Single Cell Data Analysis of Human Kidney Organoids

Mehran Piran

6/2/2020

```
library(Seurat)
library(ggplot2)
library(dplyr)
```

Importing Data

GSE108291_org_barcodes.tsv.gz, GSE108291_org_counts.csv.gz, GSE108291_org_genes.tsv.gz and GSE108291_org_matrix.mtx.gz files were downloaded from the GSE108291 GEO page. Data was imported into R using Read10X function. The function is specific for outputs of “cellranger” pipeline from 10X genomic technology.

```
org <- Read10X(data.dir = "F:/Projects/GSE108291/Data")
org <- CreateSeuratObject(counts = org, project = "pbmc3k", min.cells = 5,
min.features = 100)
org

## An object of class Seurat
## 20927 features across 17086 samples within 1 assay
## Active assay: RNA (20927 features, 0 variable features)

head(org[[]] , 10)

##                    orig.ident nCount_RNA nFeature_RNA
## AAACCTGAGAGTACAT-1     pbmc3k      11225         3209
## AAACCTGAGTTTCCTT-1     pbmc3k        115          106
## AAACCTGGTATAGTAG-1     pbmc3k        142          117
## AAACCTGGTGCGATAG-1     pbmc3k        121          101
## AAACCTGGTTAAAGAC-1     pbmc3k       9551         2703
## AAACCTGTCAACACGT-1     pbmc3k        115          109
## AAACCTGTCTCGATGA-1     pbmc3k       2559         1252
## AAACGGGAGGACAGCT-1     pbmc3k      13115         3315
## AAACGGGAGTGCAAGC-1     pbmc3k      14599         3383
## AAACGGGCAAGCCGCT-1     pbmc3k       5230         1738
```

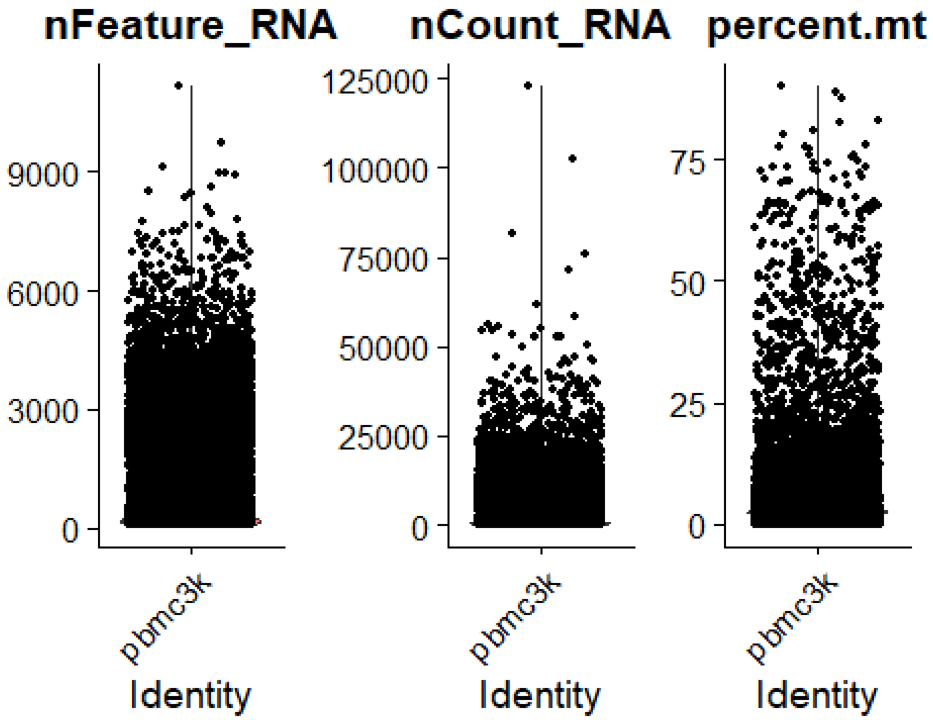
removing low quality cells

Based on two metrices , the percentage of mitochondrial genes and the number of genes in each cell, low quality cells were removed.

```
org[["percent.mt"]] = PercentageFeatureSet(org, pattern = "^MT-")
```

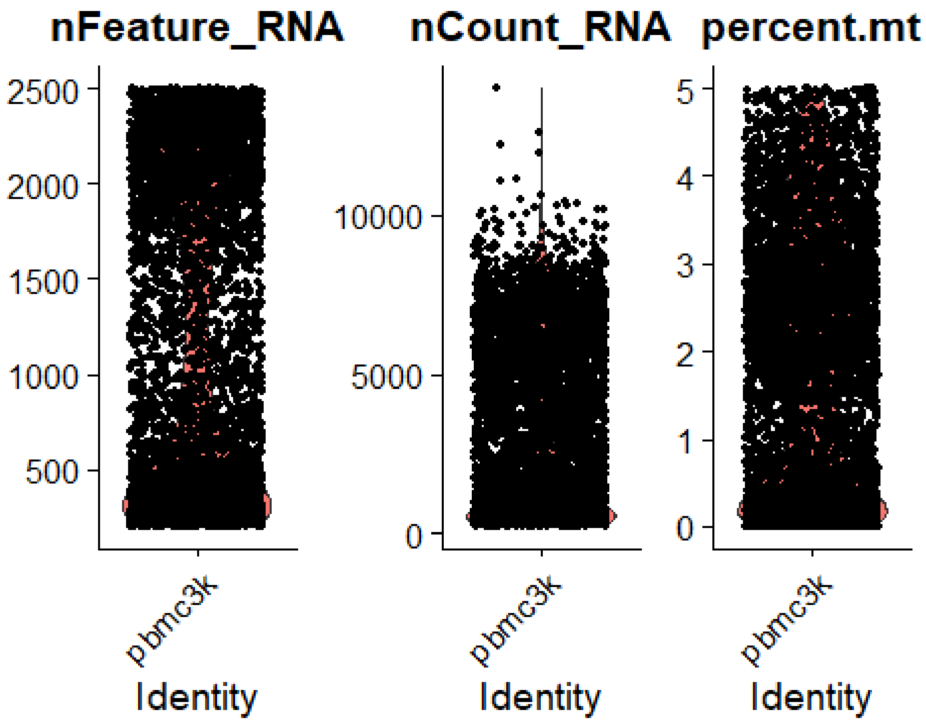
Violin plot before outlier removal

```
VlnPlot(org, features = c("nFeature_RNA", "nCount_RNA" , "percent.mt"), ncol
= 3)
```



