scRNA-seq : visualization

Bastien Job, Gustave Roussy, Villejuif

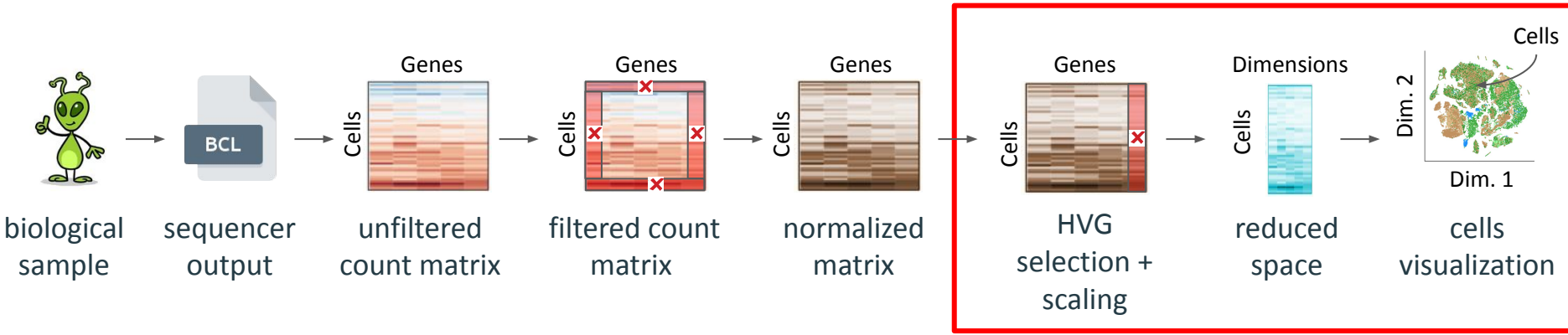
Lilia Younsi, Institut Cochin

Nathalie Lehmann, Institut Pasteur, Paris

Audrey Onfroy, Institut Mondor, Créteil

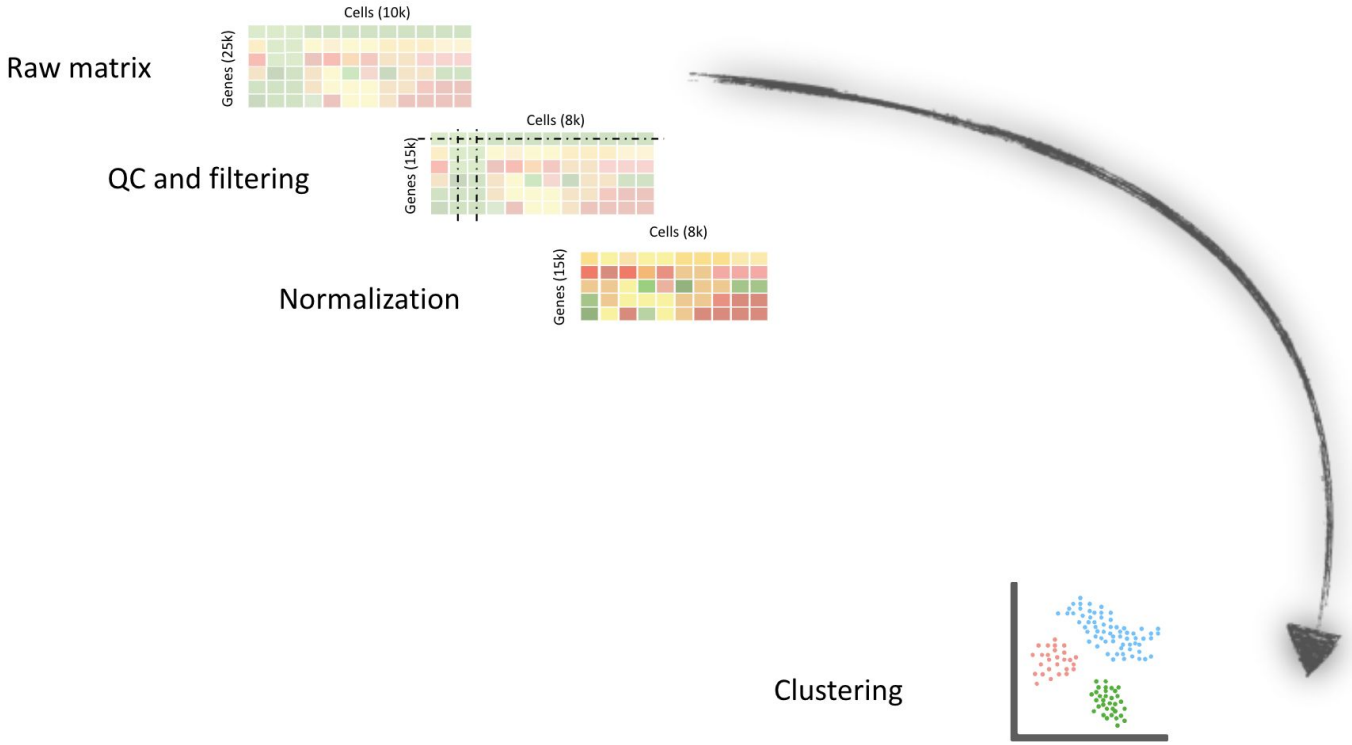
École de bioinformatique EB3I 2025

scRNA-Seq pipeline overview

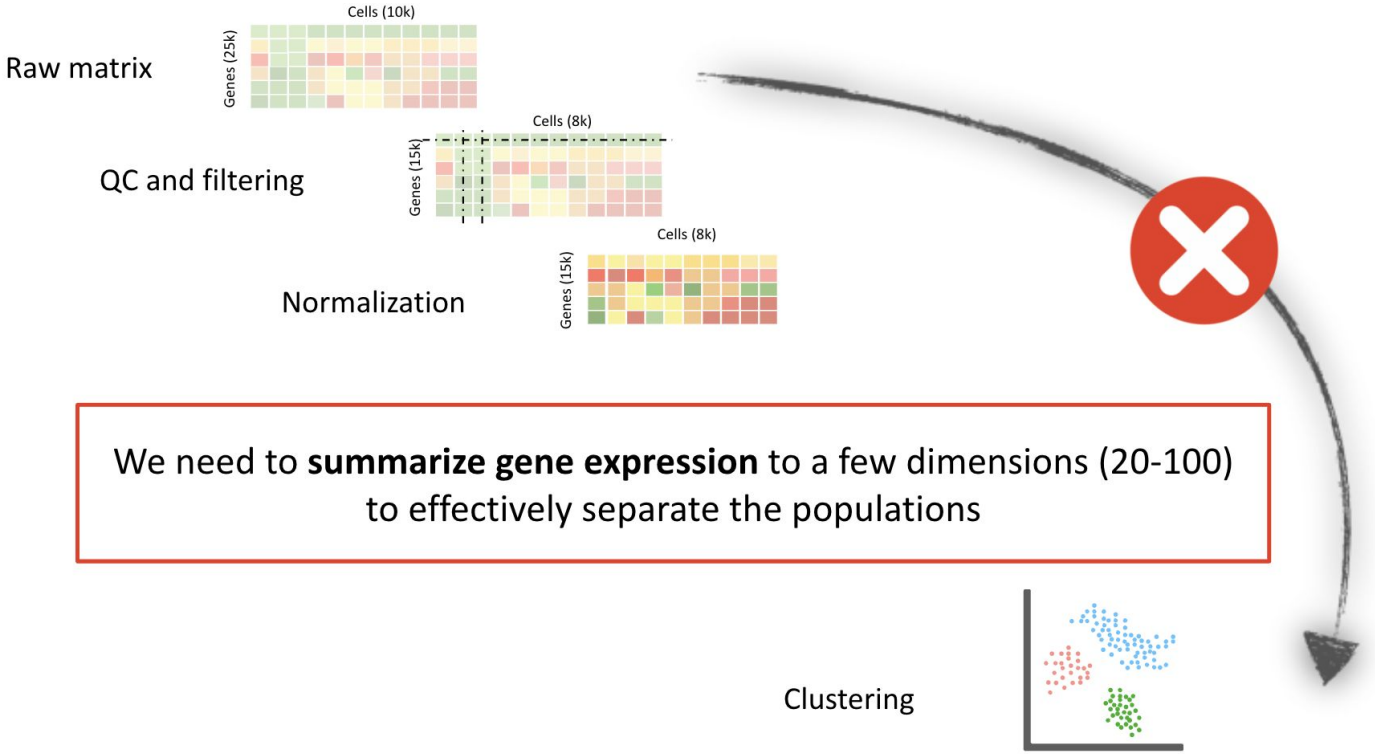


We want a visual summary of thousands cells' gene expression.

How do we get to data visualization and clustering ?

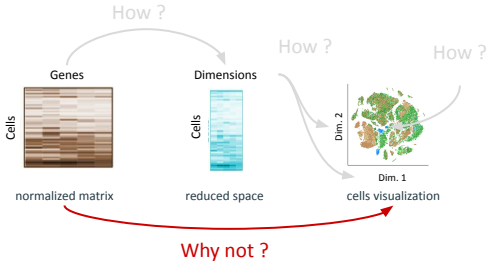


