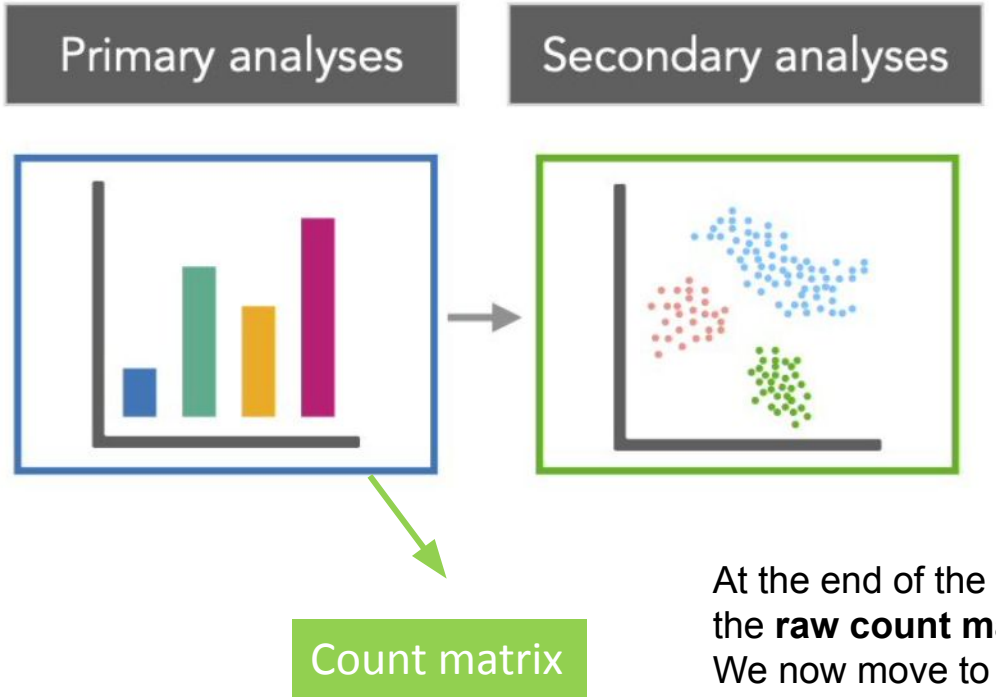
Preprocessing

Prepping the count matrix

Bastien Job, INSERM / Gustave Roussy, Villejuif
Lilia Younsi, Institut Cochin
Sébastien MELLA, Institut Pasteur
Rémi Montagne, Institut Curie, Paris
Nathalie Lehmann, Institut Pasteur

École de bioinformatique EB3I 2025

Organization of this session



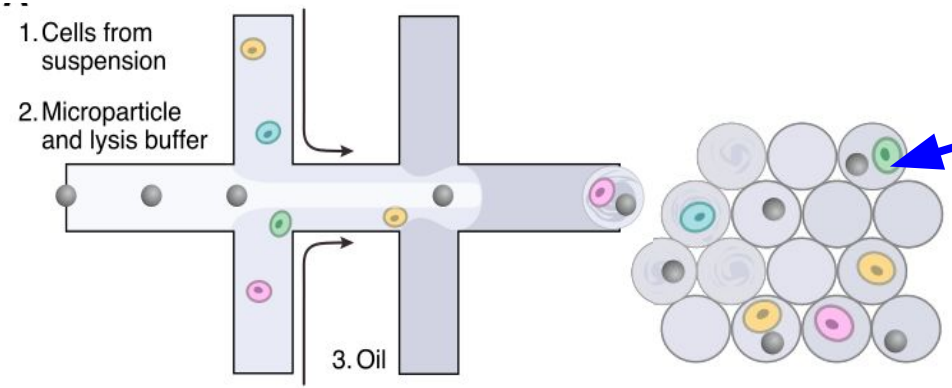
At the end of the primary analysis, the **raw count matrix** is produced. We now move to the downstream analysis

Organization of this session

- Prepping the raw counts matrix
  - Filtering out low quality / “empty” droplets
  - Assessing, removing ambient RNA
  - Generating technical and signature metrics (counts, %mito, %ribo, cell cycle phase, …)
    - Filtering out extreme “bad” cells on these
    - Lowering bias *(later)*
  - Identifying, filtering out cell doublets

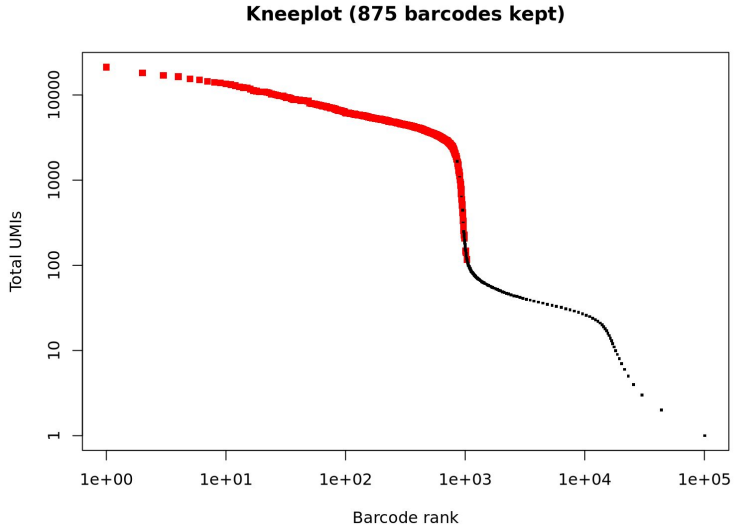
- Your input : a [**feature**] -by- [**droplet barcode**] count matrix