Violin plot after outlier removal

```
org = subset(org, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 &
percent.mt < 5)
VlnPlot(org, features = c("nFeature_RNA", "nCount_RNA" , "percent.mt"), ncol
= 3)
```



```
# There are less than 500 genes and moleculaes for a great number of cells while the distribution of
# mitochondrial genes is more uniform among the all cells.
```

Data Normalization

Data were normalized using LogNormalize method.

```
org = NormalizeData(org)
```

Identifying the tope variable genes

The tope 2000 variable genes were selected to be used in the downstream analyses. Therefore, non-significant genes can no longer impact the results.

```
org = FindVariableFeatures(org, selection.method = "vst", nfeatures = 2000)
```

Scaling Data before applying dimention reduction methods

```
org = ScaleData(org, features = rownames(org))
```

Effects of cell cycle statuses on clustering results

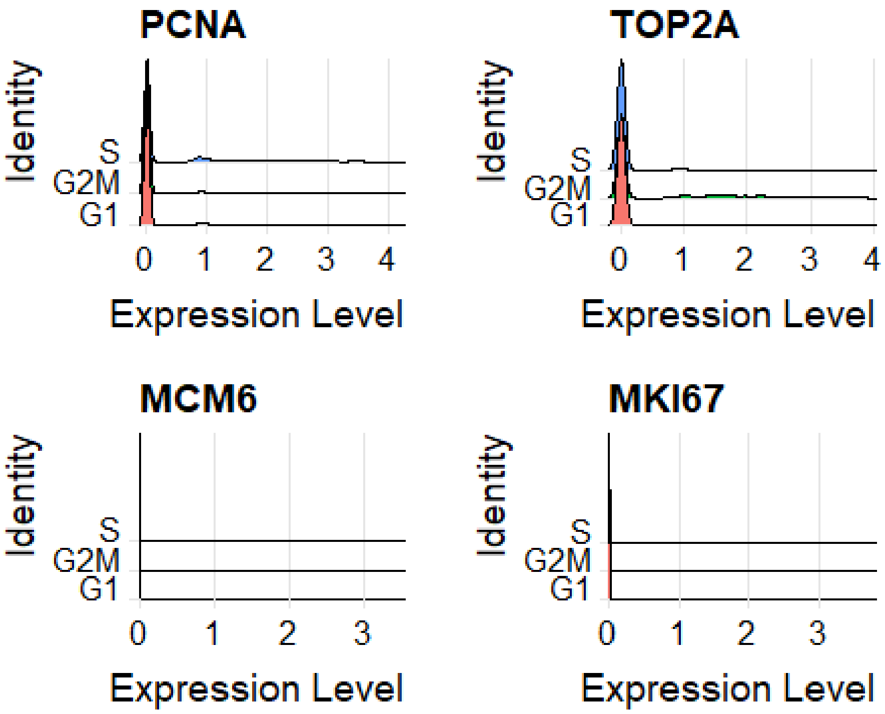
To see how much cell cycle status affect the distribution of cells, expression values for cell cycle specific genes were used to give each cell a score

```
s.genes = cc.genes$s.genes
g2m.genes = cc.genes$g2m.genes
org = CellCycleScoring(org, s.features = s.genes, g2m.features = g2m.genes, 
set.ident = TRUE)
head(org[[]])
```

```
##                      orig.ident nCount_RNA nFeature_RNA percent.mt
S.Score
## AAACCTGTCTCGATGA-1     pbmc3k       2559         1252  0.5470887 -
0.01595561
## AAACGGGCAAGCCGCT-1     pbmc3k       5230         1738  2.3900574 -
0.08361473
## AAACGGGCAGCGTCCA-1     pbmc3k        326          230  0.6134969 -
0.01220314
## AAACGGGTCAGGCAAG-1     pbmc3k       6141         2032  2.2471910 -
0.04658905
## AAACGGGTCCACTGGG-1     pbmc3k       5509         1580  0.1270648  
0.05586209
## AAAGATGAGCCTCGTG-1     pbmc3k       6186         1942  3.0876172 -
0.08662457
##                        G2M.Score Phase old.ident
## AAACCTGTCTCGATGA-1 -0.07440106    G1    pbmc3k
## AAACGGGCAAGCCGCT-1 -0.11173627    G1    pbmc3k
## AAACGGGCAGCGTCCA-1  0.02147969   G2M    pbmc3k
## AAACGGGTCAGGCAAG-1  0.06531156   G2M    pbmc3k
## AAACGGGTCCACTGGG-1  0.09008073   G2M    pbmc3k
## AAAGATGAGCCTCGTG-1 -0.08655577    G1    pbmc3k
```

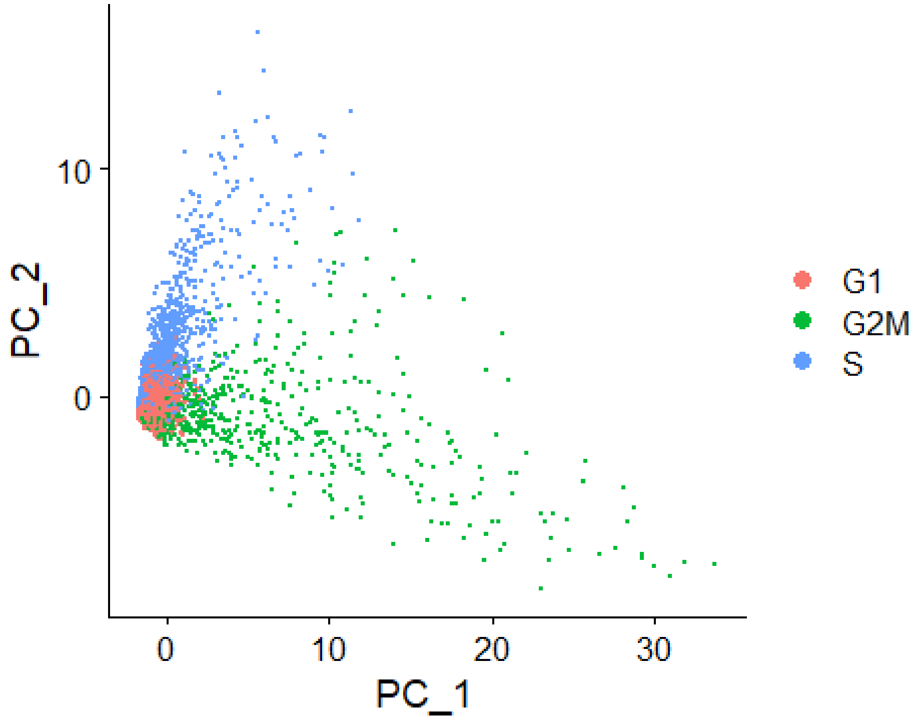
To Visualize the distribution of cell cycle markers across all cells, four genes which are expressed during cell cycle progression were used. There is no expression for MCM6 and MKI67 genes. Apparently cell-cycle markers are not expressed considerably

```
RidgePlot(org, features = c("PCNA", "TOP2A", "MCM6", "MKI67"), ncol = 2)
```



