Lecture 18: Visualization & Protein-RNA/DNA

November 7, 2025 (Fri)

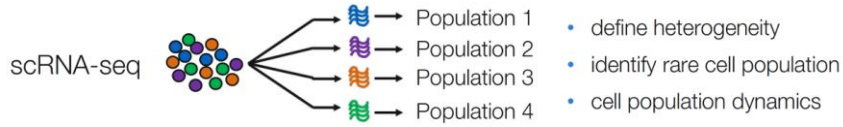
Lecturer: Prof. Li Yu Scribe: Chen Yujia

Recap from last lecture

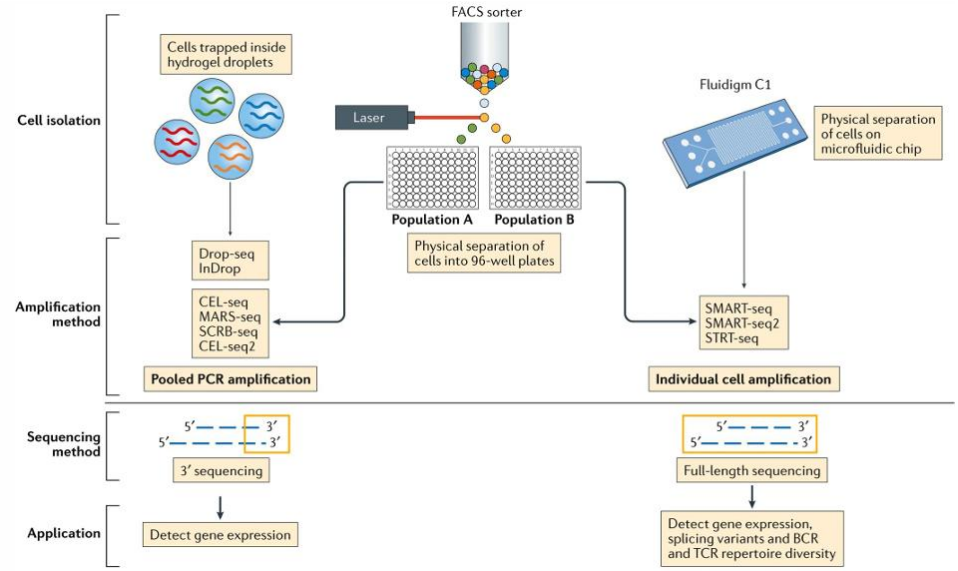
Single-cell analysis

- What is single-cell analysis?

  Single cell sequencing examines the sequence information from individual cells with optimized next-generation sequencing (NGS) technologies, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment.

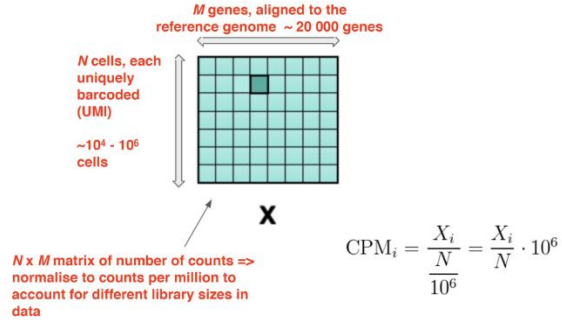


- How to do single-cell sequencing?



- The gene expression matrix

  It is expected to have a very sparse gene expression matrix and thus counts per million is used for different library sizes in data.



