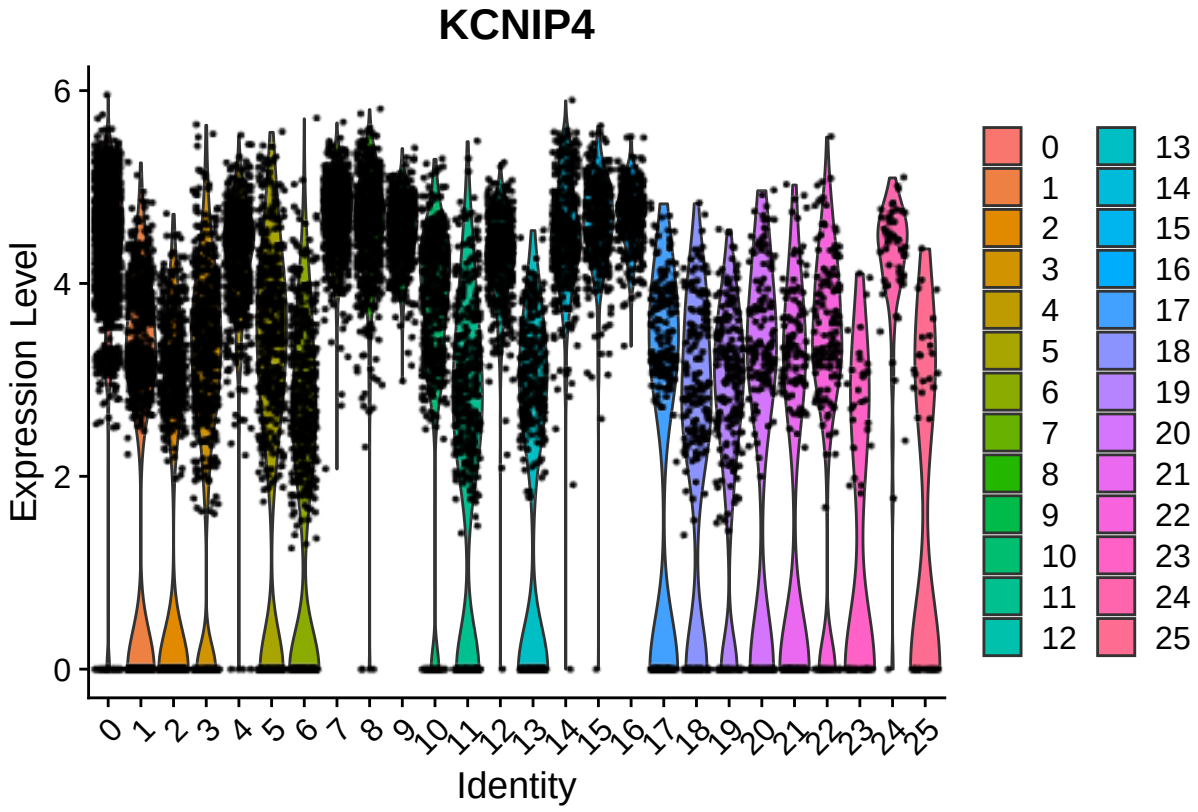


Figure 13. Expression of main marker genes for the different clusters from the human frontal cortex.

