scRNA-seq : visualization

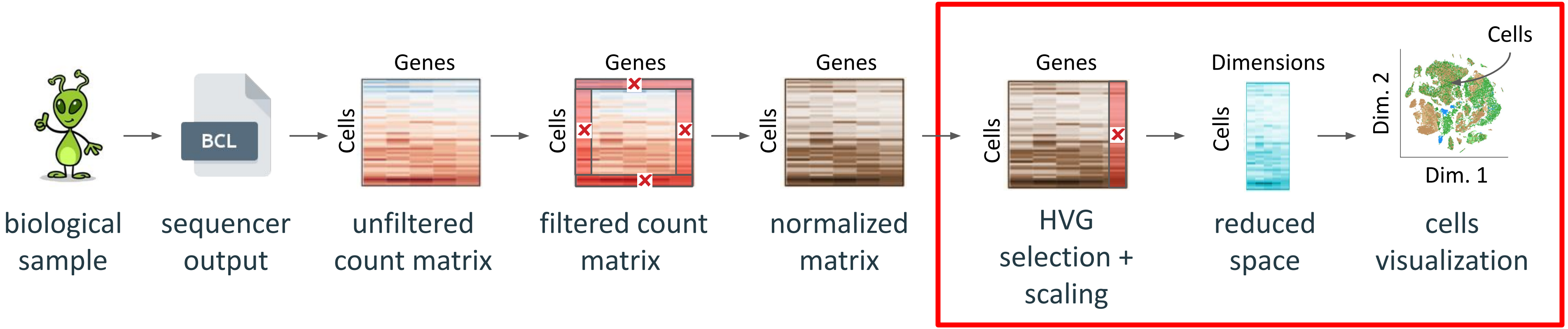
Bastien Job, Gustave Roussy, Villejuif

Nathalie Lehmann, Institut Pasteur, Paris

Audrey Onfroy, Institut Mondor, Créteil

École de bioinformatique AVIESAN-IFB-INSERM 2024

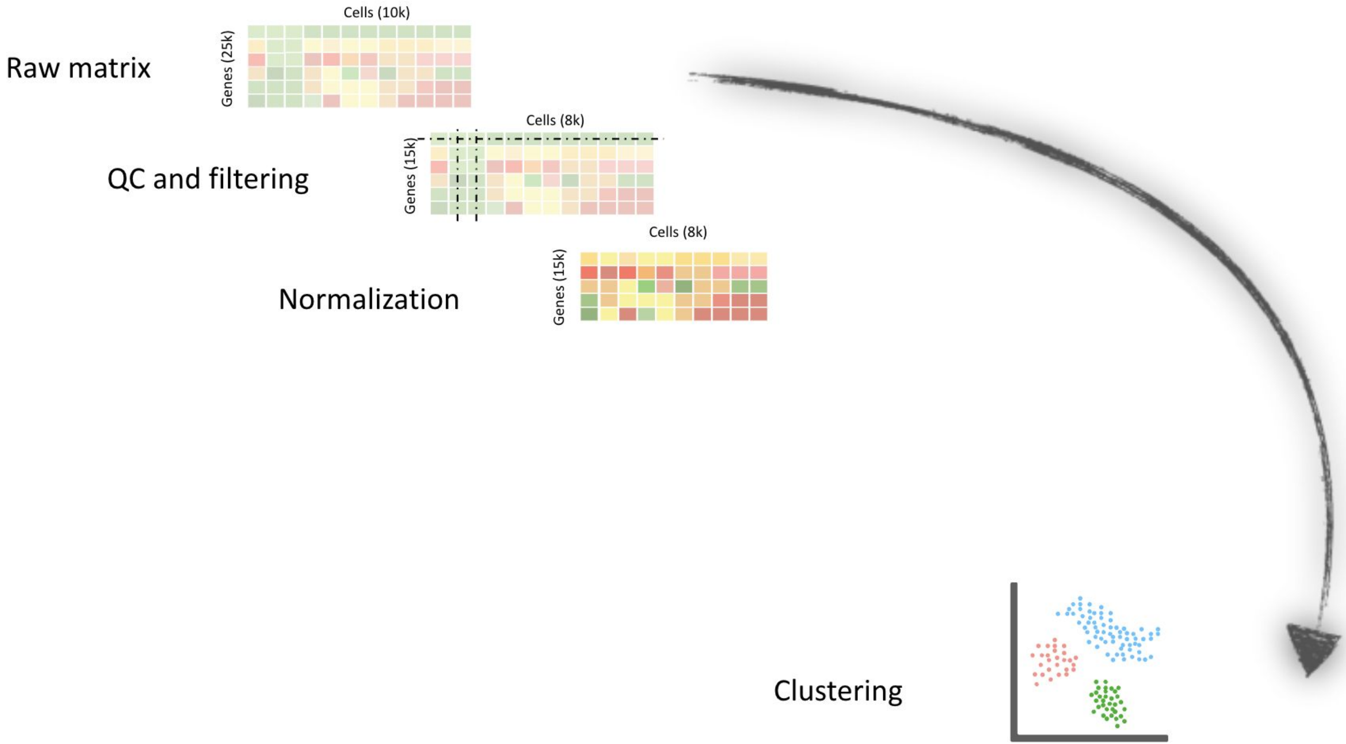
scRNA-Seq pipeline overview



biological sample → sequencer output → unfiltered count matrix → filtered count matrix → normalized matrix → HVG selection + scaling → reduced space → cells visualization

We want a visual summary of thousands cells’ gene expression.

How do we get to data visualization and clustering ?



