scRNA_Smart-Seq2

Saleem Mansour

2023-02-19

Introduction

This Analysis involves annotated plate-based mice tissues from Figshare.

The site offers:

1. Annotation file
2. The data archive file (containing csv files of each tissue)
3. Metadata

So I Choose **Brain_Neurons** tissue. My objectives include:

I. Apply the standard pre-processing pipeline Using Seurat Object.

II. Compare DimPlots of generated clusters against gender of mice

III. Visualize the distribution percentages of each subtissue among genders

IV. Find most deferentially expressed genes among two subtissues of my choice and perform proper visualization (including a heatmap)

Workflow

I. Apply standard pre-processing

load needed libraries

```
library(Seurat)
library(ggplot2)
library(plotly)
library(tidyverse)
library(patchwork)
```

And assign needed data:

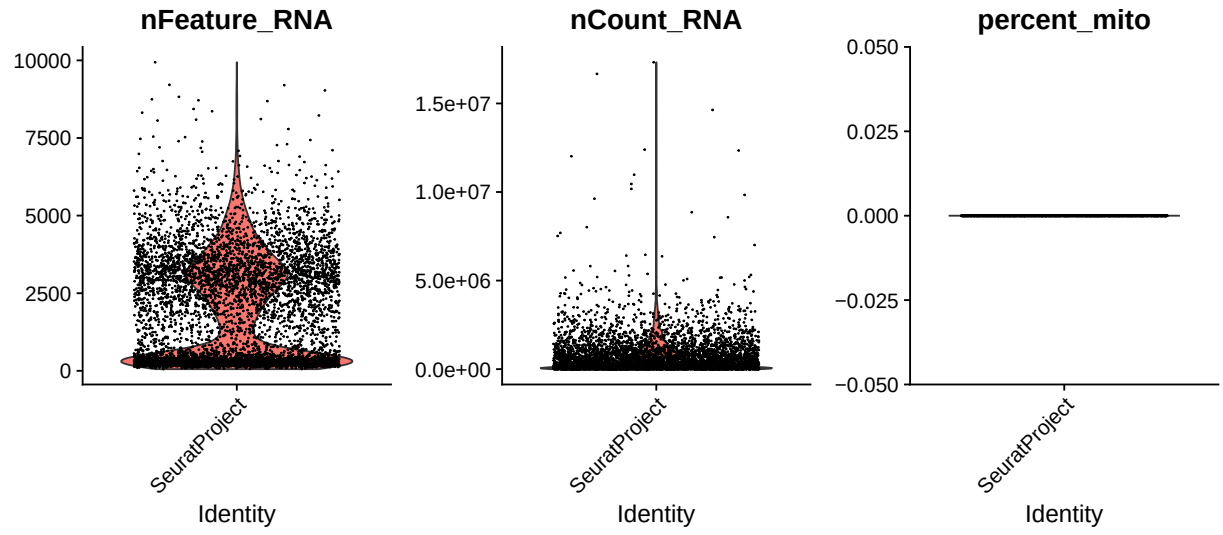
```
data <- read.csv("Seurat/Brain_Neurons-counts.csv", row.names = 1)
meta <- read.csv("Seurat/metadata_FACS.csv")
anno <- read.csv("Seurat/annotations_FACS.csv")
```

I need some wrangling at start (filtering annotation and meta files for my chosen tissue, adding mitochondrial genes percentage column, adding IDs as separate column), and create the Seurat object:

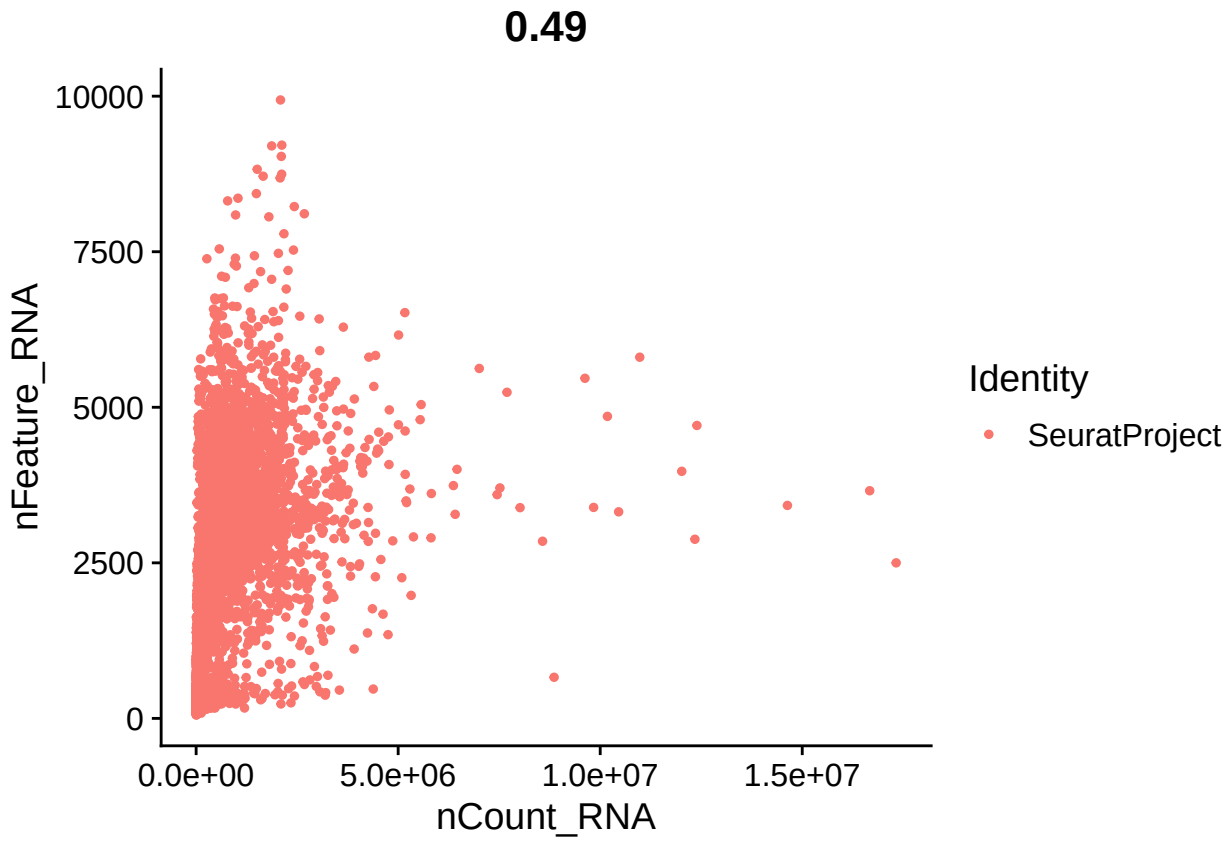
```
anno_filtered <- filter(anno, tissue == "Brain_Neurons")
meta_filtered <- filter(meta, tissue == "Brain_Neurons")

seur <- CreateSeuratObject(data)
seur[["percent_mito"]] = PercentageFeatureSet(seur, pattern = "^MT-")
seur@meta.data$IDs <- rownames(seur@meta.data)
```

A first glance at data:

```
VlnPlot(seur, features = c("nFeature_RNA", "nCount_RNA", "percent_mito"), ncol = 3, )
```




```
FeatureScatter(seur, feature1 = "nCount_RNA", feature2 = "nFeature_RNA")
```



For QC, Seems that cells with mitochondrial genes are already filtered, so I won't include them.