Empty droplets filtering



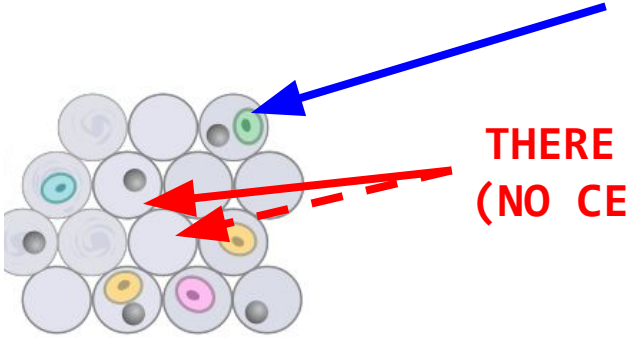
THERE IS RNA HERE
(CELL IN GEM RNA)

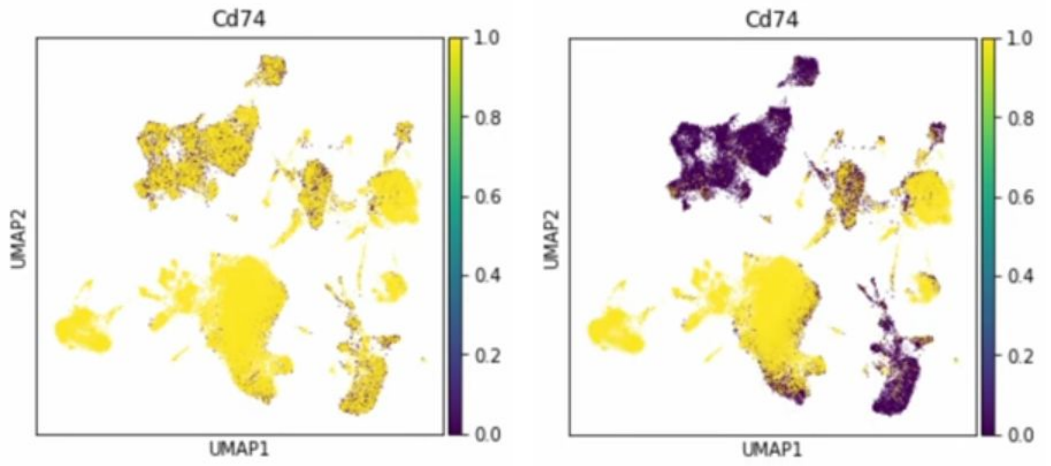
- The (double) "**kneeplot**"
- Counts = f(ranked.dropplets)
- "steep cliff" => best transition from "true cells" to empty droplets
- Actually a bit more complex …
- R tool : DropletUtils::*emptyDrops*



Ambient RNA filtering (SoupX)

- emptyDrops : removed empty droplets (contained only ambient RNA)
- ***BUT*** *non-empty droplets* ***ALSO*** *have ambient RNA !*
- ***soupX*** determines the amount of ambient RNA in counts, removes it





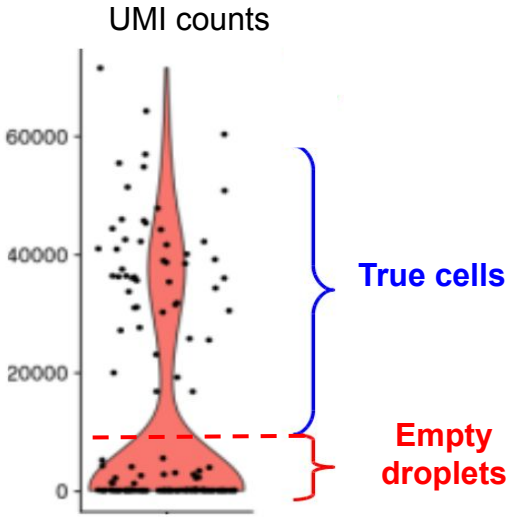
Young et al. GigaScience (2020)

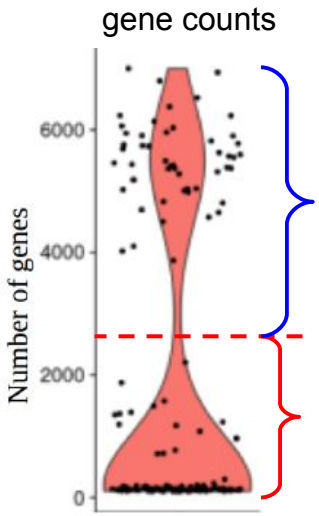

To Be Continued

QC and filtering (counts, features)

