# scRNA-Seq Data Analysis Report

## Roll number:

1068

## Word Count:

1757

## Abstract

To figure out the effect of a specific diet on the mouse mammary gland, two scRNA-seq datasets from the control and treatment group were analysed using *R* and packages like *Seurat* and *SingleR*. After the annotation, sixteen cell types were identified. The numbers of macrophages and fibroblasts are higher, and the numbers of adipocytes and luminal cells are lower in the treatment group compared to the control group. Rn18s-rs5, mt-Rnr1, mt-Rnr2, and Scd1 are suppressed in adipocytes, basal cells, and luminal cells from the treatment group, while the Gpc6 expression is stimulated in basal and luminal cells. According to the results, the specific diet may inhibit protein synthesis and lipid metabolism and promote cell growth and proliferation. It also can have a potential for mammary lipomatosis therapy.

## Introduction

The mammary gland is a unique organ that distinguishes mammals from all other animals, which is fully developed only after birth and undergoes many changes during a female's lifetime. The cell fate specification is regulated by hormonal cues, paracellular interactions, and microenvironment composition (Inman *et al.*, 2015). A specific diet has been discovered recently for altering the development of the mammary gland, whose mechanisms are still unclear. In order to uncover the detailed cellular and molecular changes, scRNA-seq was performed in the mammary gland from normal mice (Control) and mice fed with a specific diet (Treatment). The sequence data were analysed to represent the difference in cell composition and gene expression between the two groups.

## Methods

*R* version 4.1.3 (2022-03-10) with packages *Seurat* version 4.3.0, *SingleR* version 1.8.1, *dplyr* version 1.1.2, *ggplot2* version 3.4.2, and *pheatmap* version 1.0.12 was used to analyse the scRNA-seq data. *Seurat* is a popular and powerful tool for scRNA-seq data analysis, which is the principal R package used in the analysis. It provides a comprehensive suite of functions and methods, enabling researchers to explore and gain insights into cellular heterogeneity and gene expression patterns at a single-cell level. This R package is widely used and highly recommended by many researchers. Besides, the provided raw data per sample is separated into three files: barcodes.tsv.gz, features.tsv.gz, and matrix.mtx.gz, which are the output files of Cell Ranger upstream analysis, and can be easily imported and constitute Seurat objects. For the other packages, *SingleR* was used to identify and characterize cell types in the dataset, *dplyr* helped for better data manipulation, and *ggplot2* and *pheatmap* offered flexibility in visualizing data. The overall workflow is shown in Figure 1.

Datasets "Control" and "Treatment" were loaded and constructed respectively as two Seurat objects. Quality control (QC) was performed to remove the data with low quality. Genes detected in less than 3 cells and cells with less than 200 genes were excluded when creating the objects. Genes starting with "mt-" were characterized as mitochondrial RNA, whose percentage was determined for each cell and stored in a new column of metadata. The total UMI, gene, and mitochondrial percentage numbers were represented through violin plots, while the correlations of the three features were represented by scatter plots. Cells with detected genes less than 300 or more than 9000 or with mitochondrial percentage more than 5 were excluded during QC.

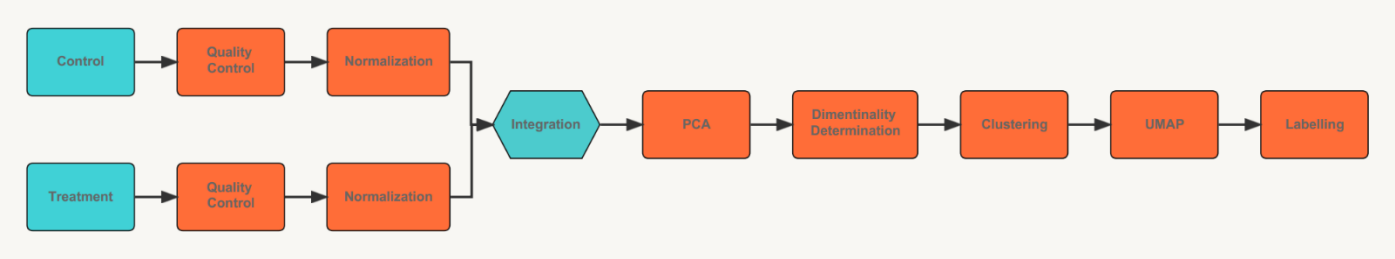



**Figure 1.** Overall workflow of the scRNA-seq data analysis.

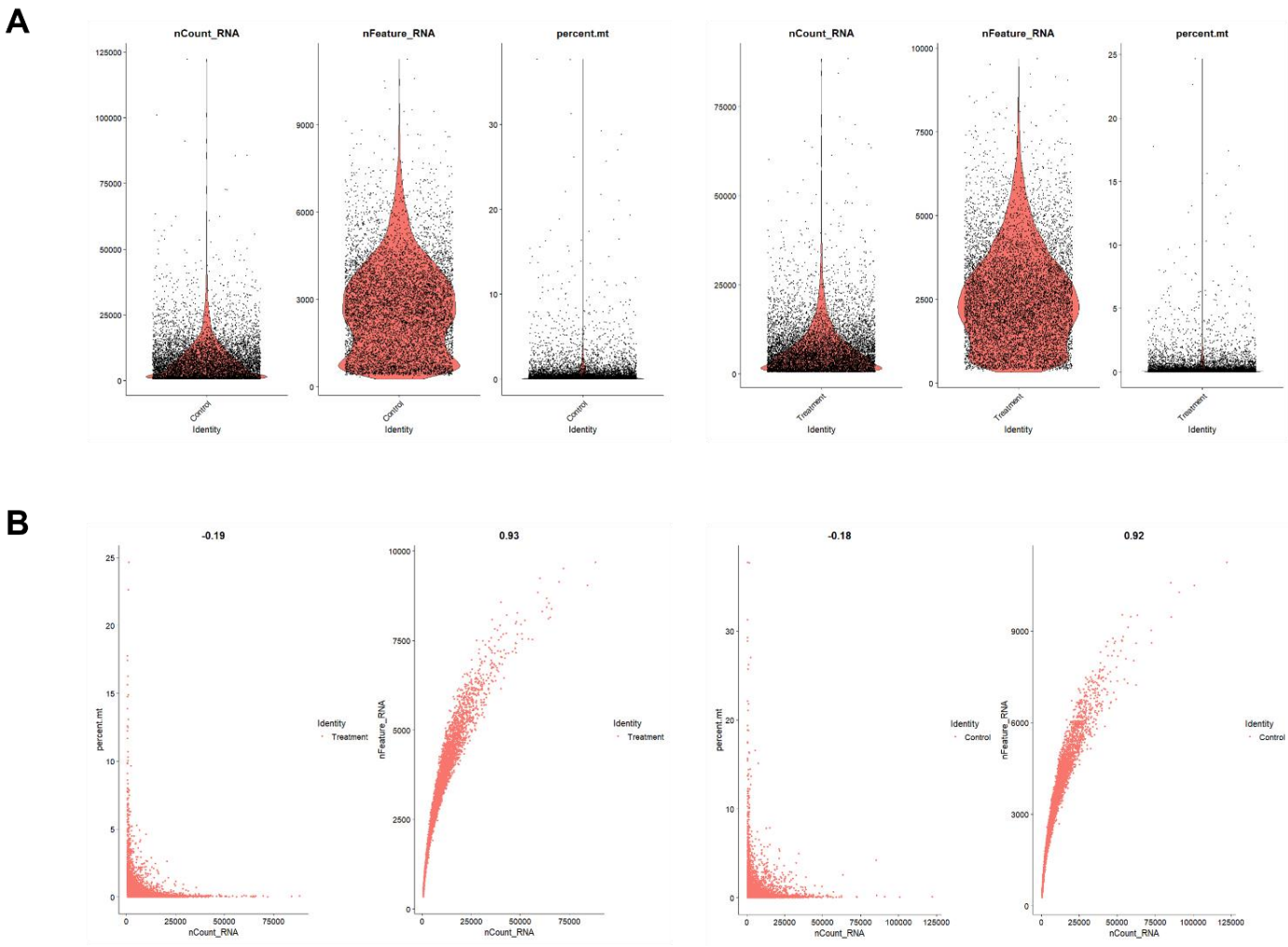



**Figure 2.** Insight of the raw datasets (Left-Control, Right-Treatment)
(A) Violin plots for QC metrics. (B) Correlations between different features.

