Single-cell RNA-seq data analysis in R with Seurat library

Réalisé par Tidjani CISSE

Sommaire

Introduction à la scRNA-seq et outils principaux

- scRNA-seq allow
  - individual cell characterization (cell heterogenity).
  - uncover regulatory relationship between genes.
  - track the trajectories of distinct cell lineages in development.

- Multiple toolkits and frameworks are develpped to facilitate scRNA-seq analysis (seurat, scanpy, etc...)

- In this tutorial, it is assumed that all data sequencing preprocessing steps have been done (including base calling, mapping and read counting).

Les données

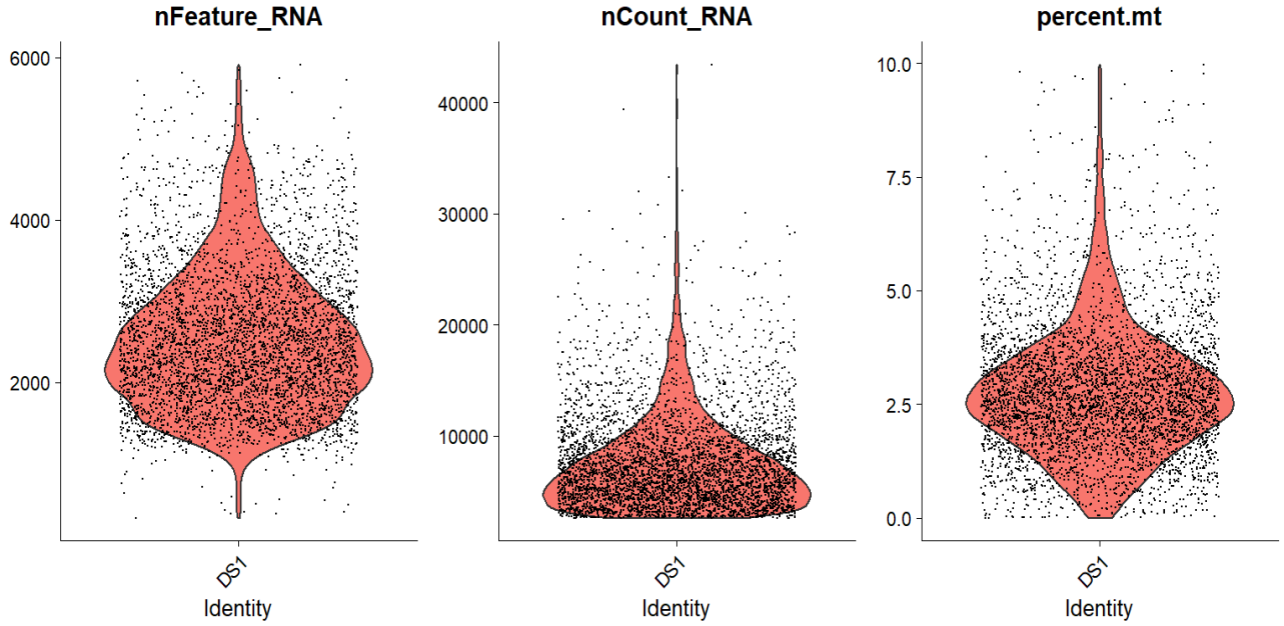
- Two scRNA-seq data sets (DS1 and DS2) are used and come from human cerebral organoids.

- They are generated using 10X Genomics Chromium and processed using Cell Ranger.

- After the Cell Ranger pipeline a raw count matrix is generated and consist of n x N (n *for number of genes in the lines* and N for number of cells in the columns to represent cout/expression matrix of scRNA-seq data).