How do we get to data visualization and clustering ?



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

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

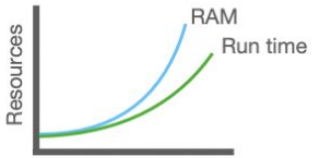
http://cmdlinetips.com/wp-content/uploads/2018/03/Sparse_Matrix.png

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

Data are noisy

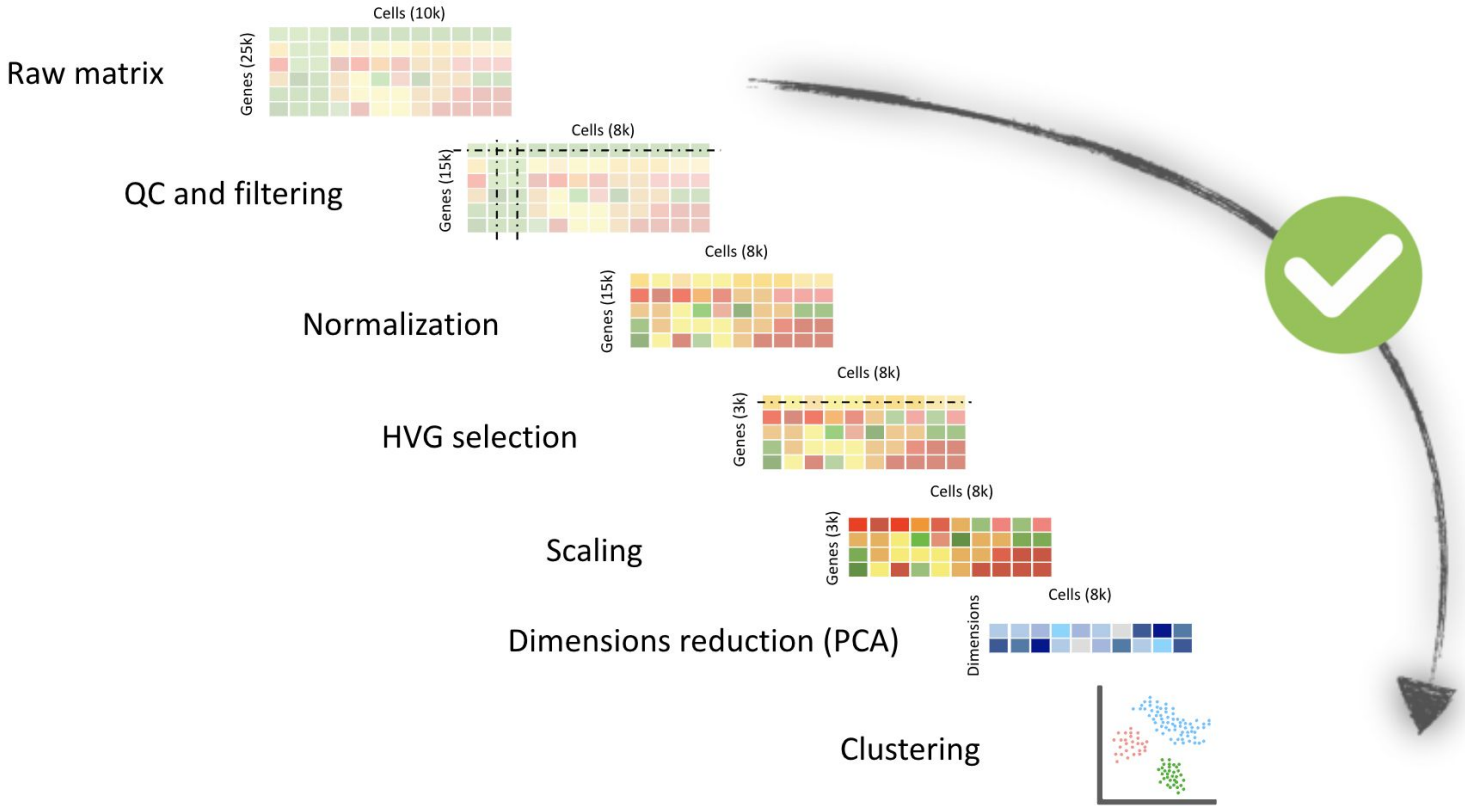
Some genes are more informative than some other.
There is **biological / technical noise** in gene expression.

