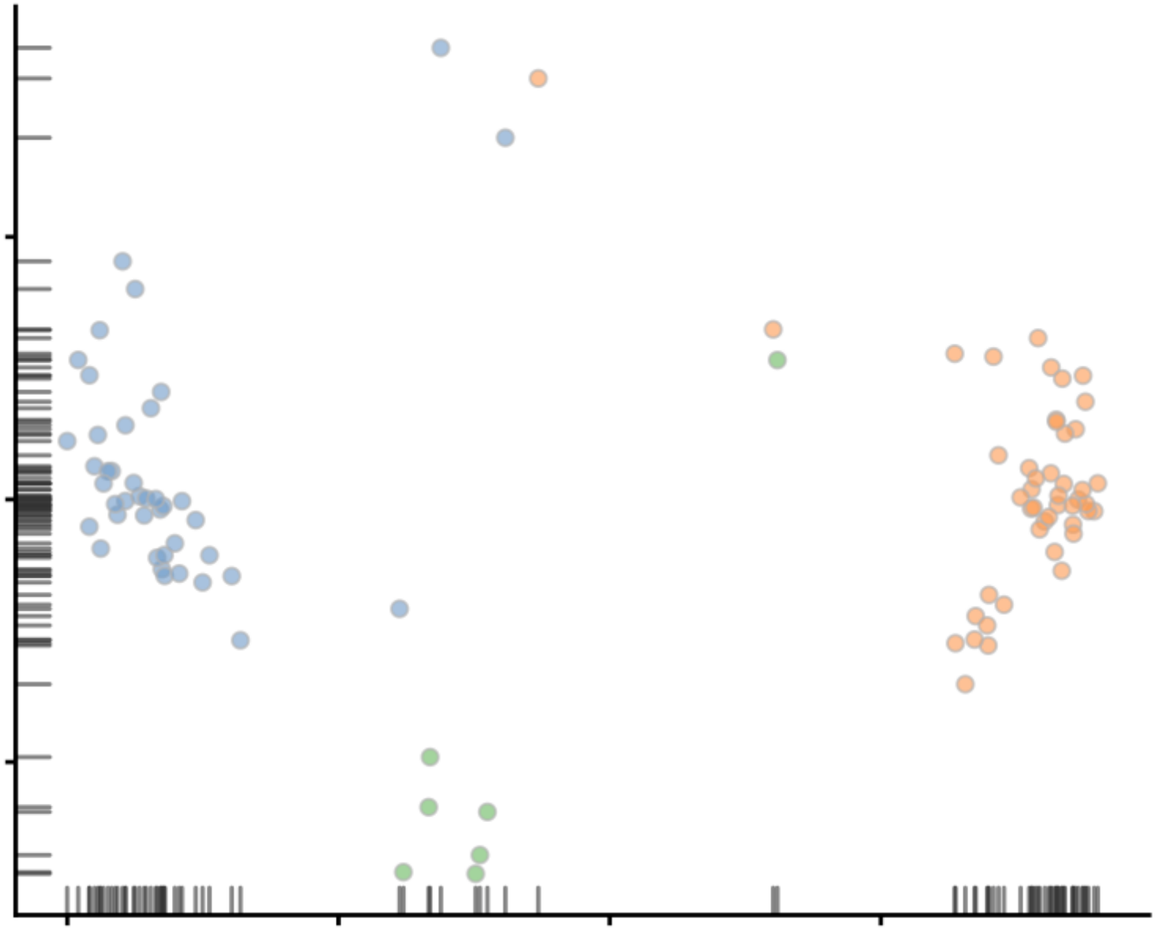
Clustering methods for scRNA-Seq

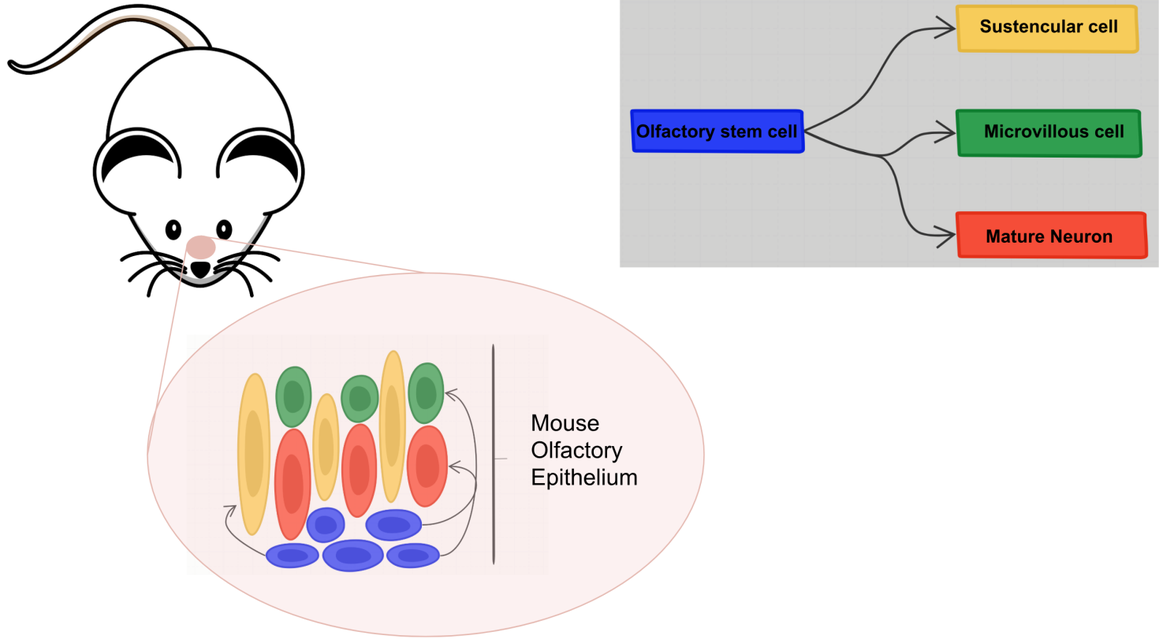
SINGLE-CELL RNA-SEQ WITH BIOCONDUCTOR IN R

Fanny Perraudeau
Senior Data Scientist, Whole Biome

datacamp

Mouse epithelium dataset





[1] Cell Stem Cell, Fletcher et al, Deconstructing Olfactory Stem Cell Trajectories at Single [2] Cell Resolution

Clustering methods

- hierarchical clustering
- k-means clustering

Challenges

- What is the number of clusters?
- What is a cell type or the expected granularity?
