Single-cell RNA-seq data analysis

using R
and
command line tools

21.9.2018

Bishwa Ghimire

Seurat Alignment

- Data
  - 10X
  - Seqwell
- Format
  - Expression matrix

Data Import

- Import gene expression matrix.

```
tenx.data <- read.table(gzfile(file.path(tenx.data.path,"dge.csv.gz")),
                        sep="\t",
                        header=T,
                        row.names = 1)
```

- Create Seurat object
  - min.cells : Include genes with detected expression in at least 3 cells.
  - Min.genes : Include cells where at least 200 genes are detected.

```
tenx <- CreateSeuratObject(raw.data = tenx.data,
                           names.delim="-",
                           names.field=2,
                           project="10X",
                           min.cells = 3,
                           min.genes = 200)
```

Setting identity class

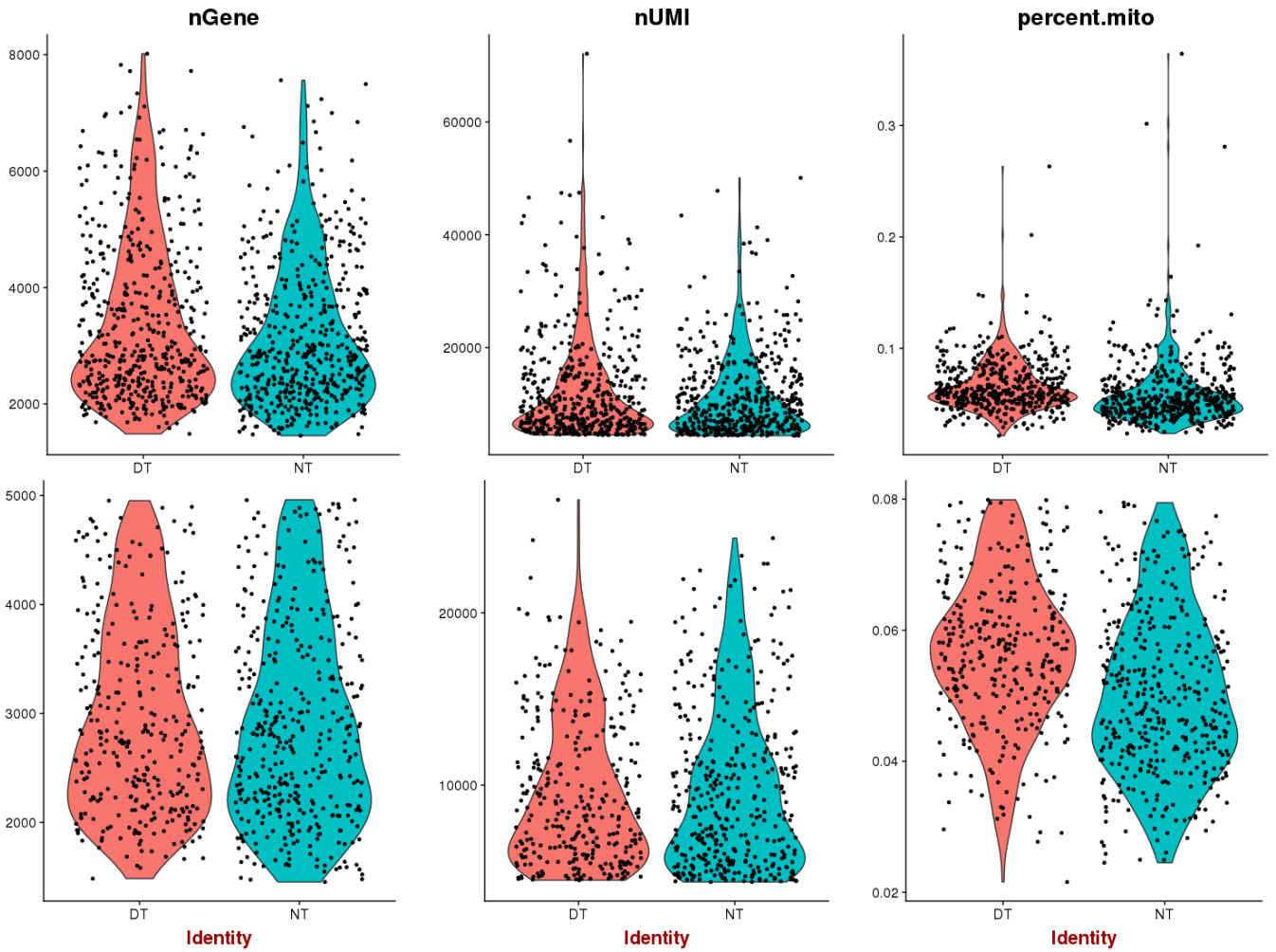
- We want to identify cells by NT and DT in out analysis.
- The cell identity class can changed at any time during the analysis according to the need.

```
## Check orig.ident
Head(tenx@meta.data)


## take only NT and DT from metadata rownames
tenx.ident <- gsub(".*\\.","", rownames(tenx@meta.data))
tenx.ident[tenx.ident==1] <- "NT"
tenx.ident[tenx.ident==2] <- "DT"

tenx@meta.data$orig.ident <- factor(tenx.ident, levels = unique(tenx.ident))
tenx <- SetAllIdent(object = tenx, id = "orig.ident")
```