However, based on the PCA on expression of S and G2M genes, there are distinct cluster of cells affected by the cell cycle states.

```r
org = RunPCA(org, features = c(s.genes, g2m.genes), nfeatures.print = 5)
DimPlot(org)
```



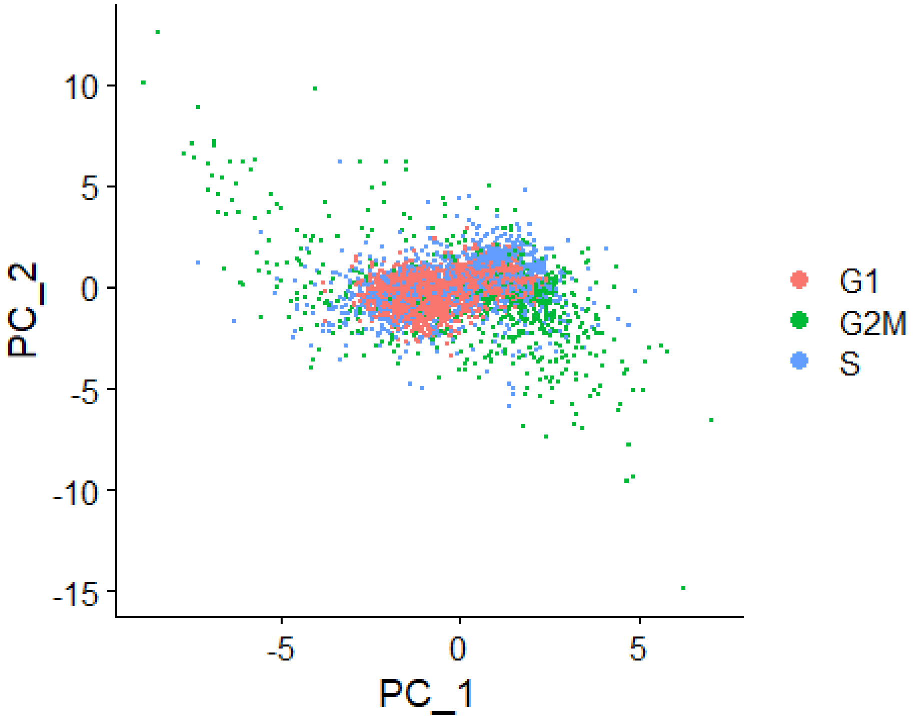
Cell cycle regression

Regressing out cell cycle scores during data scaling was done as follows.

```r
org = ScaleData(org, vars.to.regress = c("S.Score", "G2M.Score"), features = rownames(org))
```

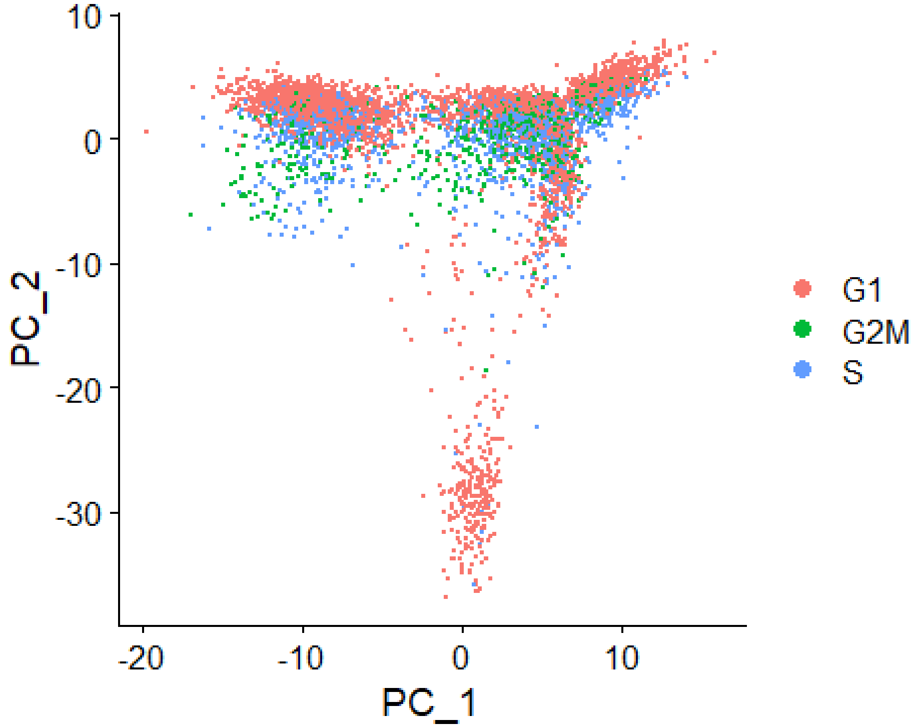
The given figure presents the PCA plot for cell cycle genes after regression.

```r
org <- RunPCA(org, features = c(s.genes, g2m.genes) , nfeatures.print = 5)
DimPlot(org)
```



plotting PCA for most variable genes after regression was done.

```
org <- RunPCA(org, features = VariableFeatures(org), nfeatures.print = 5)
DimPlot(org)
```

