Data analytics for personalized genomics and precision medicine

Lecturer: Yu LI(CSE)

Email: liyu@cse.cuhk.edu.hk

Lecture 18: Visualization and Protein-RNA/DNA

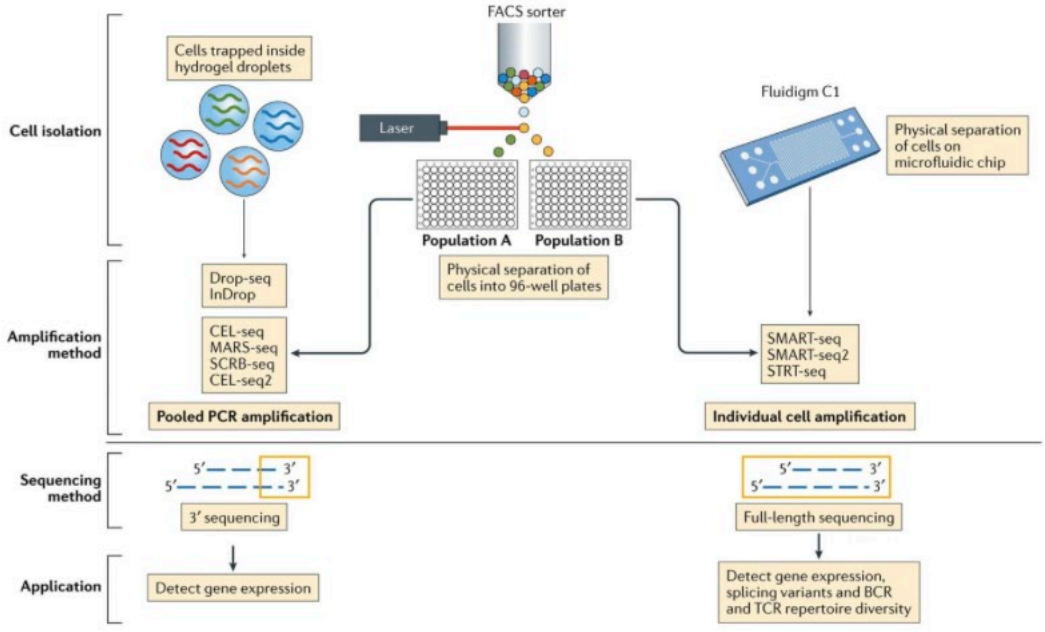
Friday, November 7, 2025

What is single-cell analysis?

*Single cell sequencing -> sequence information from individual cell with optimized NGS technologies

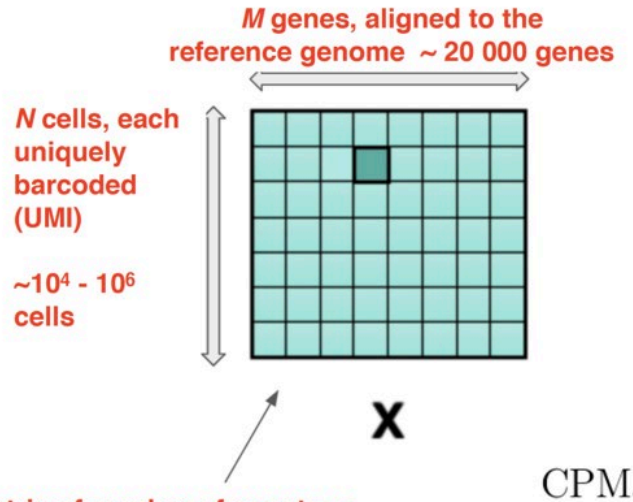
Result : provide a high resolution of cellular differences