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

- Challenges in single-cell data analysis
  - Noise
  - Doublet
  - Dropout
  - Batch effect

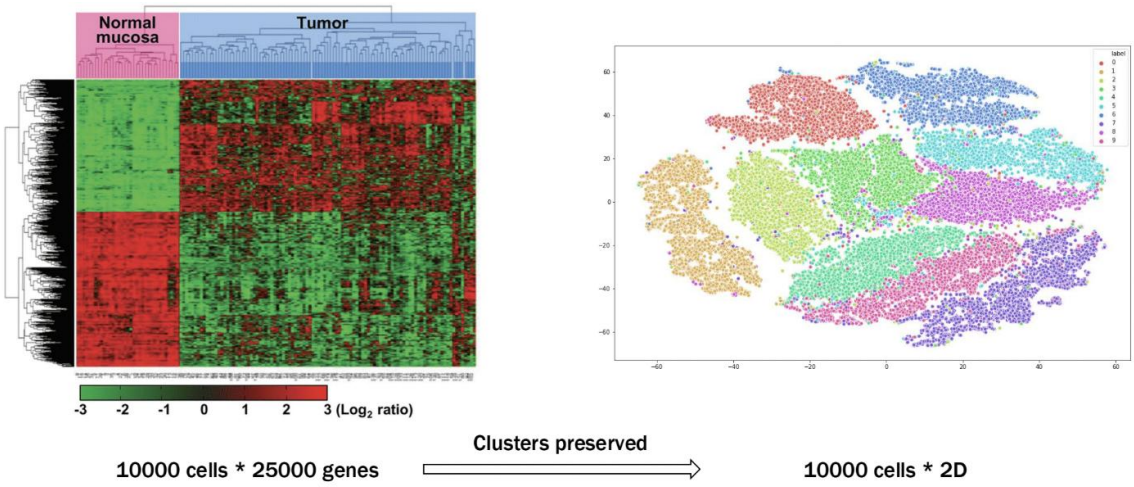
Outline

- Visualization: T-SNE
- Protein-RNA/DNA interaction: motif analysis

Visualization

- Visualize gene expression data in 2D

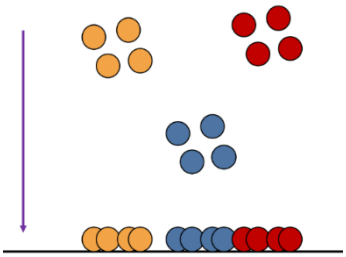
  Visualization is an important step for data analysis.



  - To visualize the data in 2D by dimension reduction using PCA?

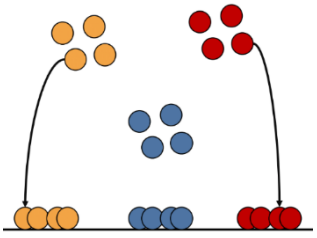
    PCA: By projecting the data to the direction with the highest variance, we preserve as much information as possible.

    Problem of PCA: The original clusters are <u>not</u> preserved. For example, in the picture on the right-hand side, suppose we are mapping the 2D data into a 1D space. After PCA and being mapped to the 1D space, the blue and red groups are close to each other, and it might be considered that only two clusters are there (yellow and blue & red). This could be more problematic in a higher space.



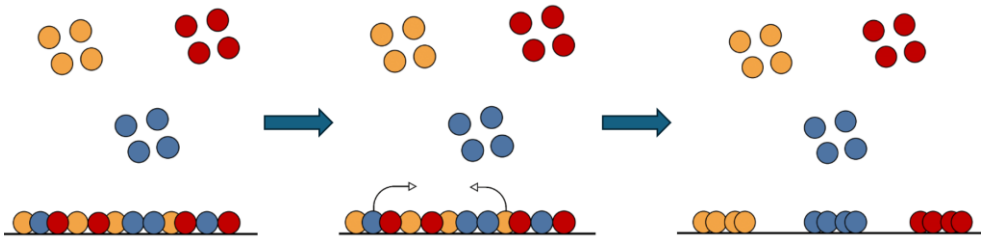
  - To preserve the clusters, something more complex than project is required.