Filtering of empty / bad quality cells

- Visualize data and **deduce** thresholds
- Possible visualization: **Violin** Plot : Distribution of a cell feature. Can add points to visualize cells exactly (1 point = 1 cell) // **Histogram**
- Ideal distribution should be normal. In practice, it is **bimodal**



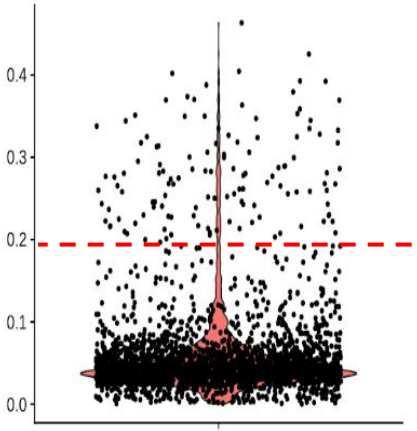


QC and filtering (%mito, %ribo, …)

Filtering of empty / bad quality cells

- Visualize data and **deduce** thresholds
- Possible visualization: **Violin** Plot : Distribution of a cell feature. Can add points to visualize cells exactly (1 point = 1 cell) // **Histogram**
- Distribution of features that **capture** a large part of expression (mito genes, riboproteins, …)

Mitochondrial (mt) genes expression

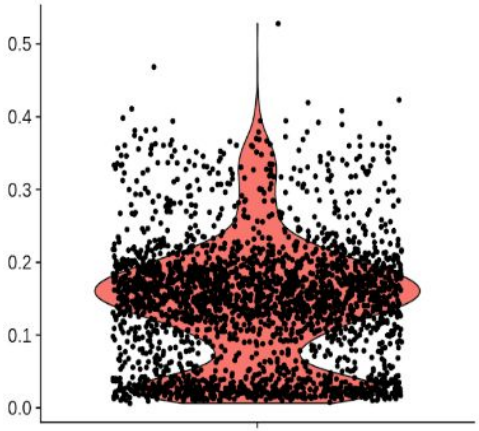


High % of mt genes may be due to apoptotic/dead or over-lyzed cells

Here the distribution has a long right tail.

Depending on dataset, remove cells > 5, 10, 20, 25% mtRNAs… (tl,tb)

Ribosomal protein genes expression ?



Reflects cell stress or cellular activity? Cell cycle?

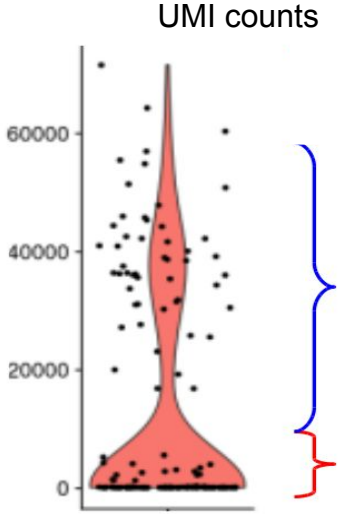
Is it a good marker: community debate.

+ Mechanical stress

QC and filtering

Filtering of empty / bad quality cells

- Visualize data and deduce thresholds
- Possible visualization: Violin Plot : Distribution of a cell feature. Can add points to visualize cells exactly (1 point = 1 cell)