Computational time and ressources

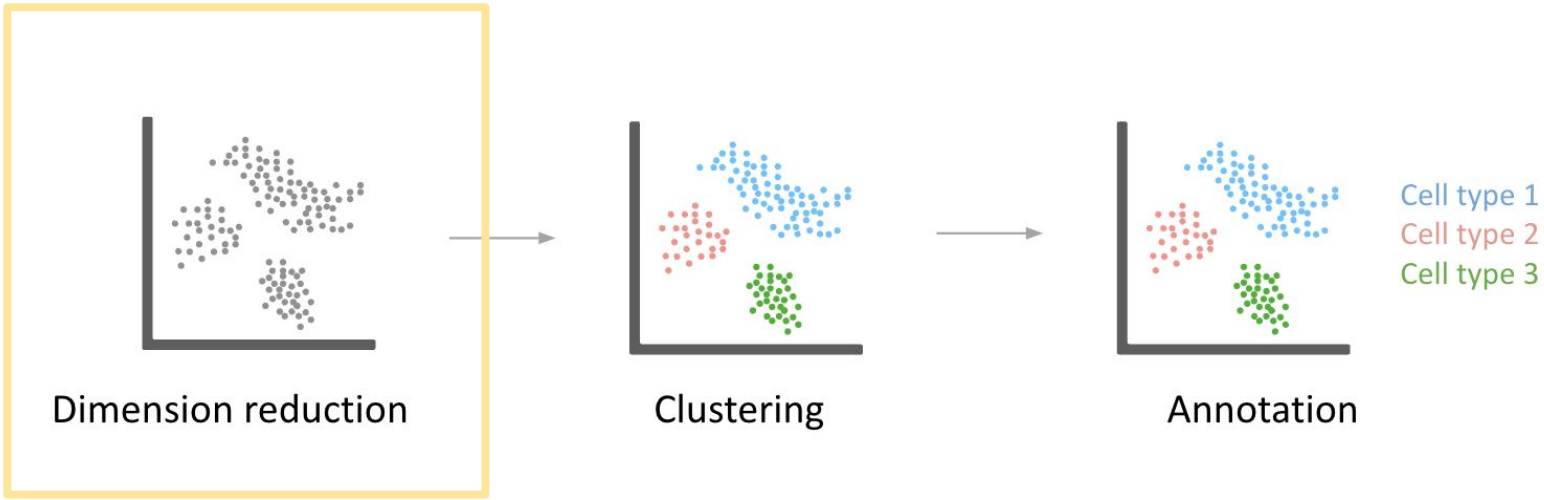


We will summarize genes expression in few dimensions, before building the 2D projection.

The right way to get to data visualization and clustering

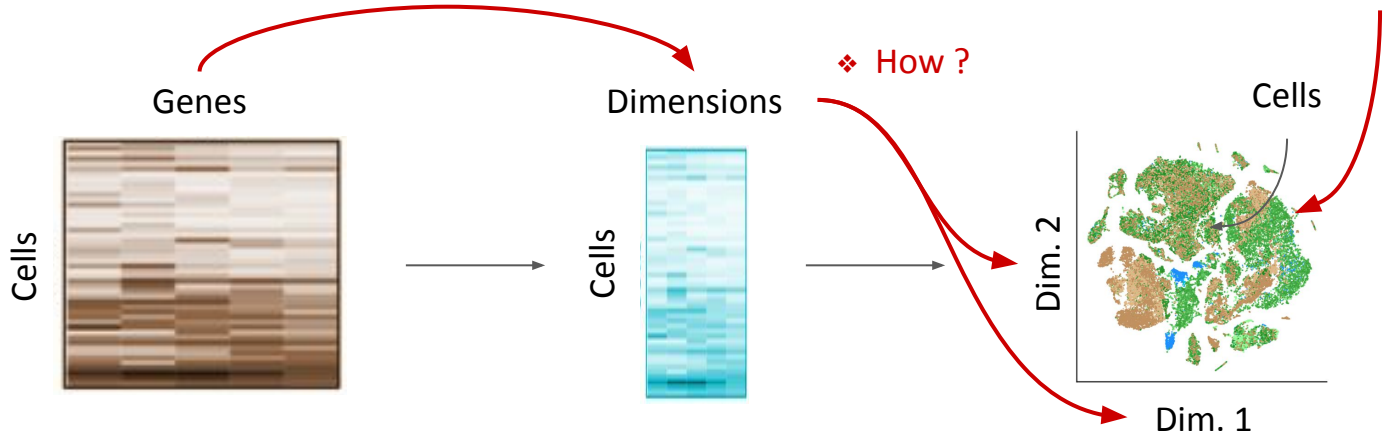


Our analyses goals



Challenges

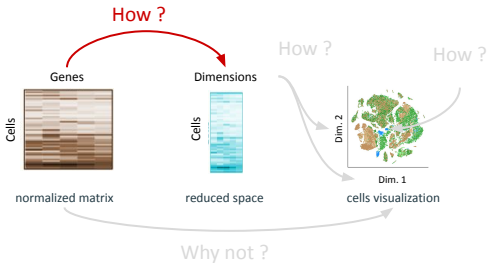
- How to reduce the number of dimensions ?
- How many dimensions ?



