Introduction to single-cell transcriptomics Exercise series 2

Exercise 1 (2D-projection)

Load the table contained in the file “SDC-bisque-scores.txt”:

```
sdc <- read.csv("SDC-bisque-scores.txt", sep="\t", stringsAsFactors=FALSE,
                check.names=FALSE)
```

Then, adapt the pbmc.R code example to perform PCA and t-SNE 2D-projections of the data. (You only need the projections, no filtering, etc.) For t-SNE, you will need to reduce the default perplexity parameter since there are too few data points (use `Rtsne(t(sdc), perplexity=7)`).

The data used in this exercise represent the estimated abundance of different cell populations in salivary duct carcinomas (SDC).

Exercise 2 (dendrogram)

Compute the hierarchical clustering (functions `dist()` and then `hclust()`) of the samples. Plot the result.

Convert the hierarchical clustering object into a dendrogram with the function `as.dendrogram()`. Plot the dendrogram. What difference do you see? Heard about ultrametric trees?

Repeat these operation, but in `hclust()` set the parameter `method` to `"ward.D"`. What do you observe?

Given a hierarchical clustering, the function `cutree()` enable us to specify a number of clusters and to get as a result a named vector assigning each sample to a cluster. The height at which the dendrogram is cut is determined by `cutree()` automatically.

Exercise 3 (clustering)

Install the Bioconductor libraries ComplexHeatmap and circlize. Then, you can get a nice heatmap of your sdc data like this:

```
library(ComplexHeatmap)
library(circlize)
color.scale <- colorRamp2(breaks=c(min(sdc), 0, max(sdc)),
                        colors=c("royalblue3", "white", "orange"))
Heatmap(sdc, col=color.scale)
```

The function Heatmap allows you to compute your own hierarchical clusterings of the rows and the columns, and to pass them as parameters. Compute row hierarchical clustering with ward.D in a variable `h.cells` and try:

```
Heatmap(sdc, cluster_rows=h.cells, cluster_columns=h2.samples,
        col=color.scale, column_split=3)
```

This gives you a visual control (the heatmap) to understand why the hierarchical clustering put specific samples or rows together or apart from each other.

Get the sample cluster numbers with `cutree()` (3 clusters) and use this information to color code the samples in the PCA and t-SNE projections of Exercise 1.

Exercise 4 (application to breast cancers)

Read TCGA breast invasive carcinoma (BRCA) data from the file BRCA-most-variable-genes.txt or BRCA-most-variable-genes-small-dataset.txt if your computer is not powerful enough.

Those files contain BRCA primary tumor transcriptomes reduced to the 264 more variable genes to limit the size of data.

Start by creating a heatmap as in Exercise 3 imposing your own sample and gene dendrograms computed with ward.D.

Define 7 clusters with `cutree()` and project in 2D with PCA and t-SNE color coding based on the cluster numbers.

Can you add to the heatmap a color code showing the clusters?

The output should look like the slides 7 & 8 of the lecture!