**Table 1.** Summary of QC

| Group | Before/After | Features | Cells |
|---|---|---|---|
| Control | Before | 55357 | 7930 |
| | After | 28360 | 7738 |
| Treatment | Before | 55357 | 8624 |
| | After | 28199 | 8545 |

Data normalization and feature selection were both done by well-set functions NormalizeData() and FindVariableFeatures() in *Seurat* with default parameters. The control and treatment group data were normalized by the method “Log Normalize” with a scale factor of 10000. Next, two thousand hypervariable genes were found using the “vst” method in each dataset. The features were saved within the Seurat objects and visualized in expression-variance coordinates.

Data integration of the control and treatment group facilitates the downstream analysis and comparison of cell-type composition and gene expression. Cross-dataset pairs of cells that are in a matched biological state were identified as anchors and applied to the merge of two datasets. The integrated dataset was then used for the following standard workflow.

For data scaling, a linear transformation was applied to the datasets. To prepare for dimensional reduction techniques, the mean expression and variance across cells were shifted to 0 and 1 respectively. Function ScaleData() with default parameters was used.

The dimensionality of the data was reduced by principal component analysis (PCA), which is a linear dimensional reduction technique commonly used for primary dimensional reduction. It facilitated computational efficiency, noise reduction, and data visualization in scRNA-seq analysis. The results of PCA were shown by VizDimLoadings(), DimPlot(), and DimHeatmap().

Some principal components (PCs) calculated by PCA are still not tidy and reliable, so the number of included components needs to be determined. Here, the function ElbowPlot() was performed, and 15 PCs were kept according to the plot.

Next, the cells were clustered by FindNeighbors() and FindClusters(), whose functionality is based on the K-nearest neighbour (KNN) graph. In a Seurat tutorial, it is recommended that setting the “resolution” parameter of FindClusters() between 0.4-1.2 typically returns good results for single-cell datasets of around 3K cells (*Seurat - Guided Clustering Tutorial*, no date). For the current large joint dataset, this parameter was set to 1.2.

Running a non-linear dimensional reduction like UMAP and tSNE can help to clarify the overlapping data points in PCA. UMAP was applied to the first to fifteenth PCA dimensions of the current dataset, forming a scatter plot with appropriate data point distribution.

Then, different cell types were identified and characterized based on the UMAP result. The cell-type annotation was performed through SingleR() in package SingleR according to the biomarkers in the MouseRNAseqData database. The result labels were saved in the metadata of the combined Seurat object. Furthermore, the annotation quality was verified by score heatmap and delta distribution. The labelled clusters were compared with the Seurat clusters as well.

The cell number and quality of SingleR-labelled epithelial cells were both too low, and basal and luminal cells were not distinguished, so two biomarkers, *Procr* and *Elf5* were retrieved from CellMarker to re-label them respectively. According to the feature plots of the two genes, cells in Seurat clusters 7 and 9 were re-identified as basal cells, while cells in clusters 5 and 22 were re-identified as luminal cells.

Finally, the cell composition change and gene expression alternations could be represented. A scatter plot and a bar plot showed the cell composition comparison. The top 10 alterations of adipocytes, basal cells, and luminal cells were found by FindMarkers() with its default Wilcox Rank Sum test and visualized through dot plots violin plots.

All R codes are attached to the supplementary materials.

## Results

The overall quality of the two scRNA-seq raw datasets was good (Figure 2). The violin plots showed that the feature numbers of each cell are around 3000, the read counts are mostly lower than 20000, and the mitochondrial gene percentages are low. The scatter plots also showed low mitochondrial percentages and nice feature-count correlations. Features detected in less than 3 cells and cells with detected genes less than 300 or more than 9000 or with mitochondrial percentage more than 5 were excluded through QC. The result of QC is summarized in Table 1.

According to the PCA dimensional heatmap (Supplementary Figure 2) and the elbow plot, the majority of true signals were captured in the first 15 components (Figure 3). The vertical and horizontal cross lines are comparatively clear before PC 15, likewise, the “elbow” of the elbow plot can be observed around PC 14 and 15. Thus, the first 15 PCs were kept for subsequent analysis.

Among the joint dataset, thirty clusters were distinguished by *Seurat*, and sixteen cell types were identified by SingleR referring to the database MouseRNAseqData. Adipocytes had the highest cell number, while the proportions of fibroblasts and endothelial cells, which build the basic construction of the mammary gland, were also at a high level. Many immune cells were detected and annotated, like macrophages, T cells, NK cells, monocytes, and B cells, as well as neurons, astrocytes, and microglia. Basal cells and luminal cells were annotated as two clusters by hand using markers *Procr* and *Elf5.* The scatter plot and the bar plot suggest that adipocytes and luminal cells in the control group are more than the treatment group, and the treatment group has more macrophages and fibroblasts than the control group. The number of cell type in the mammary gland stays the same (Figure 4).

There are over a hundred significant alterations of gene expression determined by the Wilcox Rank Sum test in each cell type of adipocytes, basal cells, and luminal cells. Among the ten most significant alterations, the expression level of the control group tends to be higher. Rn18s-rs5, mt-Rnr1, mt-Rnr2, and Scd1 are suppressed in each cell type from the treatment group. The expression changes in basal cells and luminal cells look similar, they both have promoted Gpc6 expression (Figure 5).

