SinCellTE 2022

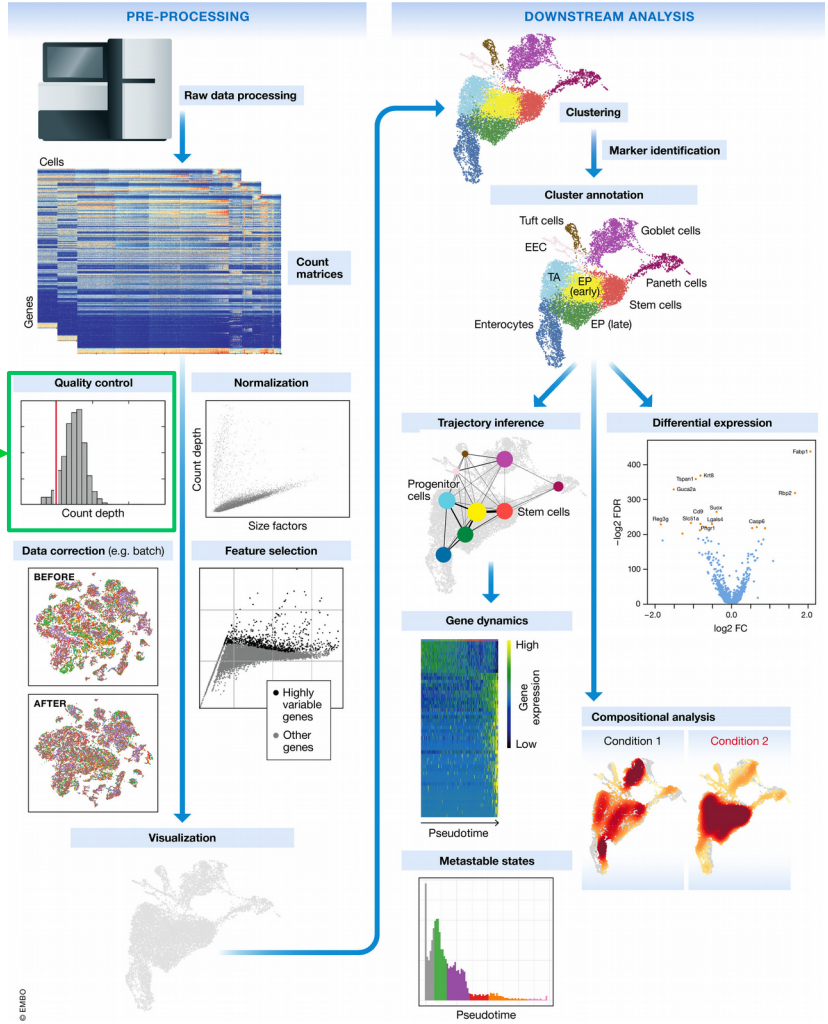
Practice : Primary Analysis

2022-01-10

Marine AGLAVE, Bioinformatics Core Facility, Gustave Roussy
Rémi MONTAGNE, Bioinformatics Core Facility, Curie Institute
Agnès PAQUET, Bioinformatics Core Facility, Curie Institute

Main steps of single cell data processing

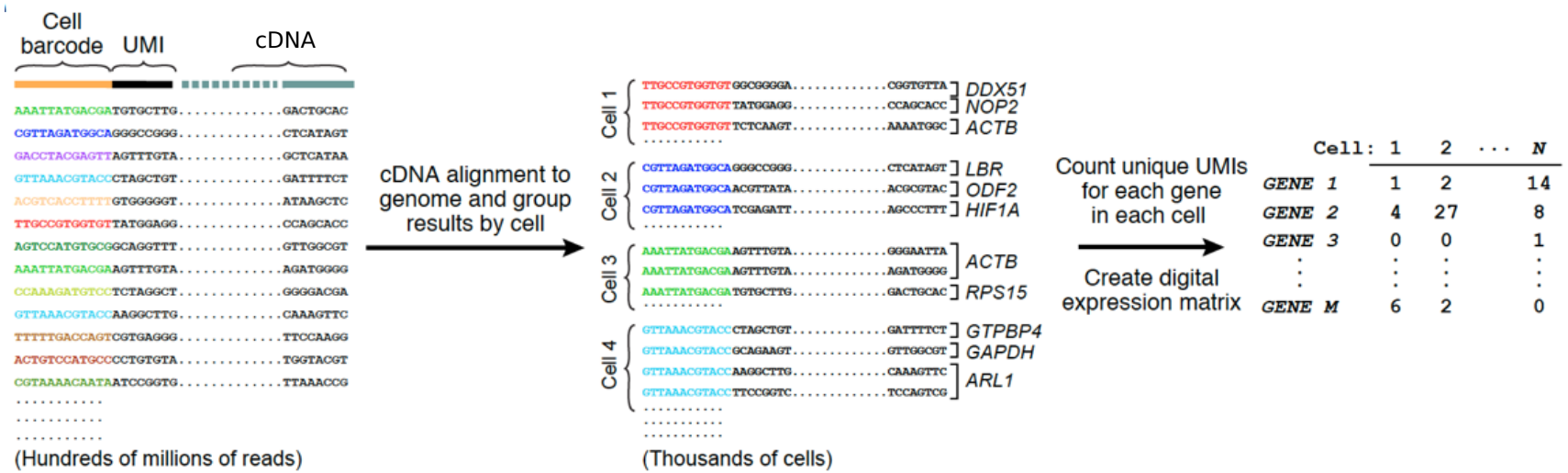
This practical session



Malte D Luecken & Fabian J Theis
Molecular Systems Biology (2019)