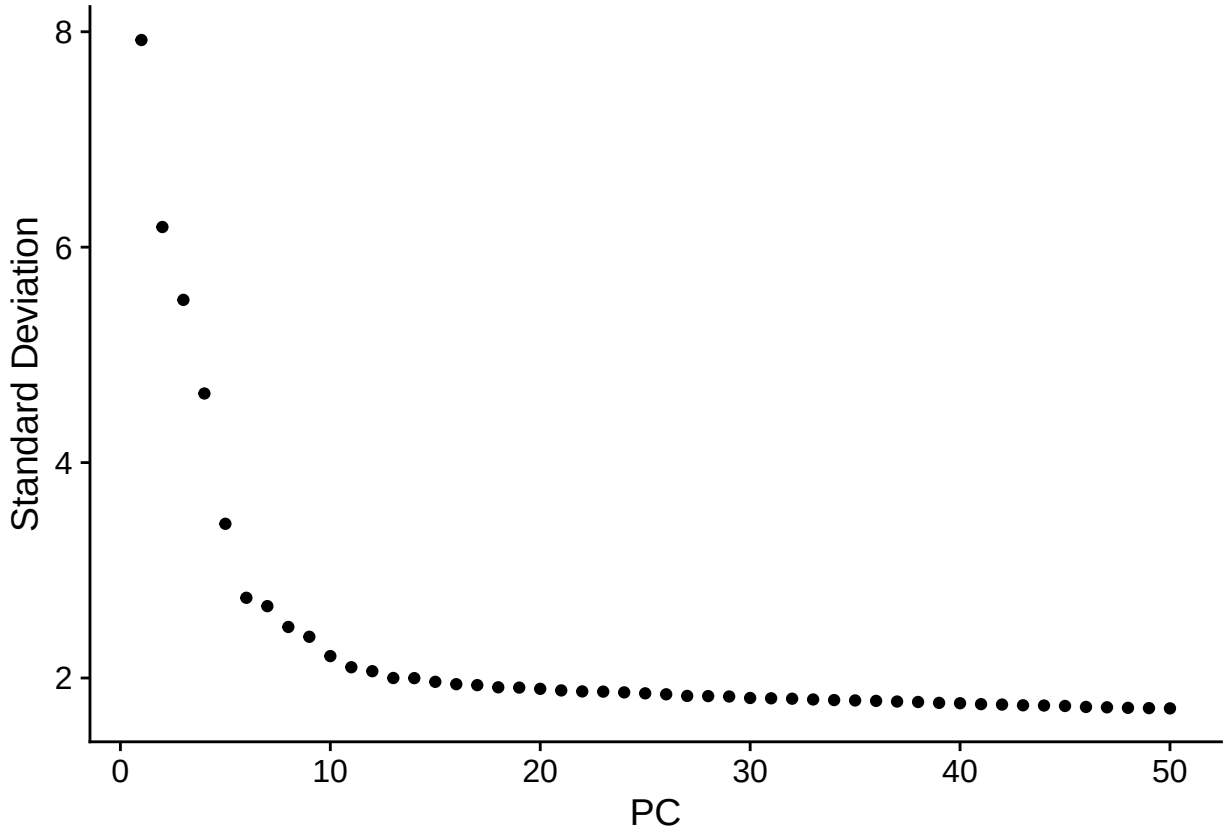
We can conclude from the previous plots (high features and low counts) that probably the sequencing depth (coverage) is not high.

here I construct my main Seurat obj (seur_mod) that results from pre-processing, Dim linear reduction:

```
genes <- rownames(seur)

seur_mod <- seur %>%
  subset(subset = nFeature_RNA > 200 & nFeature_RNA < 2500) %>%
  NormalizeData() %>%
  FindVariableFeatures(selection.method = "vst", nfeatures = 2000) %>%
  ScaleData(features = genes) %>%
  RunPCA()
rm(genes)
```

We can check best dimensionality after which variation insignificant; using Elbow plot (It seems that dims 1:30 hold the majority of variation, as so adjust later steps):

```
ElbowPlot(seur_mod, ndims = 50)
```




```
seur_mod <- seur_mod %>%
  RunUMAP(dims = 1:30) %>%
  FindNeighbors(dims = 1:30) %>%
  FindClusters(resolution = 0.25)
```

II. DimPlots of gender against clusters

-Before commencing to generate plots, we need to create a column for gender in metadata of seur_mod, and another resembles plate.barcode (they are both derivated from IDs, using mutate with regex).

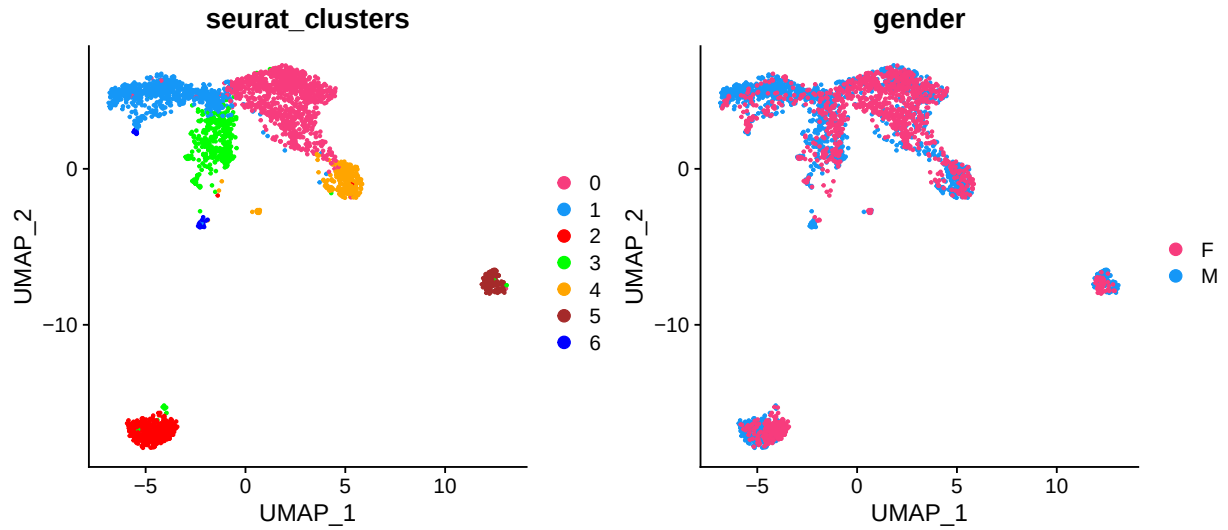
-we create a df containing only plate barcodes and subtissues from meta_filtered