    T-SNE: T-distributed stochastic neighbor embedding. It is a <u>nonlinear</u> dimensionality reduction technique well-suited for embedding <u>high-dimensional</u> data for visualization in a <u>low-dimensional</u> space of two or

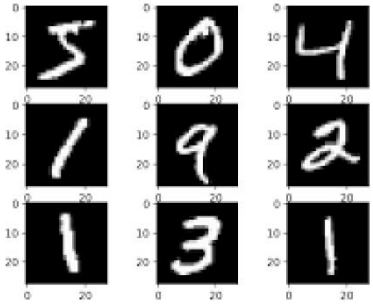


three dimensions. Through this iterative process, similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.

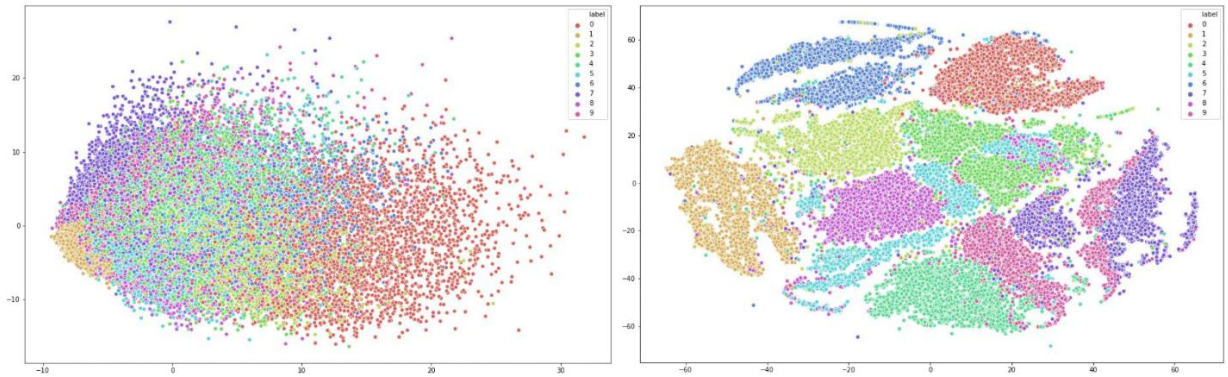
- The process of T-SNE
  - Random initialization
  - For each point, update the position a little bit: Compare the clusters to the original cluster, the points from the same cluster will attract each other, while the points from different clusters will push apart each other.
  - Repeat the second step until no more update.



- PCA vs T-SNE



Each data point has 28 * 28 = 784 dimensions.



The first 2D from PCA **t-SNE**

- After PCA, the data points from different clusters are mixed together, but through T-SNE, the original clusters are preserved.
- The x and y axes in the PCA result do have physical meanings, while that in the T-SNE result have no physical meanings and there is no direct relationship between the data points and the original data.

- Disadvantages of T-SNE

- Iterative: longer running time
- Non-deterministic: different runs may have different results
- Noisy pattern
- The original distance is not precisely preserved as T-SNE is only a way for visualization.
- UMAP could be an alternative (similar to T-SNE, only different in implementation)