Raw matrix

QC and filtering

Normalization

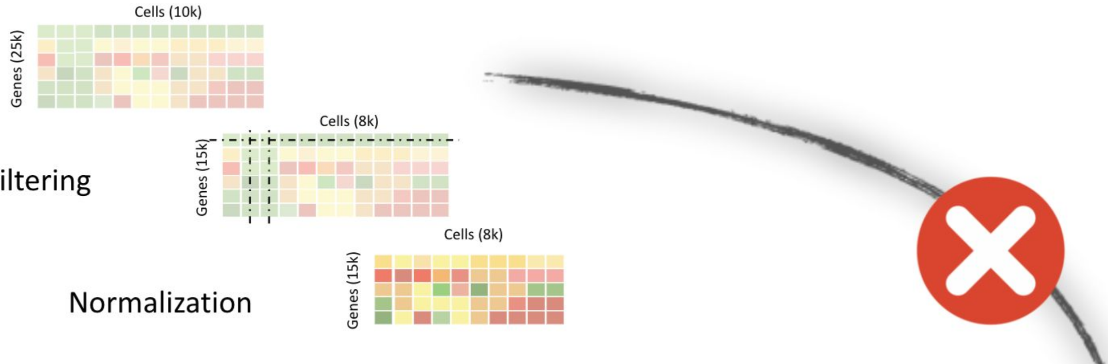
Clustering

How do we get to data visualization and clustering ?

Raw matrix

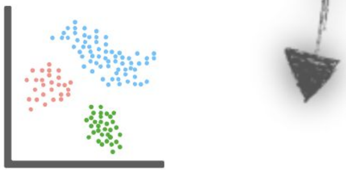
QC and filtering

Normalization

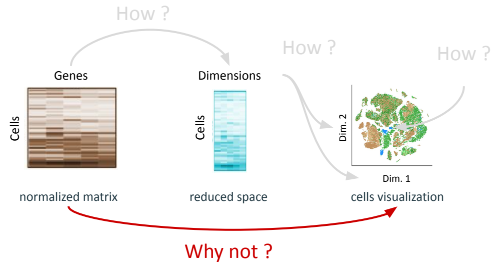


We need to **summarize gene expression** to a few dimensions (20-100) to effectively separate the populations

Clustering

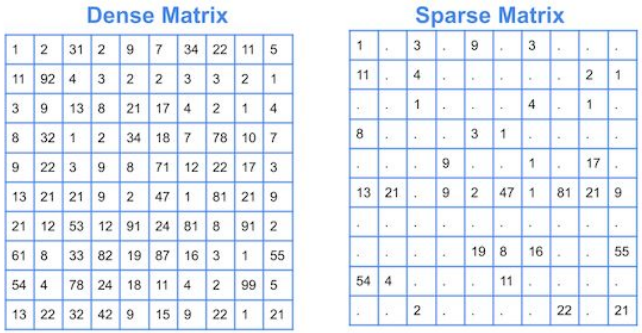


Why an intermediary step is necessary ?



scRNA-Seq data are sparse

> 70 % of the expression matrix is 0 : **not very informative**



Dense Matrix

1	2	31	2	9	7	34	22	11	5
11	92	4	3	2	2	3	3	2	1
3	9	13	8	21	17	4	2	1	4
8	32	1	2	34	18	7	78	10	7
9	22	3	9	8	71	12	22	17	3
13	21	21	9	2	47	1	81	21	9
21	12	53	12	91	24	81	8	91	2
61	8	33	82	19	87	16	3	1	55
54	4	78	24	18	11	4	2	99	5
13	22	32	42	9	15	9	22	1	21

Sparse Matrix

1	.	3	.	9	.	3	.	.	.
11	.	4	.	.	.	.	.	2	1
.	.	1	.	.	.	4	.	1	.
8	.	.	.	3	1	.	.	.	.
.	.	.	9	.	.	1	.	17	.
13	21	.	9	2	47	1	81	21	9
.	.	.	.	.	.	.	.	.	.
.	.	.	.	19	8	16	.	.	55
54	4	.	.	.	11	.	.	.	.
.	.	2	.	.	.	.	22	.	21

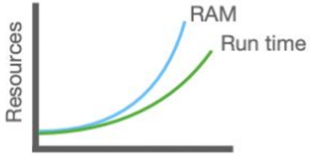
http://cmdlinetips.com/wp-content/uploads/2018/03/Sparse_Matrix.png

```
prop(expr_mat == 0)
```

Data are noisy

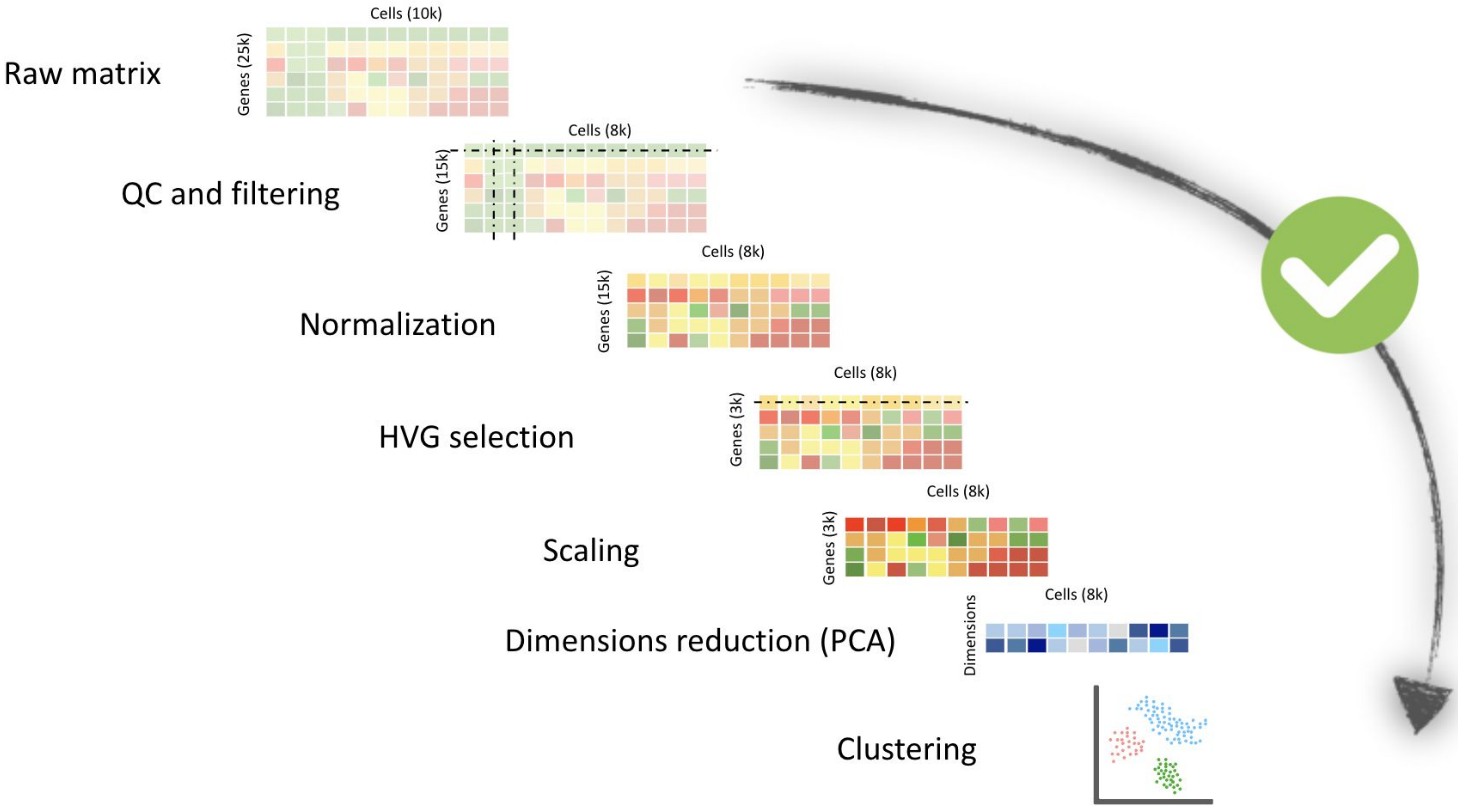
Some genes are more informative than some other.
There is **biological / technical noise** in gene expression.