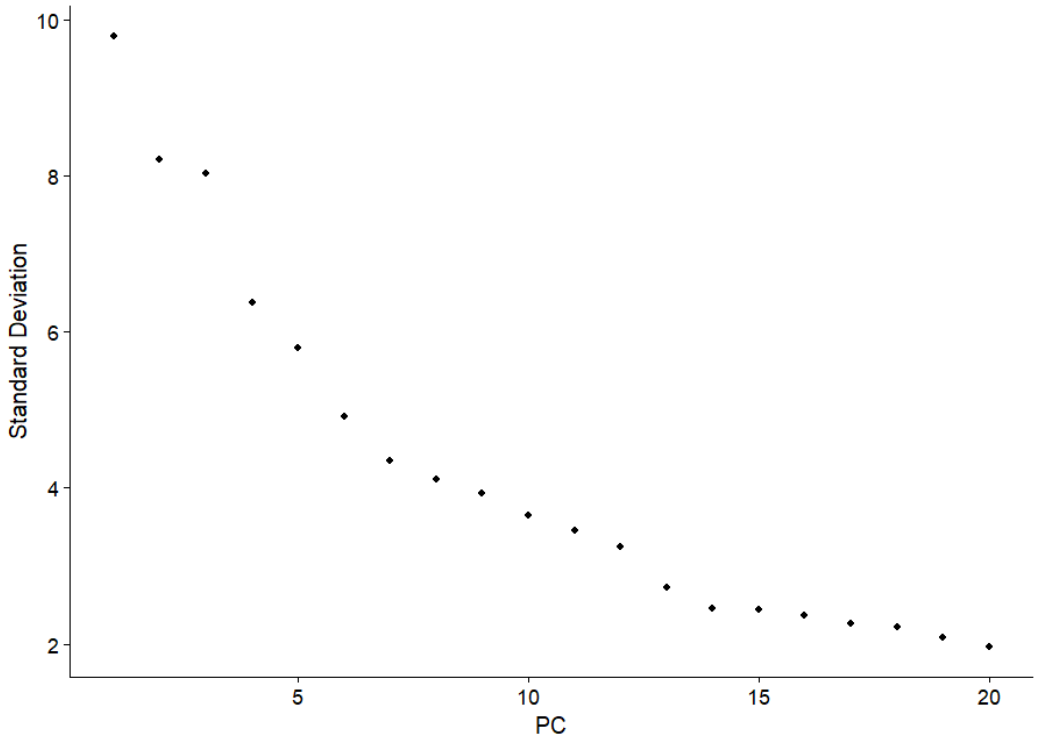



**Figure 3.** Elbow plot for PCA.

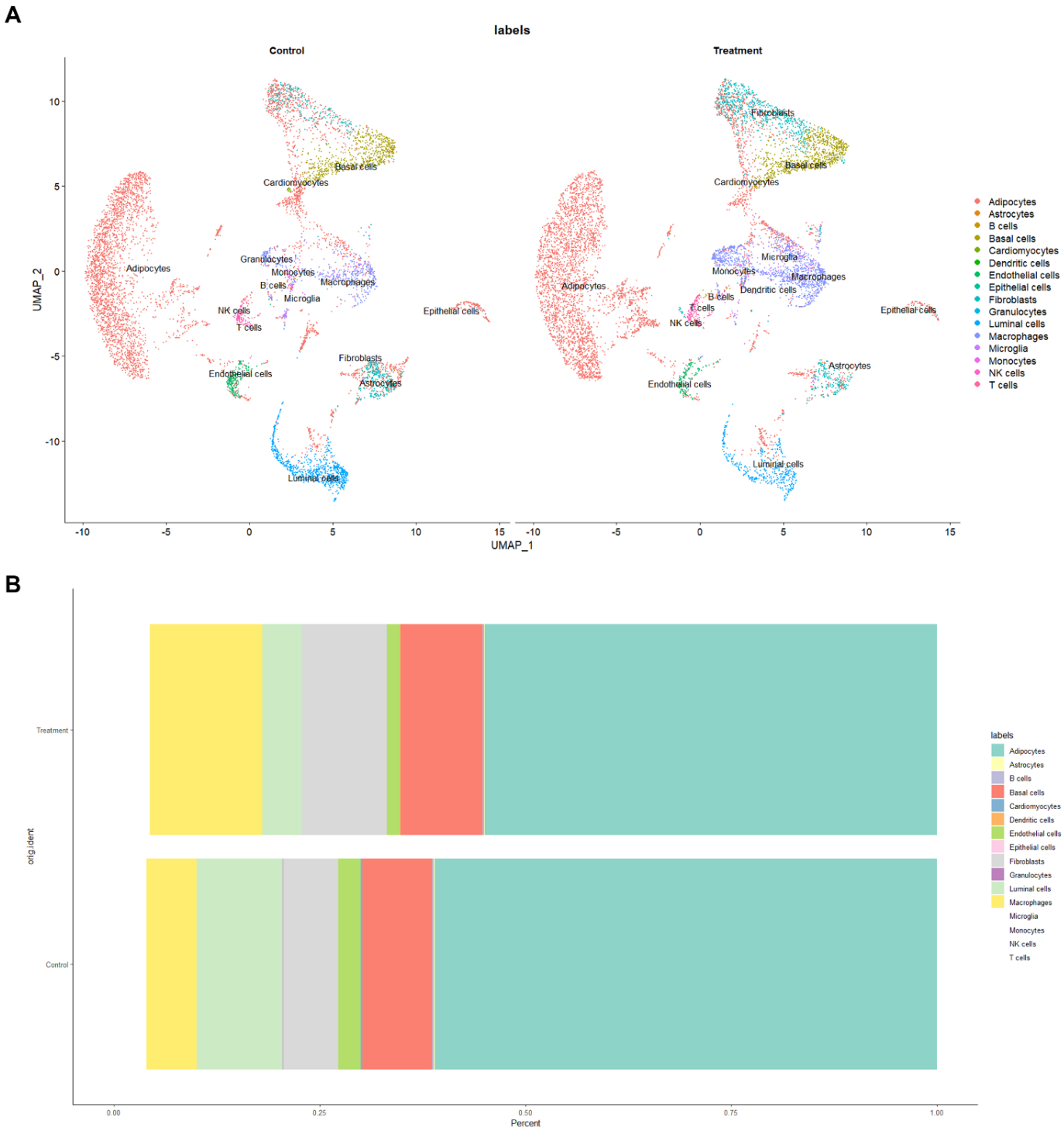



**Figure 4.** Cell-type compositions of the control and treatment group.
(A) DimPlot from *Seruat*. (B) Barplot from *ggplot2*.