How to do the single-cell sequencing



- From cell isolation to application

Gene expression matrix



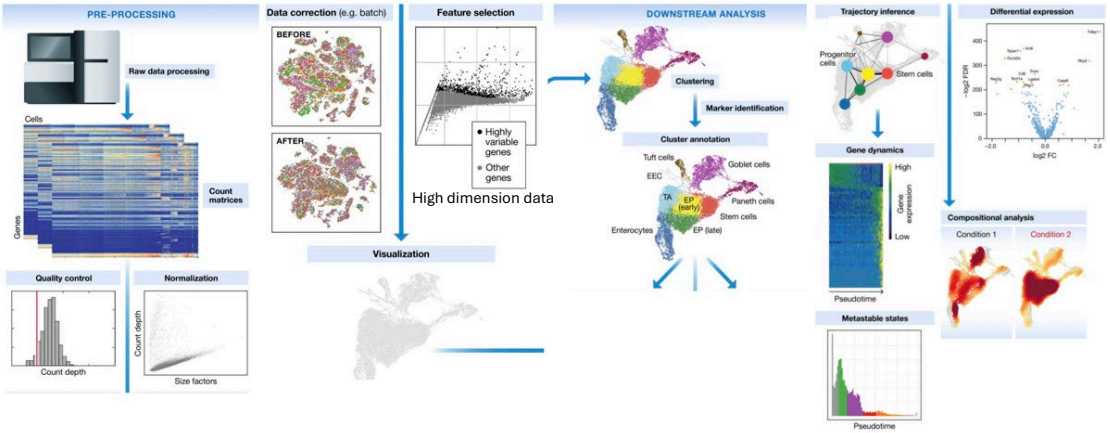
$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

- N means total number

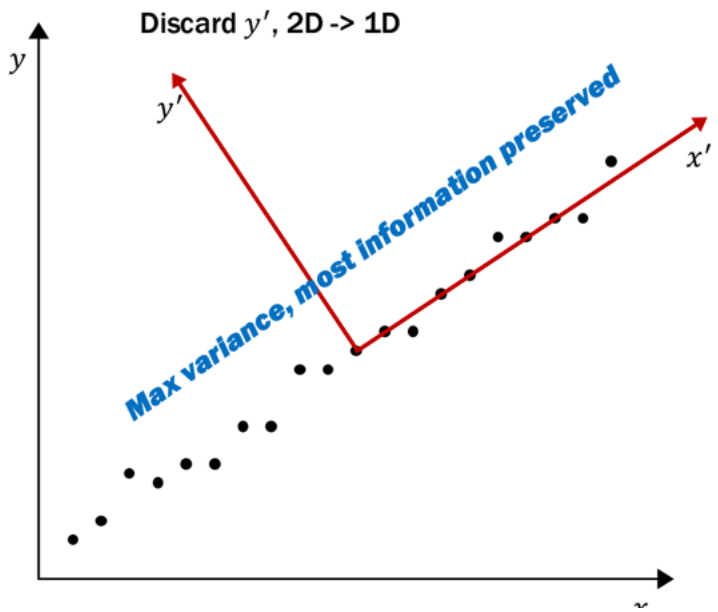
Challenges in single-cell data analytics

- Noise
- Doublet
- Dropout
- Batch effect (non-biological signal)

Single-cell RNA-seq analysis



Dimension reduction-- PCA



* Data to the direction with highest variance -> we will preserve (the 2 higher value)

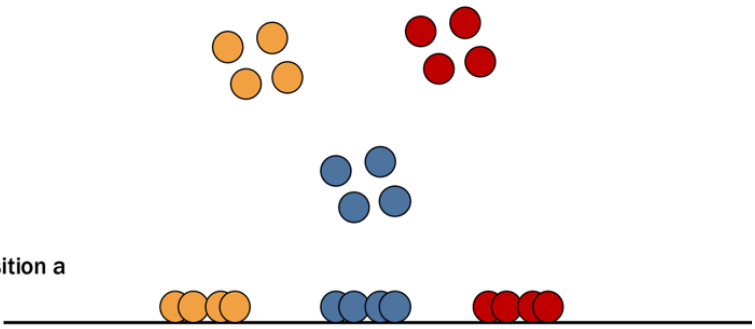
Problem of PCA

- original clusters are not preserved

**use T-SNE (t-distributed stochastic neighbor embedding)

- A nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions
- Similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability

Process of t-SNE



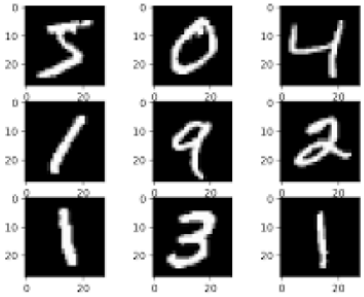
1. Random initialization
2. For each point, update the position a little bit
3. ...
4. Until no more update

Compare the cluster to the original cluster. The points from the same cluster attract each other. The points from different clusters push apart each other In low-dimension space