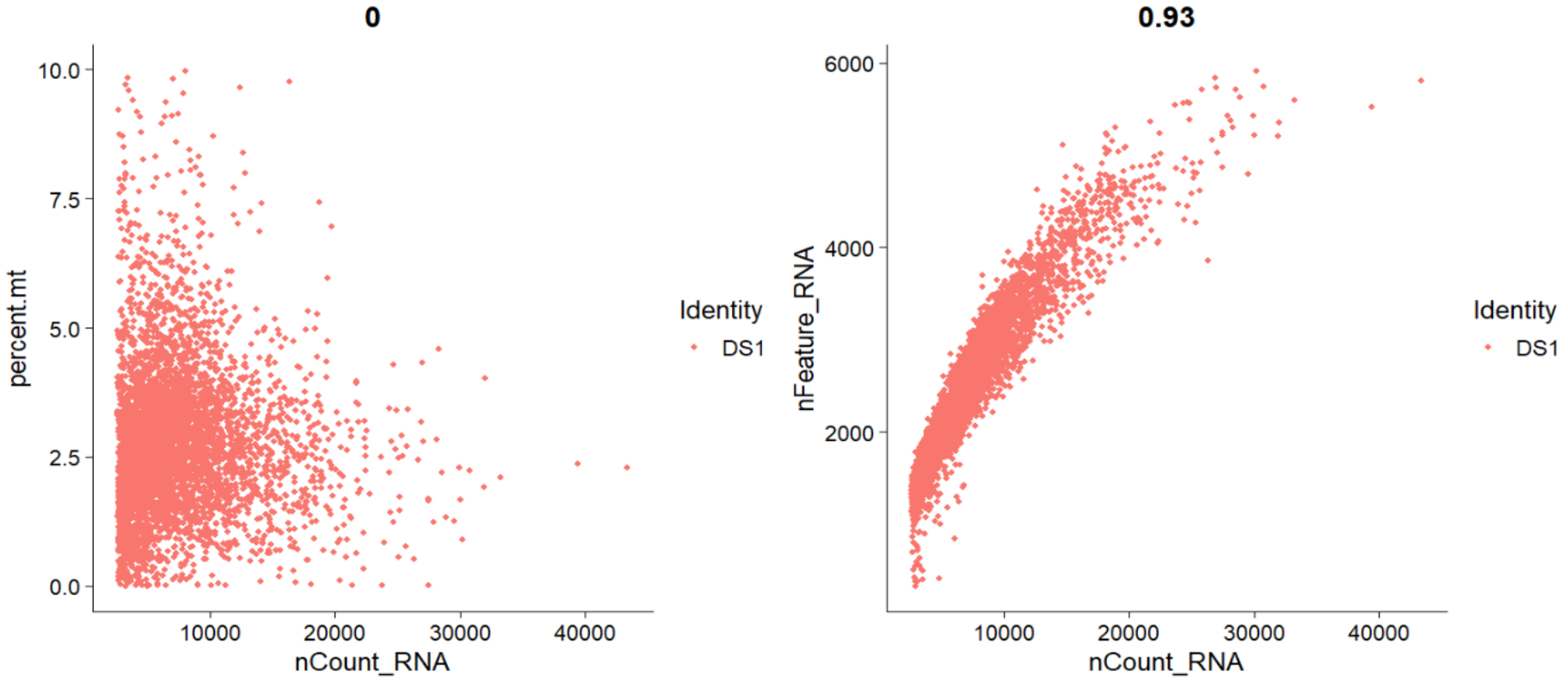
Contrôle de qualité(QC)

- QC consist in the filtering out:
  - Cells with too few or too many genes detected and Cells with high mitochondrial transcript percentage.
  - nFeature_RNA is the number of detected genes cells
  - nCount_RNA is the number of transcripts in cells
  - Percent.mt is the rate of mitochondrial transcripts.



Contrôle de qualité(QC)

- Correlation between number nFeatures_RNA and nCount_RNA,
- Not between then and percent.mt
- Our data set contain healthy cells



Contrôle de qualité(QC)