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

Dimensionality reduction

Overview

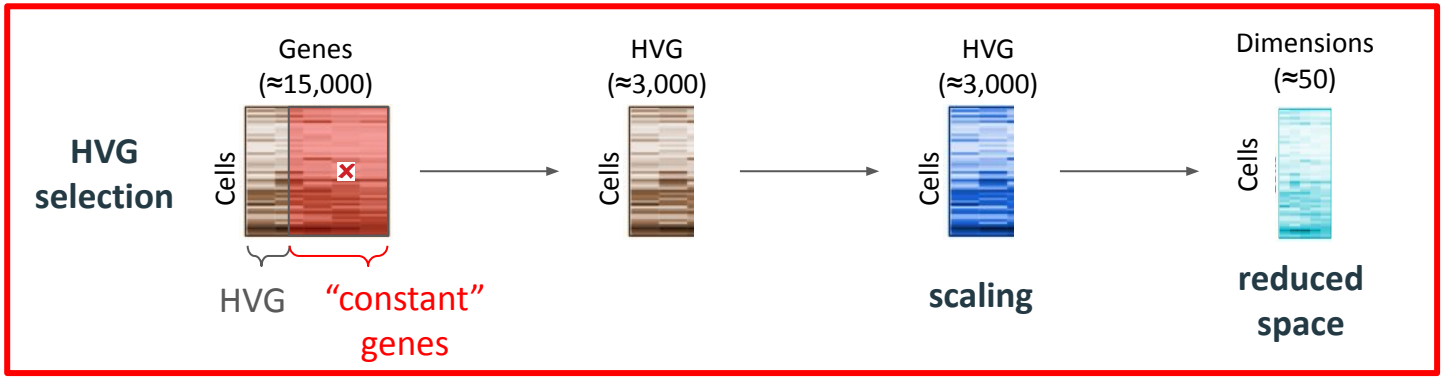


Commonly used **dimensionality reduction methods**

- **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
- …

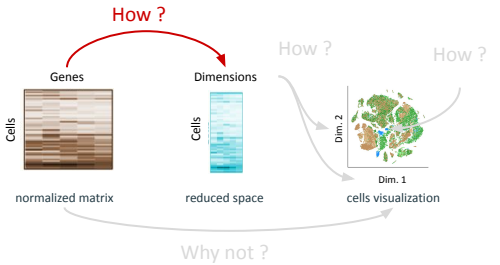
Important parameters

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



Dimensionality reduction

PCA introduction



1000 wine bottles

Features that will vary from one bottle to another :

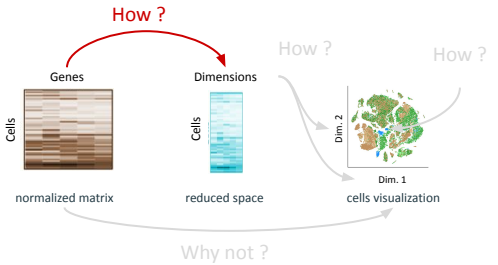
- Acidity
- Tannins
- Alcohol level
- Aroma
- Color
- Clarity
- Color intensity
- Freshness (acidity driven)
- …

Which features explain the big differences between my bottles ?

How can I sum up this data ?

Dimensionality reduction

PCA introduction



1000 wine bottles

Features that will vary from one bottle to another :

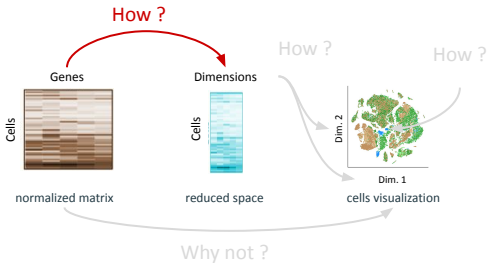
- Acidity
- Tannins
- Alcohol level
- Aroma
- Color
- Clarity
- Color intensity
- Freshness (acidity driven)
- …

TOO MUCH INFO !

Impossible to compare all these variables 1 to 1 for all bottles without getting lost.

Dimensionality reduction

PCA introduction



1000 wine bottles

Features that will vary from one bottle to another :

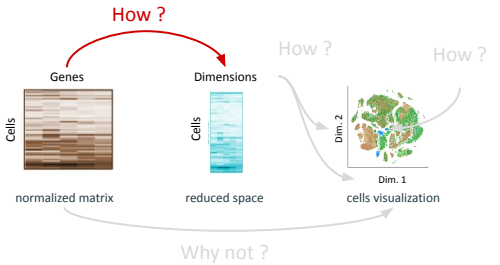
- **Acidity**
- Tannins
- Alcohol level
- Aroma
- Color
- Clarity
- Color intensity
- **Freshness** (acidity driven)
- …

REDUNDANT INFORMATION

Acidity ⇔ Freshness

Dimensionality reduction

PCA introduction



1000 wine bottles

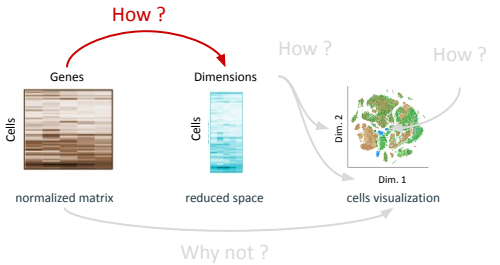
Features that will vary from one bottle to another :

- Acidity
- Tannins
- Alcohol level
- Aroma
- **Color**
- **Clarity**
- **Color intensity**
- Freshness (acidity driven)
- …

→ Features can be combined in one dimension : “**robe du vin**” (or *wine apprerance*).