Computational time and ressources



We will summarize genes expression in few dimensions, before building the 2D projection.

The right way to get to data visualization and clustering



Raw matrix

QC and filtering

Normalization

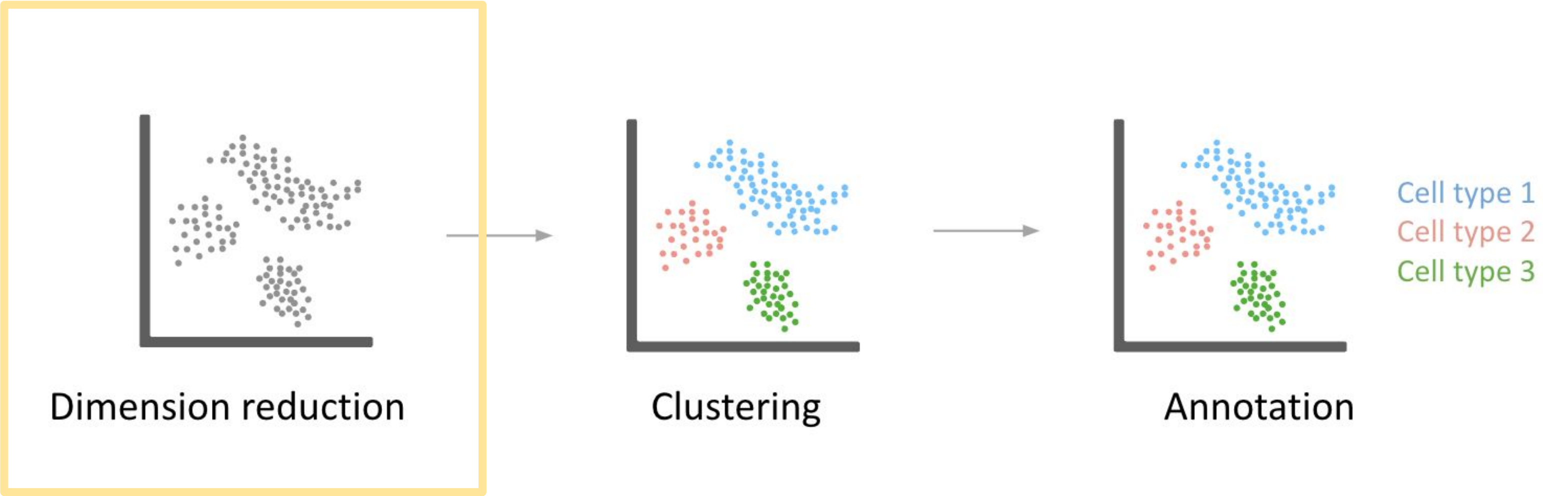
HVG selection

Scaling

Dimensions reduction (PCA)

Clustering

Our analyses goals



Dimension reduction

Clustering

Annotation

Cell type 1

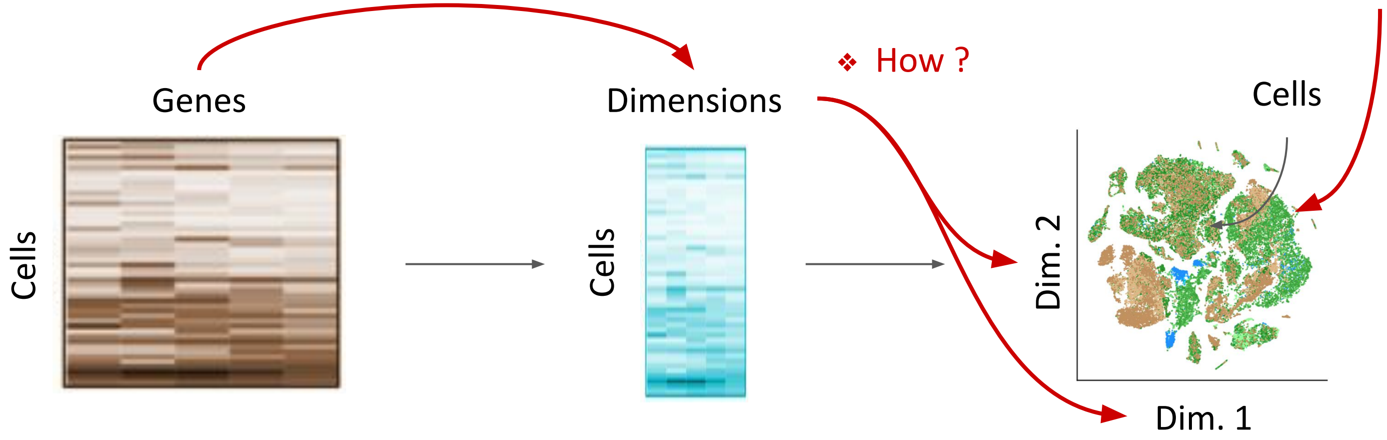
Cell type 2

Cell type 3

Challenges

- ❖ How to reduce the number of dimensions ?
- ❖ How many dimensions ?