- Scalability: in scRNA-Seq experiments the number of cells could be millions, tools developed for single-cell data don't scale well.

Create Seurat object

```
library(Seurat)
library(SingleCellExperiment)

seuset <- CreateSeuratObject(
    raw.data = assay(sce),
    normalization.method = "LogNormalize",
    scale.factor = 10000,
    meta.data = as.data.frame(colData(sce))
)
seuset <- ScaleData(object = seuset)
seuset
```

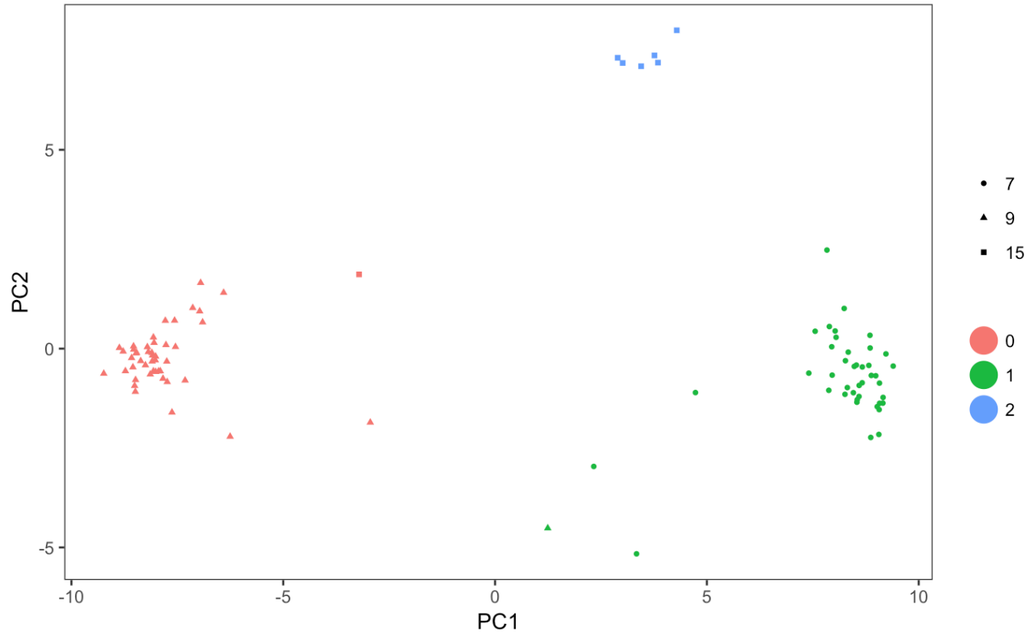
```
An object of class seurat in project SeuratProject
 100 genes across 94 samples.
```