-We merge it with seur_mod metadata by plate barcode, so we have subtissues in it.

```
subtiss <- meta_filtered %>% select(c(1,4))

seur_mod@meta.data <- seur_mod@meta.data  %>% mutate(gender = sub(pattern = "\\w\\d+.\\w+\\d+.\\d+_\\d+
  mutate(gender = sub(pattern = ".1.1", "", x = gender))

seur_mod@meta.data <- seur_mod@meta.data  %>% mutate(onto = sub(pattern = "\\w\\d+.", "",x = IDs)) %>%
  mutate(onto = sub(pattern = ".\\d+_\\d+_\\w.1.1", "", x = onto))

seur_mod@meta.data <- merge(seur_mod@meta.data, subtiss, by.x = "onto", by.y = "plate.barcode")
rownames(seur_mod@meta.data) <- seur_mod@meta.data$IDs
rm(subtiss)
```

DimPlot of gender vs seurat clusters (at res = 0.25):

```
cols = c("#F73C7D", "#1597F7", "red", "green", "orange", "brown", "blue", "darkgray", "black", "pink")

D1 <- DimPlot(seur_mod, reduction = "umap", label = F, group.by = "seurat_clusters")+
  scale_color_manual(values =  cols)
D2 <- DimPlot(seur_mod, reduction = "umap", group.by = "gender", label = F)+
  scale_color_manual(values =  cols)
D1+D2
```

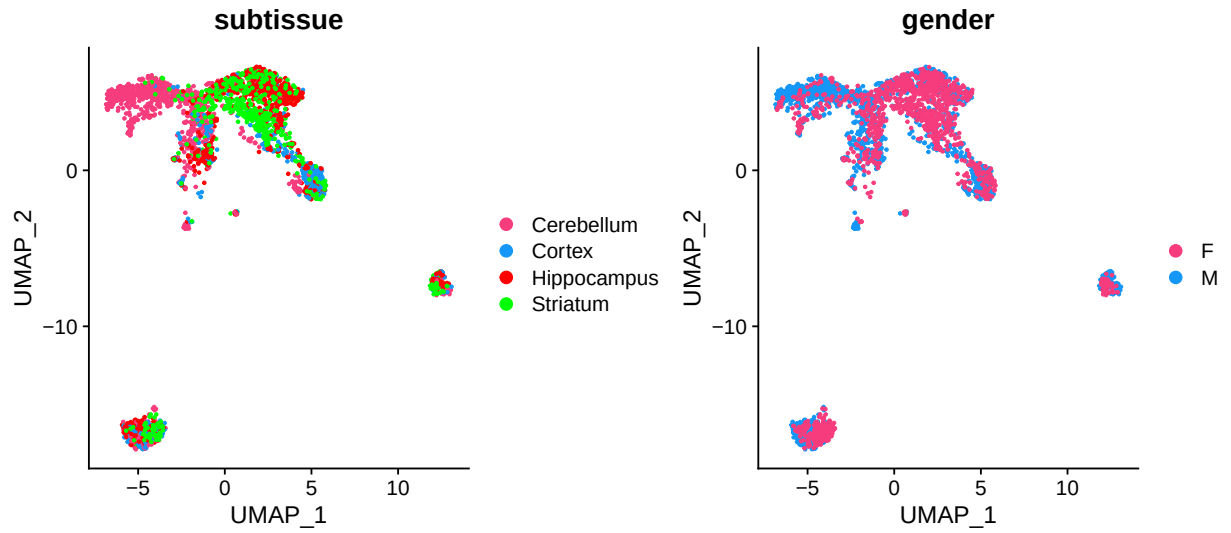



```
rm(D1,D2)
```

We can notice that cluster 0 consists of more female-originated cells, while cluster 1 has more male-originated cells

However, we are also interested in gender distribution in subtissues, so we can make another DimPlot:

```
D1 <- DimPlot(seur_mod, reduction = "umap", label = F, group.by = "subtissue")+
  scale_color_manual(values =  cols)
D2 <- DimPlot(seur_mod, reduction = "umap", group.by = "gender", label = F)+
  scale_color_manual(values =  cols)
D1+D2
```




```
rm(D1,D2)
```

We notice that seurat clustering was close enough to partially distinguish between some subtissues.

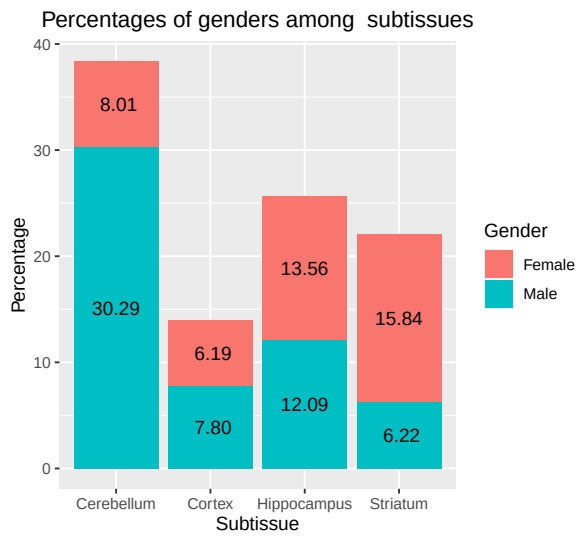
We can see the Cerebellum (position of cluster 1) has majority male-originated cells, while Striatum (position of cluster 0) has its majority of female-originated cells.