❖ How to identify cell populations ?



❖ How ?

Genes

Cells

Dimensions

Cells

Cells

Dim. 2

Dim. 1

normalized matrix

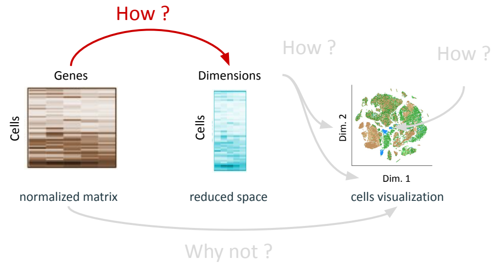
reduced space

cells visualization

We want a visual summary of thousands cells’ gene expression.

Dimensionality reduction

Overview

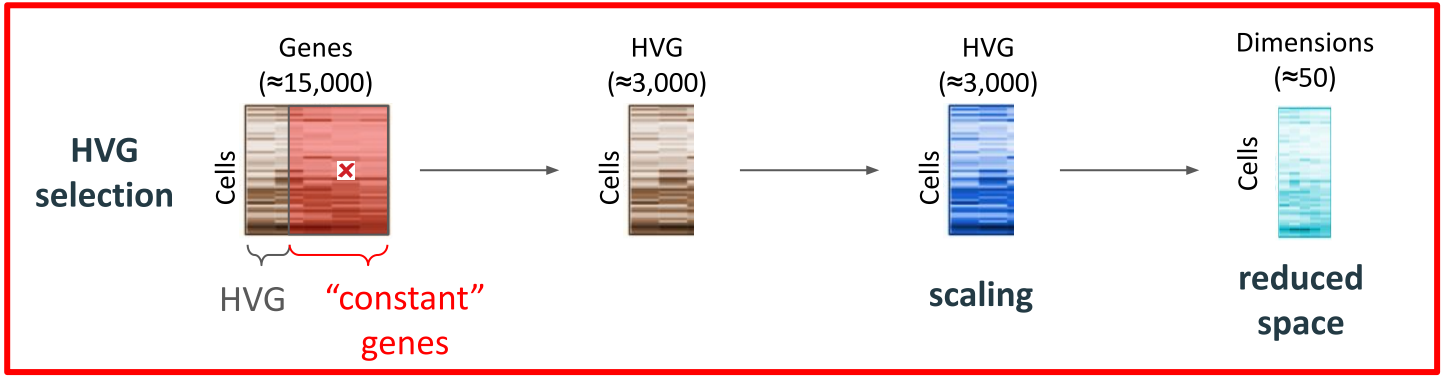


Commonly used dimensionality reduction methods

PCA	**P**rincipal **C**omponent **A**nalysis
BFA	**B**inary **F**actor **A**nalysis
ICA	**I**ndependent **C**omponent **A**nalysis
LSI	**L**atent **S**emantic **I**ndexing
LDA	**L**inear **D**iscriminant **A**nalysis
…	

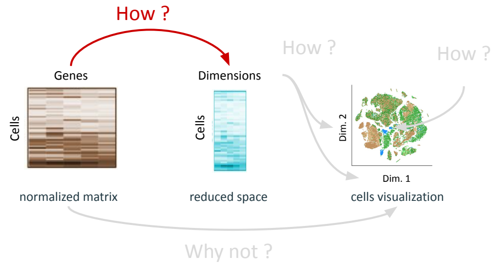
Important parameters

- **information** : number of variable genes (HVG)
- number of **dimensions** to generate (signal / noise)
- *randomness : random seed*
- *convergence criteria*

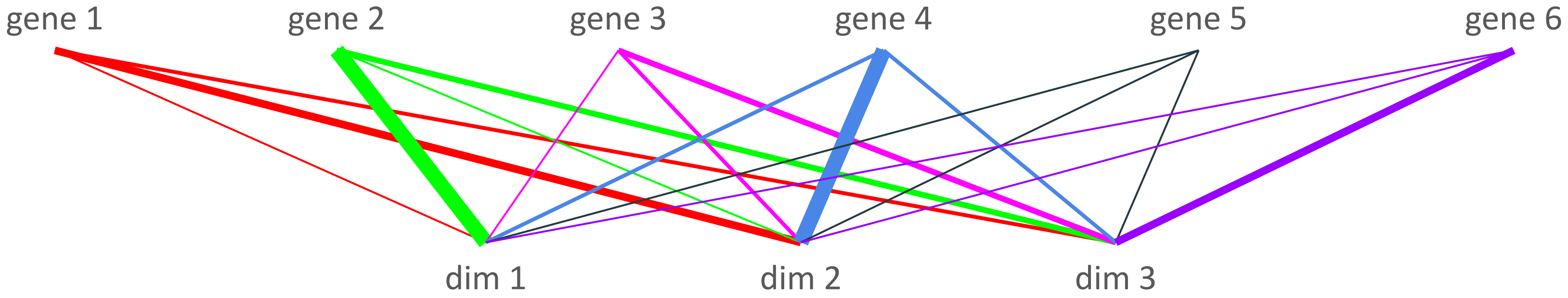


Dimensionality reduction

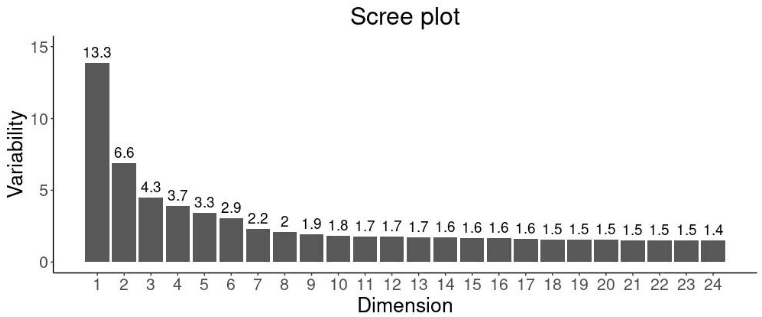
Principal Component Analysis - principle



- Input : **X** (≈ 2 000 - 5 000) HVG with **scaled** expression levels
- Goal : Group genes by dimensions when they have similar expression across cells

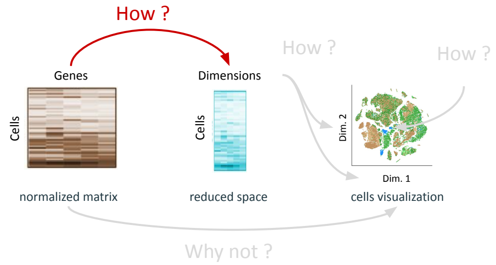


- Output : **Z** (≈ 50 - 100) dimensions “Principal Component”
- Each PC summarizes a certain amount of the input data variability
  - First PC recapitulates the most part of information
  - Last PC can be considered as noise



Dimensionality reduction

Principal Component Analysis - visualization



How ?