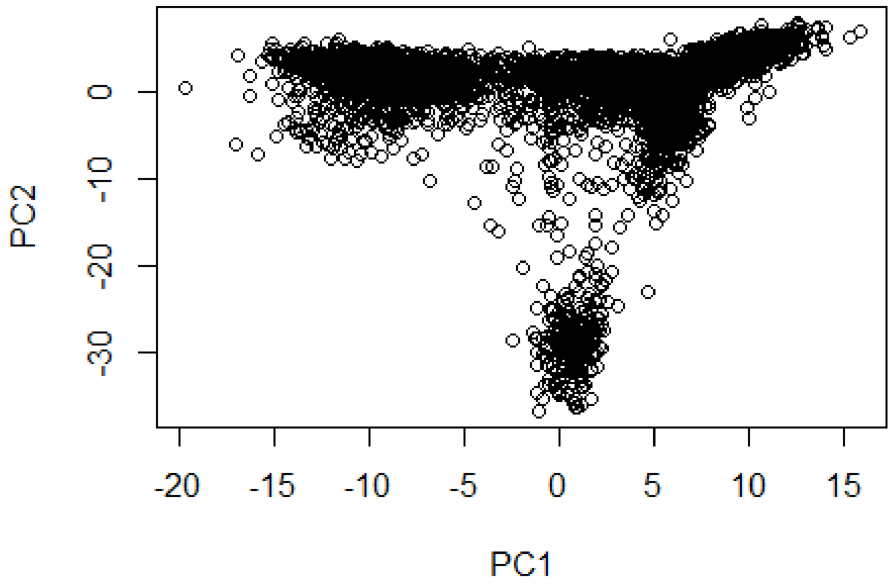


```
# the majority of cells are in G1 phase. In addition, no cluster related to
specific cell cycle
# status has been emerged
```

Determine the ‘dimensionality’ of the dataset

Here we find which eigenvector best explain the variability which are present in the dataset. In statistics, the distance in PC1 explain how far observations are from eachother.

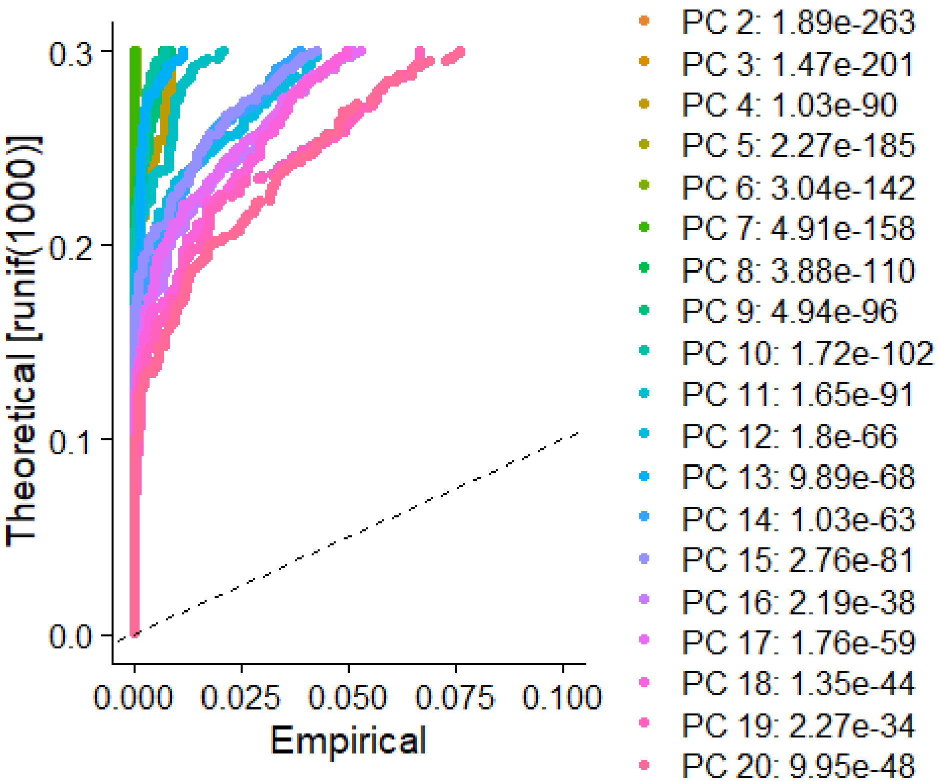
```
plot(org[["pca"]]@cell.embeddings[,1] , org[["pca"]]@cell.embeddings[,2] ,
xlab = "PC1" ,
     ylab = "PC2" , cex = 1)
```



```
dim(org[["pca"]]@cell.embeddings)

## [1] 5440   50

names(org)

## [1] "RNA" "pca"

org[["pca"]]

## A dimensional reduction object with key PC_
##  Number of dimensions: 50
##  Projected dimensional reduction calculated:  FALSE
##  Jackstraw run: FALSE
##  Computed using assay: RNA
```

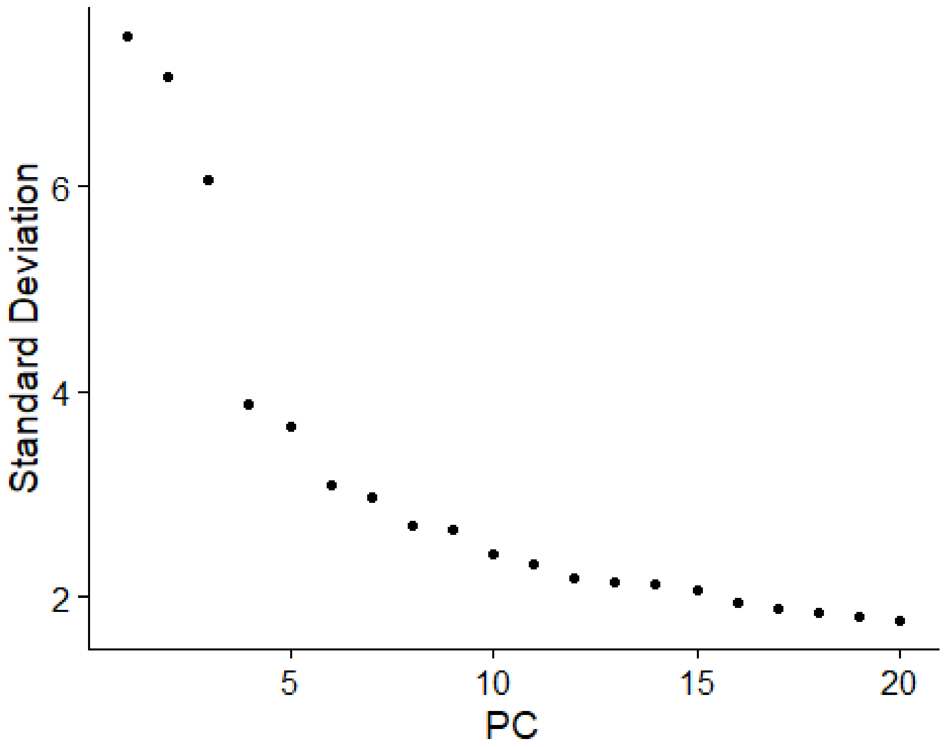
pc1 and pc2 have the smallest p-value and define the majority of variation in the datasets.

```
org <- JackStraw(org, num.replicate = 100)
org <- ScoreJackStraw(org, dims = 1:20)
JackStrawPlot(org, dims = 1:20)
```