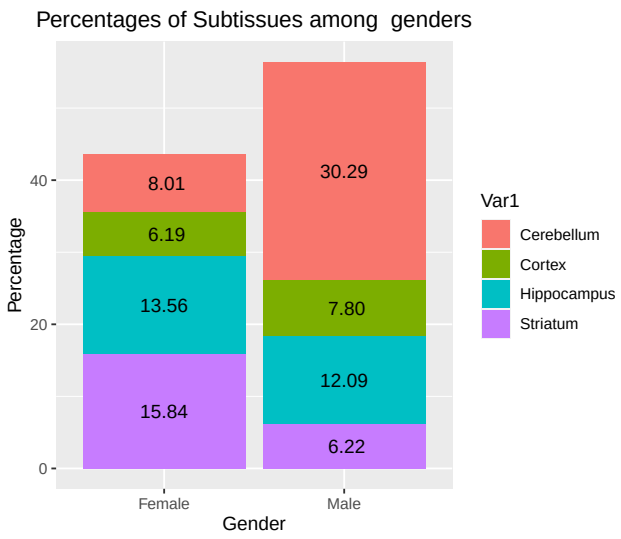
Barplot visualization

We can make a column plot with percentages of genders and subtissues to emphasis our previous observations:

I create a table of frequencies and wrangle it slightly before plotting column plot.

```
t1 <- as.data.frame(table(seur_mod@meta.data$subtissue,seur_mod@meta.data$gender))
t1 <- mutate(t1, perc = Freq/sum(Freq)*100)
t1$Var2 <- recode(t1$Var2, 'F' = "Female", 'M' = "Male"  )
t1 <- rename(t1, 'Gender'="Var2")

p1 <- ggplot(t1, aes(x = Var1, y = perc, fill = Gender ))+
  geom_col()+
  labs(y = "Percentage", x = "Subtissue", title = "Percentages of genders among  subtissues")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_text(aes(label = format(round(perc,2))), position = position_stack(vjust = 0.5))

p2 <- ggplot(t1, aes(x = Gender, y = perc, fill = Var1 ))+
  geom_col()+
  labs(y = "Percentage", x = "Gender", title = "Percentages of Subtissues among  genders")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_text(aes(label = format(round(perc,2))), position = position_stack(vjust = 0.5))
p1 + p2
```






```
rm(t1, p1,p2)
```

IV. Finding markers for DE genes

We noticed previously in the DimPlot (by subtissues) that Cerebellum and Cortex clusters are the least overlapping and also held interesting specifications according to gender, therefore, I choose **Cortex** and **Cerebellum** for further analysis of differentially expressed genes.

First we subset from our project:

```
subset_seur <- seur_mod %>% subset(subset = subtissue == "Cortex" |subtissue == "Cerebellum" )
Idents(subset_seur) <- subset_seur@meta.data$subtissue
```

We can get DE genes by Finding top markers for each cluster !(Following instructions, as sample is not UMI-based, we can't use poisson or negative binom tests)!

We can subset top 10 for further inspection (sorted by log2FC average).

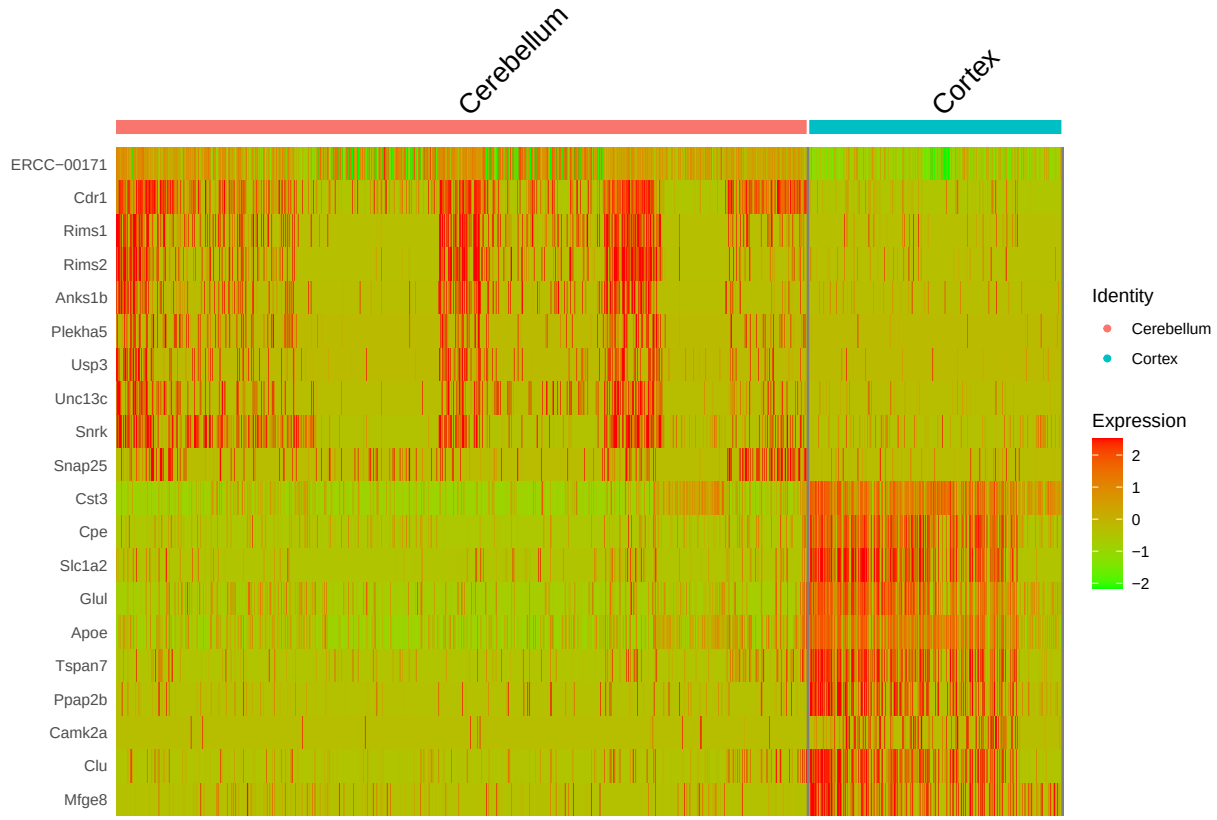
```
cluster_all <- subset_seur %>% FindAllMarkers(min.pct = 0.1, logfc.threshold = 0.25, only.pos = T, test
top10 <- cluster_all %>%
    group_by(cluster) %>%
    top_n(n = 10, wt = avg_log2FC)
top10
```

```
## # A tibble: 20 x 7
## # Groups:   cluster [2]
##        p_val avg_log2FC pct.1 pct.2 p_val_adj cluster    gene
##        <dbl>      <dbl> <dbl> <dbl>     <dbl> <fct>      <chr>
##  1 2.56e- 56       2.45 0.941 0.959  6.01e-52 Cerebellum ERCC-00171
##  2 1.81e- 39       3.45 0.53  0.184  4.24e-35 Cerebellum Cdr1
##  3 1.29e- 20       2.97 0.308 0.079  3.02e-16 Cerebellum Rims1
##  4 1.06e- 16       2.83 0.289 0.092  2.49e-12 Cerebellum Rims2
##  5 6.68e- 14       2.48 0.234 0.066  1.57e- 9 Cerebellum Anks1b
##  6 4.06e- 13       2.41 0.179 0.033  9.51e- 9 Cerebellum Plekha5
```

```
##  7 1.96e- 12       2.39 0.184 0.041  4.59e- 8 Cerebellum Usp3
##  8 1.83e- 11       2.74 0.231 0.087  4.29e- 7 Cerebellum Unc13c
##  9 1.37e-  9       3.01 0.365 0.24   3.22e- 5 Cerebellum Snrk
## 10 1.05e-  4       2.33 0.187 0.113  1   e+ 0 Cerebellum Snap25
## 11 4.96e-102       3.16 0.931 0.503  1.16e-97 Cortex     Cst3
## 12 7.00e- 83       3.78 0.673 0.202  1.64e-78 Cortex     Cpe
## 13 1.37e- 76       3.36 0.555 0.123  3.20e-72 Cortex     Slc1a2
## 14 3.52e- 71       2.51 0.739 0.314  8.24e-67 Cortex     Glul
## 15 2.57e- 64       2.71 0.913 0.702  6.03e-60 Cortex     Apoe
## 16 1.17e- 63       2.74 0.558 0.152  2.74e-59 Cortex     Tspan7
## 17 6.36e- 57       2.65 0.491 0.127  1.49e-52 Cortex     Ppap2b
## 18 1.23e- 52       3.49 0.266 0.017  2.88e-48 Cortex     Camk2a
## 19 1.49e- 43       3.35 0.545 0.236  3.50e-39 Cortex     Clu
## 20 9.17e- 39       3.09 0.494 0.206  2.15e-34 Cortex     Mfge8
```