- x reverse

PCA vs t-SNE

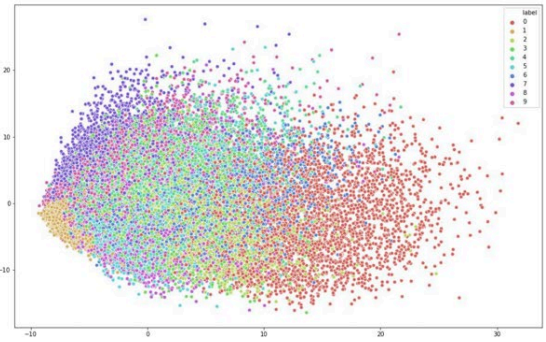
From 0-9



28*28=784D

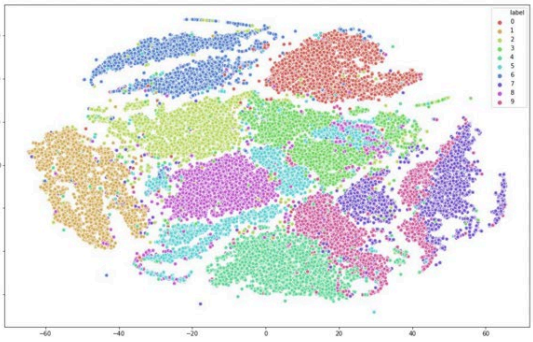
- 28*28 means each image dimension

x,y -axis coordination has meaning



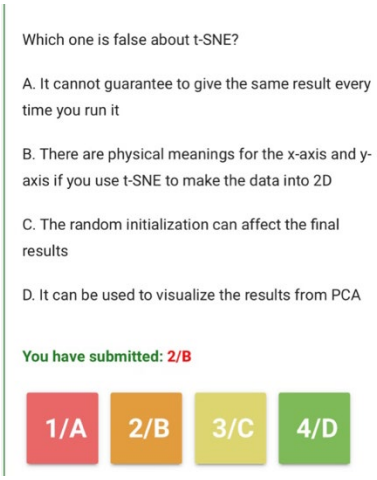
The first 2D from PCA

No meaning of their coordination



t-SNE

- t-SNE much better than PCA in visualization



T-SNE

- No physical meaning
- No physical distance
- Just realize data

Disadvantages of t-SNE

❖Iterative: longer running time

❖Non-deterministic: different runs may have different results

❖Noisy patterns

❖The original distance is not precisely preserved

❖UMAP could be an alternative

- For UMAP is similar with t-SNE

T-SNE/UMAP in Python

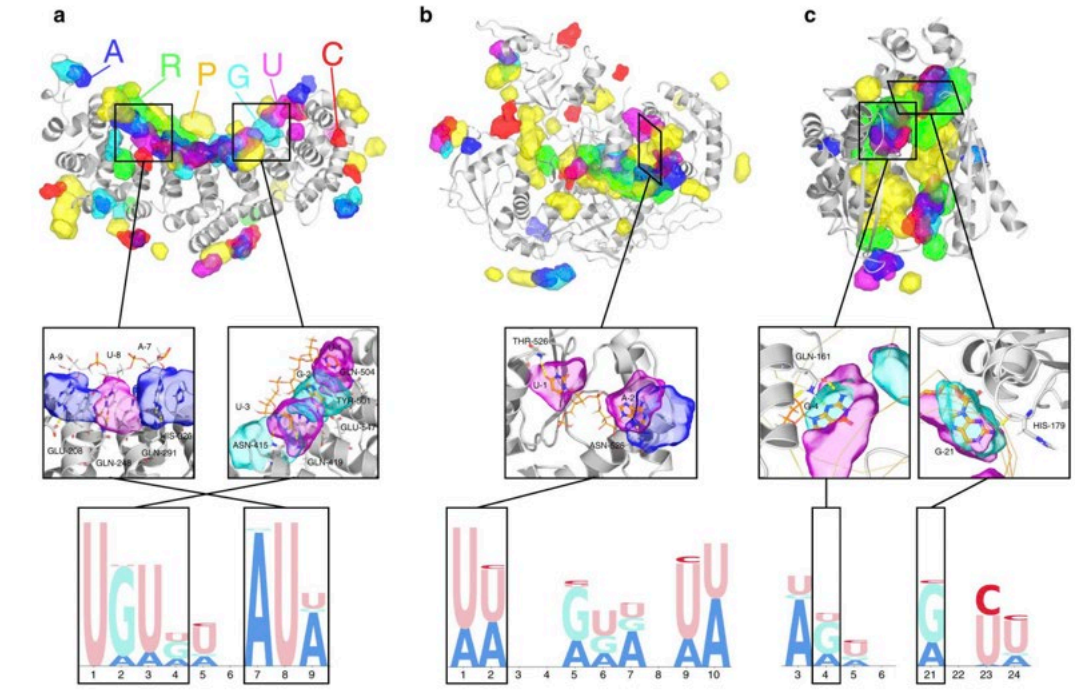
Examples

```
>>> import numpy as np
>>> from sklearn.manifold import TSNE
>>> X = np.array([[0, 0, 0], [0, 1, 1], [1, 0, 1], [1, 1, 1]])
>>> X_embedded = TSNE(n_components=2, learning_rate='auto',
...                   init='random').fit_transform(X)
>>> X_embedded.shape
(4, 2)
```