Genes

Dimensions

How ?

How ?

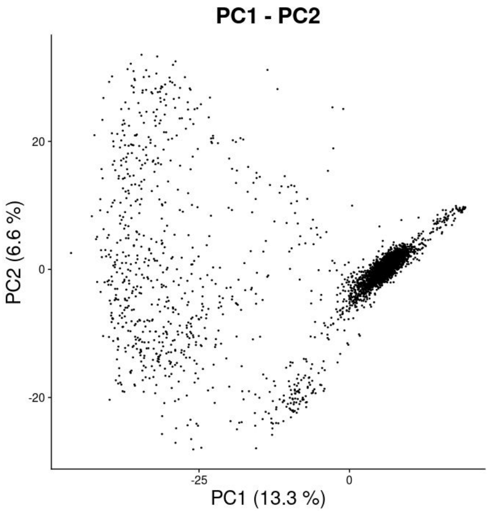
normalized matrix

reduced space

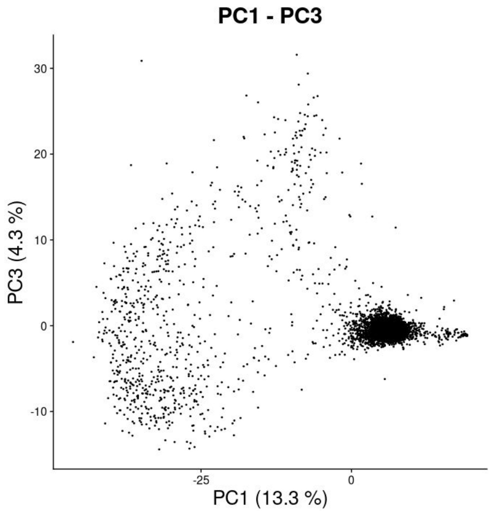
cells visualization

Why not ?

- Input : **X** most variable genes
- Goal : Group genes by dimensions when they have similar expression across cells
- Output : **Z** dimensions “Principal Component”
- Each PC summarizes a certain amount of the input data variability



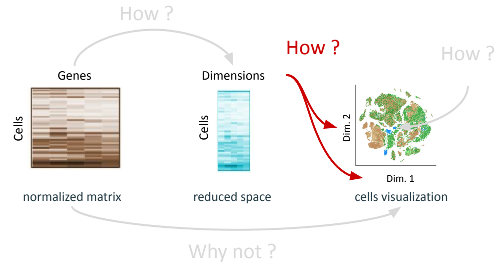
PC1 - PC2



PC1 - PC3

Now, we will use the reduced space to make a 2D representation.

2D space for cells visualization

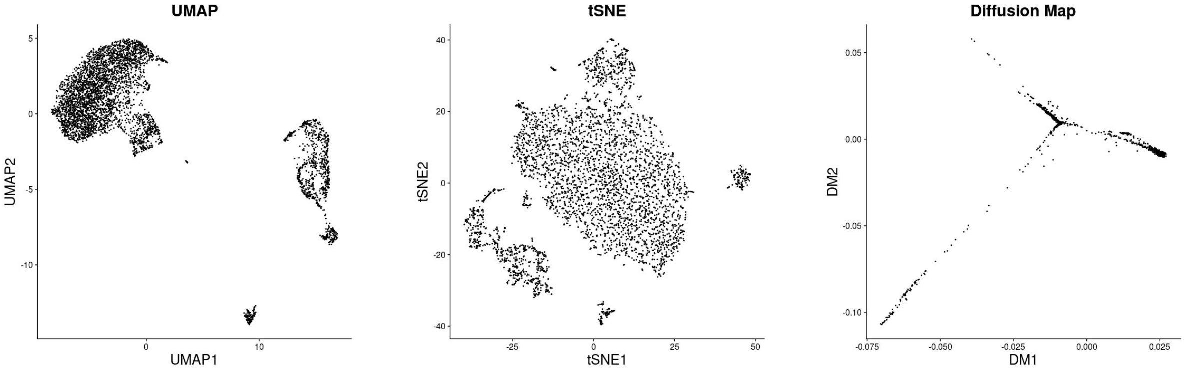


Commonly used 2D space

- **UMAP**
- **tSNE**
- **Diffusion Map**
- …

Important parameters

- **input information** : number of dimensions
- cells **neighborhood** : number of neighbors, perplexity, distance method, …

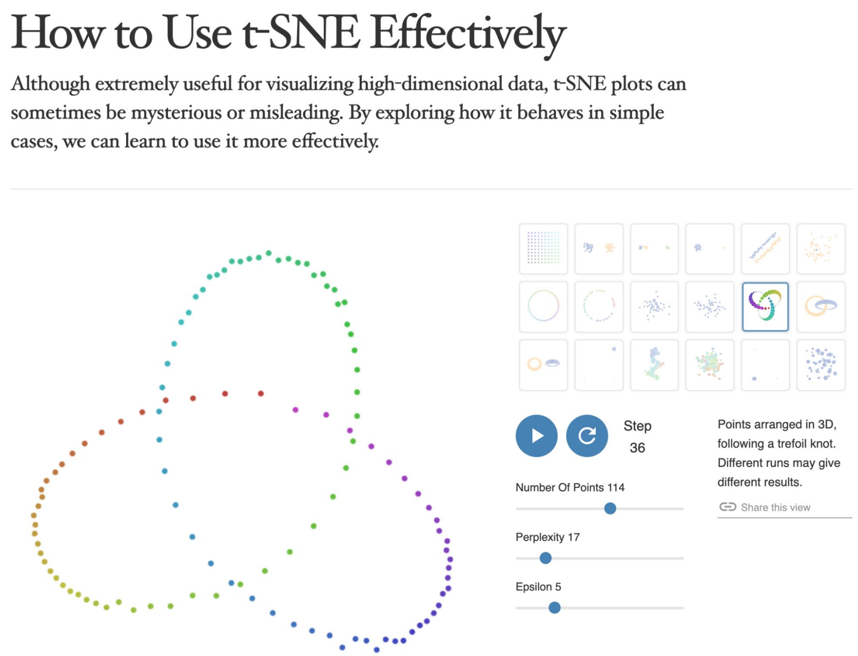


The same cells can be represented using **different 2D spaces**.