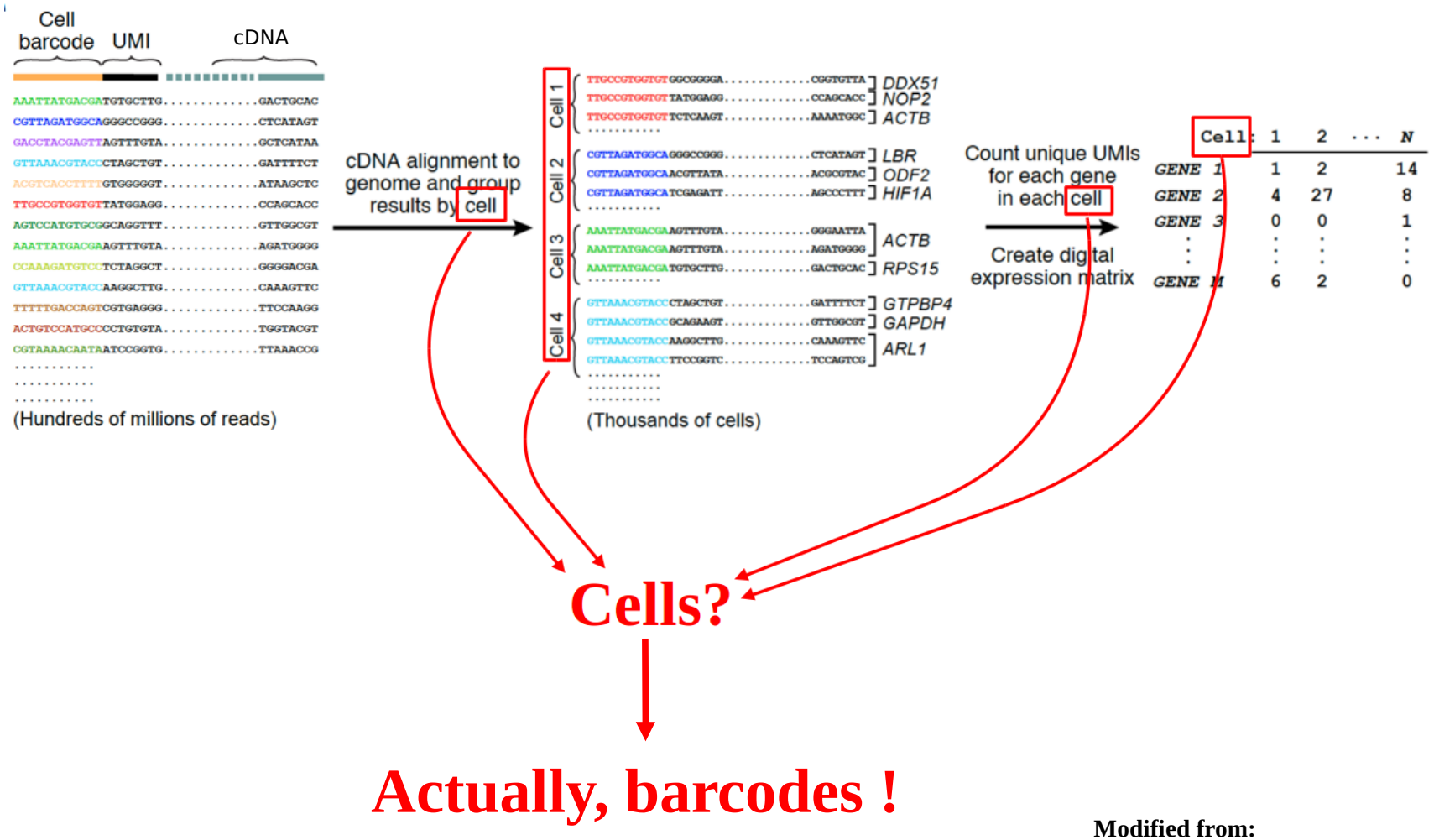
Theoretical Part

Alignment



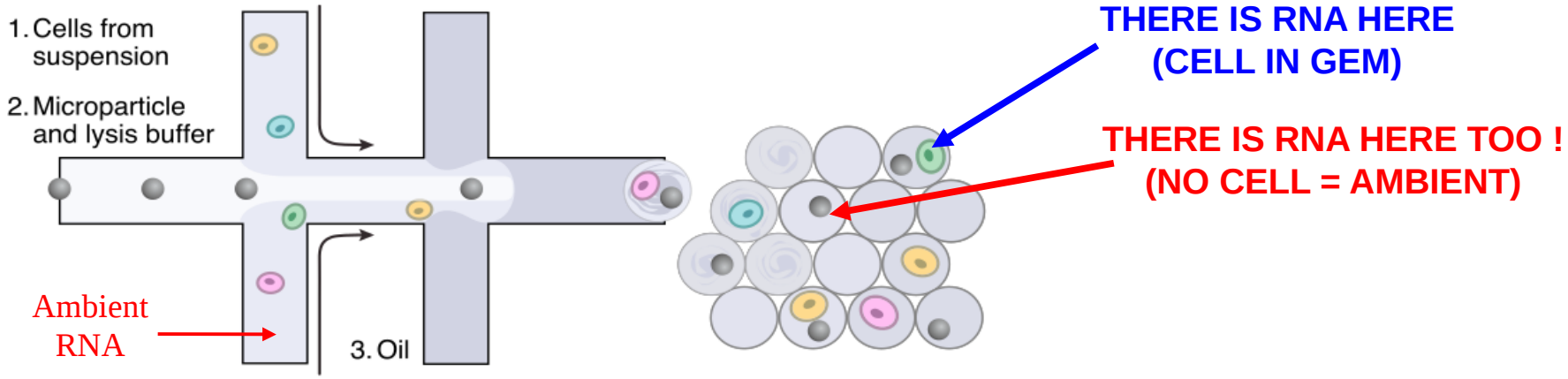
Modified from:
http://mccarrolllab.org/dropseq/