Perform Clustering

```
seuset <- FindClusters(
    object = seuset,
    reduction.type = "pca",
    dims.use = 1:10,
    resolution = 1.8,
    print.output = FALSE
)
```

Plot Clusters

```
PCAPlot(
    object = seuset,
    group.by = "ident",
    pt.shape = "publishedClusters")
```



Let's practice!

SINGLE-CELL RNA-SEQ WITH BIOCONDUCTOR IN R

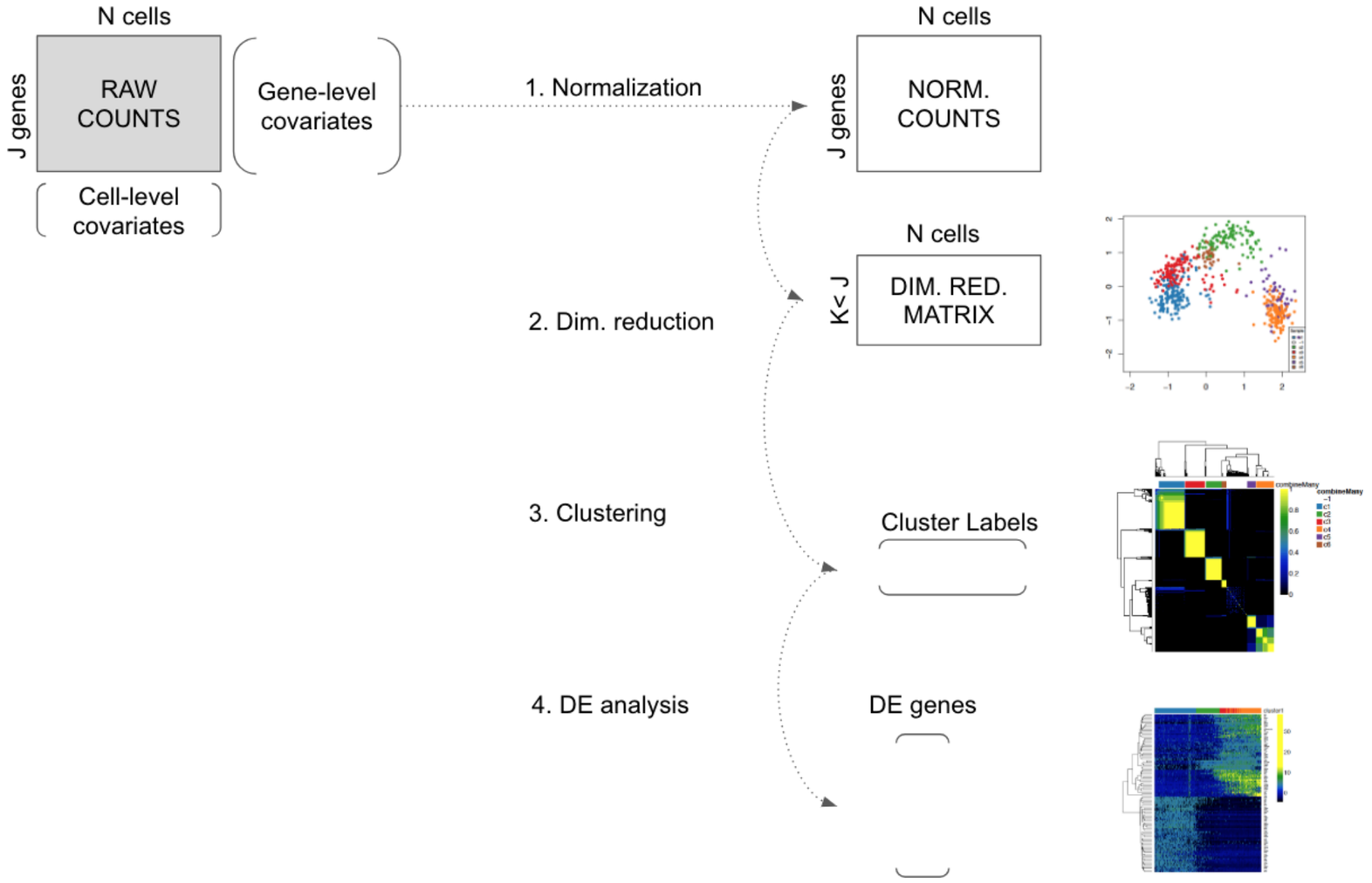
Differential expression analysis

SINGLE-CELL RNA-SEQ WITH BIOCONDUCTOR IN R

Fanny Perraudeau
Senior Data Scientist, Whole Biome

datacamp

Typical workflow



DE methods

Bulk RNA-Seq methods:

- edgeR
- DESeq2

Single-cell methods:

- Single Cell Differential Expression (SCDE)
- Model-based Analysis of Single-cell Transcriptomics (MAST)

Fit zero-inflated regression using MAST

```
library(MAST)
zlm <- zlm(~ celltype + cngeneson, sce)
summary <- summary(zlm, doLRT = "celltype9")
summary
```

```
Fitted zlm with top 2 genes per contrast:
( log fold change Z-score )
 primerid celltype9 celltype15 cngeneson
 Cyp2a5    -14.5      -7.6*       1.5
 Gap43      21.9*      2.3        3.0
 Ncam1      15.4       1.3        4.7*