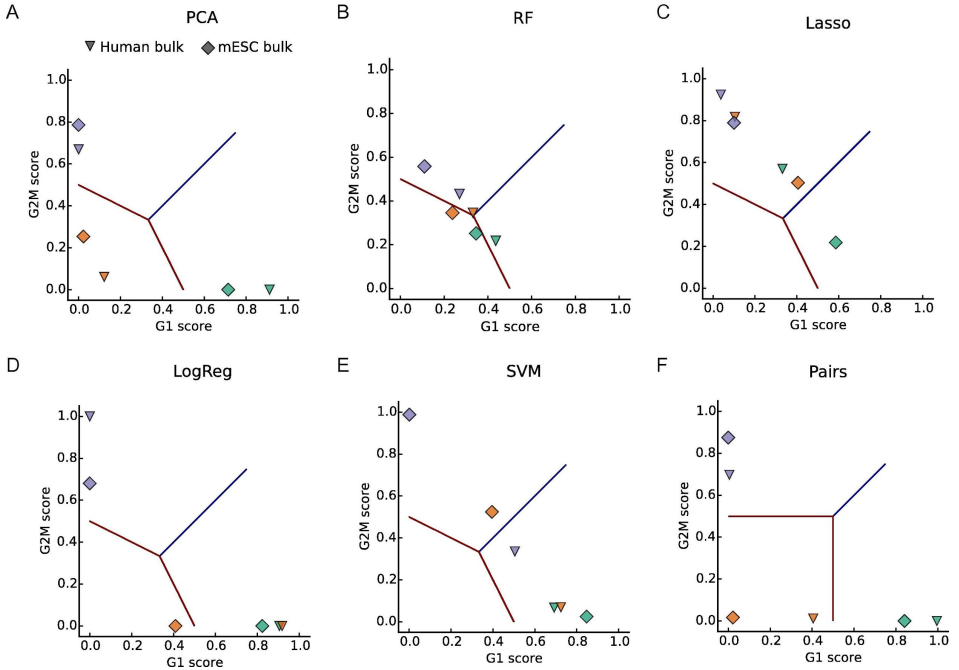


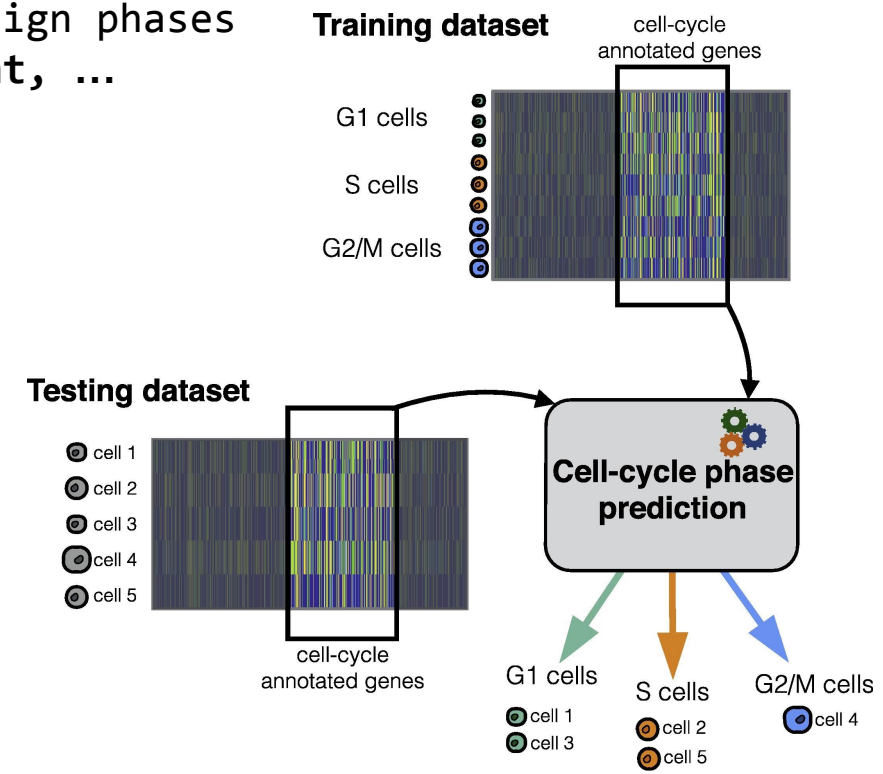
Select the thresholds carefully if you expect a population with a **small transcriptome**: e.g. immune cells (B especially), stem cells, …


To Be Continued

Cell cycle phase estimation

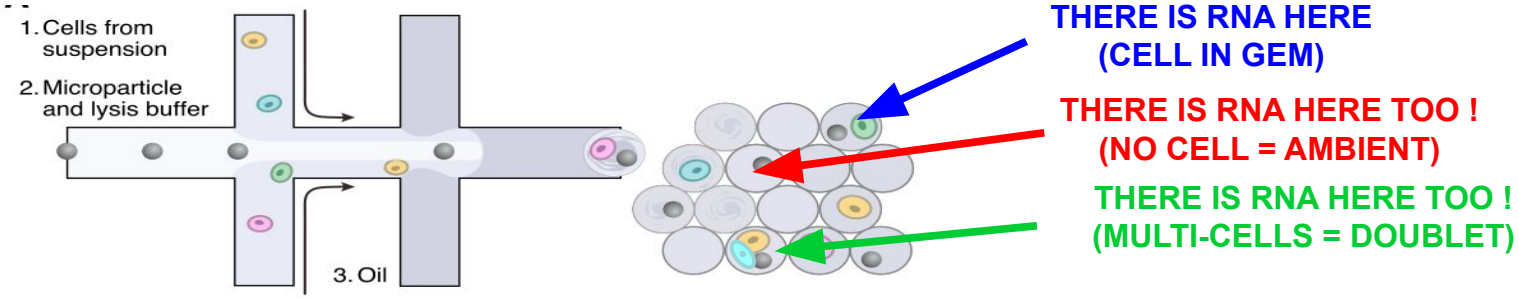
- Variational expression due to cell phase may be strong !
- Training on reference set with the 3 phases identified
- Use pairs of differential genes
- Apply model pairs to new dataset, assign phases
- Implemented in **cyclone** *(scran)*, **Seurat**, …