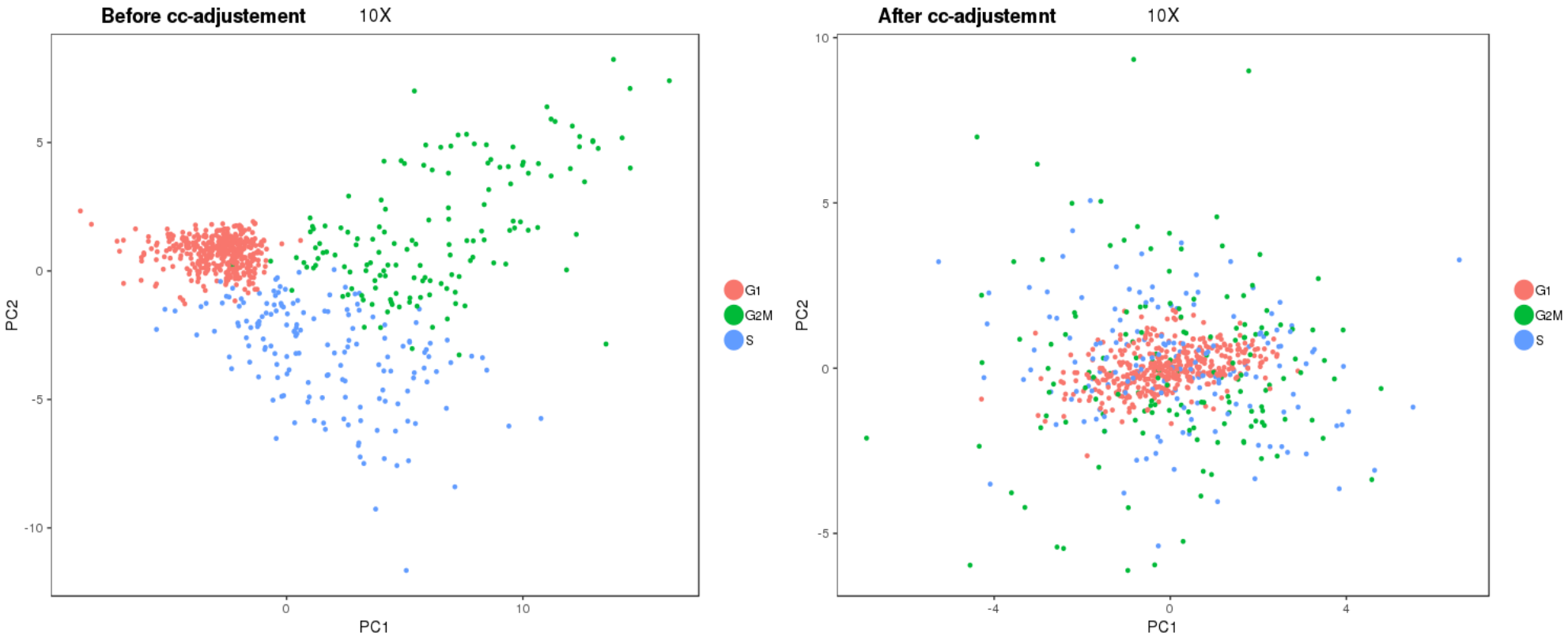
Quality control

- Quality matrices
  - Percentage of mitocondrial genes/cell
  - Number of genes per cell