Elbow plot shows that most of the variation in dataset is defined by PC1 and PC2

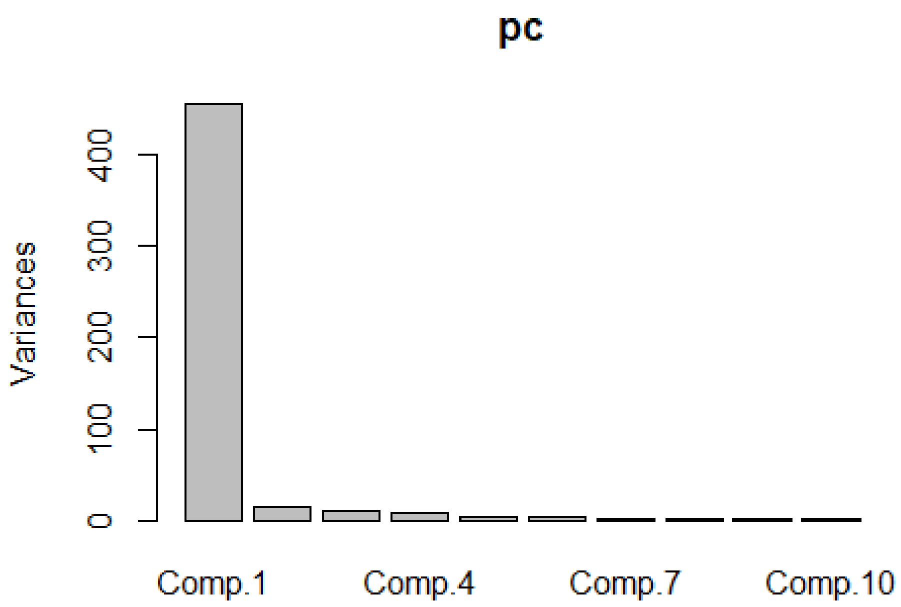
```
ElbowPlot(org)
```



```
pc = princomp(as.data.frame(GetAssayData(object = org, slot = "data")))
```

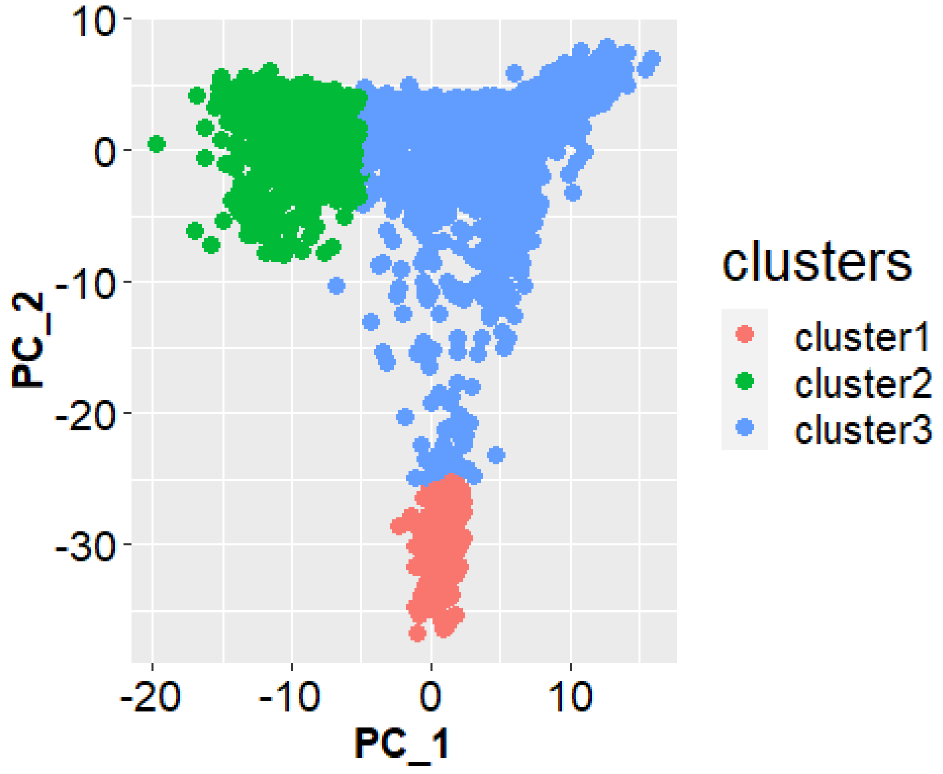
Based on the following plot, PC1 (Comp.1) and PC2 (Comp.2) define most of the variabilty between cells.

```
plot(pc)
```