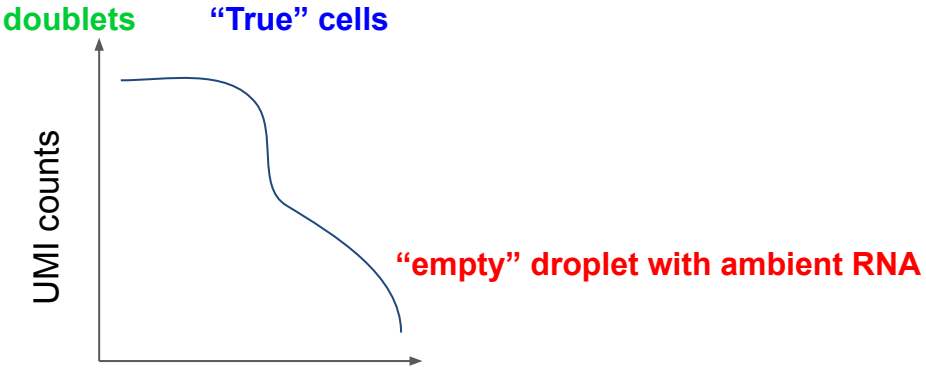




Scialdone et al., Methods (2015)

Filtered matrix composition : Doublets

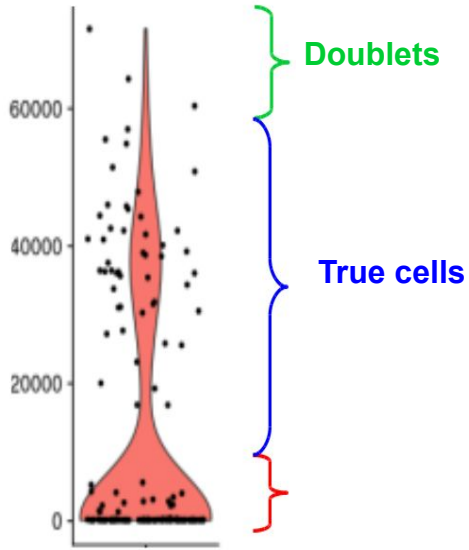




- True cells
- Empty, low quality droplets
- Doublets:
  - 1% for 1000 cells
  - 5% for 10 000 cells

Doublets detection

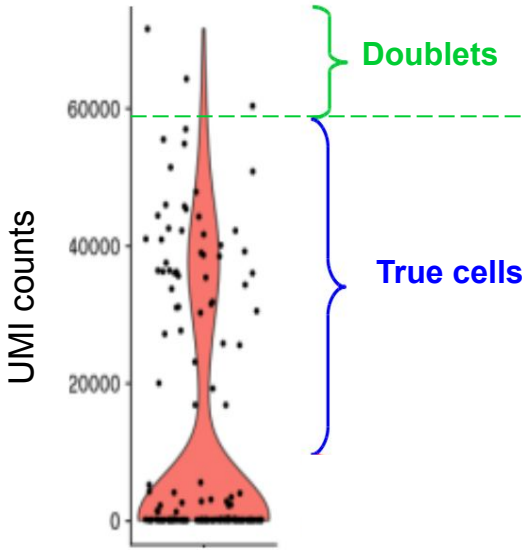
- Visualize nb UMIs (nCount) as a Violin Plot and set a threshold



- Doublets harbor a non-natural expression :
  - Higher level but same profile for doublets of the same cell type *(homotypic)*
  - Artificial profile for doublets of different cell types *(heterotypic)*
- This may have a **major impact** on the structure of signal in the data

Doublets detection

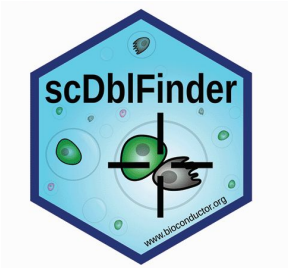
- Visualize nb UMIs as a Violin Plot and set a threshold



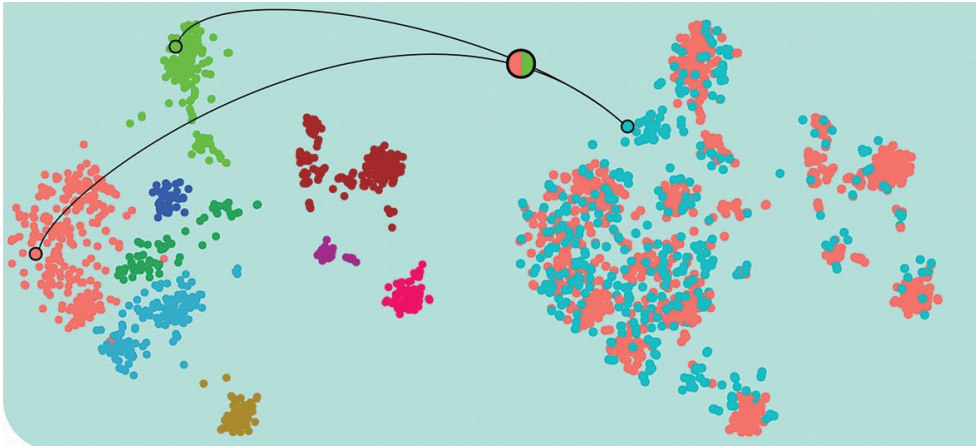
This threshold might be hard to tell
+
There may be a true cell subpopulation with higher expression ?