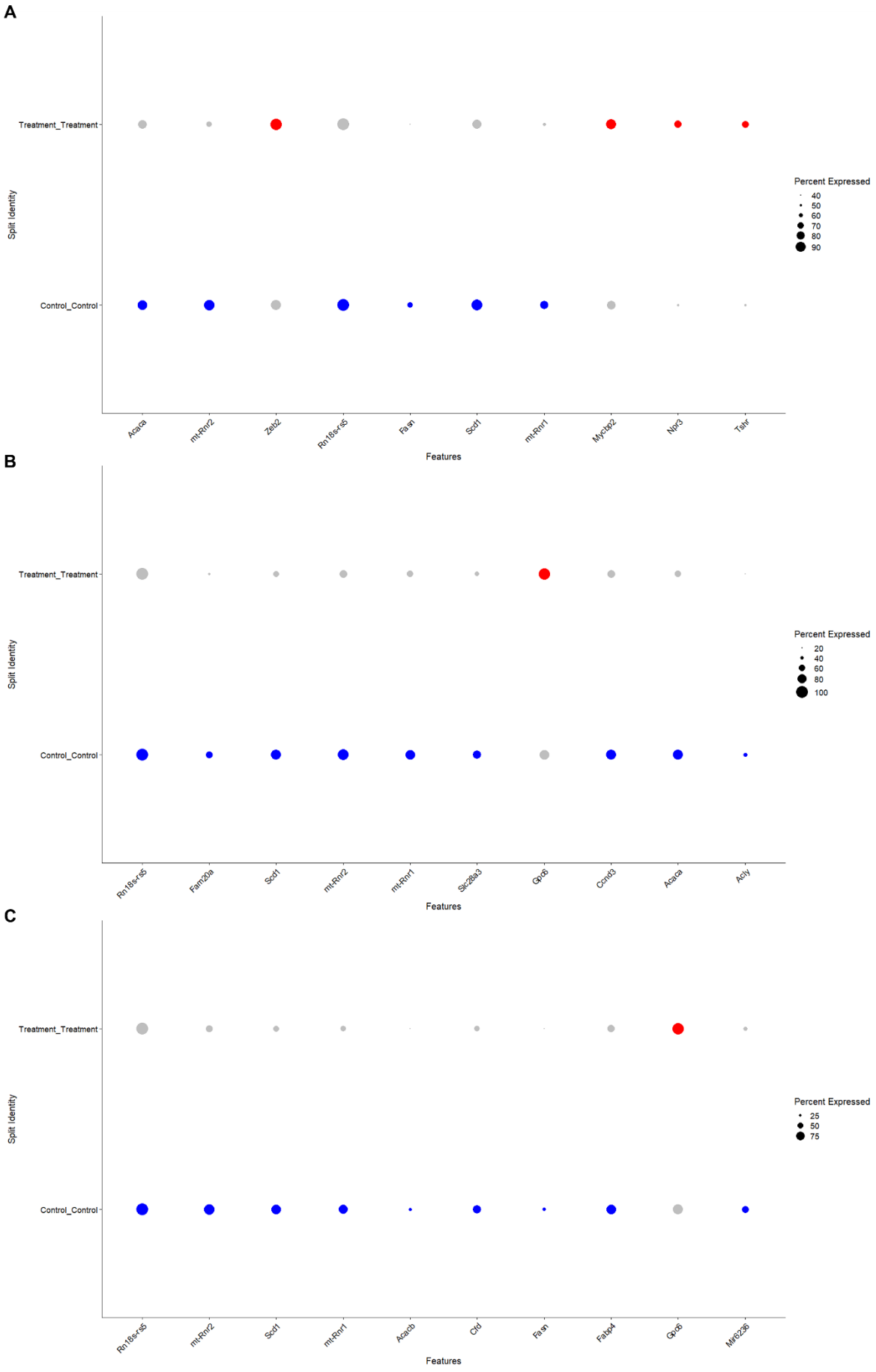



**Figure 5.** Gene expression alterations in adipocytes, basal cells, and luminal cells.
(A) Adipocytes. (B) Basal cells. (C) Luminal cells.

## Discussion

According to the results of the scRNA-seq dataset analysis, the mammary gland has decreased adipocytes and luminal cells and increased macrophages and fibroblasts, which indicates that the specific diet may prevent and cure mammary lipomatosis and enhance the immune response in mammary glands. Basal cells and luminal cells are both epithelial cells, they are likely to have similar receptors and biological pathways, which can be the reason why they have many expression alterations in common under the treatment condition. Some altered genes express RNA, Rn18s-rs5 is an 18s ribosome RNA (rRNA), mt-Rnr1 and mt-Rnr2 are 12s and 16s mitochondrial rRNA, whose low expressions may reduce the protein synthesis in cytoplasm and mitochondria. Scd1 expresses stearoyl-Coenzyme A desaturase 1, which catalyses the synthesis of monounsaturated fatty acids and is vital in lipid metabolism (Tian *et al.*, 2022). Its suppression might explain the decrease of adipocytes in the treatment group. Glypican 6 is a heparan sulphate proteoglycan expressed by Gpc6, it modulates signalling of numerous growth factors and control cell growth and division (Kim *et al.*, 2011). Other cells possibly grow and proliferate faster due to high Gpc6 in basal and luminal cells.

After dimensionality determination, 15 PCs were kept for downstream analysis. However, the position of “elbow” in the elbow plot is not correct to 15, so more or fewer PCs can be included and perform the analysis to see if there would be any change in results. Cardiomyocytes and hepatocytes are not mammary cells, but they were still labelled by *SingeR*. This may be caused by sampling problems in the wet experiments and the bias of *SingleR* algorithm and database. Applying multiple databases for annotation is an optional way heading to higher accuracy. Furthermore, using one marker each to label basal and luminal cells can bring errors, more genes should be employed in future research.

## Reference

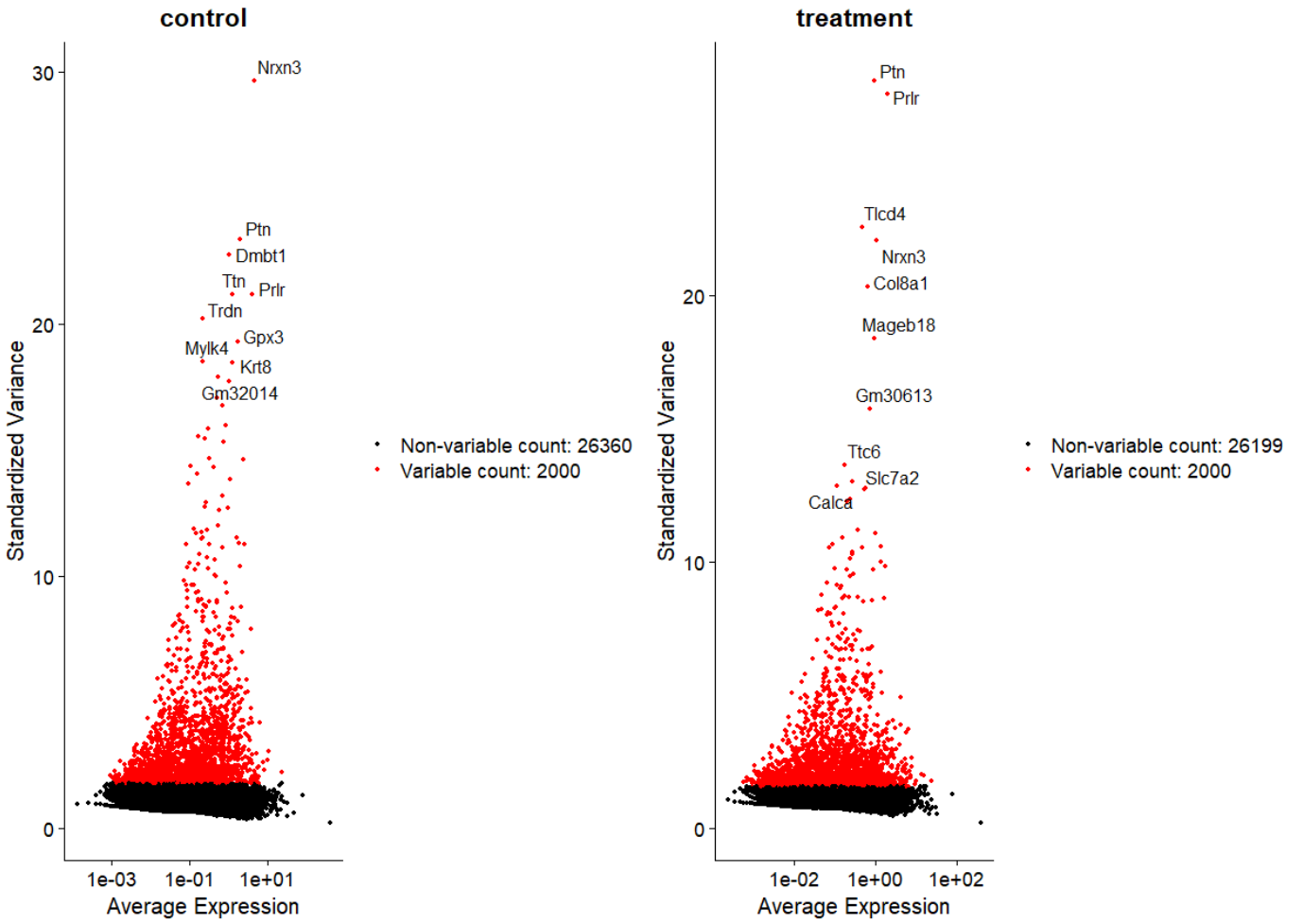

Inman, J.L. *et al.* (2015) ‘Mammary gland development: cell fate specification, stem cells and the microenvironment’, *Development*, 142(6), pp. 1028–1042. Available at: https://doi.org/10.1242/dev.087643.

Kim, M.-S. *et al.* (2011) ‘Structure of the protein core of the glypican Dally-like and localization of a region important for hedgehog signaling’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(32), pp. 13112–13117. Available at: https://doi.org/10.1073/pnas.1109877108.