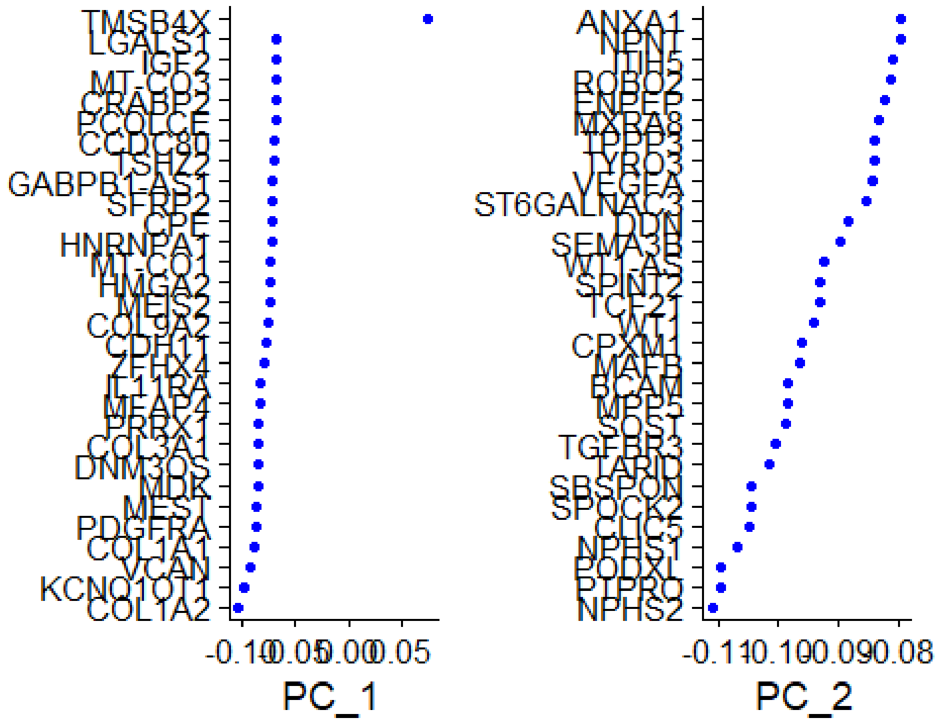


###clustering based the on pc1 and pc2

```
pc = org[["pca"]]@cell.embeddings
pc = as.data.frame(pc)

cluster1 = rownames(pc)[pc$PC_1 > -5 & pc$PC_2 < -25 ]
cluster2 = rownames(pc)[pc$PC_1 < -5 & pc$PC_2 > -10 ]


pc$clusters = "cluster3"
pc$clusters[which(rownames(pc) %in% cluster1)] = "cluster1"
pc$clusters[which(rownames(pc) %in% cluster2)] = "cluster2"
pc$clusters = factor(pc$clusters)

ggplot(pc , aes(PC_1 , PC_2 , color = clusters)) + geom_point(size = 3) +
  theme(axis.title=element_text(size=15 , face  = "bold") ,
        axis.text=element_text(size=16 , colour = "black") , legend.title =
element_text(size = 20) ,
        legend.text = element_text(size = 15))
```



```r
VizDimLoadings(org, dims = 1:2, reduction = "pca")
```



K-nearest neighbor (KNN) graph-based clustering

Applying Jaccard similarity was done.

```
org <- FindNeighbors(org, dims = 1:20)

names(org)

## [1] "RNA"     "RNA_nn"  "RNA_snn" "pca"

org[["RNA"]]

## Assay data with 20927 features for 5440 cells
## Top 10 variable features:
##  CTGF, STMN2, IGFBP7, CPB1, ACTC1, PLCG2, DLK1, HES6, ACTA2, PODXL

head(org[["RNA_nn"]])[,1:100]

## 6 x 100 sparse Matrix of class "dgCMatrix"
##
## AAACCTGTCTCGATGA-1 1 . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGCAAGCCGCT-1 . 1 . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGCAGCGTCCA-1 . . 1 . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCAGGCAAG-1 . . . 1 . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCCACTGGG-1 . . . . 1 . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAAGATGAGCCTCGTG-1 . . . . . 1 . . . . . . . . . . . . . . . . . . . . . .
. .
##
## AAACCTGTCTCGATGA-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGCAAGCCGCT-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGCAGCGTCCA-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCAGGCAAG-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCCACTGGG-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAAGATGAGCCTCGTG-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
##
## AAACCTGTCTCGATGA-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```

```
. .
## AAACGGGCAAGCCGCT-1 . . . . . . . 1 . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGCAGCGTCCA-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCAGGCAAG-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCCACTGGG-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAAGATGAGCCTCGTG-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
##
## AAACCTGTCTCGATGA-1 . . . . . . . . . .
## AAACGGGCAAGCCGCT-1 . . . . . . . . . .
## AAACGGGCAGCGTCCA-1 . . . . . . . . . .
## AAACGGGTCAGGCAAG-1 . . . . . . . . . .
## AAACGGGTCCACTGGG-1 . . . . . . . . . .
## AAAGATGAGCCTCGTG-1 . . . . . . . . . .
```

```r
head(org[["RNA_snn"]])[,1:100]
```

```
## 6 x 100 sparse Matrix of class "dgCMatrix"
##
## AAACCTGTCTCGATGA-1 1 . . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGCAAGCCGCT-1 . 1 . . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGCAGCGTCCA-1 . . 1 . . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCAGGCAAG-1 . . . 1 . . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCCACTGGG-1 . . . . 1 . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAAGATGAGCCTCGTG-1 . . . . . 1 . . . . . . . . . . . . . . . . . . . . . .
. .
##
## AAACCTGTCTCGATGA-1 . . . .          . .    . . .          . . . . .
.
## AAACGGGCAAGCCGCT-1 . . . .          . .    . . .          . . . . .
.
## AAACGGGCAGCGTCCA-1 . . . .          . .    . . 0.08108108 . . . . .
.
## AAACGGGTCAGGCAAG-1 . . . .          . .    . . .          . . . . .
.
## AAACGGGTCCACTGGG-1 . . . 0.08108108 . 0.25 . . .          . . . . .
.
## AAAGATGAGCCTCGTG-1 . . . .          . .    . . .          . . . .
0.08108108 .
##
## AAACCTGTCTCGATGA-1 . .         . . . . . . . . . . . . . . . . . . . .
## AAACGGGCAAGCCGCT-1 . .         . . . . . . . . . . . . . . . . . . . .
## AAACGGGCAGCGTCCA-1 . 0.1111111 . . . . . . . . . . . . . . . . . . . .
## AAACGGGTCAGGCAAG-1 . .         . . . . . . . . . . . . . . . . . . . .
## AAACGGGTCCACTGGG-1 . .         . . . . . . . . . . . . . . . . . . . .
## AAAGATGAGCCTCGTG-1 . .         . . . . . . . . . . . . . . . . . . . .
##
## AAACCTGTCTCGATGA-1 .         . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGCAAGCCGCT-1 0.3793103 . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGCAGCGTCCA-1 .         . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCAGGCAAG-1 .         . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAACGGGTCCACTGGG-1 .         . . . . . . . . . . . . . . . . . . . . . . .
. .
## AAAGATGAGCCTCGTG-1 .         . . . . . . . . . . . . . . . . . . . . . . .
. .
##
## AAACCTGTCTCGATGA-1 . . . . . . .
## AAACGGGCAAGCCGCT-1 . . . . . . .
## AAACGGGCAGCGTCCA-1 . . . . . . .
## AAACGGGTCAGGCAAG-1 . . . . . . .
## AAACGGGTCCACTGGG-1 . . . . . . .
## AAAGATGAGCCTCGTG-1 . . . . . . .
```