We can notice how the two subtissues have different fill pattern in the heatmap:

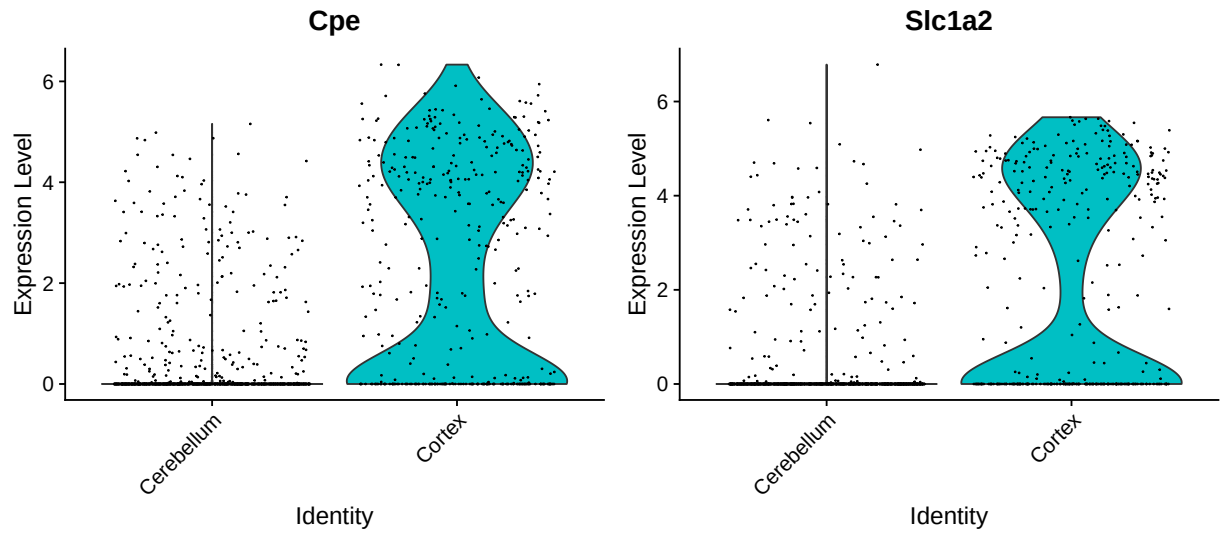
```
DoHeatmap(subset_seur, features = top10$gene, group.by = "subtissue") + scale_fill_gradient(low = "gree
```



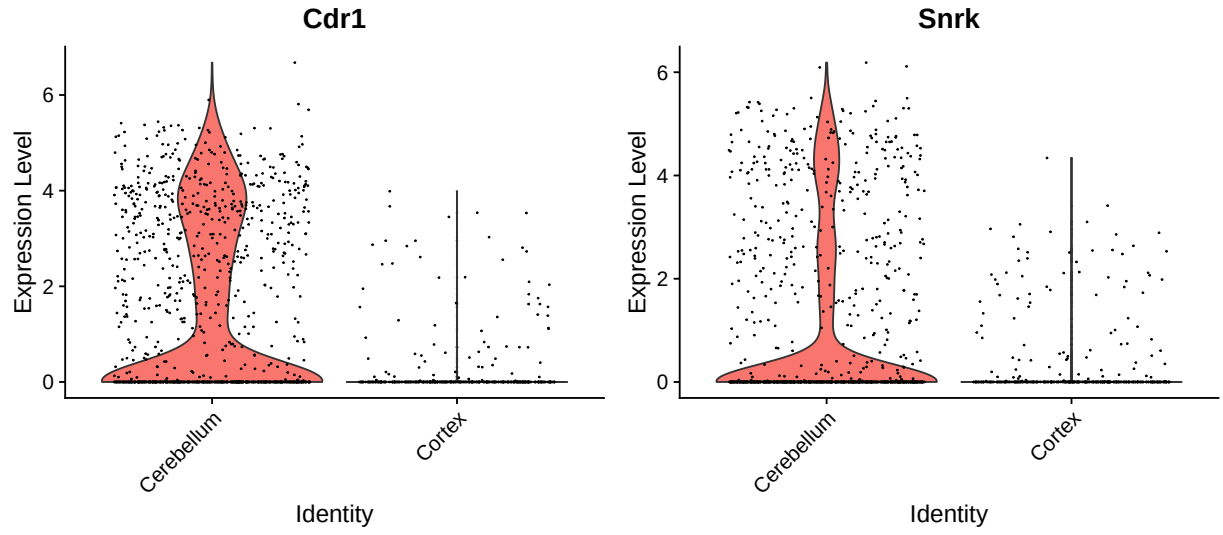
Furthermore, we can choose two convenient genes from top10 for the sake of visualization:

The genes (Cpe, Slc1a2) are upregulated in Cortex, whereas genes (Cdr1, Snrk) have higher expression in Cerebellum:

```
VlnPlot(subset_seur, features = c("Cpe","Slc1a2"))
```