➢ T-SNE/UMAP in Python

T-SNE: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

UMAP: https://umap-learn.readthedocs.io/en/latest/

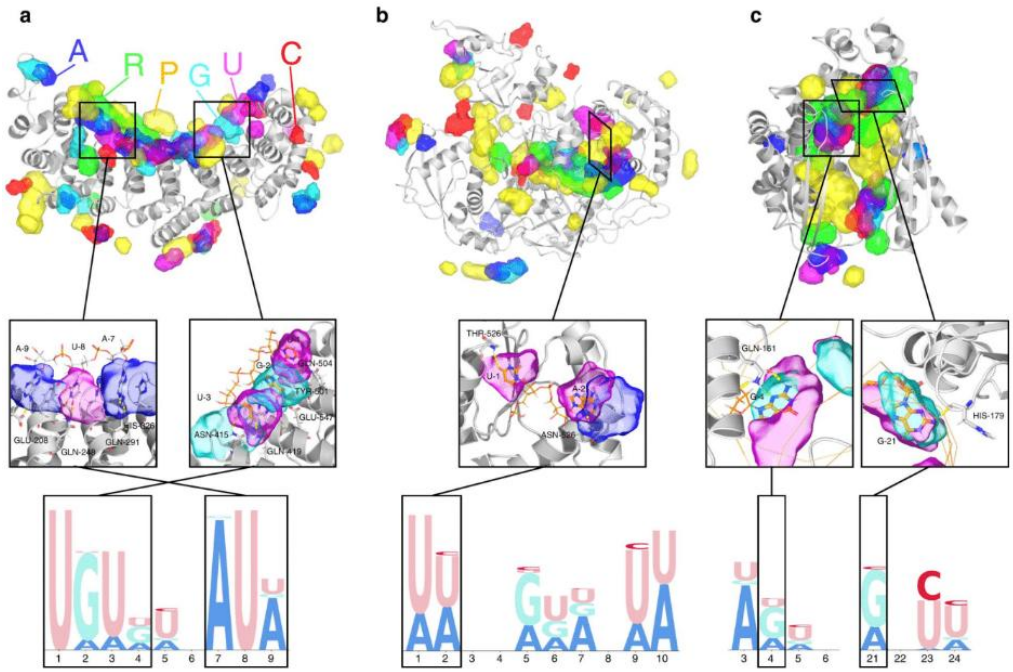
```
>>> import numpy as np
>>> from sklearn.manifold import TSNE
>>> X = np.array([[0, 0, 0], [0, 1, 1], [1, 0, 1], [1, 1, 1]])
>>> X_embedded = TSNE(n_components=2, learning_rate='auto',
...                   init='random').fit_transform(X)
>>> X_embedded.shape
(4, 2)
```

Protein-RNA/DNA interaction

➢ Protein binding has preference

The picture shows the RNA binding proteins (RBPs) and the patterns of their binding preferences known as motifs.

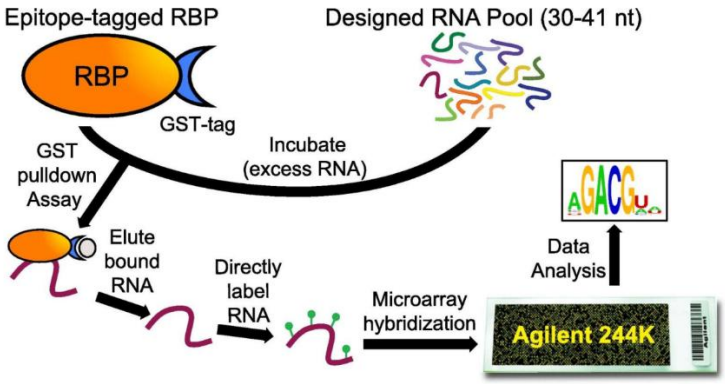
Motif is a kind of sequence pattern of DNA, RNA, or protein that is repetitive and appears frequently in certain conditions.



➢ How to get the binding motif by experiments?

The Epitope-tagged RBP with a GST-tag is incubated with excess RNA form a designed RNA pool which may contain millions of RNA fragments. Subsequently, GST pulldown assay is applied to obtain

the RBP and RBP-bound RNA. The RNA will undergo microarray hybridization or sequencing. After data analysis and identification of the binding motif, more similar sequences can be generated and put into the RNA pool. The experiment can be done iteratively to make the pattern clear.



- From aligned sequences to motif

Notice that the sequences should be aligned before converting into motif, or it will be most likely to get a uniform distribution of 0.25 for each of A, T, C, G in each position.

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

An example of unaligned sequences is shown below:

AAA**ATCG**AAAAA
GGGG**ATCG**GGGG
CCCCC**ATCG**CCC
TTTTTT**ATCG**TT

In this example, although the motif of ATCG exists in all sequences, each of A, T, C, G has a probability of 0.25 in each position, and the motif can't be observed in this way.