```
org[["pca"]]
```

```
## A dimensional reduction object with key PC_
##  Number of dimensions: 50
##  Projected dimensional reduction calculated:  FALSE
##  Jackstraw run: TRUE
##  Computed using assay: RNA
```

Finding clusters

```
org <- FindClusters(org, resolution = 10)
```

```
## Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck
##
## Number of nodes: 5440
## Number of edges: 194999
##
## Running Louvain algorithm...
## Maximum modularity in 10 random starts: 0.5412
## Number of communities: 71
## Elapsed time: 0 seconds
```

```
head(Idents(org) , 10)
```

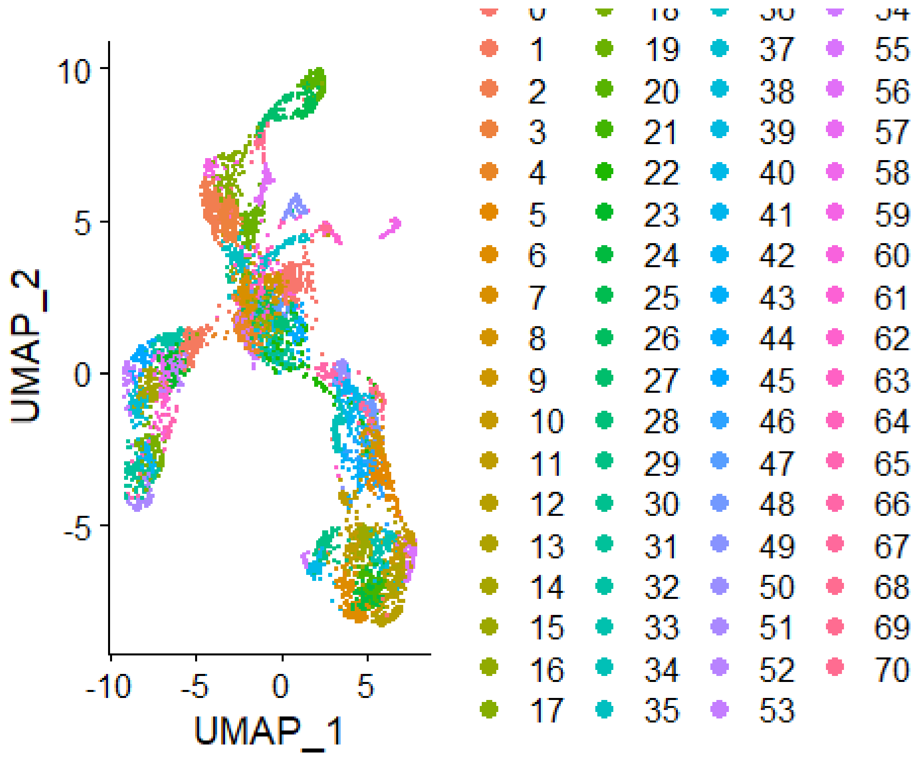
```
## AAACCTGTCTCGATGA-1 AAACGGGCAAGCCGCT-1 AAACGGGCAGCGTCCA-1 AAACGGGTCAGGCAAG-
1
##                 62                 66                 64
29
## AAACGGGTCCACTGGG-1 AAAGATGAGCCTCGTG-1 AAAGATGCACGACGAA-1 AAAGATGGTCAGAGGT-
1
##                 41                 13                 34
31
## AAAGATGGTCTCCACT-1 AAAGATGTCGGCGGTT-1
##                 61                 39
## 71 Levels: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
... 70
```

```
head(org[[]])
```

```
##                    orig.ident nCount_RNA nFeature_RNA percent.mt
S.Score
## AAACCTGTCTCGATGA-1     pbmc3k       2559         1252  0.5470887 -
0.01595561
## AAACGGGCAAGCCGCT-1     pbmc3k       5230         1738  2.3900574 -
0.08361473
## AAACGGGCAGCGTCCA-1     pbmc3k        326          230  0.6134969 -
0.01220314
## AAACGGGTCAGGCAAG-1     pbmc3k       6141         2032  2.2471910 -
0.04658905
## AAACGGGTCCACTGGG-1     pbmc3k       5509         1580  0.1270648
0.05586209
## AAAGATGAGCCTCGTG-1     pbmc3k       6186         1942  3.0876172 -
0.08662457
##                      G2M.Score Phase old.ident RNA_snn_res.10
seurat_clusters
## AAACCTGTCTCGATGA-1 -0.07440106    G1        G1             62
62
## AAACGGGCAAGCCGCT-1 -0.11173627    G1        G1             66
66
## AAACGGGCAGCGTCCA-1  0.02147969   G2M       G2M             64
64
## AAACGGGTCAGGCAAG-1  0.06531156   G2M       G2M             29
29
## AAACGGGTCCACTGGG-1  0.09008073   G2M       G2M             41
41
## AAAGATGAGCCTCGTG-1 -0.08655577    G1        G1             13
13
```

Run non-linear dimensional reduction (UMAP/tSNE)

```
org <- RunUMAP(org, dims = 1:20 , label = T)
DimPlot(org, reduction = "umap")
```



Finding differentially expressed features (cluster biomarkers)

```
# Find all markers of cluster 1
cluster1.markers = FindMarkers(org, ident.1 = 1, min.pct = 0.25)
head(cluster1.markers, 10)