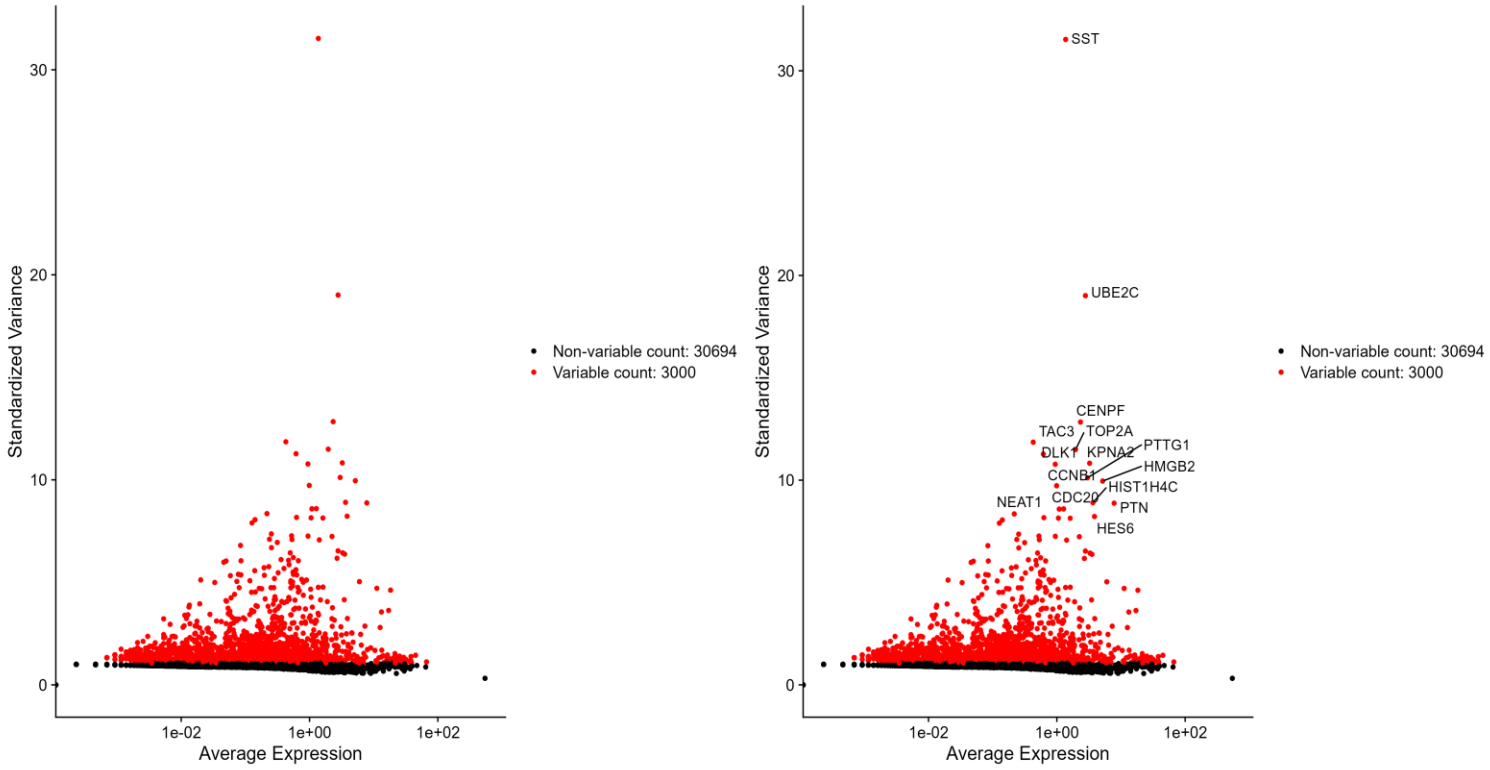
According to the QC, we decide to continue with detected gene number between 500 and 5000, and a mitochondrial transcript percentage lower than 5%.

Below , the nFeature_RNA and nCount_RNA before and after the cutoff.