Doublets detection

- doublet detection by simulation



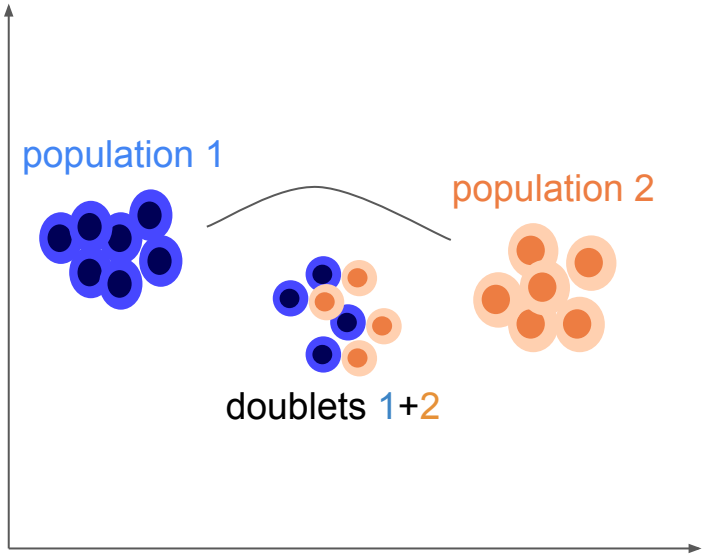
Pick 2 random cells ⇨ Generate artificial doublets ⇨ train doublet classifier ⇨ exclude doublet

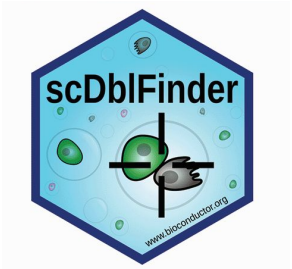


if a droplet looks like many artificial doublets, it has a high doublet score

Doublets detection

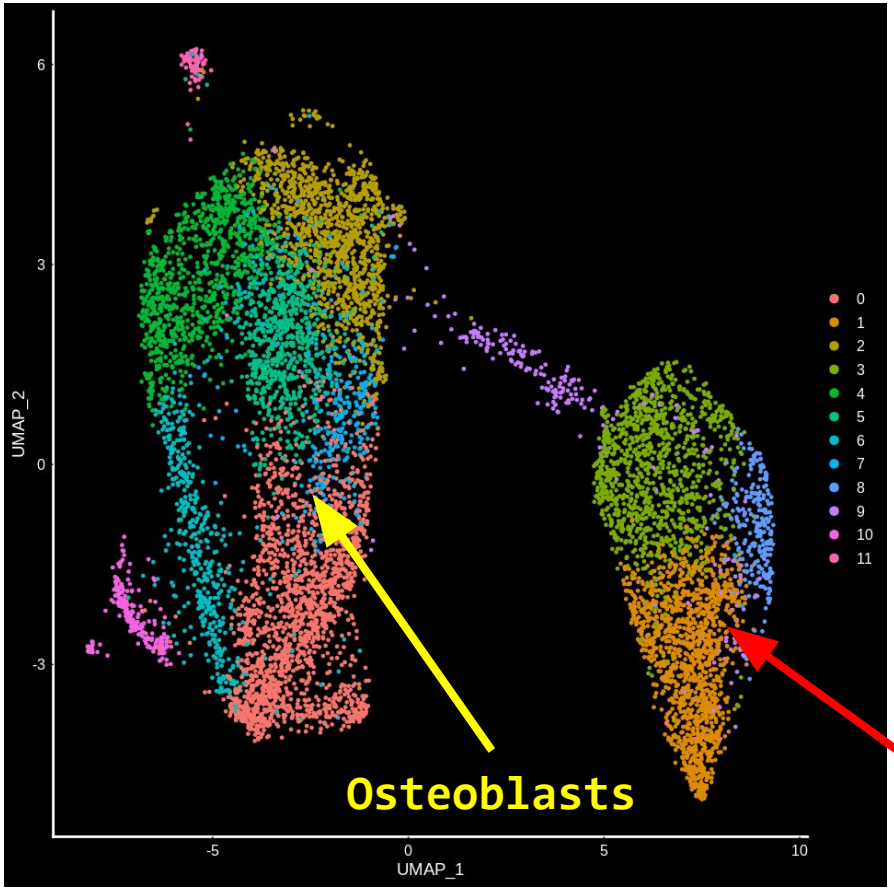
- doublet detection by clustering:
  - doublets composed of two cell types cluster between these cell types
  - check differentially expressed genes between putative doublets cluster and pop1 + pop2: there should not be many