##                 p_val  avg_logFC pct.1 pct.2    p_val_adj
## TMSB4X  7.438552e-66  1.3332725 1.000 0.985 1.556666e-61
## TMSB10  3.472702e-46  0.8969394 0.986 0.931 7.267323e-42
## GNG11   2.034436e-37  1.3954352 0.767 0.380 4.257463e-33
## ACTB    1.883798e-32  0.7899811 0.979 0.927 3.942224e-28
## HNRNPA1 1.853607e-29 -1.0061947 0.479 0.882 3.879043e-25
## RPLP1   5.201774e-28  0.6021736 0.959 0.973 1.088575e-23
## FTL     6.486577e-26  0.8159029 0.890 0.903 1.357446e-21
## EIF4A2  1.750311e-25 -0.9287762 0.041 0.539 3.662876e-21
## RPS19   2.370841e-25  0.5431039 0.966 0.960 4.961459e-21
## EEF1A1  4.543897e-24 -0.5455612 0.863 0.984 9.509013e-20
```

```
# Find all markers distinguishing cluster 1 from clusters 2 and 3
cluster5.markers = FindMarkers(org, ident.1 = 1, ident.2 = c(2, 3), min.pct =
0.25)
head(cluster5.markers, n = 10)
```

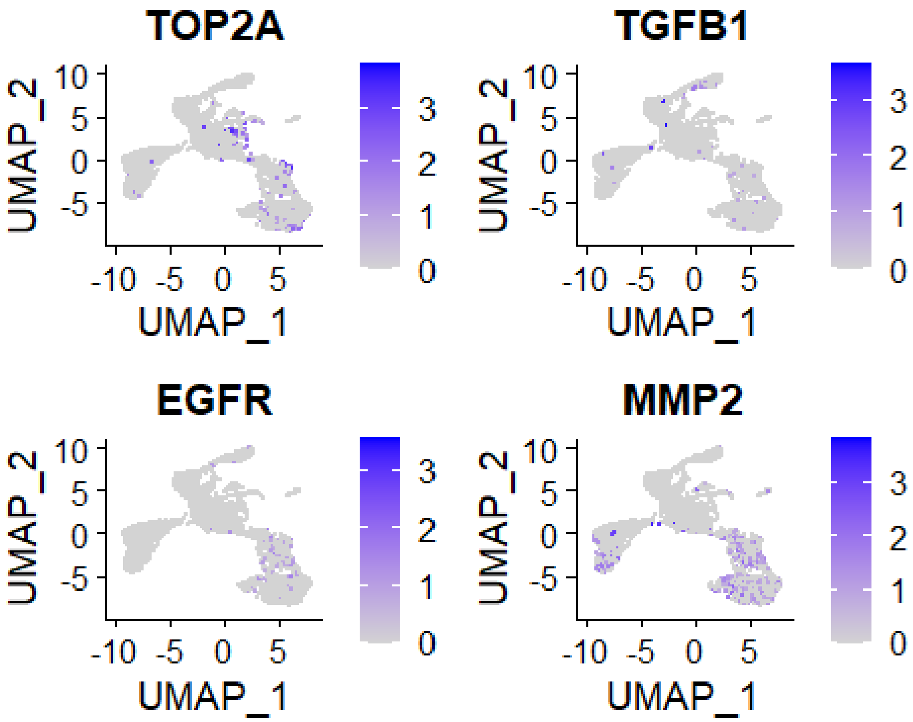
```
##                p_val avg_logFC pct.1 pct.2    p_val_adj
## TMSB4X 1.948869e-59  1.581852 1.000 0.985 4.078398e-55
## TMSB10 1.678282e-55  1.898561 0.986 0.567 3.512142e-51
## GNG11  3.656587e-52  2.723763 0.767 0.051 7.652140e-48
## ACTB   2.540437e-35  1.242755 0.979 0.633 5.316373e-31
## RPS19  1.813209e-32  1.004749 0.966 0.753 3.794503e-28
## GYPC   4.352352e-24  2.003708 0.452 0.051 9.108166e-20
## AIF1   7.397500e-23 -2.934831 0.007 0.487 1.548075e-18
## EGFL7  1.681150e-22  2.289790 0.322 0.004 3.518142e-18
## TUBA1A 6.674352e-22  2.100479 0.411 0.044 1.396742e-17
## FKBP1A 1.112164e-21  1.506513 0.630 0.211 2.327425e-17
```

```
# Find markers for every cluster compared to all remaining cells.
org.markers = FindAllMarkers(org, min.pct = 0.25, logfc.threshold = 0.25)

org.markers %>% group_by(cluster) %>% top_n(n = 2, wt = avg_logFC)

## # A tibble: 142 x 7
## # Groups:   cluster [71]
##       p_val avg_logFC pct.1 pct.2 p_val_adj cluster gene
##       <dbl>     <dbl> <dbl> <dbl>     <dbl> <fct>   <chr>
##  1 5.05e-96     2.07  0.583 0.09   1.06e-91 0       UBE2C
##  2 7.40e-89     2.12  0.801 0.216  1.55e-84 0       HMGB2
##  3 7.44e-66     1.33  1     0.985  1.56e-61 1       TMSB4X
##  4 2.03e-37     1.40  0.767 0.38   4.26e-33 1       GNG11
##  5 3.03e-44     2.05  0.529 0.148  6.33e-40 2       DUSP23
##  6 6.65e-44     2.13  0.42  0.091  1.39e-39 2       TCF21
##  7 7.32e-22     1.55  0.387 0.132  1.53e-17 3       VAMP8
##  8 3.08e- 8     1.66  0.394 0.266  6.45e- 4 3       AIF1
##  9 1.64e-13     0.499 0.919 0.941  3.44e- 9 4       RPS27
## 10 8.72e-12     0.522 0.83  0.713  1.82e- 7 4       MT-CO2
## # ... with 132 more rows
```

```
cluster1.markers <- FindMarkers(org, ident.1 = 0, logfc.threshold = 0.25,
test.use = "roc")
head(cluster1.markers)

##        myAUC  avg_diff power pct.1 pct.2
## TUBA1B 0.867 1.3519141 0.734 0.954 0.652
## HMGB2  0.852 2.1211405 0.704 0.801 0.216
## H2AFZ  0.845 1.1387391 0.690 0.934 0.656
## HMGB1  0.842 1.0908293 0.684 0.967 0.746
## HMGN2  0.831 1.0868828 0.662 0.921 0.682
## TUBB   0.813 0.9053725 0.626 0.967 0.803
```

Identifying the cells expressing specific genes.

```r
FeaturePlot(org, features = c("TOP2A", "TGFB1", "EGFR" , "MMP2"))
```



Assigning cell type identity to clusters