 Stmn2      16.7       2.2        5.0*
 Stmn3      25.8*      1.1        1.9
 Ugt2a1    -14.7      -8.0*       2.8
```

Fit zero-inflated regression using MAST

```
# get summary table
fit <- summary$datatable

# pvalues and logFC
fit <- merge(fit[contrast=='celltype9' & component=='H',
                 .(primerid, `Pr(>Chisq)`)],
             fit[contrast=='celltype9' & component=='logFC',
                 .(primerid, coef)],
             by='primerid')
```

Adjusted p-values

```
# adjusted p-values
fit[, padjusted:=p.adjust(`Pr(>Chisq)`, 'fdr')]
```

```
# result table
res = data.frame(gene = fit$primerid,
                 pvalue = fit[,'Pr(>Chisq)'],
                 padjusted = fit$padj,
                 logFC = fit$coef)
head(res)
```

```
           gene   Pr..Chisq.    padjusted      logFC
1 1810011O10Rik 7.256038e-15 1.422753e-14 -10.767717
2 5730409K12Rik 1.446961e-24 1.607735e-23  12.544552
3        Actr1b 1.753458e-07 2.112600e-07   7.043411
4           Ado 2.940357e-08 3.630071e-08   7.980018
5           Ak1 3.579921e-17 9.944224e-17  11.003118
6         Anxa1 4.423790e-10 6.059987e-10  -7.718933
```

Let's practice!