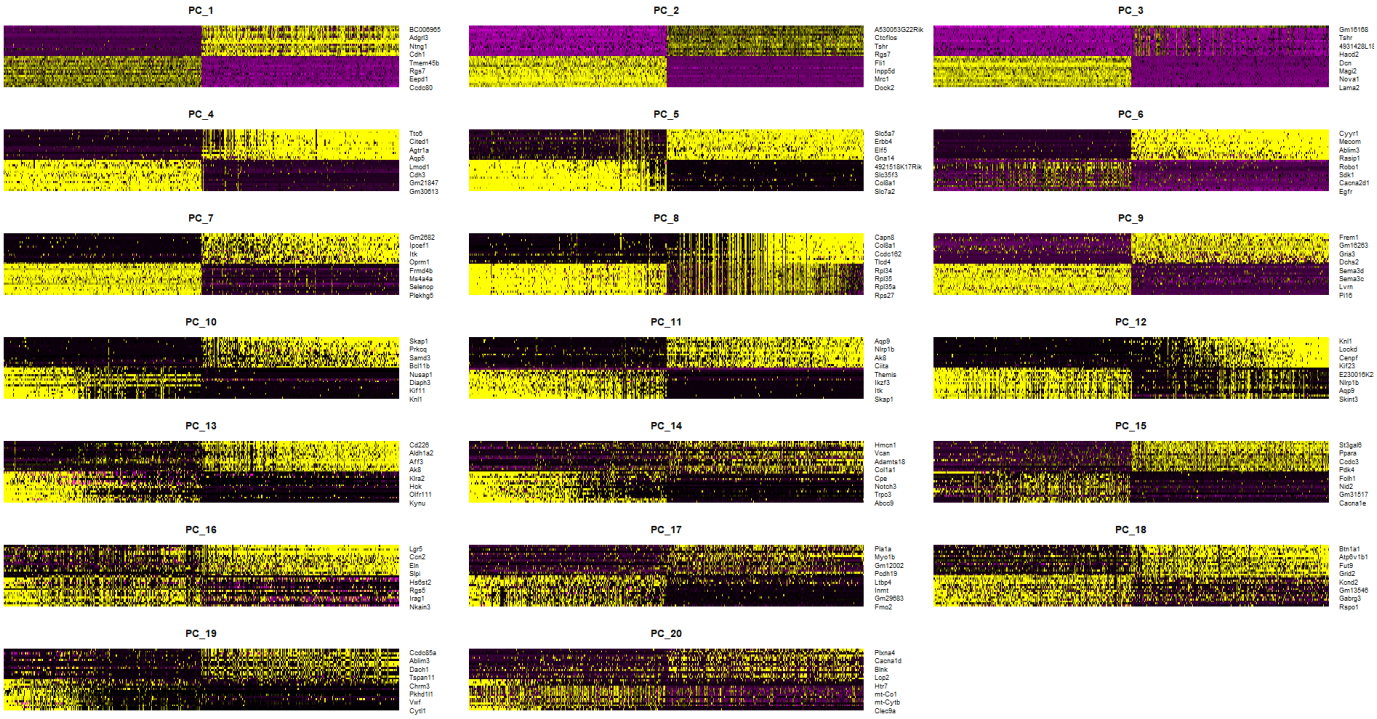
*Seurat - Guided Clustering Tutorial* (no date). Available at: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html (Accessed: 20 May 2023).

Tian, H. *et al.* (2022) ‘Effects of CRISPR/Cas9-mediated stearoyl-Coenzyme A desaturase 1 knockout on mouse embryo development and lipid synthesis’, *PeerJ*, 10, p. e13945. Available at: https://doi.org/10.7717/peerj.13945.

