BMEG3105

Lec 18 Visualization and Protein - RNA/DNA

7th November, 2025

[Recap]

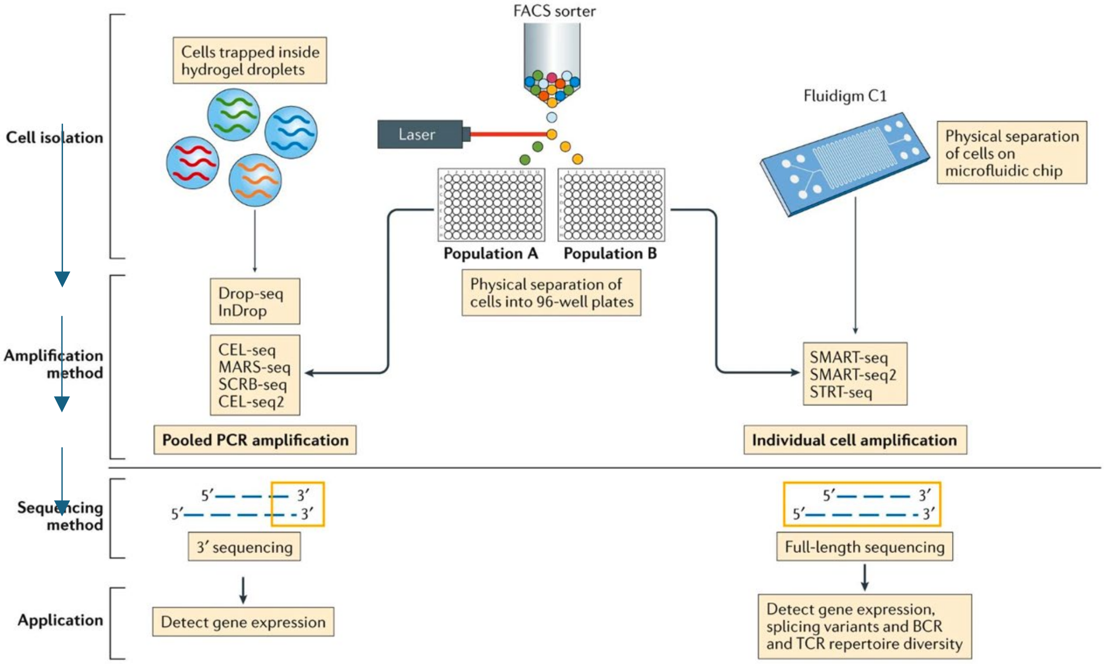
Single-cell analysis

1. Examines the sequence information from individual using optimized next-generation sequencing (NGS) technologies
2. Provide higher resolution of cellular differences & better understanding of the function of an individual cell in the context of its microenvironment

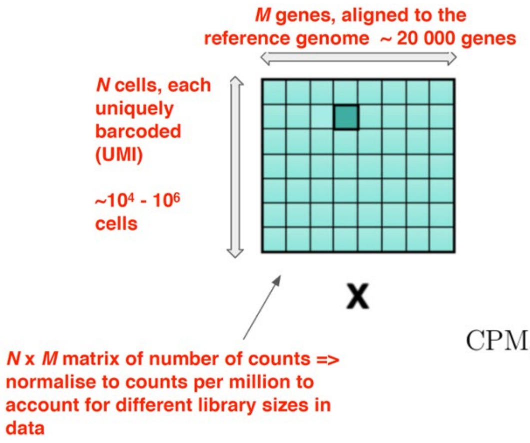
√ Define heterogeneity

√ Identify rare cell population

√ Tell cell population dynamics



Gene expression matrix


M genes, aligned to the reference genome ~ 20 000 genes
N cells, each uniquely barcoded (UMI)
~10^4 - 10^6 cells
X
N x *M* matrix of number of counts => normalise to counts per million to account 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 analysis

1. Nosie
2. Doublet
3. Dropout
4. Batch effect (= non-biological effect)

[New]