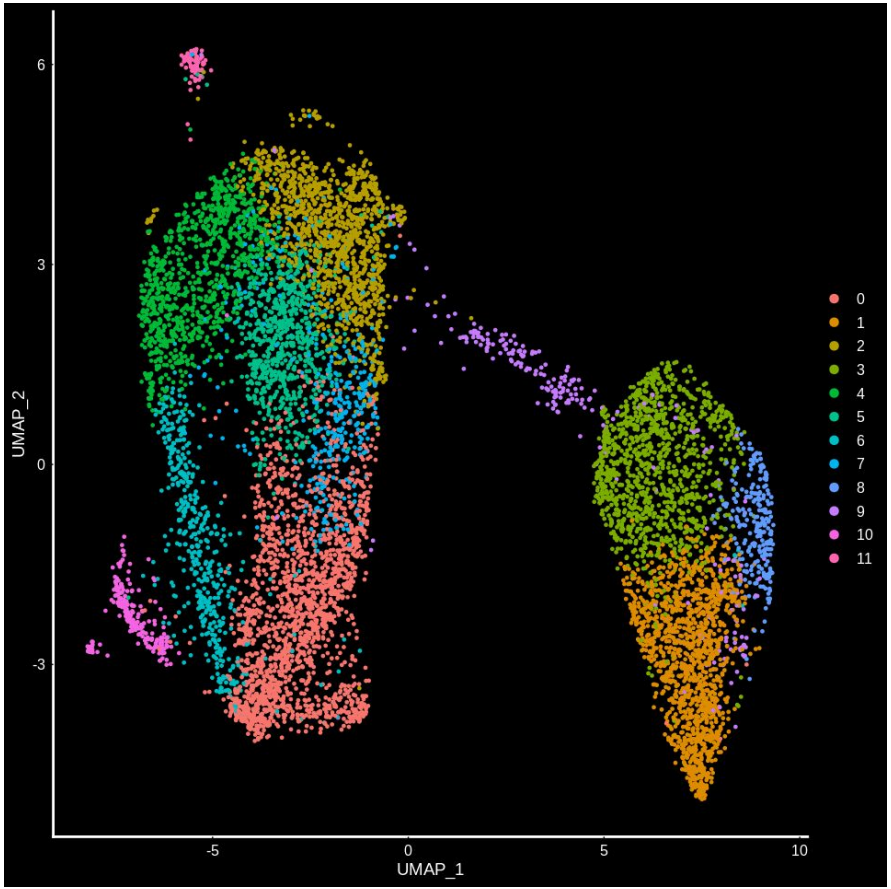
`findDoubletClusters()`

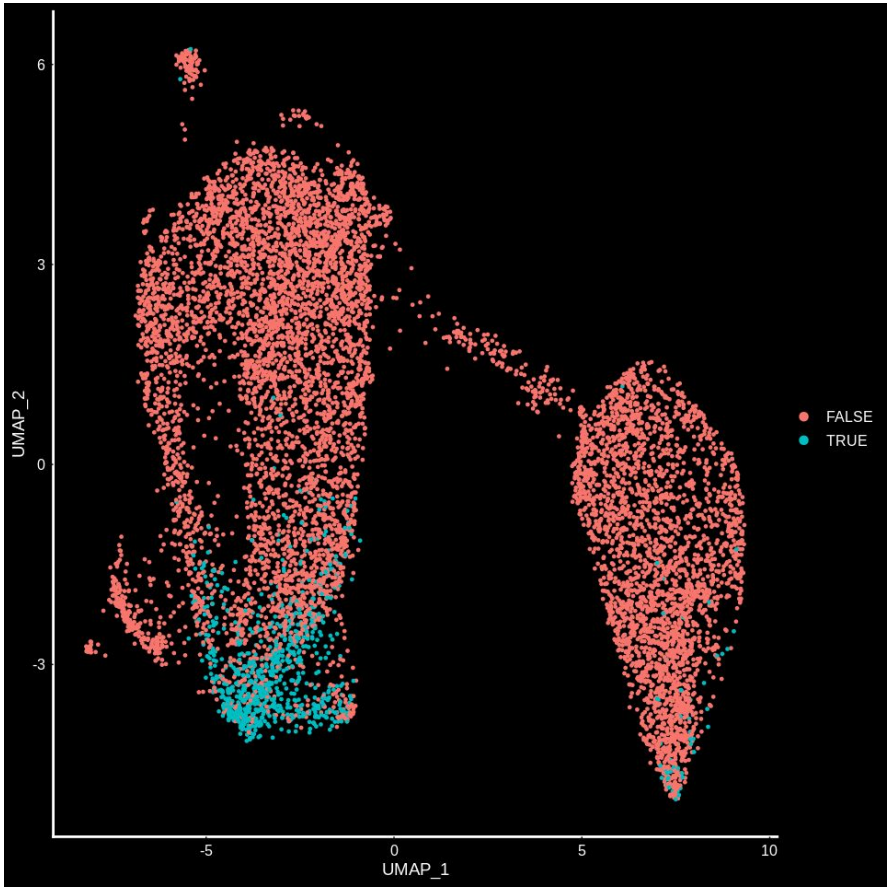
Visualization : a real-life example



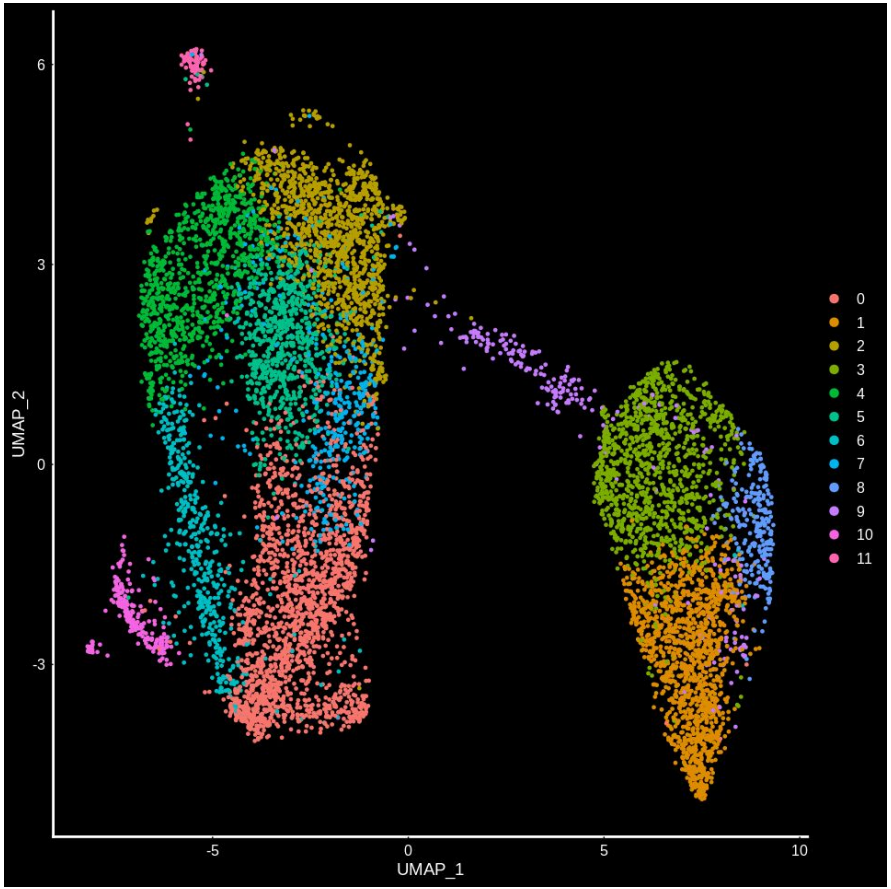
- 10X 3’ scRNAseq v2
- Osteosarcoma metastasis
- 8911 cells x 18613 genes