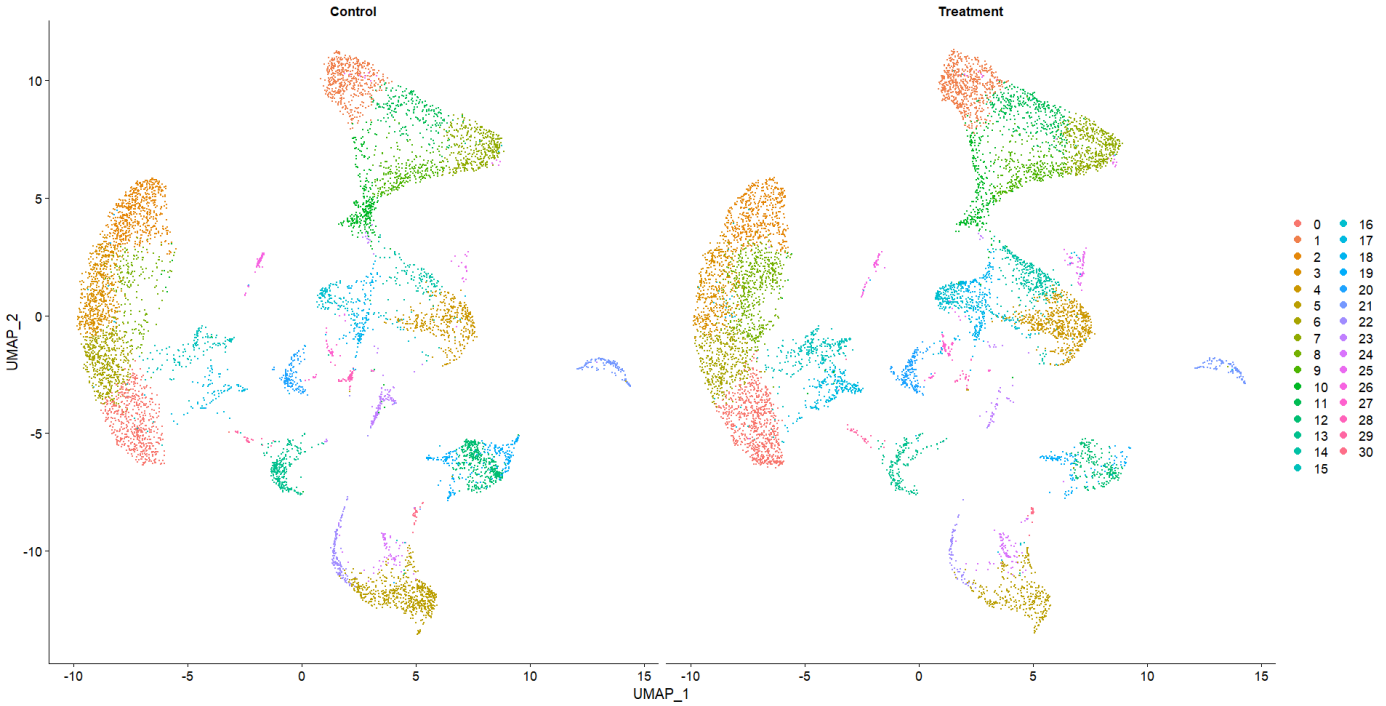
## Supplementary Material



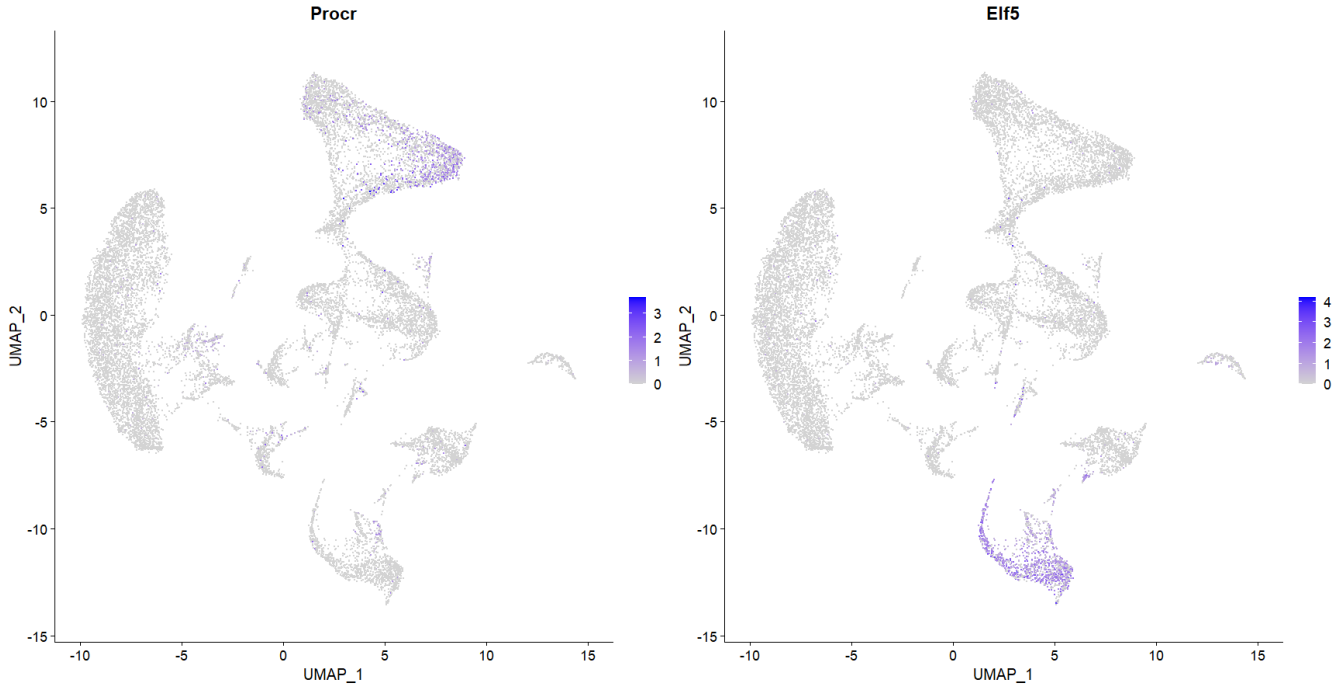

**Supplementary Figure 1.** The hypervariable genes of the control and treatment group.




**Supplementary Figure 2.** PCA Dimensional heatmaps.




**Supplementary Figure 3.** The Seurat clusters of the control and treatment group.



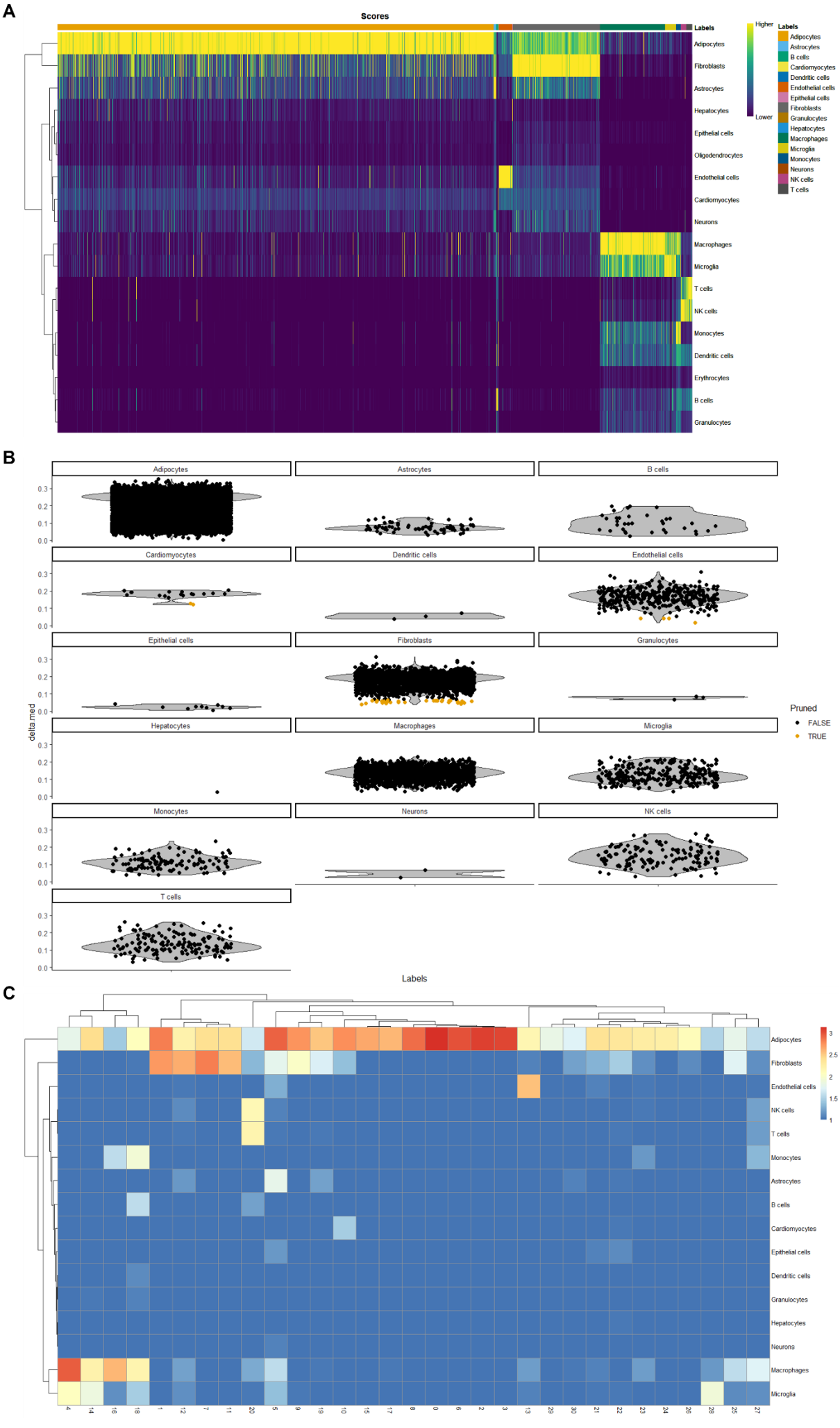

**Supplementary Figure 4.** Distributions of the cells expressing *Procr* and *Elf5*.

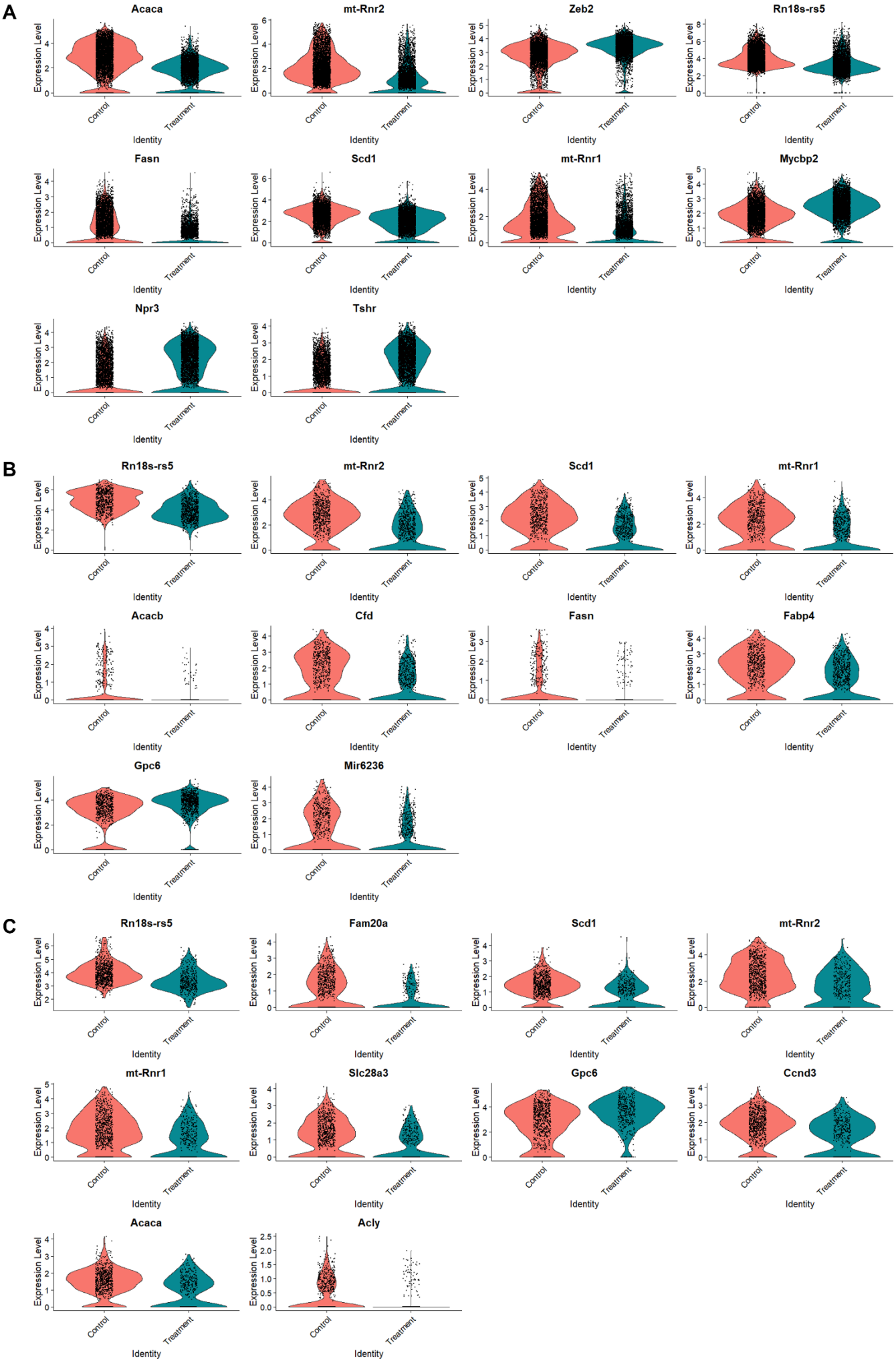
**Supplementary Figure 5.** Validation of SingleR labelling. (Next page)

(A) The heatmap of the SingleR assignment scores across all cell-label combinations.

(B) The distribution of deltas across cells assigned to each reference label.

(C) Comparison between Seurat clusters and SingleR labelling.


A
Scores
Labels
Adipocytes
Fibroblasts
Astrocytes
Hepatocytes
Epithelial cells
Oligodendrocytes
Endothelial cells
Cardiomyocytes
Neurons
Macrophages
Microglia
T cells
NK cells
Monocytes
Dendritic cells
Erythrocytes
B cells
Granulocytes
Higher
Lower
Labels
Adipocytes
Astrocytes