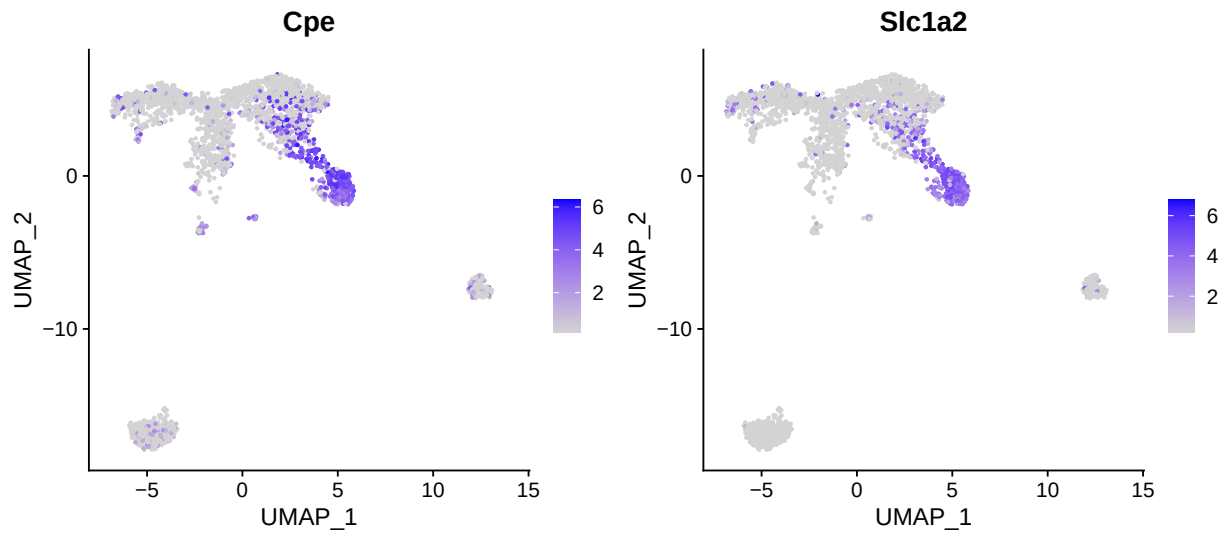



```
VlnPlot(subset_seur, features = c("Cdr1","Snrk"))
```

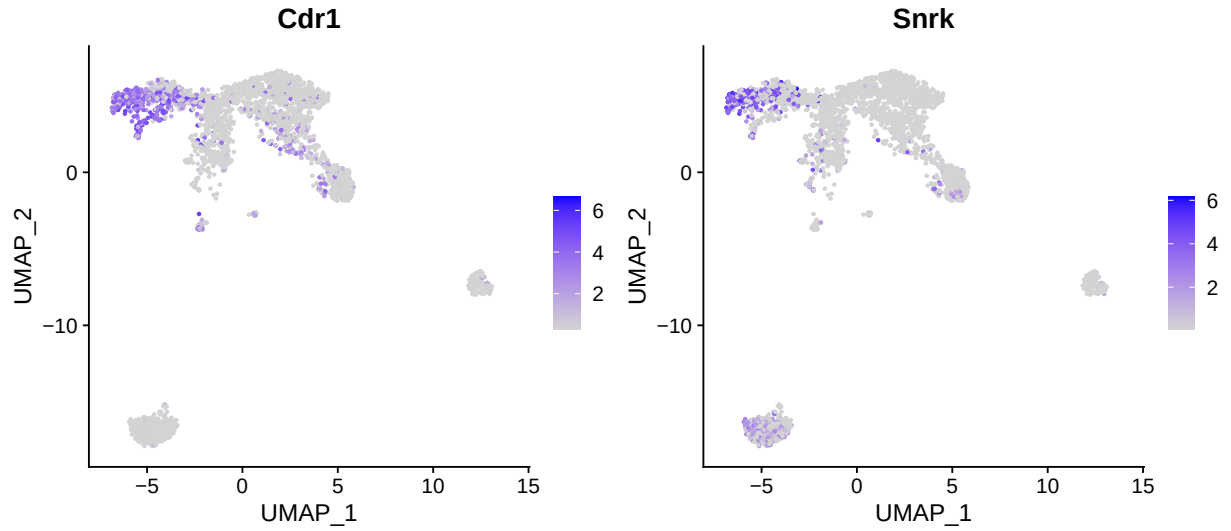


We can also return to see how unique these markers in the whole tissue:

```
FeaturePlot(seur_mod, features = c("Cpe","Slc1a2"), min.cutoff = "q10")
```




```
FeaturePlot(seur_mod, features = c("Cdr1","Snrk"), min.cutoff = "q10")
```