Do not make to many interpretations from the 2D space, it is an **over-simplified representation** of cells.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6417818/

There are an infinite way to represent our data into 2D



How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.

Step 36

Points arranged in 3D, following a trefoil knot. Different runs may give different results.

Share this view

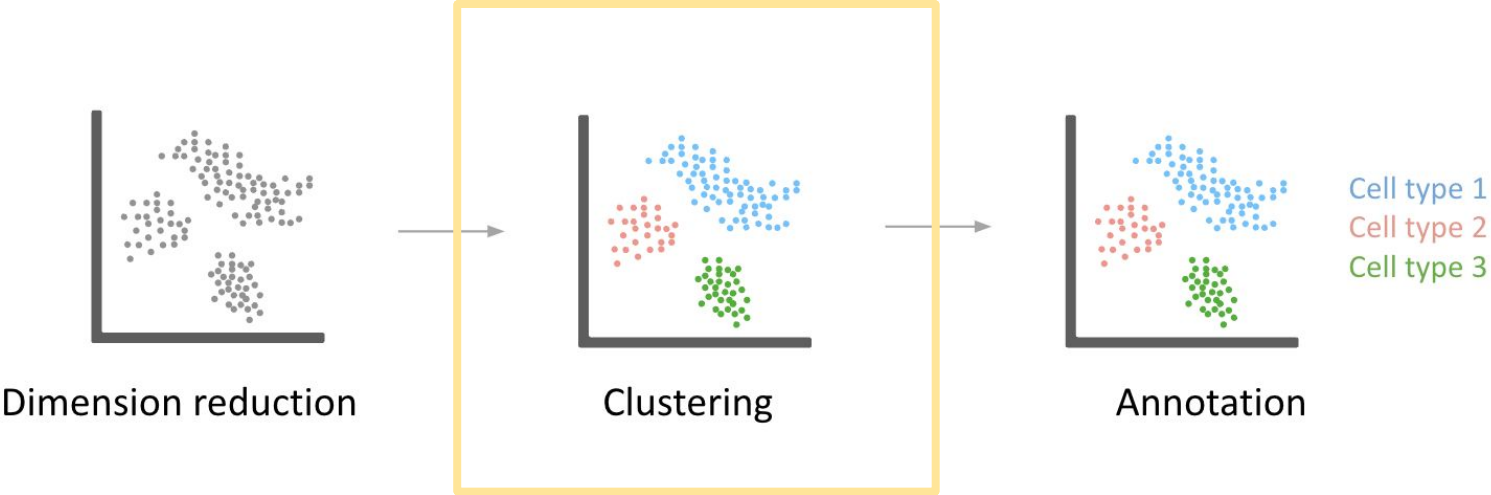
Number Of Points 114

Perplexity 17

Epsilon 5

https://distill.pub/2016/misread-tsne/

Our analyses goals



Cell type 1

Cell type 2

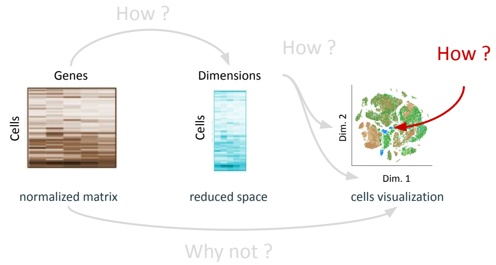
Cell type 3

Dimension reduction

Clustering

Annotation

Clustering

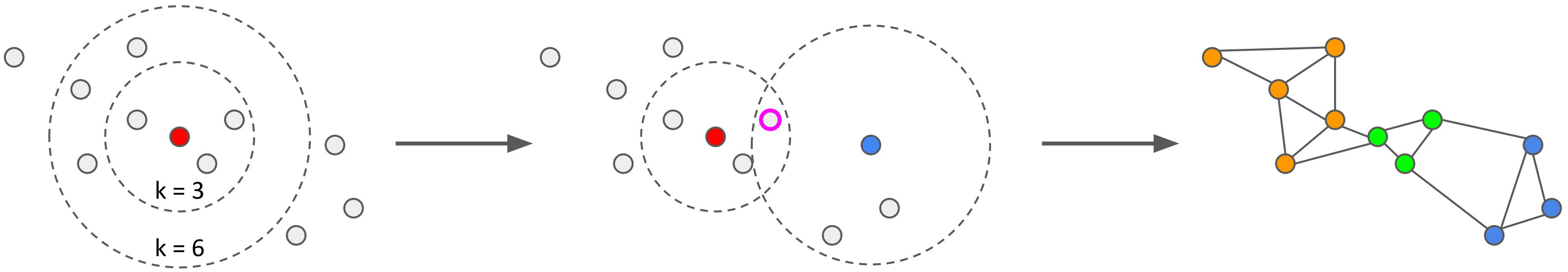


Commonly used methods

- **Louvain clustering**
- Leiden clustering
- k-means
- …

Important parameters

- **input information** : number of dimensions
- cells **neighborhood** parameters : number of neighbors, distance measurement method, **resolution**…