B cells
Cardiomyocytes
Dendritic cells
Endothelial cells
Epithelial cells
Fibroblasts
Granulocytes
Hepatocytes
Macrophages
Microglia
Monocytes
Neurons
NK cells
T cells
B
Adipocytes
Astrocytes
B cells
Cardiomyocytes
Dendritic cells
Endothelial cells
Epithelial cells
Fibroblasts
Granulocytes
Hepatocytes
Macrophages
Microglia
Monocytes
Neurons
NK cells
T cells
delta.med
0.0
0.1
0.2
0.3
Pruned
FALSE
TRUE
Labels
C
Adipocytes
Fibroblasts
Endothelial cells
NK cells
T cells
Monocytes
Astrocytes
B cells
Cardiomyocytes
Epithelial cells
Dendritic cells
Granulocytes
Hepatocytes
Neurons
Macrophages
Microglia
3
2.5
2
1.5
1
4
14
16
18
1
12
7
11
20
5
9
19
10
15
17
8
0
6
2
3
13
29
30
21
22
23
24
26
28
25
27


A
Acaca
mt-Rnr2
Zeb2
Rn18s-rs5
Fasn
Scd1
mt-Rnr1
Mycbp2
Npr3
Tshr
B
Rn18s-rs5
mt-Rnr2
Scd1
mt-Rnr1
Acacb
Cfd
Fasn
Fabp4
Gpc6
Mir6236
C
Rn18s-rs5
Fam20a
Scd1
mt-Rnr2
mt-Rnr1
Slc28a3
Gpc6
Ccnd3
Acaca
Acly
Expression Level
Control
Treatment
Identity

**Supplementary Figure 6.** Violin plots of the top 10 altering expression genes. (Previous page) (A) Adipocytes. (B) Basal cells. (C) Luminal cells.