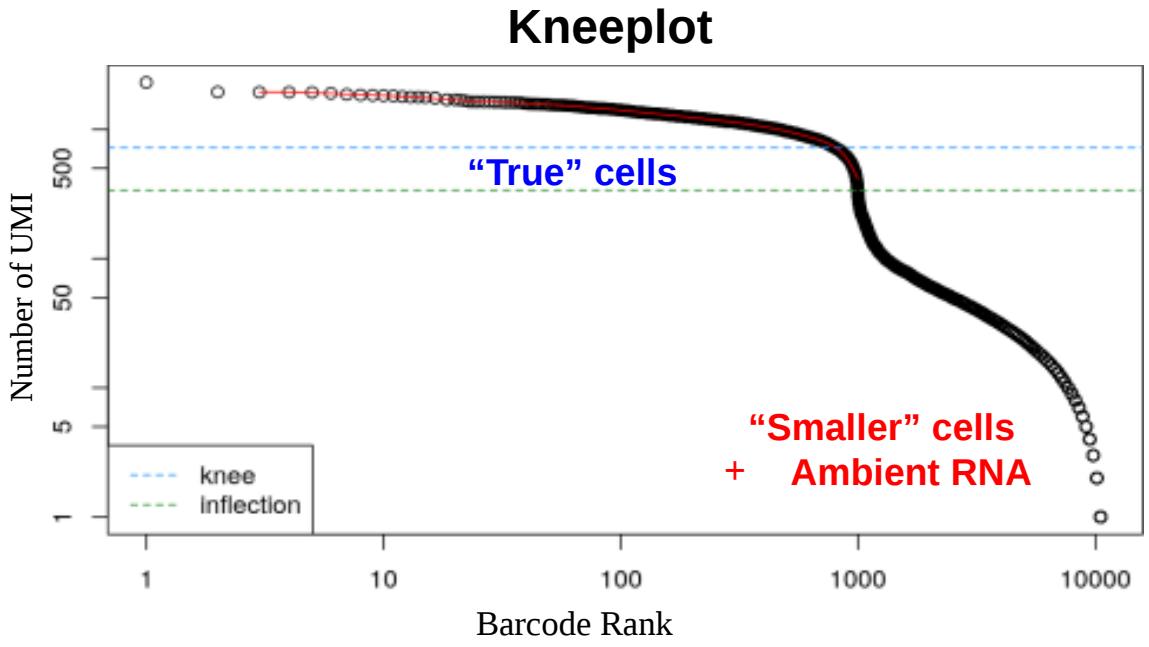
Alignment



Modified from:
http://mccarrolllab.org/dropseq/

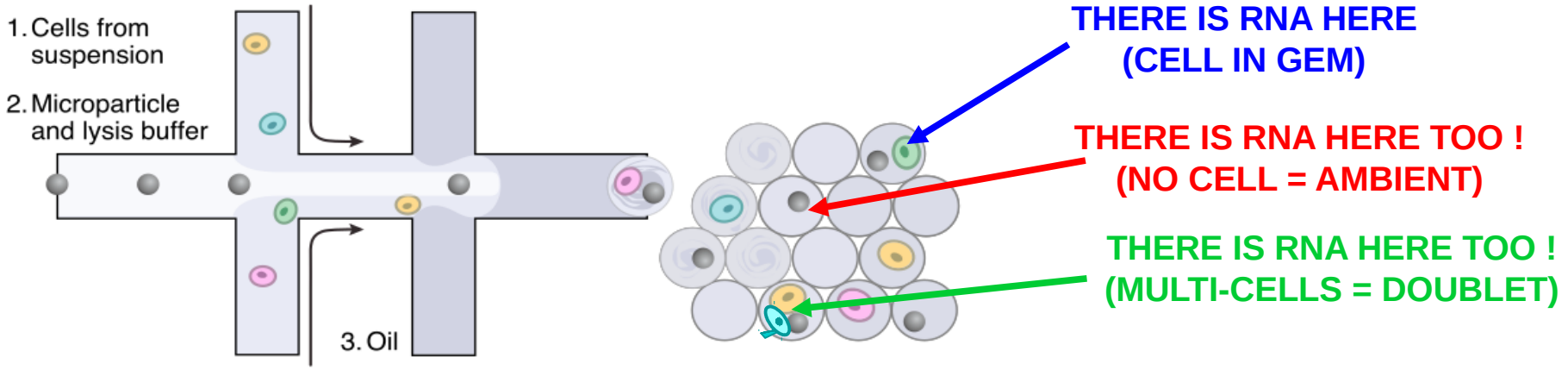
Filtering droplets: empty droplets

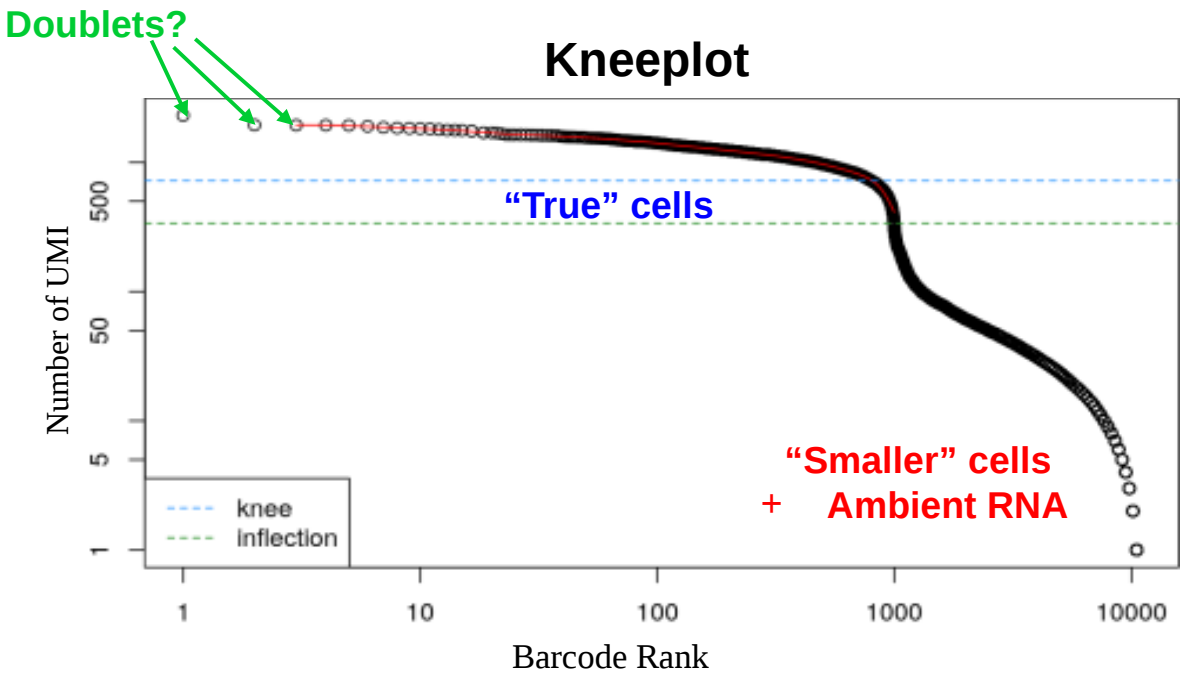