Dimensionality reduction

PCA introduction



1000 wine bottles

Features

- Acidity
- Tannins
- Alcohol level
- Aroma
- **Color**
- **Clarity**
- **Color intensity**
- Freshness (acidity driven)
- …

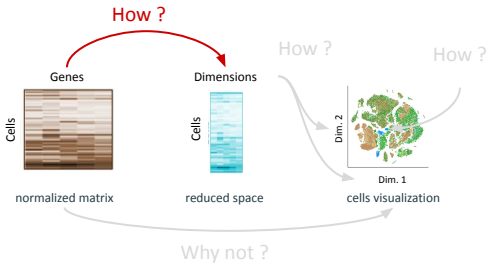
Features can be combined in one dimension : “**robe du vin**” (or *wine apprerance*).

⇔

Principal Componant (PC)

Dimensionality reduction

PCA introduction



1000 wine bottles

Features

- Acidity
- Tannins
- Alcohol level
- Aroma
- **Color**
- **Clarity**
- **Color intensity**
- Freshness (acidity driven)
- …

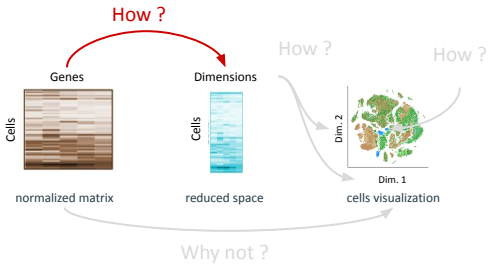
Features can be combined in one dimension : “**robe du vin**” (or *wine apprerance*).

⇔

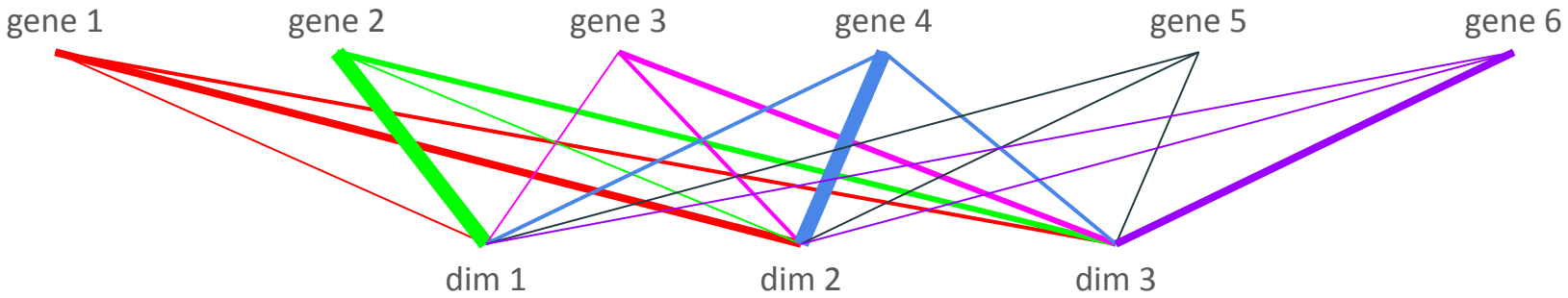
Principal Componant (PC)

Dimensionality reduction

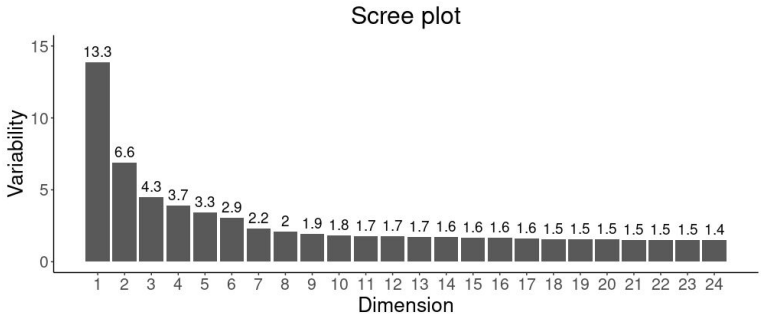
Principal Component Analysis - principle



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

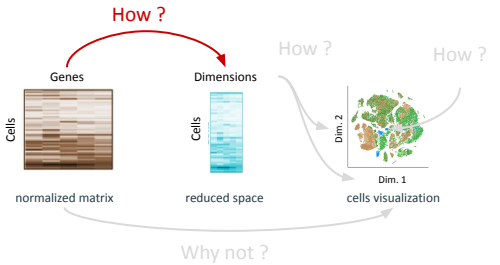


- Output : **Z** (≈ 50 - 100) dimensions “Principal Component”
- Each PC summarizes a certain amount of the input data variability
  - First PC recapitulates the most part of information
  - Last PC can be considered as noise