- n_components=2 means 2D place

- (4,2) = (no. of different data point, no. of dimension)

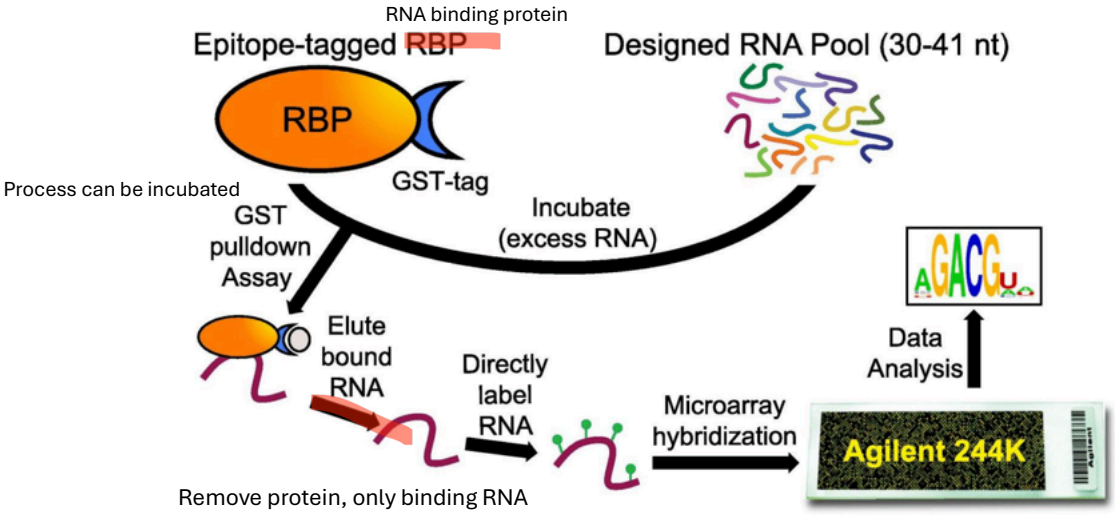
Protein binding has preference



RNA binding protein sequence

- Different binding protein -> different binding pattern

Method of get the binding motif by experiments



From aligned sequences to motif

❖Notice that the sequences should be aligned before converting into motif

Table 1: Starting sequences.

#	Sequence
1	AAGAAT
2	ATCATA
3	AAGTAA
4	AACAAA
5	ATTAAA
6	AAGAAT

Table 2: Position Count Matrix.

Position	1	2	3	4	5	6
A	6	4	0	5	5	4
C	0	0	2	0	0	0
G	0	0	3	0	0	0
T	0	2	1	1	1	2

- Aligned means the first thing do alignment
- If don't aligned -> will have different sequence -> doesn't match

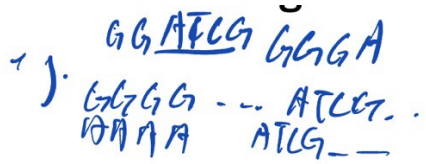


Table 2: Position Count Matrix.

Position	1	2	3	4	5	6
A	6	4	0	5	5	4
C	0	0	2	0	0	0
G	0	0	3	0	0	0
T	0	2	1	1	1	2

Table 3: Position Probability Matrix.

Position	1	2	3	4	5	6
A	1.00	0.67	0.00	0.83	0.83	0.66
C	0.00	0.00	0.33	0.00	0.00	0.00
G	0.00	0.00	0.50	0.00	0.00	0.00
T	0.00	0.33	0.17	0.17	0.17	0.33

Figure 1: Sequence logo of a Position Probability Matrix