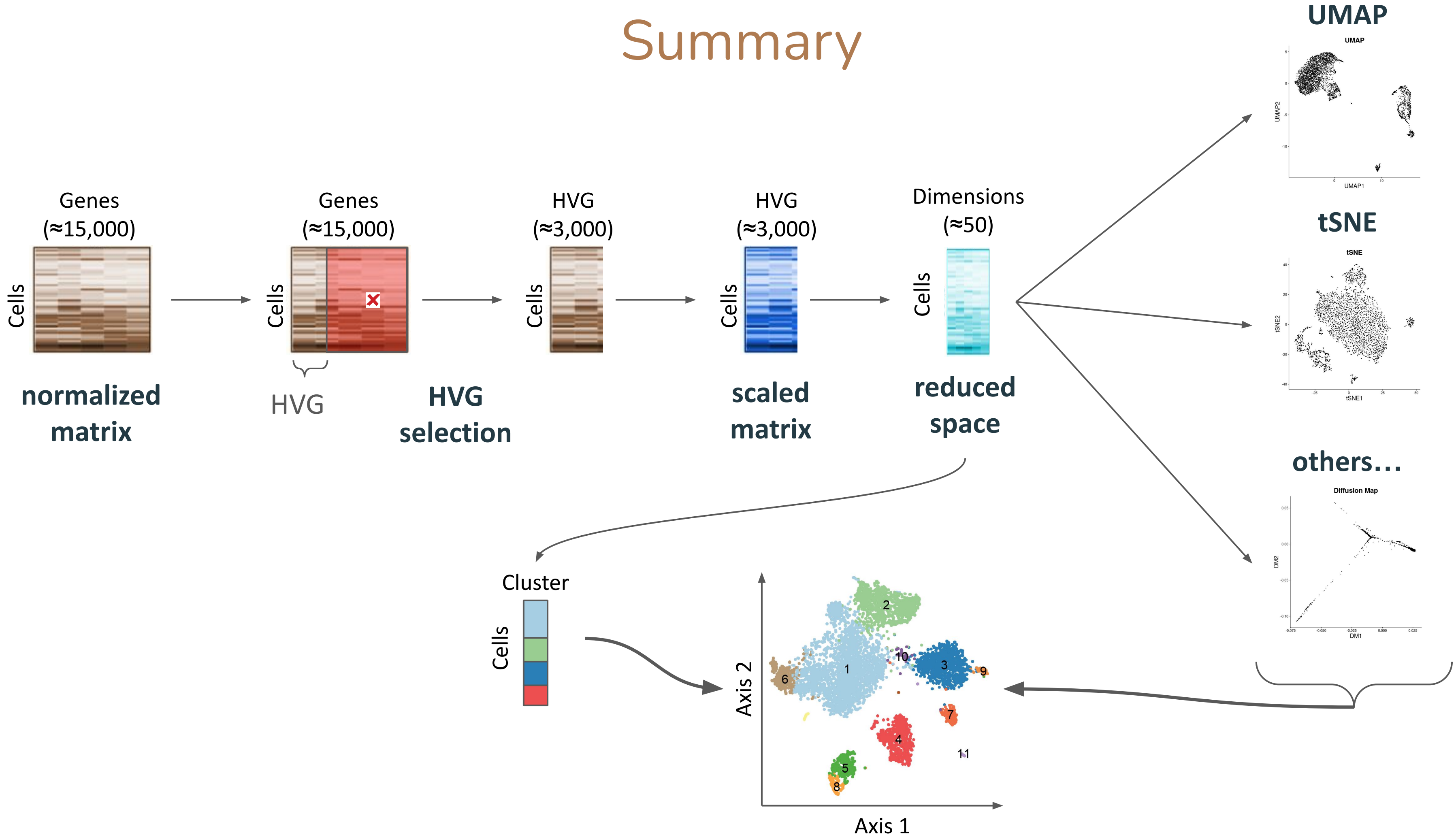
k-nearest neighbors
(kNN)

shared nearest neighbors
(SNN)

clustering
(from SNN graph)

Clustering is made on expression matrix or reduced space, <u>not</u> on the 2D projection.
The 2D projection is not a clustering. A clustering is an **annotation**.

Summary



Genes (≈15,000)

Cells

normalized matrix

Genes (≈15,000)

Cells

HVG

HVG selection

HVG (≈3,000)

Cells

HVG (≈3,000)

Cells

scaled matrix

Dimensions (≈50)

Cells

reduced space

UMAP

tSNE

others…

Cluster

Cells

Take Home Messages

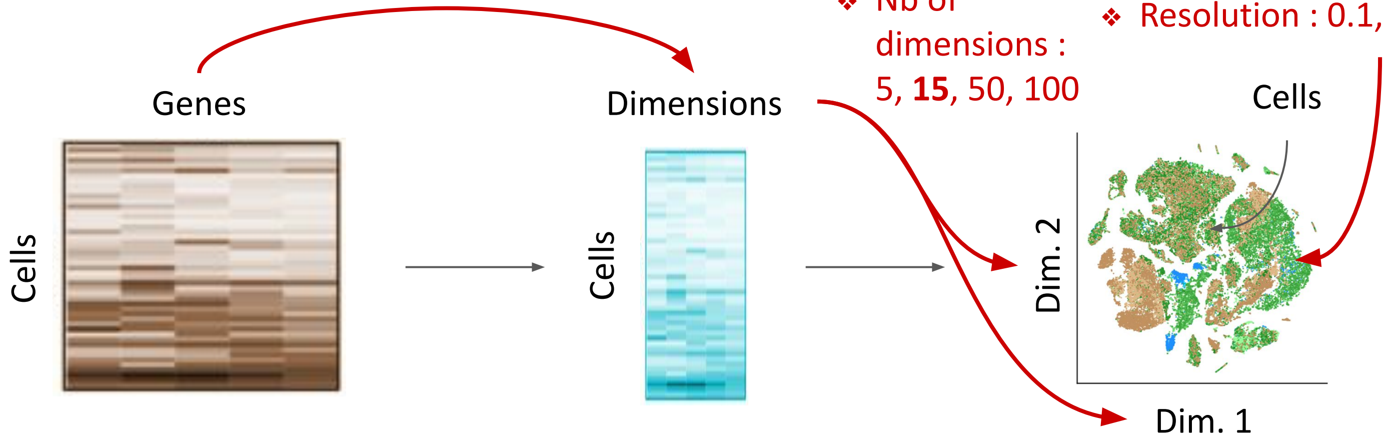
- The **number of variable genes** impact the PCA, thus the 2D space. It depends on the expected number of cell populations in the dataset.
- Number of **dimensions** = amount of information (not enough < - - > noisy data)
- **UMAP** is suited to visualize several cell types and their **global** transcriptomic profile
- **tSNE** is suited to visualize sub cell types and their **local** transcriptomic particularity
- **Diffusion Map** is suited to visualize cell **differentiation** data
- The **resolution** impacts the number of clusters : not enough clusters / not biologically interpretable clusters

Advice :

1. Make the analysis with all default settings :
   - **2000** HVG
   - **15** PC to generate a UMAP (or tSNE)
   - Resolution **1** for the clustering
2. Identify your cell populations
3. Change the settings to make the representation showing what you identified

The goal is to generate a quick representation for your cells. Run your favorite analyses and represent results on the representation. Do not make to many interpretations from the 2D representation itself.