nFeature_RNA	nCount_RNA
33 694	4 672
33 694	4 317

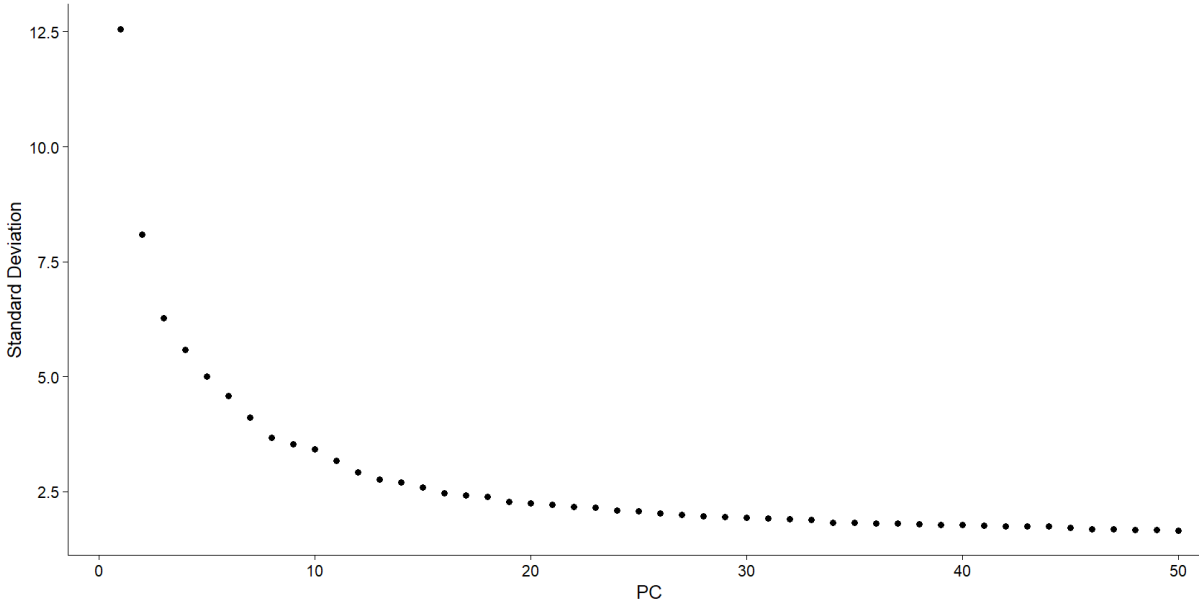
Normalisation des data et sélection des features variables

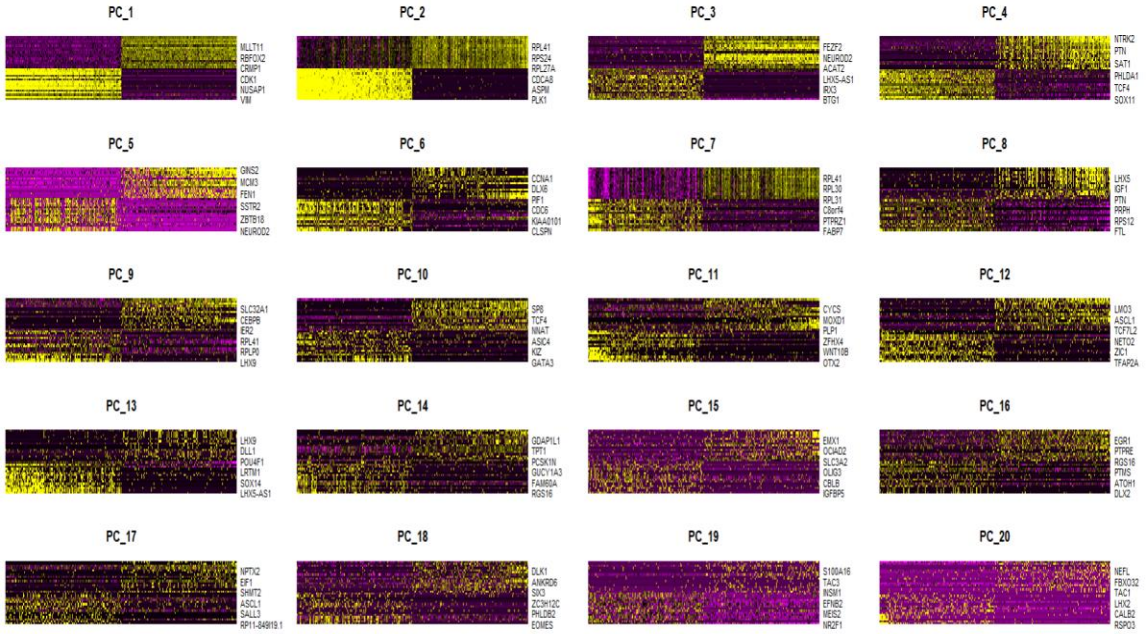
- Need nFeature_RNA and nCount_RNA normalization because of the difference of the capture RNA between cells.
- Applying of features variables selection to find the cellular heterogeneity of samples , by looking for cell groups with distinct molecular signature.