- Viability = 70%
  → 30% dead cells
  → ambient RNA
  → noise in empty droplet
  + noise in droplet with cell
- Package R: EmptyDrops
  Raw thresholds

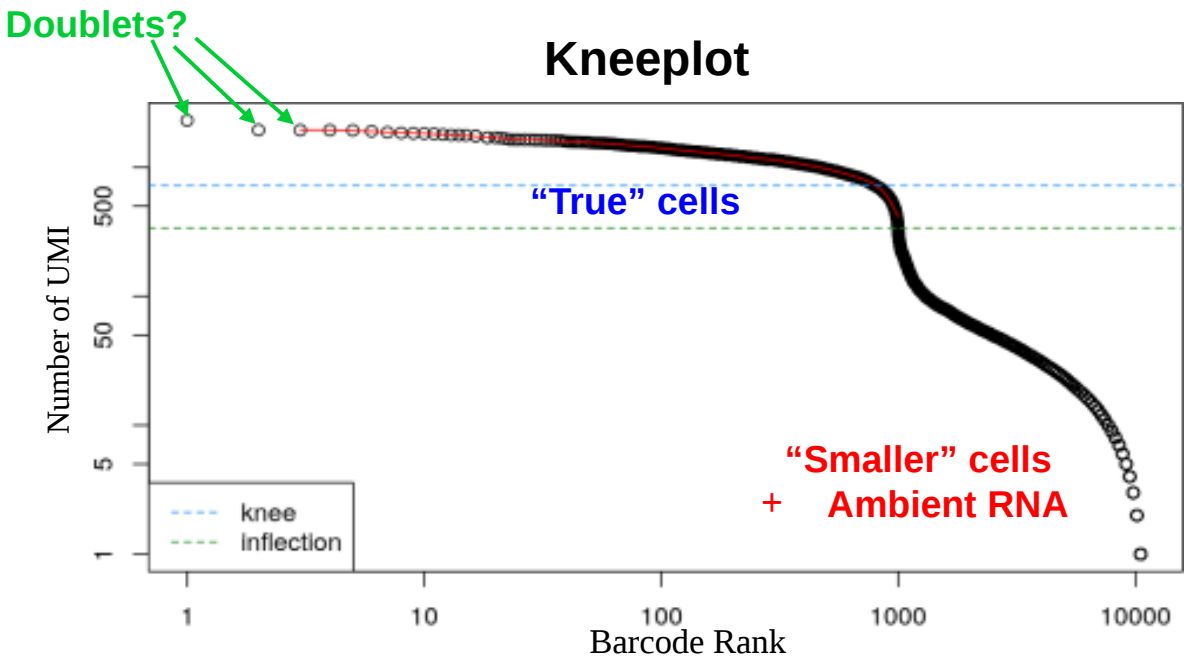
Filtering droplets: doublets or multiplets