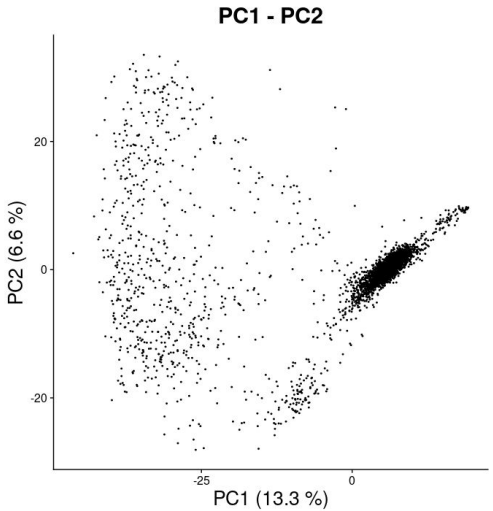


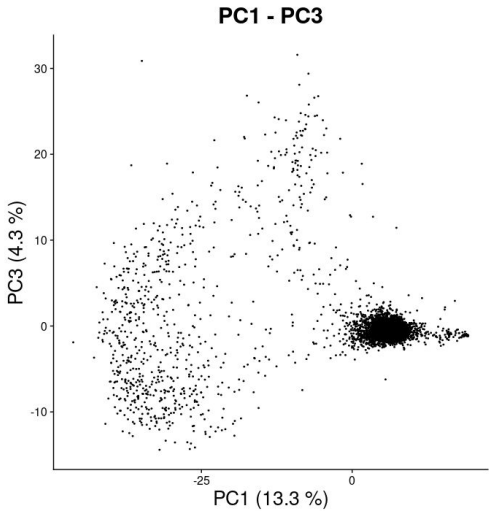
Dimensionality reduction

Principal Component Analysis - visualization



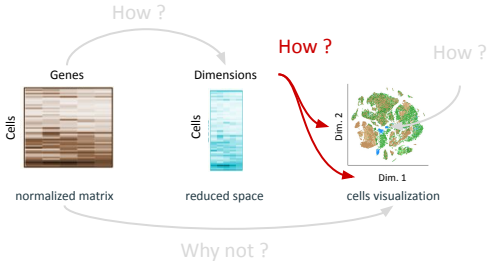
- Input : **X** most variable genes
- Goal : Group genes by dimensions when they have similar expression across cells
- Output : **Z** dimensions “Principal Component”
- Each PC summarizes a certain amount of the input data variability





Now, we will use this reduced space to build a 2D graphical representation.

2D space for cells visualization

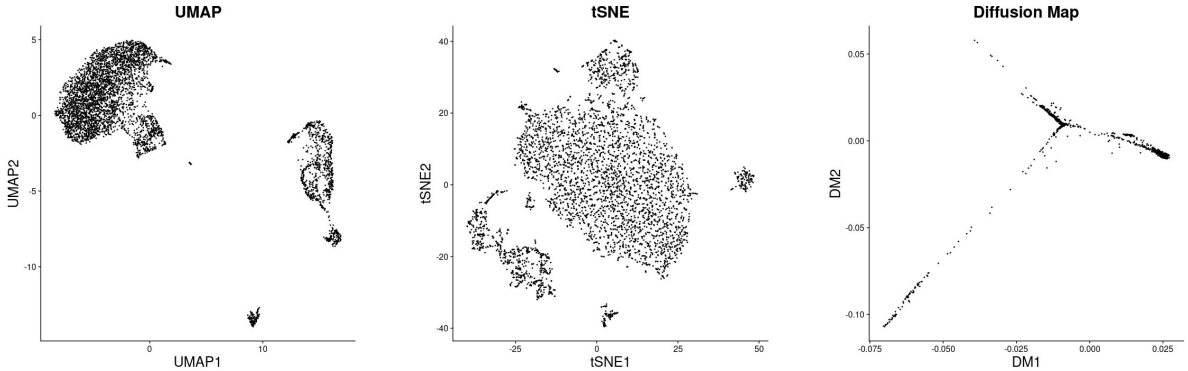


Commonly used 2D spaces

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

Important parameters

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

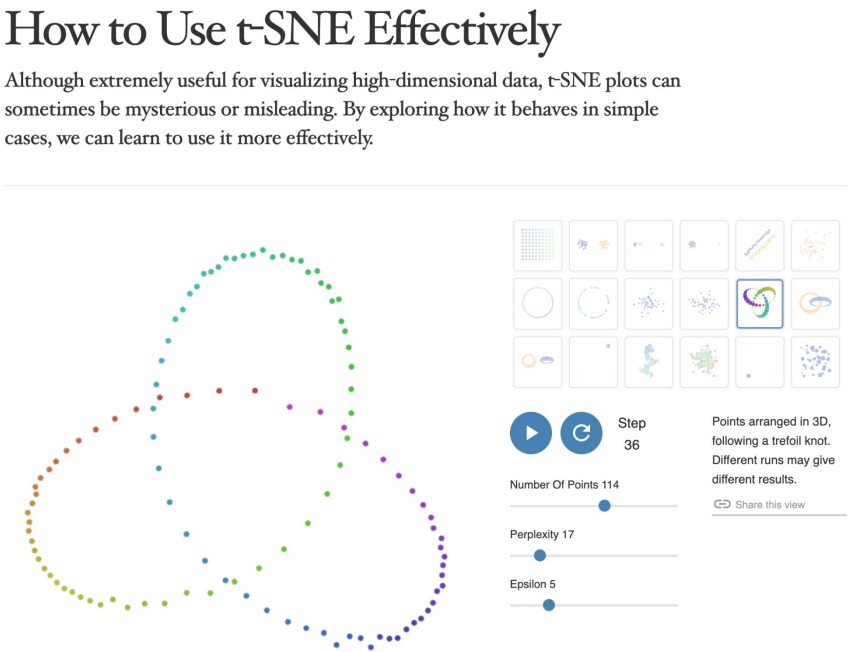


The same cells can be represented using **different 2D spaces**.

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

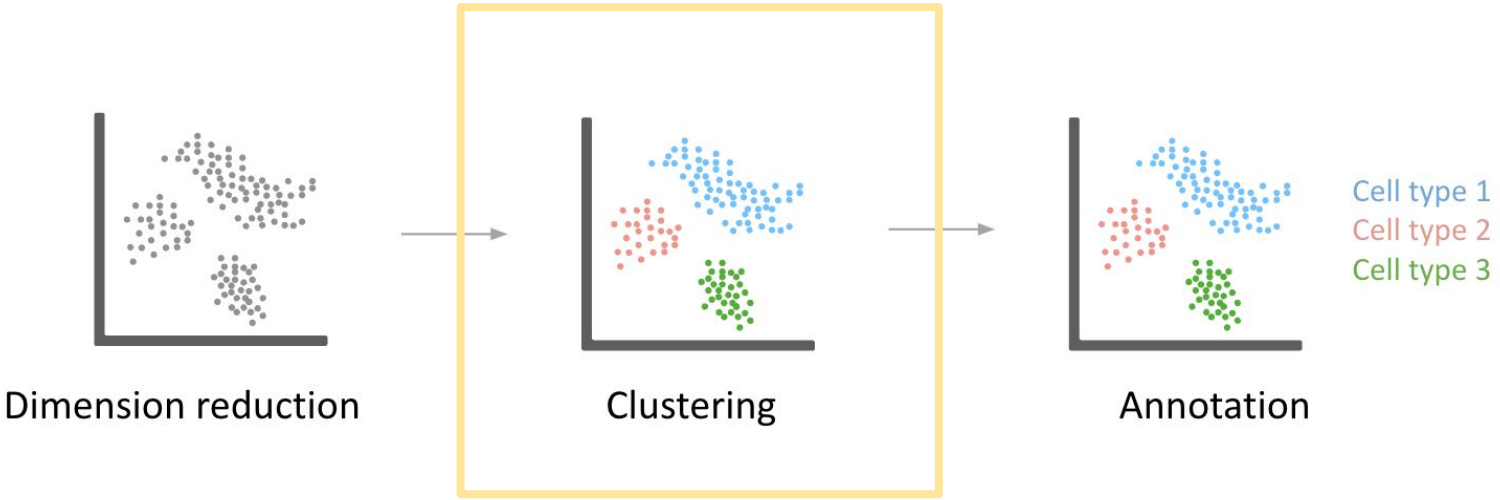
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6417818/