Let’s go to practice



- Nb of variable features : 500, **2000**, 5000
- Nb of dimensions : **50**

- Nb of dimensions : 5, **15**, 50, 100

- Nb of dims : same as UMAP
- Resolution : 0.1, 0.5, **1**, 5

normalized matrix

reduced space

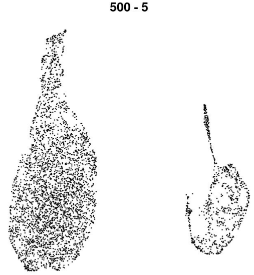
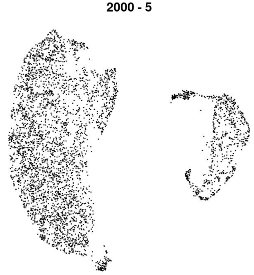
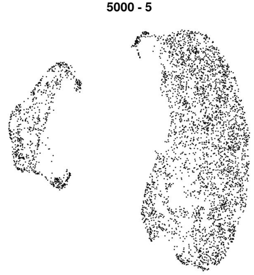
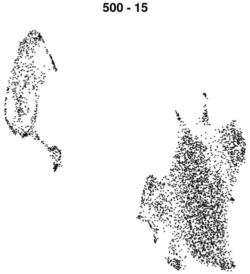
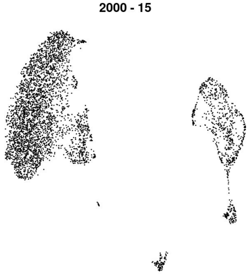
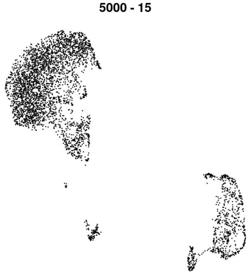
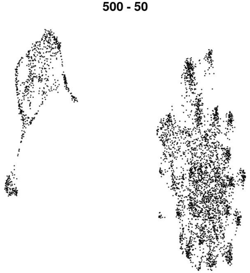
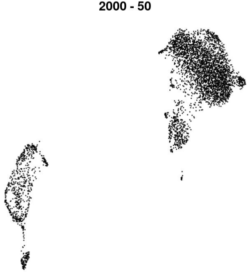
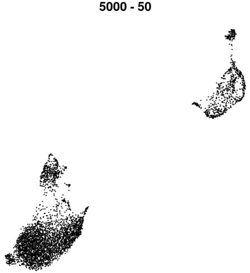
PCA

cells visualization

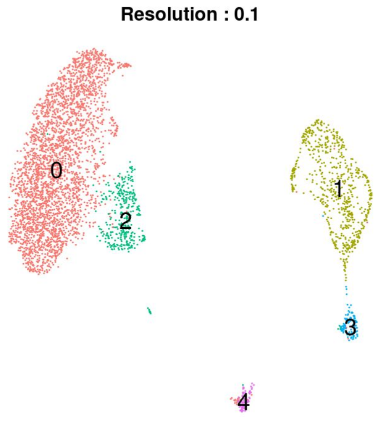
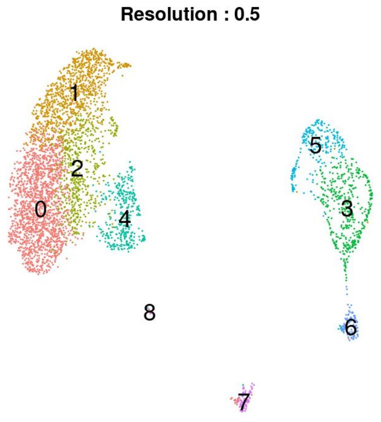
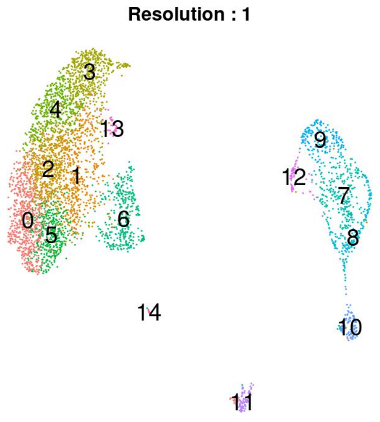
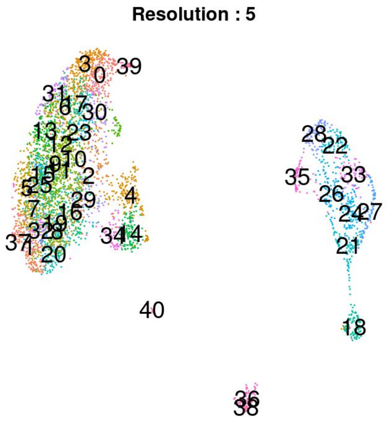
UMAP

Number of variable features

Number of PC (/50) to make the UMAP

	500	2000	5000
5	 500 - 5	 2000 - 5	 5000 - 5
15	 500 - 15	 2000 - 15	 5000 - 15
50	 500 - 50	 2000 - 50	 5000 - 50

Resolution

0.1	0.5	1	5
 Resolution : 0.1 0 1 2 3 4	 Resolution : 0.5 1 5 2 3 0 4 8 6 7	 Resolution : 1 3 4 13 9 2 1 12 7 0 6 5 8 14 10 11	 Resolution : 5 3 0 39 31 8 17 30 13 23 12 28 22 9 10 15 11 2 35 33 5 25 4 26 29 7 16 24 27 32 19 34 14 37 1 20 21 40 18 36 38