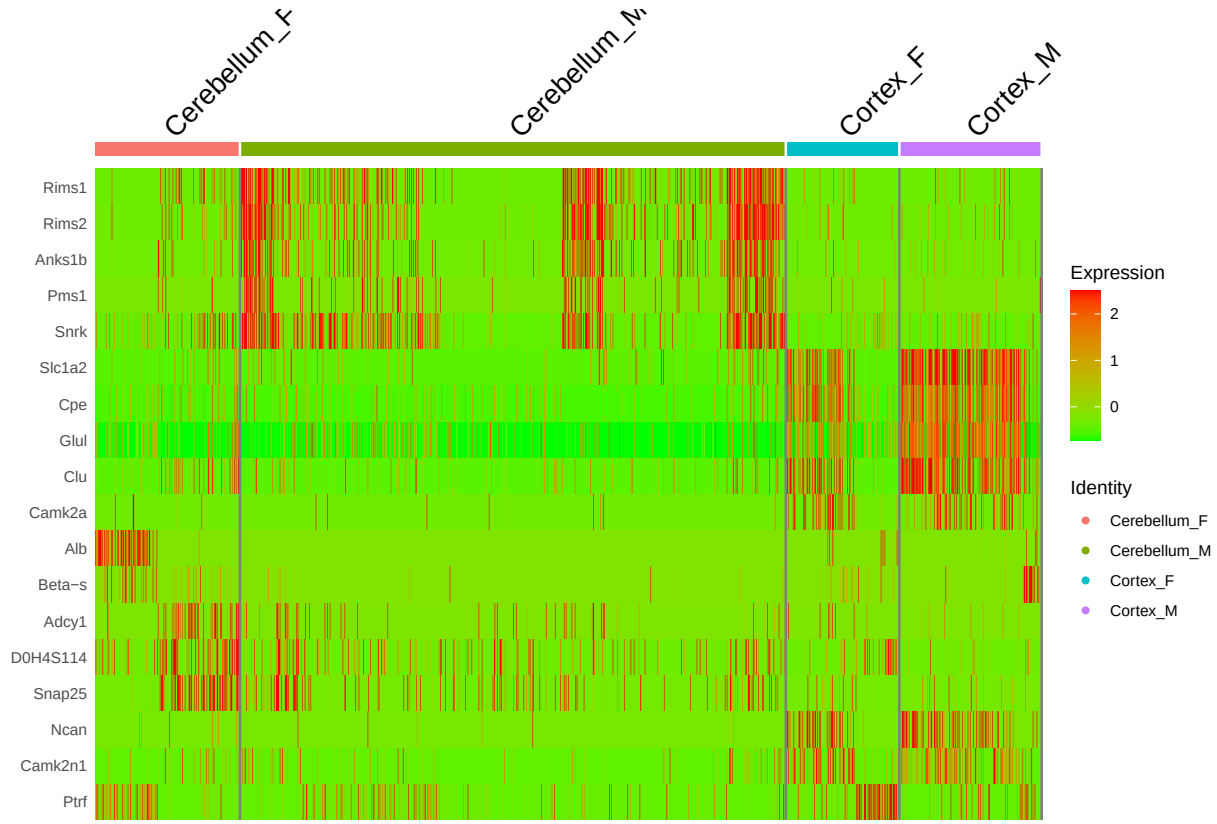


We notice some overlapping with Hippocampus cluster regardign Cpe and Slc1a2, however they seem more distinct in cortex. Cdr1 and Snrk are good markers for Cerebellum, too.

In fact, summing up previous observations sparks curiosity to further investigate those two tissues and find any connection between gene expression patterns and gender distribution. Therefore, We can create a paired column to split the clusters into four groups (of two subtissues and two genders):

```
subset_seur@meta.data <- mutate(subset_seur@meta.data, paired = "")

subset_seur@meta.data <- mutate(subset_seur@meta.data, paired =
```

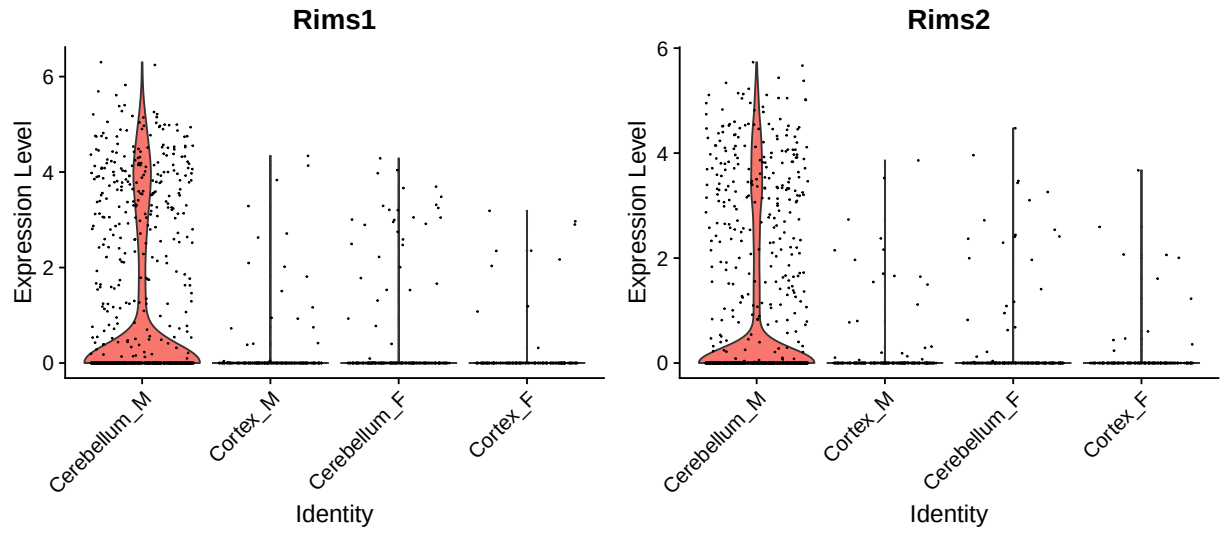
```
paste0(subset_seur@meta.data$subtissue,"_",subset_seur@meta.data$gender), x = paired)

Idents(subset_seur) <- subset_seur@meta.data$paired

cluster_all <- subset_seur %>% FindAllMarkers(min.pct = 0.1, logfc.threshold = 0.25, only.pos = T, test

top5_pair <- cluster_all %>%
    group_by(cluster) %>%
    top_n(n = 5, wt = avg_log2FC)

DoHeatmap(subset_seur, features = top5_pair$gene, group.by = "paired") + scale_fill_gradient(low = "gre
```



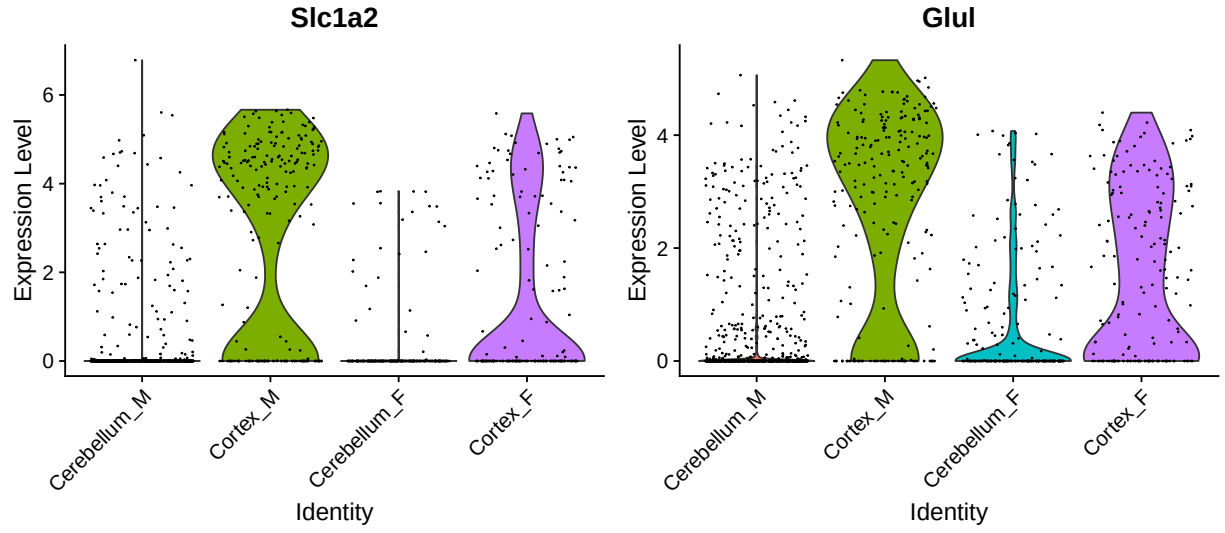
For first glance, heatmap shows subtle yet visible difference in patterns for same subtissue among genders. Top convenient genes in each cluster are:

- Cerebellum_M: Rims1, Rims2
- Cerebellum_F: D0H4S114, Snap25
- Cortex_M: Slc1a2, Glul
- Cortex_F: Ptrf, Camk2n1

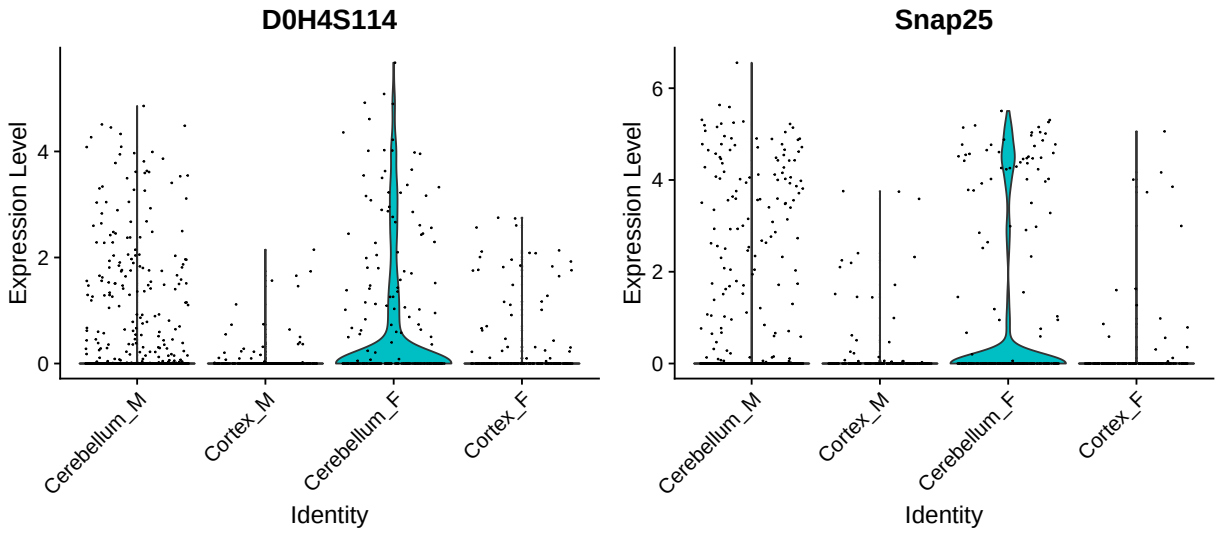
```
VlnPlot(subset_seur, features = c("Rims1", "Rims2"))
```



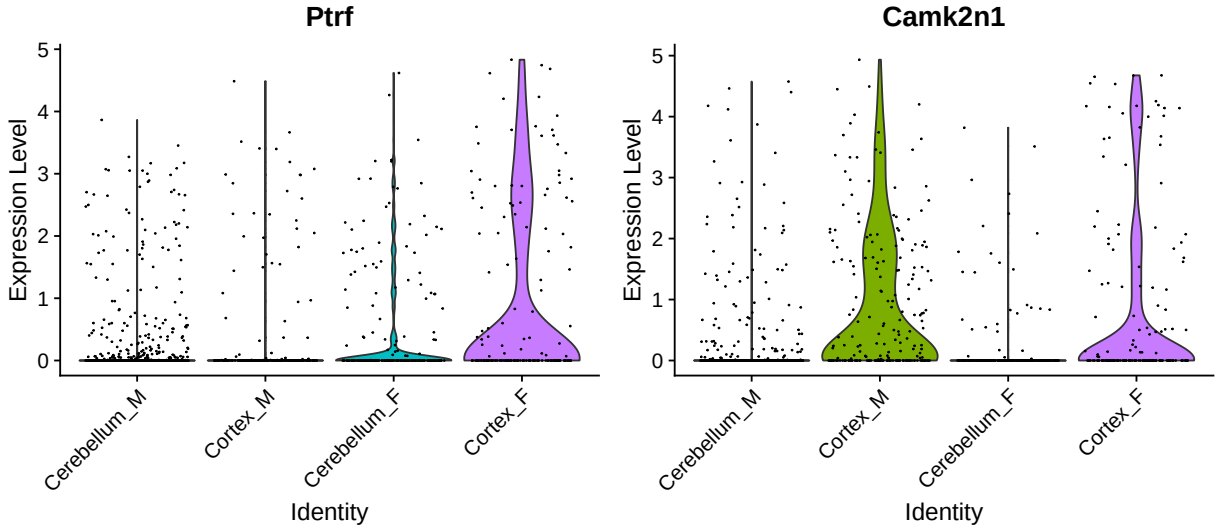

```
VlnPlot(subset_seur, features = c("Slc1a2", "Glul"))
```




```
VlnPlot(subset_seur, features = c("D0H4S114", "Snap25"))
```




```
VlnPlot(subset_seur, features = c("Ptrf", "Camk2n1"))
```



We Indeed found some true differences between males and females.

Conclusions

This analysis included standard processing and furthers simple analysis of differentially expressed genes under grouping by subtissues, gender, or both.

- The most exciting steps and procedures:

In fact, correct manipulation of data is what makes the analysis possible, therefore when I get the data ready for result generating I believe difficult part is done. Commencing additional extra steps due to emerging results is vital for satisfying curiosity and correcting assumptions. However, the most exciting step is when reaching results connected to real life. For example, After a quick internet search, I found that the genes that I deduced as markers are truly expressed with huge functionality in their corresponding subtissues.

- Primary Conclusions and biological information learnt:

  Due to this analysis, I constructed some points to view, such as the apparent difference in the proportions of cells between the genders in clusters. For example, female-derived cells make a majority in Striatum and cere, although this might be just due to preparations and sampling procedures.

  However the most important conclusion is the difference in gene expression between genders for the same subtissue. In addition, genes that differ in this matter can be further studied as backgrounds for behavior differences of the genders on the cellular level or even the systematic one.

  It can be logical to infer from dimplots that some genes are shared by subtissues, which also can be investigated as factor for similar functions or even interaction between the two subtissues. For example, we can obviously notice the large overlapping between Striatum and Hippocampus clusters.

  On the other hand, there are different isolated clusters, whose cells belong to multiple subtissues, such as clusters 2 and 5. We can assume that these cells are likely being affected by a common condition (or triggered to do a common function) a group of genes is activated (or deactivated) in a set of cells simultaneously.

  Though, technical considerations should be always taken into count, as single cell sequencing is a complex technique with many steps and gaps (like high drop-out ratio) and therefore more prone to Apophenia and fallacies.