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

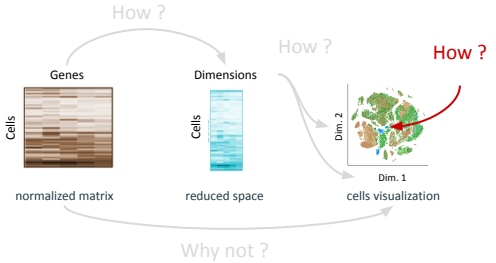


https://distill.pub/2016/misread-tsne/

Our analyses goals



Clustering

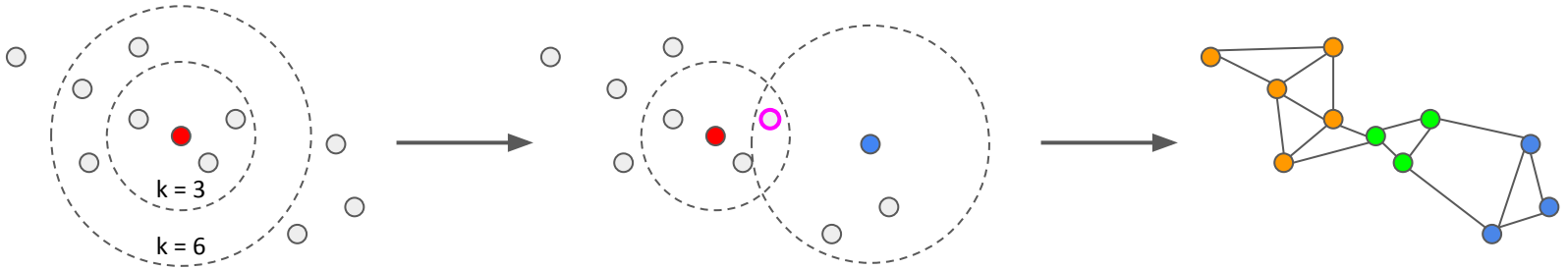


Commonly used methods

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

Important parameters

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



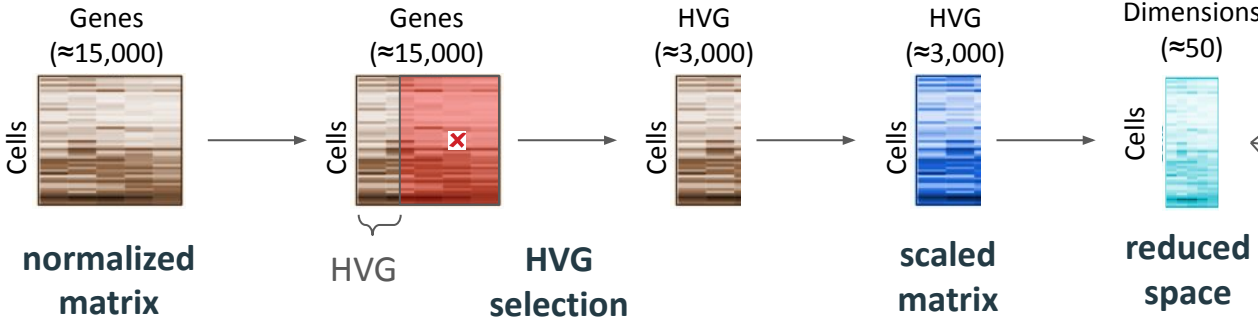
k-nearest neighbors
(kNN)

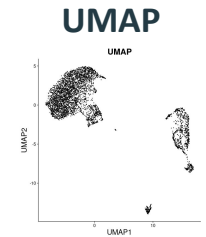
shared nearest neighbors
(SNN)

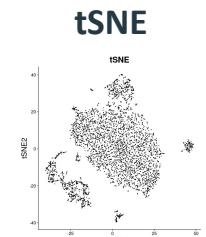
clustering
(from SNN graph)