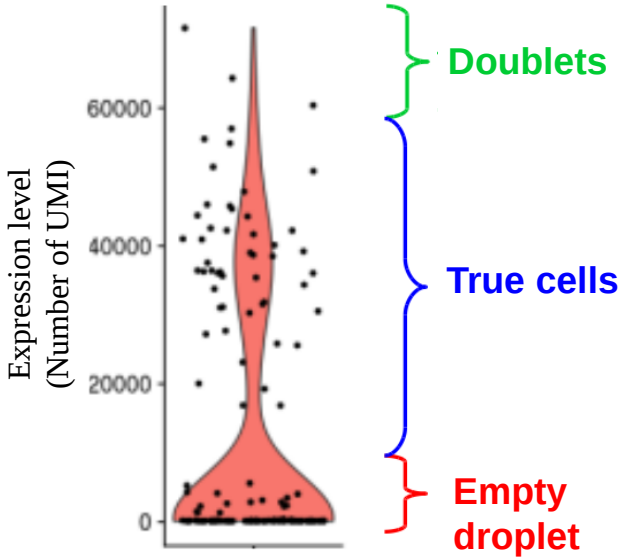


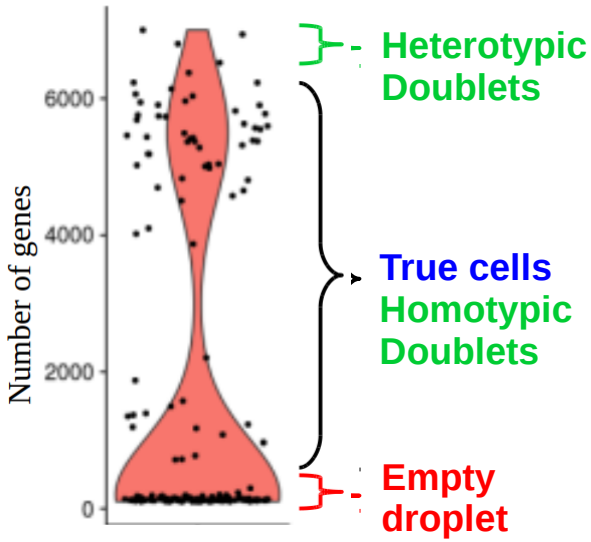


- Doublet types:
  - Homotypic: same cell type in droplet
  - Heterotypic: different cell types in droplet
- Doublet rate:
  - 1% for 1000 cells
  - 5% for 10 000 cells

Filtering droplets: doublets or multiplets







What is an expressed gene ?

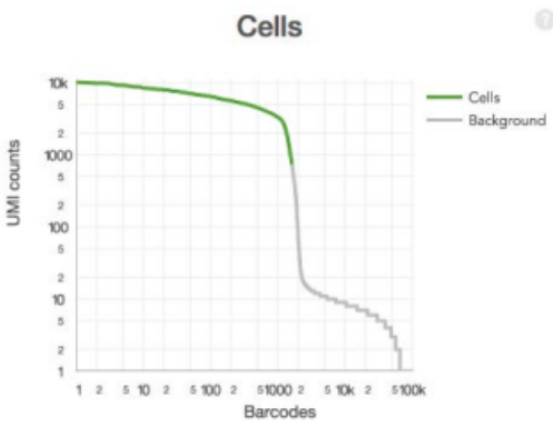
- Minimum amount of UMIs ?
  (Be careful ! Droplet => low depth!)
- Minimum amount of expressing cells ?
- Minimum gene expression level ?

How to select a threshold ?

- Kept genes : expressed in >= 5 cells
- Kept cells : counts > 0 for at least 200 genes
- Prior knowledge : expected amount of cells in the sample

Kneeplot: Diagnosis

Typical Sample Profile

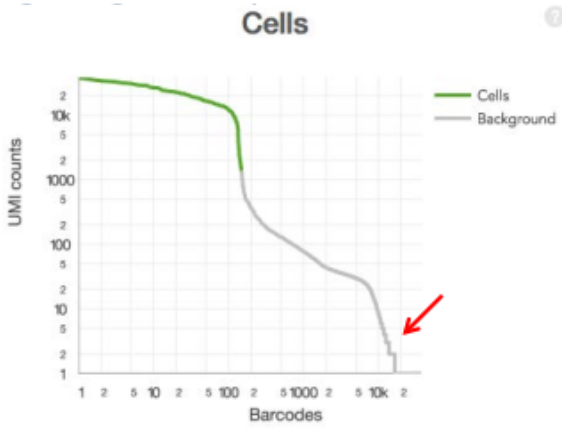


Defined cliff and knee

Metric	Value
Barcodes	> 90,000
Cell Barcodes	> 1,000
UMIs	> 10,000

Good!

Low Barcode Counts



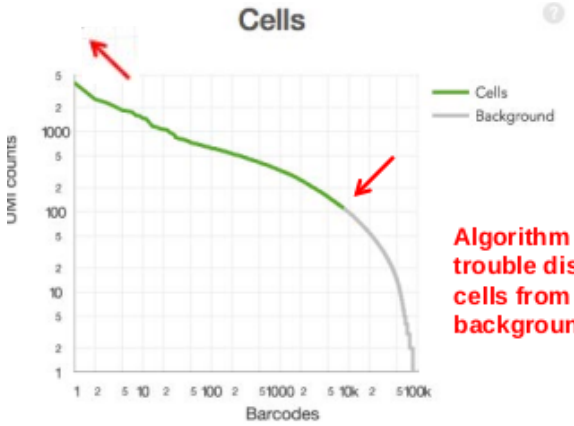
Low number of barcodes detected

Metric	Value
Barcodes	~ 15,000
Cell Barcodes	> 100
UMIs	> 10,000

Bad!

Depth is too low : although no sequencing of ambient RNA, there is almost no sequencing of genes neither !