Réduction de dimension linéaire (PCA)

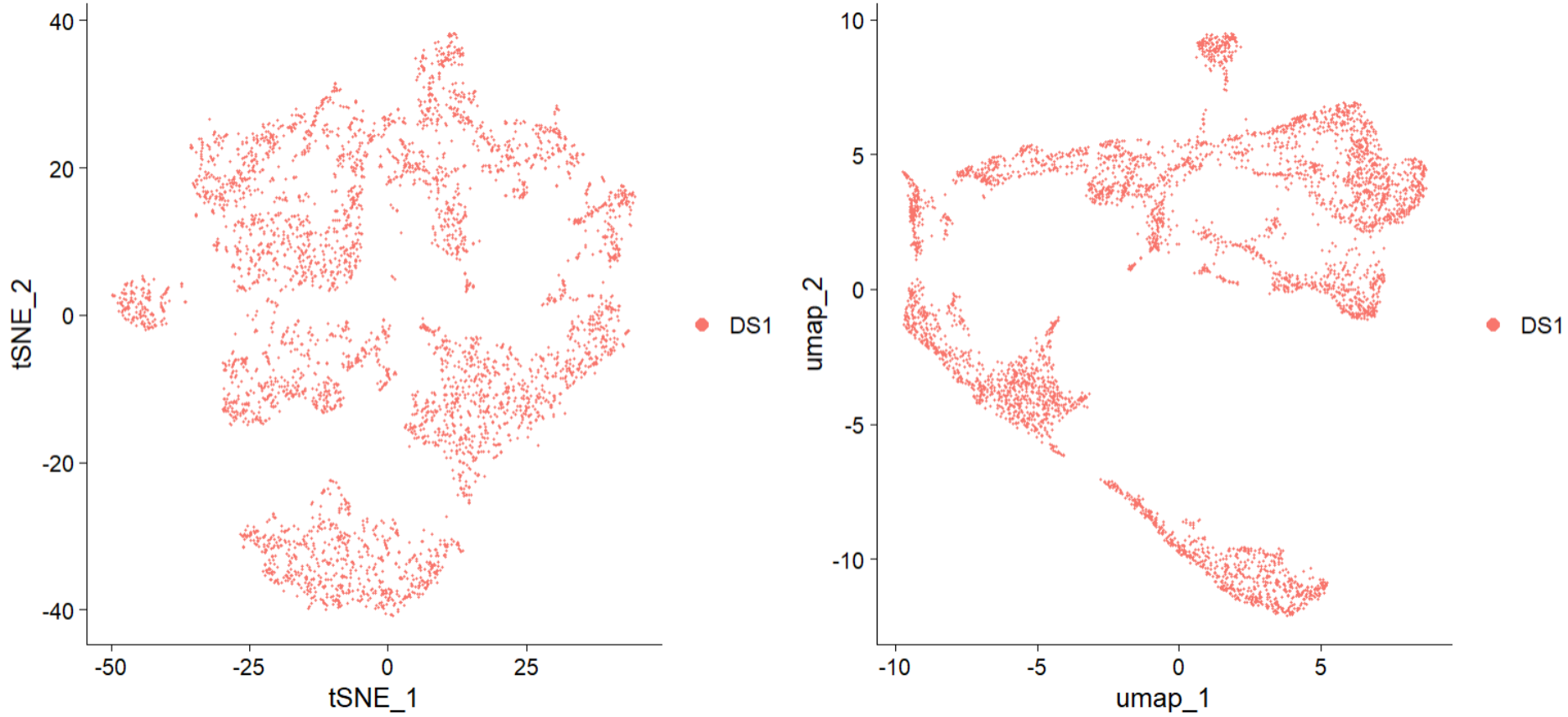
- Dimension reduction allows the data to become much more compact so that computation becomes much faster and gives signal robustness,
- Features x cells/genes of 20 first principal components (PCs).
- We can check which genes are mostly contributing to each of the top PCs