SINGLE-CELL RNA-SEQ WITH BIOCONDUCTOR IN R

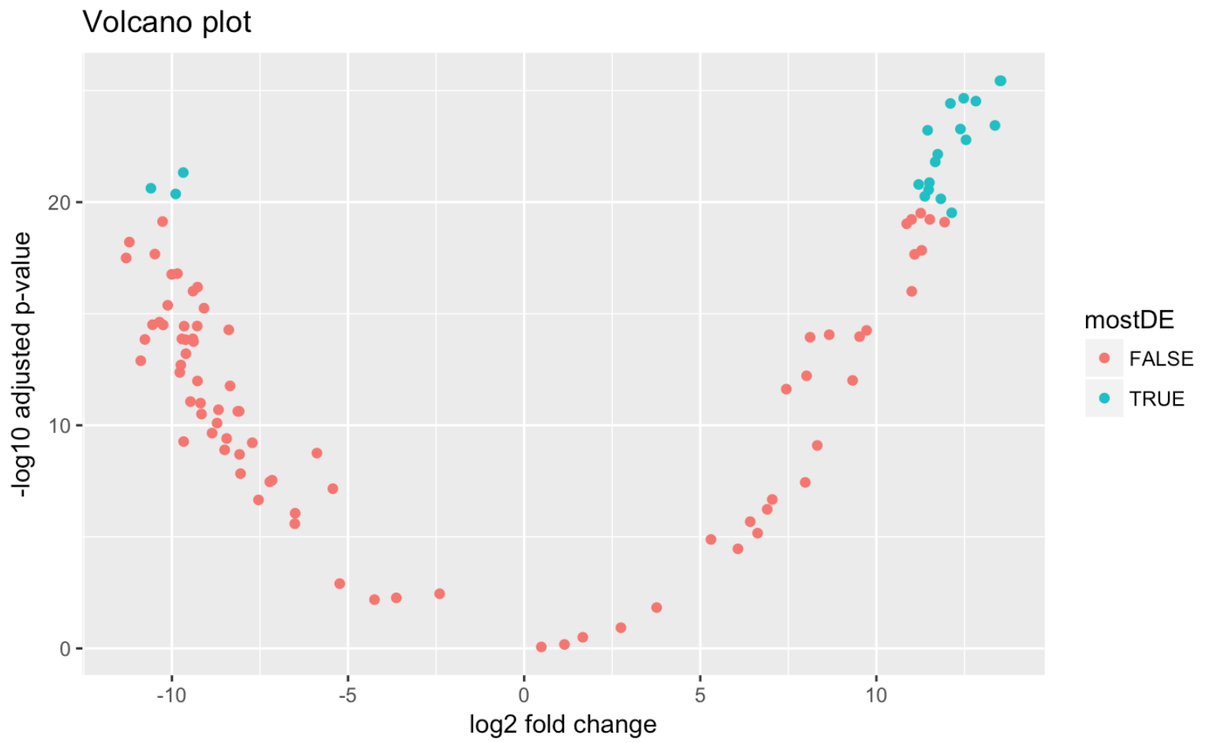
Visualization of DE genes

SINGLE-CELL RNA-SEQ WITH BIOCONDUCTOR IN R

Fanny Perraudeau
Senior Data Scientist, Whole Biome

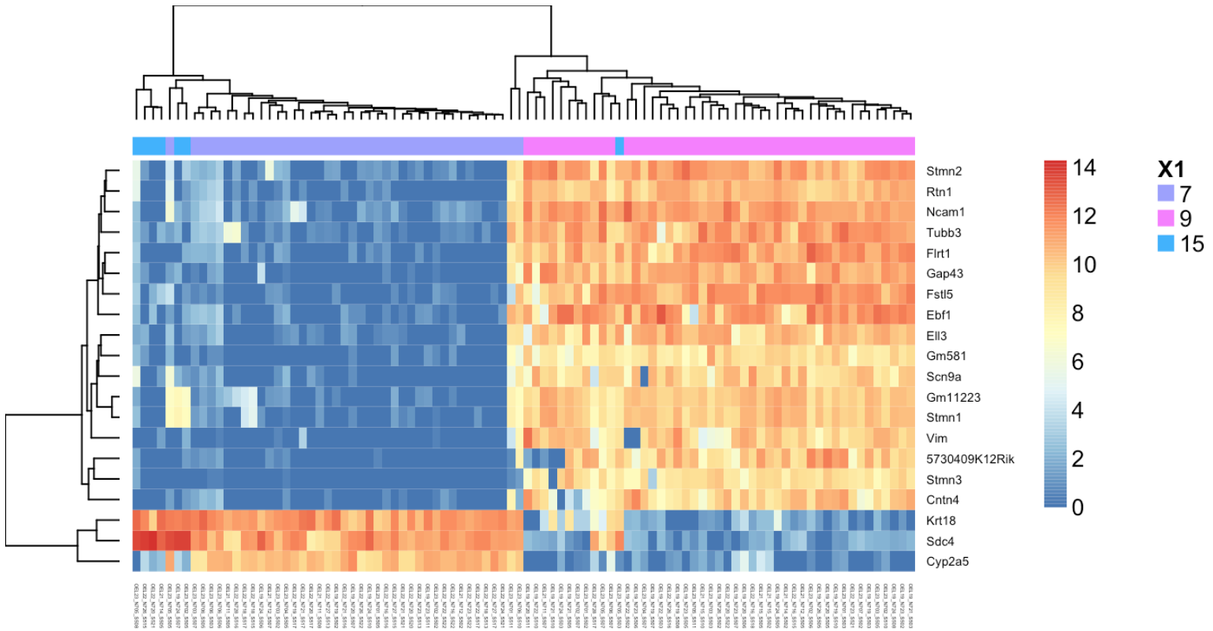
Volcano plot

```
ggplot(res, aes(x=logFC, y=-log10(padjusted), color=mostDE)) + geom_point() +
    ggtitle("Volcano") + xlab("log2 FC") + ylab("-log10 adjusted p-value")
```

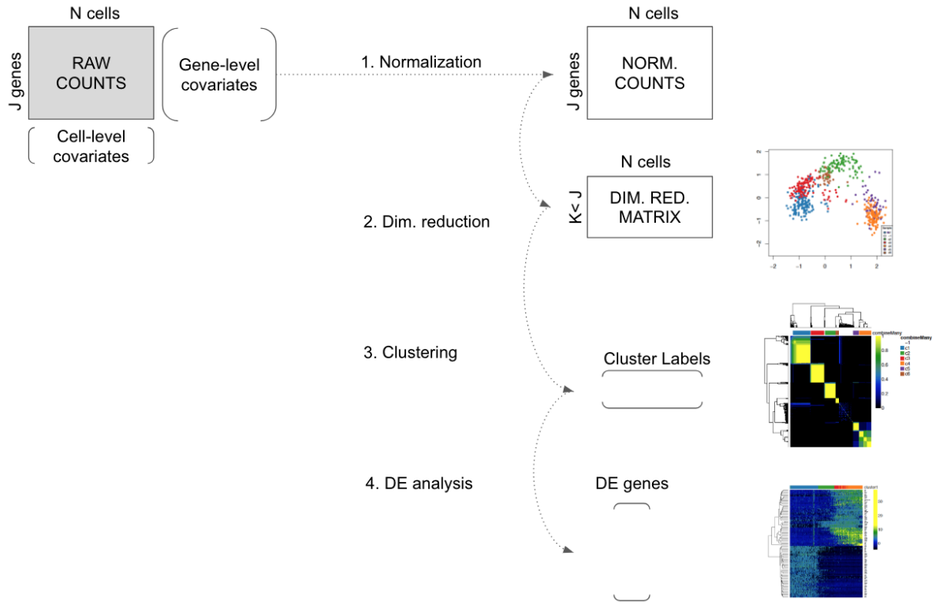


Heatmap

```
library(NMF)
norm <- assay(sce[mostDE, ], "logcounts")
norm <- as.matrix(norm)
aheatmap(norm, annCol = colData(sce)$publishedClusters)
```



Typical workflow



- https://hemberg-lab.github.io/scRNA.seq.course/index.html
- A step-by-step workflow for low-level analysis of single-cell RNA-seq data (Lun et al).
- Bioconductor workflow for single-cell RNA sequencing (Perraudeau et al).

Let's practice!

SINGLE-CELL RNA-SEQ WITH BIOCONDUCTOR IN R