Loss of Single Cell Behaviour



Algorithm has trouble discerning cells from the background

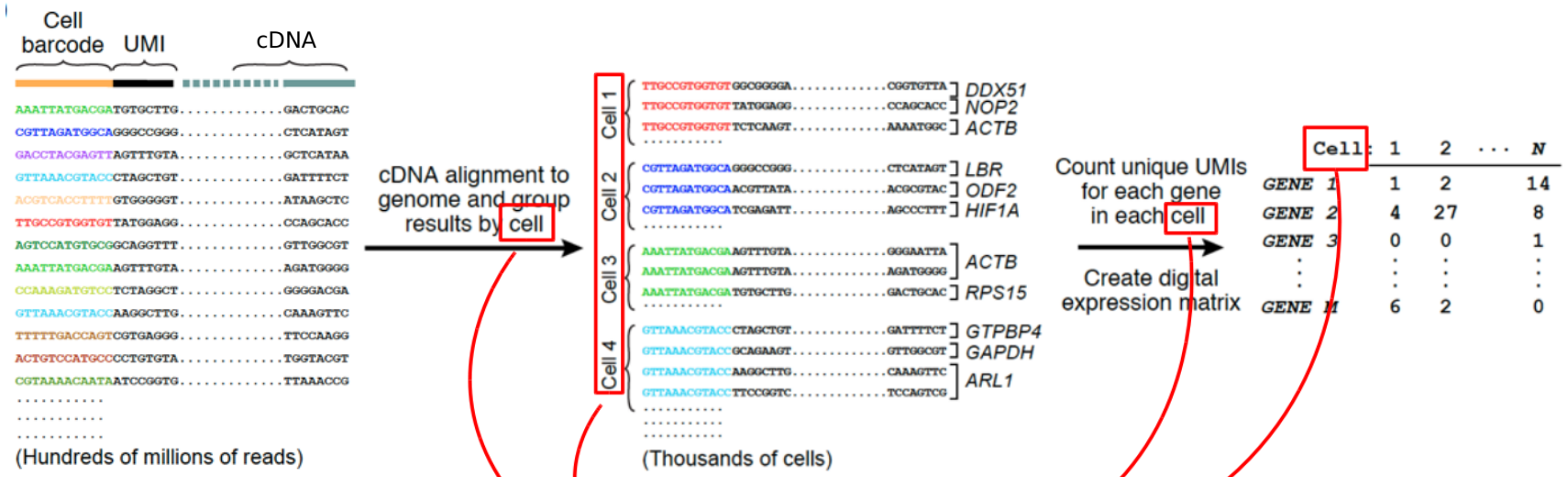
Lack of defined cliff and knee

Metric	Value
Barcodes	> 90,000
Cell Barcodes	~ 10,000
UMIs	> 10,000

Bad!

Problem in cell lysis : RNA not released into the droplet reaction mix. Almost only noise, low signal (Corresponds to the bottom knee on the first.)

Filtering droplets

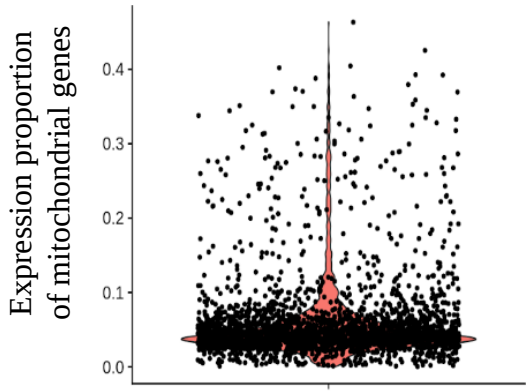


Cells?

Now, sure !

Filtering cells

Mitochondrial genes expression

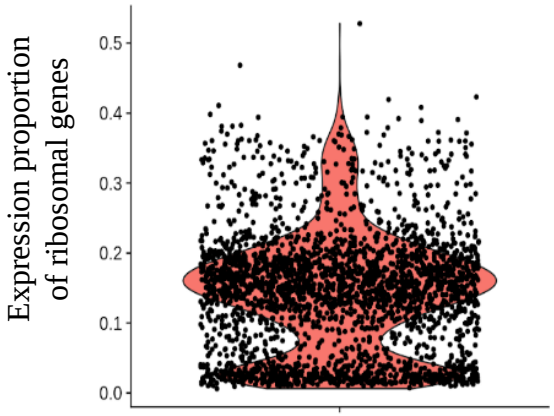


High percentage of mitochondrial genes expression may be due to apoptotic cells :

Kept cells < 5-20% mtRNAs

Gene names beginning with “MT-”.

Ribosomal protein genes expression



Linked to: cellular activity? cell cycle? Not very clear!
Community debate, hard to say if it does matter or not.

Kept cells < 25% rbRNAs ?
10% rbRNAs < Kept cells ?

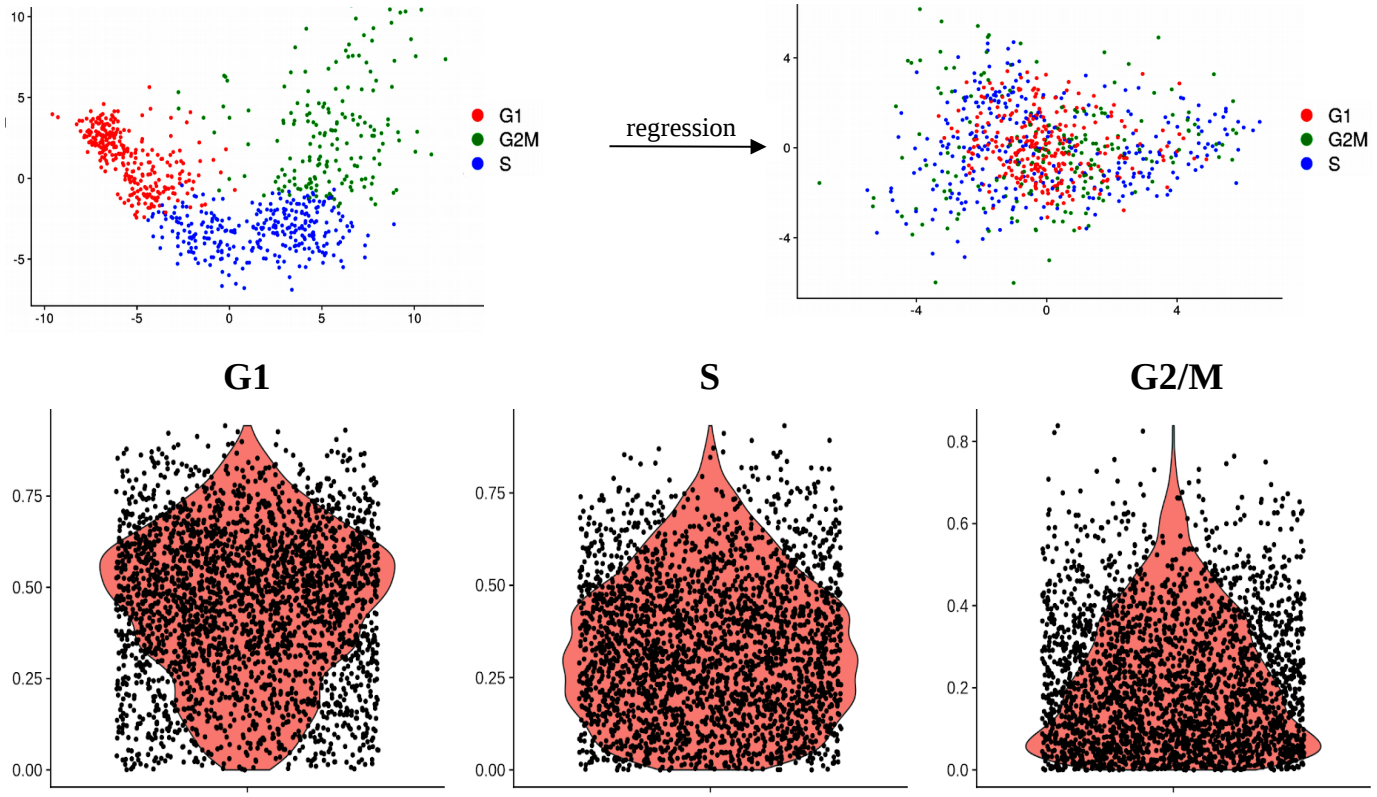
Genes names beginning with “RP-”.