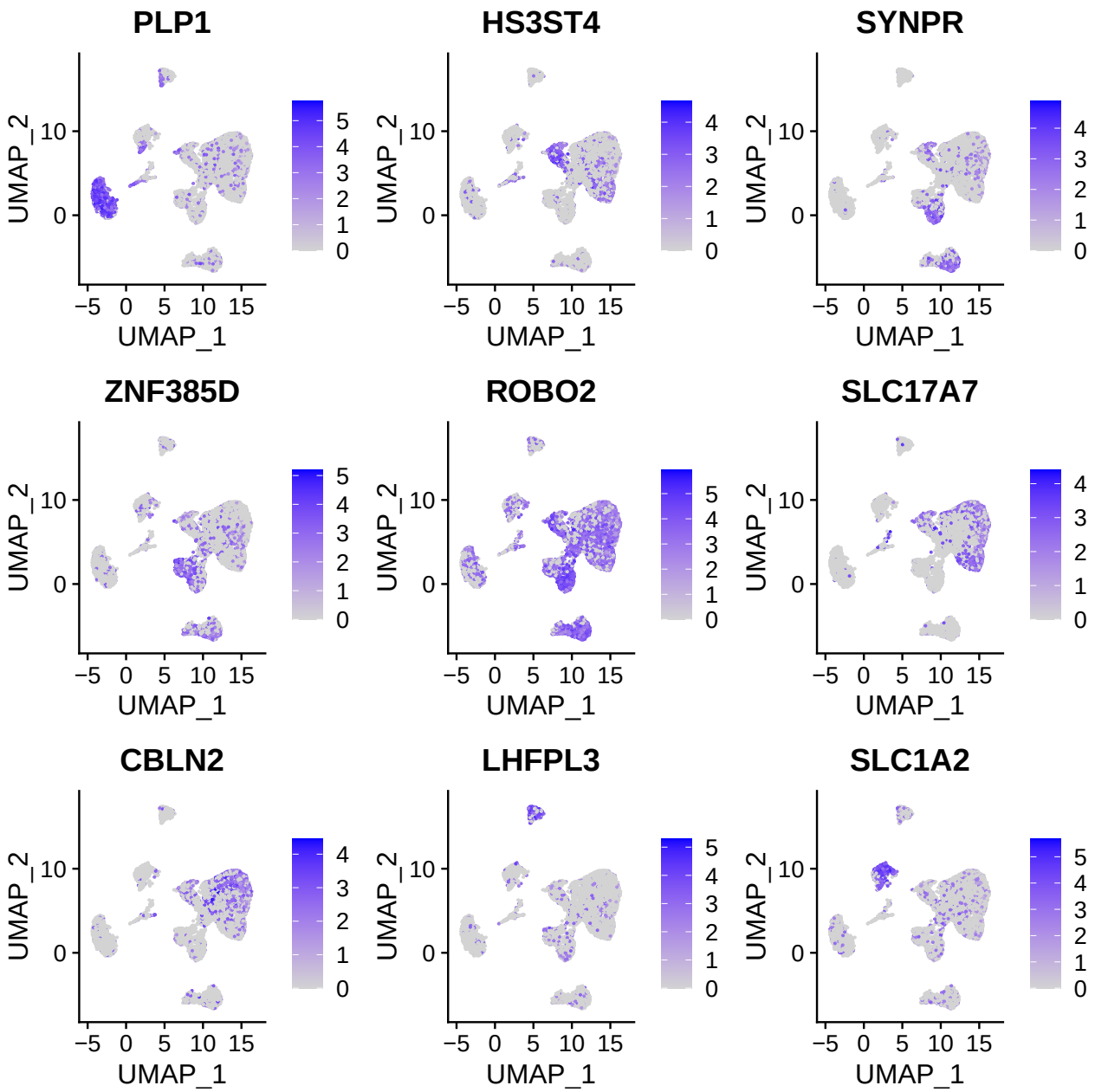
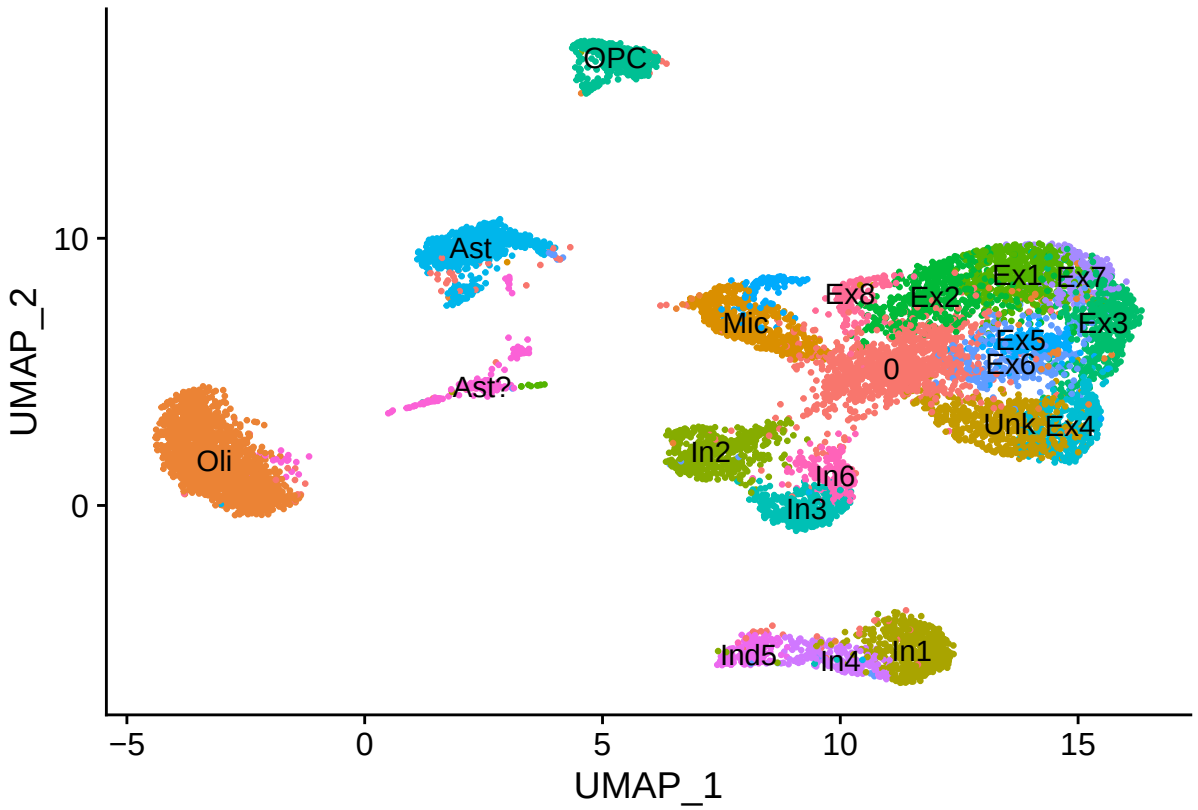


Figure 14. Annotation of the different clusters from the human frontal cortex.


OPC
Ast
Ast?
Oli
Mic
Ex8
Ex2
Ex1
Ex7
Ex3
Ex5
Ex6
0
Unk
Ex4
In2
In6
In3
Ind5
In4
In1
10
0
−5
0
5
10
15
UMAP_2
UMAP_1