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

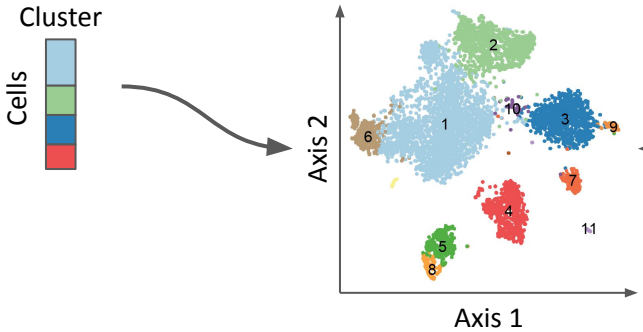
Summary


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
UMAP
UMAP2
UMAP1


tSNE
tSNE
tSNE2
tSNE1


others…
Diffusion Map
DM2
DM1


Cluster
Cells
Axis 2
Axis 1
1
2
3
4
5
6
7
8
9
10
11

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

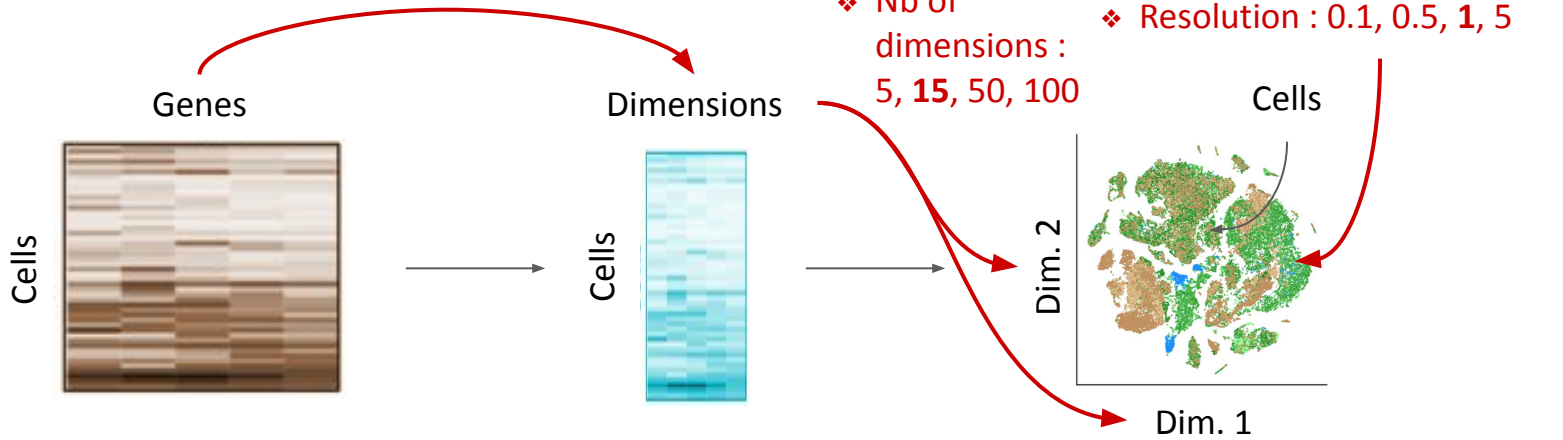
The goal is to generate a quick representation for your cells. Run your favorite analyses and represent results on the representation. Do not make to many interpretations from the 2D representation itself.

Let’s go to practice

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

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

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



normalized matrix

reduced space

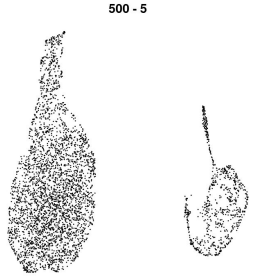
PCA

cells visualization

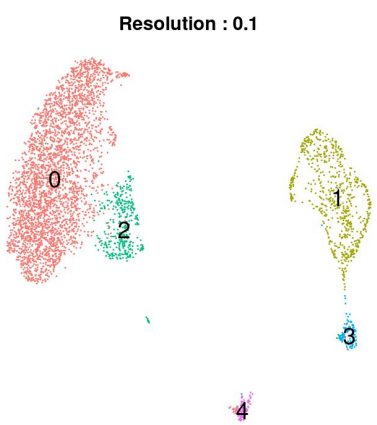
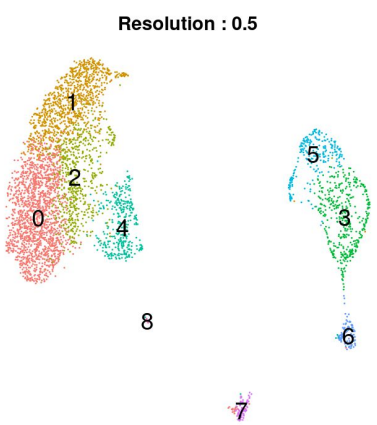
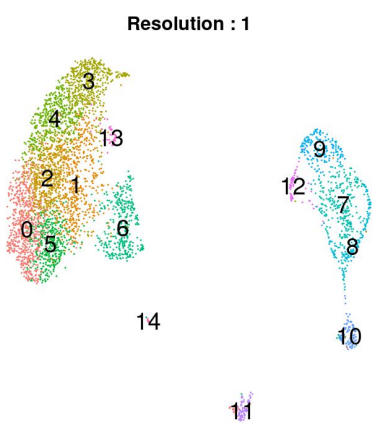
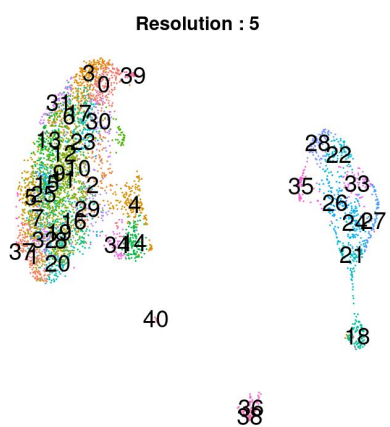
UMAP

Number of variable features

Number of PC (/50) to make the UMAP

	500	2000	5000
5			
15			
50			

Resolution

0.1	0.5	1	5
			

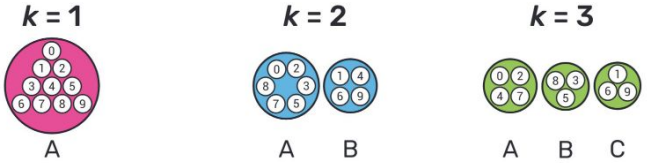
Choosing resolution wisely

- Too low resolution → losing information of populations
- Too high resolution → ***overclustering***

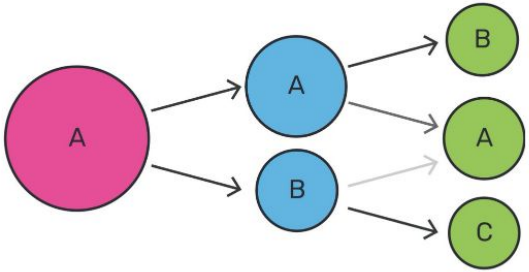
→ **Clustering trees** can be used to help choose the optimal resolution !

Step 1 - Clustering at multiple resolutions

```
res <- c(0.1,1.2,0.1)
```



Step 2 - Building and visualizing the clustering tree



Example :