These thresholds are subjective !
Needs to be adjusted according to the biological knowledge of the sample.

R packages: Scater, scRNAseq

Filtering cells

Cell cycle



Cell cycle state of a cell can affect its global gene expression (sometimes strongly), so for a defined cell type, we may observe expression variations linked to the cell cycle that mask other biological signals.

1) Estimation of a cell-cycle score of each cell, then label a cycle stage.
2) Normalization (regression of the score or stage).

R packages: Scran (function cyclone), seurat

Practical Part

Prepare your work environment

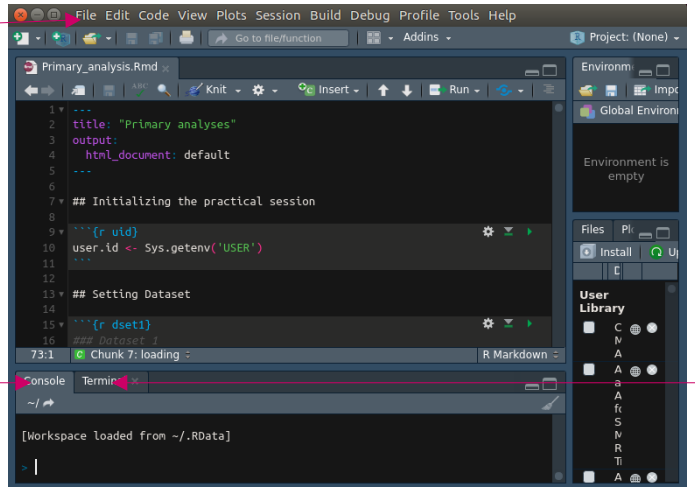
In a bash terminal :
1) Create your working folder
> mkdir -p /shared/projects/sincellte_2022/${USER}/Primary_analyses/

2) Copy scripts
> cp -r /shared/projects/sincellte_2022/Courses/Primary_analyses/scripts /shared/projects/sincellte_2022/${USER}/Primary_analyses/scripts

3) Link datasets
> ln -s /shared/projects/sincellte_2022/Courses/Primary_analyses/input /shared/projects/sincellte_2022/${USER}/Primary_analyses/input

In the RstudioServer console :
1) Go to your working directory
> user.id <- Sys.getenv('USER')
> setwd(paste0("/shared/projects/sincellte_2022/", user.id, "/Primary_analyses"))

2) Open script



Console

Terminal

Script

1) Setting Dataset, Parameters, Seed and Loading R packages
2) Loading data
3) Removing empty droplets :
    a) Automatically: emptydrops()
    b) Manually: raw threshold
4) Computing basic metrics :
    a) Percentage of mito + Percentage of ribo
    b) Identification of background genes
    c) Cell cycle prediction
    d) Identification of doublets
5) Filtering
6) Checking the effect of filtering
7) Save Results