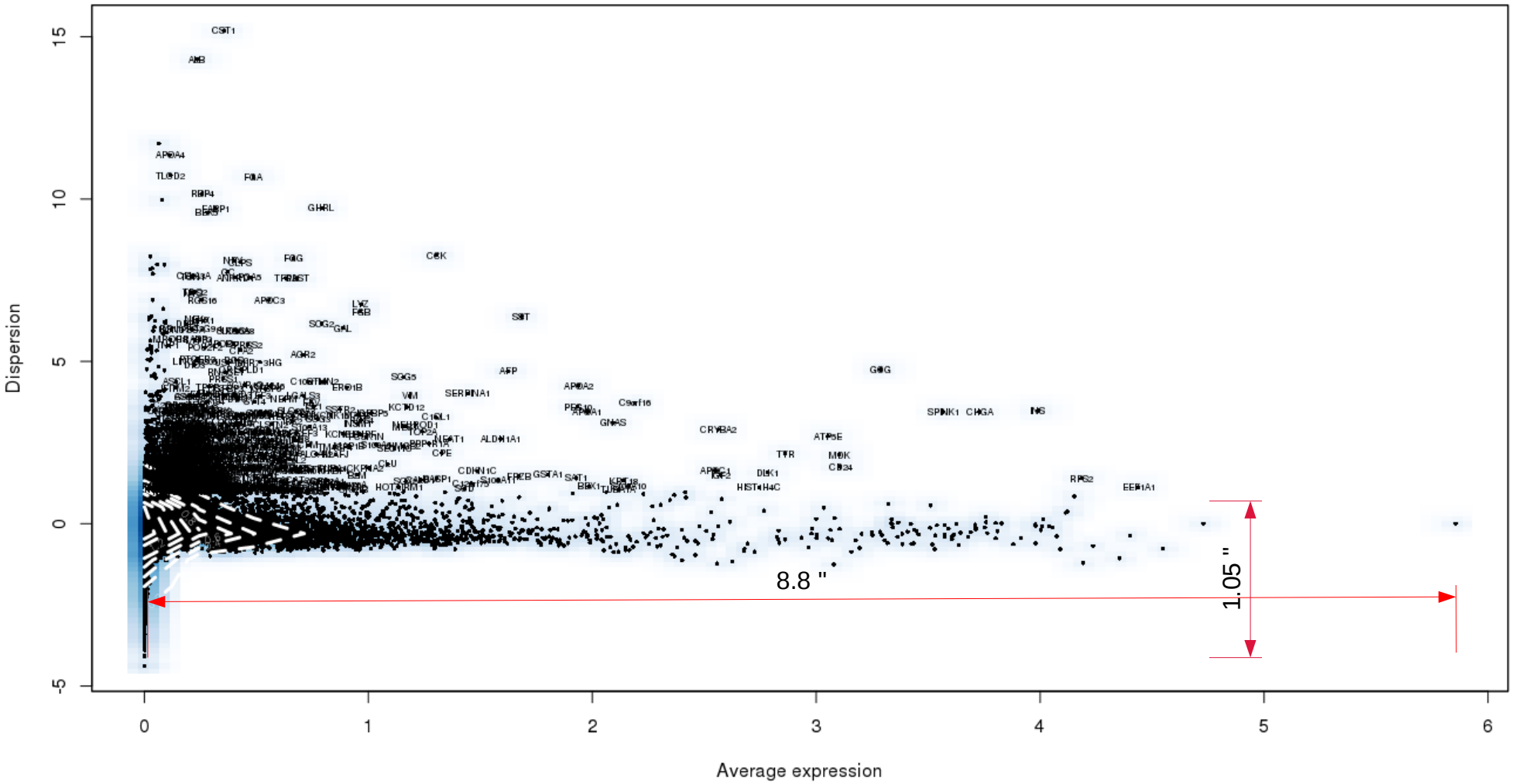
Filtering


nGene
nUMI
percent.mito
8000
6000
4000
2000
60000
40000
20000
0.3
0.2
0.1
DT
NT
5000
4000
3000
2000
20000
10000
0.08
0.06
0.04
0.02
Identity

Cell cycle effect



Finding variable genes



Canonical correlation analysis

- Identifies shared correlation between two data sets
- Lets consider 2 sets of data x and y.
  - X – vectors of p variables
  - Y – vectors of q variables
- Finds projects direction u and v in subspace of x and y respectively in such a way that u and v has maximum correlation.