- PCA (109 PCs retained)
- Louvain clustering
  - 12 clusters
- uMAP representation

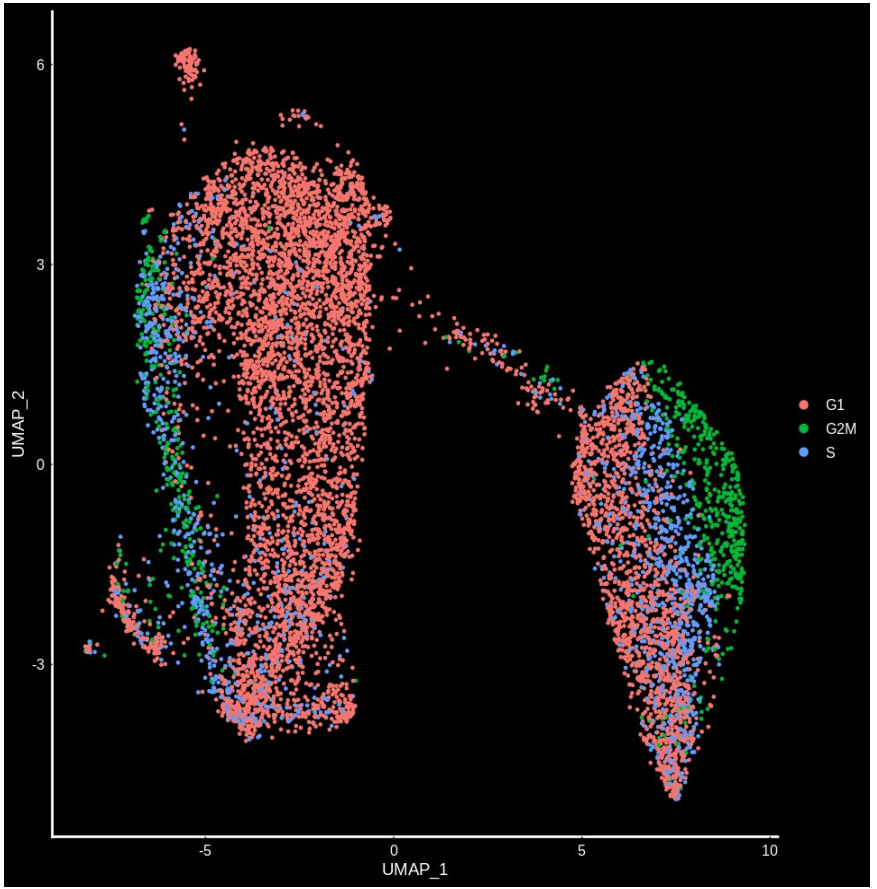
Bias : Dying cells (%mt high) status / score





Bias : Cell cycle phases / scores





Bias : Cell doublet status / score