Dataset : description

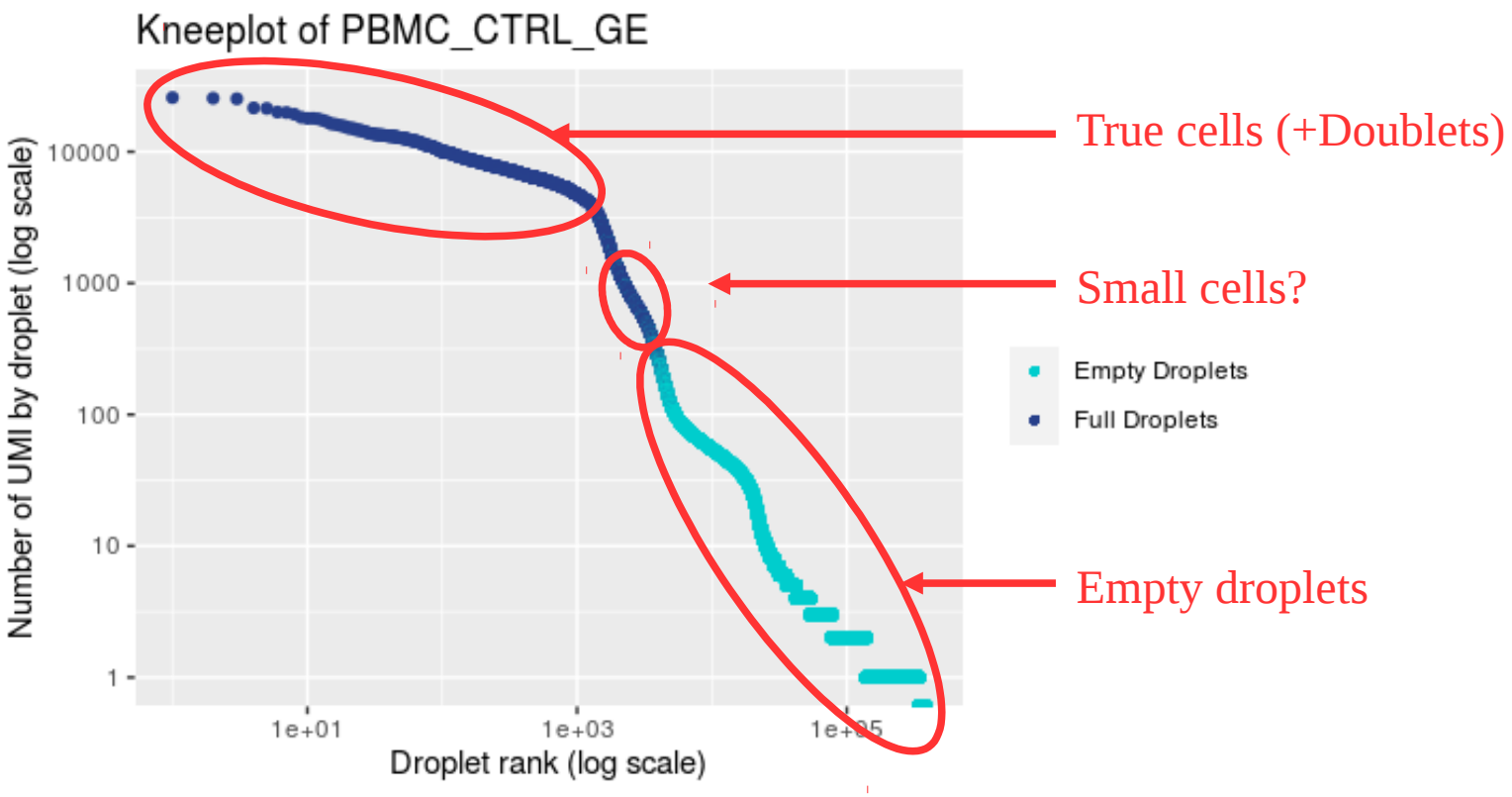
Goal:
Identify the different cell types.

Data information *:*
Organism: Human
Type of tissue collected: Peripheral blood mononuclear cells.
Origin: patient.
Cells treatment: no treatment, healthy cells.
Technology : 10X Genomics Chromium

Expected Result:
We expect about **ten** cell types.

Input type:
Raw, unfiltered counts table from CellRanger.

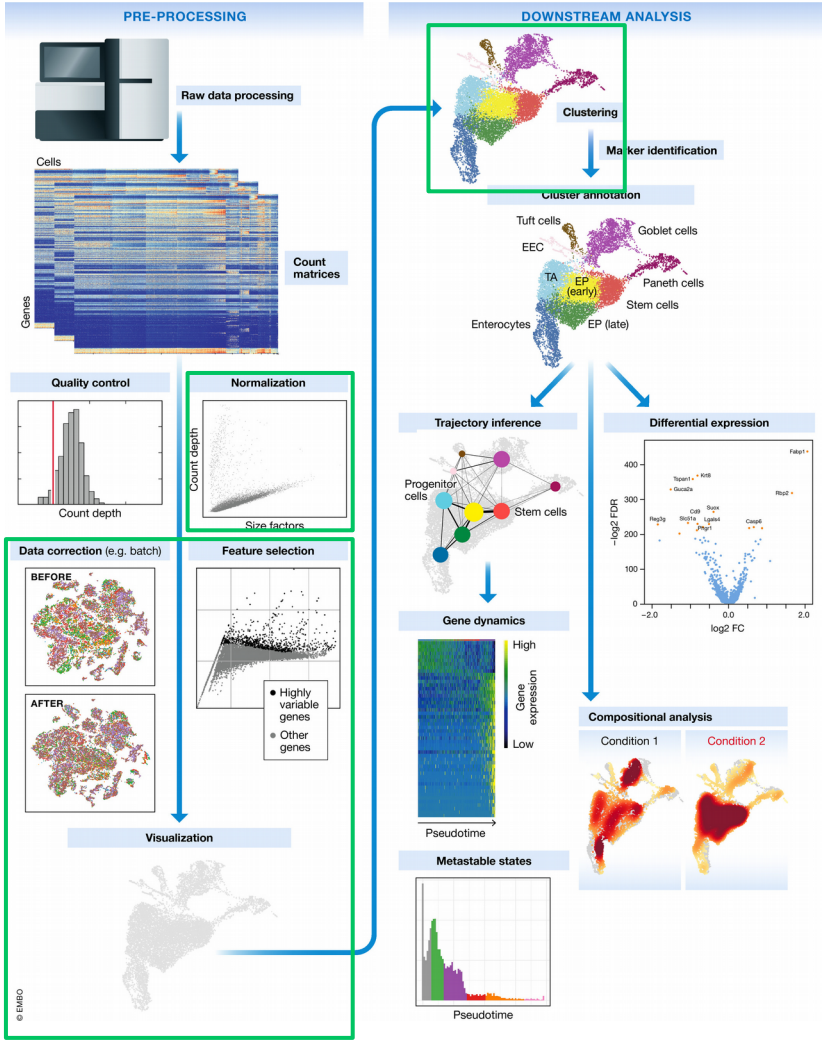
Dataset : results



Here we will make 2 different versions of the results (see html files):
- **with** this population of small cells
- **without** this population of small cells

Main steps of single cell data processing

Tomorrow morning



Malte D Luecken & Fabian J Theis
Molecular Systems Biology (2019)