## R Codes

```
library(dplyr)
library(Seurat)
library(SingleR)
library(ggplot2)
library(pheatmap)
dir_control <- "./Control_MG_10X_Matrix/"
dir_treatment <- "./Treatment_MG_10X_Matrix/"

# Import the data
counts_control <- Read10X(data.dir = dir_control)
counts_treatment <- Read10X(data.dir = dir_treatment)

dim(counts_control)
dim(counts_treatment)

# Create Seurat objects
control <- CreateSeuratObject(counts = counts_control,project = "Control",min.cells = 3,min.features = 200)
treatment <- CreateSeuratObject(counts = counts_treatment,project = "Treatment",min.cells = 3,min.features = 200)

control@assays
treatment@assays

### Quality Control

# Add a new column 'percent.mt' for each object
control[["percent.mt"]] <- PercentageFeatureSet(control,pattern = "^mt-")
treatment[["percent.mt"]] <- PercentageFeatureSet(treatment,pattern = "^mt-")

# View the meta.data
head(control@meta.data)
head(treatment@meta.data)
# nCount_RNA: number of UMIs per cell
# nFeature_RNA: number of gene detected per cell

# Visualization by violin plots
VlnPlot(control,features = c("nCount_RNA","nFeature_RNA","percent.mt"),ncol = 3)
VlnPlot(treatment,features = c("nCount_RNA","nFeature_RNA","percent.mt"),ncol = 3)

# Correlation between nCount_RNA and percent.mt, nCount_RNA and nFeature_RNA in the control group
plot_con_cp <- FeatureScatter(control,feature1 = "nCount_RNA",feature2 = "percent.mt")
plot_con_cf <- FeatureScatter(control,feature1 = "nCount_RNA",feature2 = "nFeature_RNA")
plot_con_cp+plot_con_cf
# Correlations in the treatment group
plot_treat_cp <- FeatureScatter(treatment,feature1 = "nCount_RNA",feature2 = "percent.mt")
plot_treat_cf <- FeatureScatter(treatment,feature1 = "nCount_RNA",feature2 = "nFeature_RNA")
plot_treat_cp+plot_treat_cf

# Pick the data points that have 300<nFeatures_RNA<9000 and percent.mt<5%
control <- subset(control,subset = nFeature_RNA>300 & nFeature_RNA<9000 & percent.mt<5)
treatment <- subset(treatment,subset = nFeature_RNA>300 & nFeature_RNA<9000 & percent.mt<5)

control@assays
treatment@assays

### Normalization

# The normalized data is stored in control/treatment[["RNA"]]@data
control <- NormalizeData(control)
treatment <- NormalizeData(treatment)

# Find hypervariable genes
control <- FindVariableFeatures(control)
treatment <- FindVariableFeatures(treatment)

# Plot the top10 genes with labels
top10_con <- head(VariableFeatures(control),10)
top10_treat <- head(VariableFeatures(treatment),10)
lp_control <- LabelPoints(plot = VariableFeaturePlot(control),points = top10_con,repel = T)+
  labs(title = "control")
```

```r
lp_treatment <- LabelPoints(plot = VariableFeaturePlot(treatment),points = top10_treat,repel = T)+
  labs(title = "treatment")
lp_control + lp_treatment

### Perform Integration

# Store two objects in one vector
mg.vec <- c(control,treatment)

# Select features that are repeatedly variable across datasets for integration
features <- SelectIntegrationFeatures(object.list = mg.vec)

# Create an 'integrated' data assay with anchors
mg.anchors <- FindIntegrationAnchors(object.list = mg.vec,anchor.features = features)
mg.combined <- IntegrateData(anchorset = mg.anchors)

# Specify that we will perform downstream analysis on the corrected data note that the
# original unmodified data still resides in the 'RNA' assay
DefaultAssay(mg.combined) <- "integrated"

# Run the standard workflow for visualization and clustering

### Data Scaling

# The scaled data is stored in [["RNA"]]@scale.data
mg.combined <- ScaleData(mg.combined,verbose = FALSE)

### Linear Dimensional Reduction-PCA,

# Processed data is stored in [["pca"]]
mg.combined <- RunPCA(mg.combined,npcs = 30,verbose = FALSE)
VizDimLoadings(mg.combined,dims = 1:5,reduction = "pca")
DimPlot(mg.combined,reduction = "pca")
DimHeatmap(mg.combined,dims = 1:20,cells = 500,balanced = T)

### Determine the Vital PCs

# JackStraw is not useful here, elbow plot is chosen
ElbowPlot(mg.combined)
# keep 15 PCs

### Clustering

mg.combined <- FindNeighbors(mg.combined,reduction = "pca",dims = 1:15)
mg.combined <- FindClusters(mg.combined,resolution = 1.2)

### Nonlinear Dimensional Reduction

# UMAP
mg.combined <- RunUMAP(mg.combined,reduction = "pca",dims = 1:15)

# Visualization
p1 <- DimPlot(mg.combined,reduction = "umap",group.by = "orig.ident")
p2 <- DimPlot(mg.combined,reduction = "umap",label = TRUE,repel = TRUE)
p1 + p2

# To visualize the two conditions side-by-side, we can use the split.by argument to show each condition colored by
cluster.
DimPlot(mg.combined,reduction = "umap",split.by = "orig.ident")

### Cell Type Annotation

# For performing differential expression after integration, we switch back to the original data
DefaultAssay(mg.combined) <- "RNA"

# Load the mouse annotation database
load("MouseRNAseqData.RData")

# Use SingleR to annotate cell types
data <- GetAssayData(mg.combined,slot = "data")
mg.mrd <- SingleR(test = data,ref = mouseRNA,labels = mouseRNA$label.main)
mg.combined@meta.data$labels <- mg.mrd$labels
table(mg.combined@meta.data$labels)
print(DimPlot(mg.combined,group.by = c("seurat_clusters","labels"),reduction = "umap",label = T,repel = T))
DimPlot(mg.combined,group.by = "labels",split.by = "orig.ident",reduction = "umap",label = T,repel = T)
# Modification: cardiomyocytes and hepatocytes shouldn't be here
```

```
# Check the quality of annotation
# based on "scores within cells"
print(plotScoreHeatmap(mg.mrd))
# based on per cell "deltas"
plotDeltaDistribution(mg.mrd,ncol = 3)

# Compare with clusters
tab <- table(label = mg.mrd$labels,cluster = mg.combined@meta.data$seurat_clusters)
pheatmap(log10(tab + 10))

# The quality of the epithelial cells is poor, so just find the markers for basal and luminal cells directly

# Basal marker is Procr, luminal marker is Elf5
# (retrieved at CellMarker: http://xteam.xbio.top/CellMarker/)
FeaturePlot(mg.combined,features = c("Procr","Elf5"),reduction = "umap")

# Then we change the labels of the cells expressing the two genes
mg.combined@meta.data[mg.combined@meta.data$seurat_clusters==7|mg.combined@meta.data$seurat_clusters==9,]$labels     <-
"Basal cells"
mg.combined@meta.data[mg.combined@meta.data$seurat_clusters==5|mg.combined@meta.data$seurat_clusters==22,]$labels    <-
"Luminal cells"

### Cell Composition Change

DimPlot(mg.combined,group.by = "labels",split.by = "orig.ident",reduction = "umap",label = T,repel = T)

mg.combined@meta.data %>%
  ggplot(aes(x = orig.ident,fill = labels))+
  geom_bar(position = position_fill())+
  scale_fill_brewer(palette = 'Set3')+
  theme_classic()+
  labs(y = 'Percent')+
  coord_flip()

### Gene Expression Alternations

# adipocytes
adipocytes <- subset(mg.combined,subset = labels=="Adipocytes")
Idents(adipocytes) <- "orig.ident"
adipocytes.markers <- FindMarkers(adipocytes,ident.1 = "Control",ident.2 =  "Treatment",min.pct = 0.25)
adipocytes_signif.markers <- row.names(adipocytes.markers[adipocytes.markers$p_val_adj<0.05,])
adipocytes_alters <- length(adipocytes_signif.markers)
adipocytes_markers.to.plot <- head(adipocytes_signif.markers,10)
DotPlot(adipocytes,features = adipocytes_markers.to.plot,cols = c("blue","red"),dot.scale = 8,split.by = "orig.ident")
+
  RotatedAxis()
VlnPlot(adipocytes,features = adipocytes_markers.to.plot,split.by = "orig.ident")

# basal cells
basal_cells <- subset(mg.combined,subset = labels=="Basal cells")
Idents(basal_cells) <- "orig.ident"
basal.markers <- FindMarkers(basal_cells,ident.1 = "Control",ident.2 =  "Treatment",min.pct = 0.25)
basal_signif.markers <- row.names(basal.markers[basal.markers$p_val_adj<0.05,])
basal_alters <- length(basal_signif.markers)
basal_markers.to.plot <- head(basal_signif.markers,10)
DotPlot(basal_cells,features = basal_markers.to.plot,cols = c("blue","red"),dot.scale = 8,split.by = "orig.ident") +
  RotatedAxis()
VlnPlot(basal_cells,features = basal_markers.to.plot,split.by = "orig.ident")

# luminal cells
luminal_cells <- subset(mg.combined,subset = labels=="Luminal cells")
Idents(luminal_cells) <- "orig.ident"
luminal.markers <- FindMarkers(luminal_cells,ident.1 = "Control",ident.2 =  "Treatment",min.pct = 0.25)
luminal_signif.markers <- row.names(luminal.markers[luminal.markers$p_val_adj<0.05,])
luminal_alters <- length(luminal_signif.markers)
luminal_markers.to.plot <- head(luminal_signif.markers,10)
DotPlot(luminal_cells,features = luminal_markers.to.plot,cols = c("blue","red"),dot.scale = 8,split.by = "orig.ident")
+
  RotatedAxis()
VlnPlot(luminal_cells,features = luminal_markers.to.plot,split.by = "orig.ident")

# Significant alterations of the three cell types
barplot(c(adipocytes_alters,basal_alters,luminal_alters),names.arg = c("Adipocytes","Basal cells","Luminal cells"),col
= c("orangered","light green","light blue"))
```