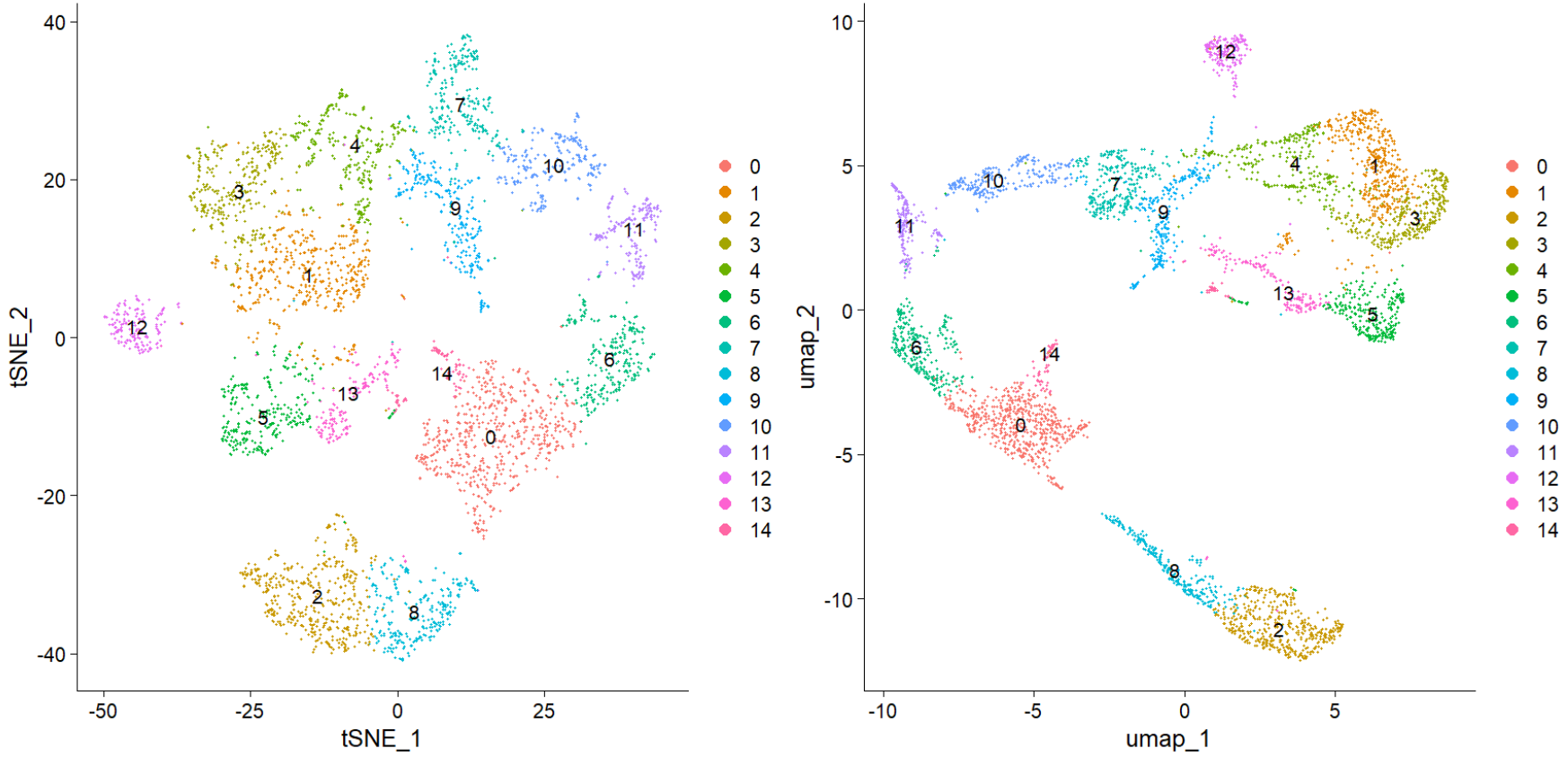
Réduction de dimension non-linéaire (tSNE, UMAP)

- PCA method is good for interpration of PCs but not for visualization over three dimension.
- Therefore tSNE and UMAP allow to visualize distinct cell groups and cell state change during development and differentiation.



Clustering des cellules

- We identify 14 cell groups in the data set after the clustering.
- we can assume that cells with the same label are similar therefore can be seen to be of the same cell type or cell state.



Clustering des cellules

- Which exact cell types or cell states these cell clusters are representing ?
- Several options exist for trying to resolve this question.
  - Check the expression of canonical cell type and cell state markers in these clusters.
  - Identify signature genes, or marker genes, of each identified cell cluster and do literature search, enrichment analysis or do experiment (or ask people around) for annotation.
  - For each cluster, compare its gene expression profile with existing reference data.

For our data set (cerebral organoid), some of the markers are known.

Annotation des clusters