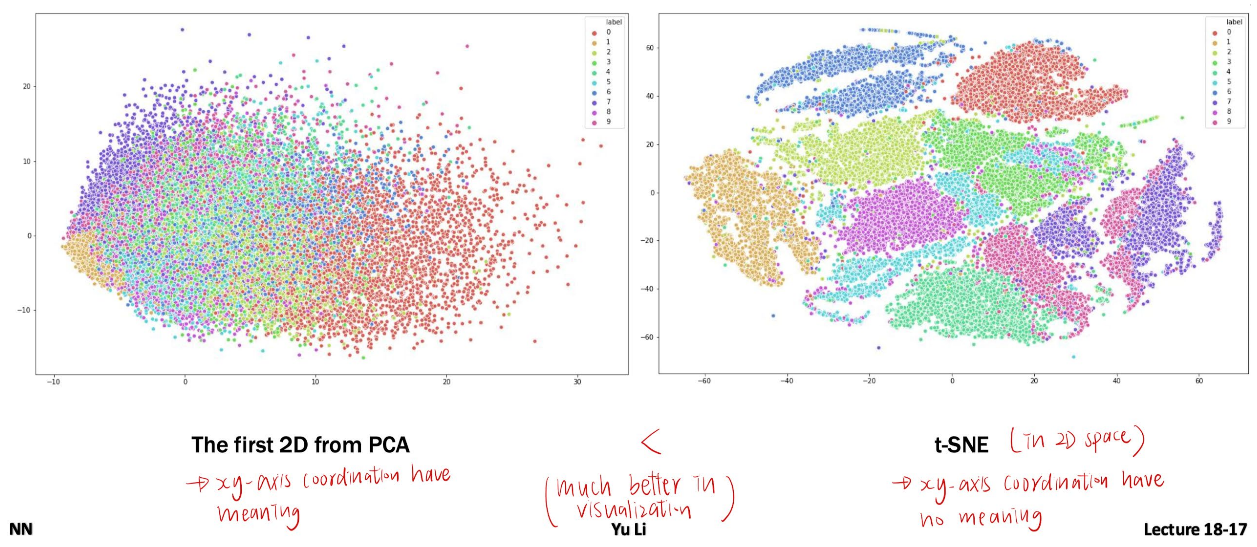
T-SNE

- T-distributed stochastic neighbor embedding
- Nonlinear dimensionality reduction technique for embedding high-dimensional data for visualization in a low-dimensional space of 2/3 dimensions
- Model similar object by nearby points + dissimilar distant object with high probability
- Iterative process

**Process:*

1. Random initialization
2. Update the position for each point – compare the cluster to the original cluster: pinots from same cluster attract each other; points from different clusters push apart each other
3. Continue update
4. Until no more update

PCA vs T-SNE



The first 2D from PCA

→ xy-axis coordination have meaning

<

(much better in visualization)

t-SNE (in 2D space)

→ xy-axis coordination have no meaning

NN

Yu Li

Lecture 18-17

Disadvantage 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

Motif = sequence pattern

From aligned sequence to motif

- Sequences should be aligned before converting into motif
- If not aligned -> have different sequences -> cannot pair the sequence up

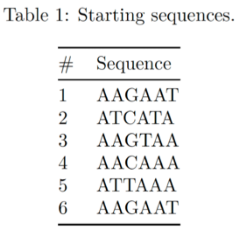


Table 1: Starting sequences.

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

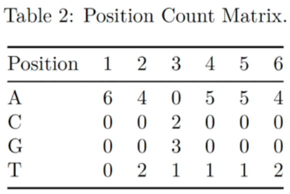


Table 2: Position Count Matrix.

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

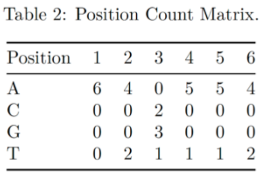


Table 2: Position Count Matrix.

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

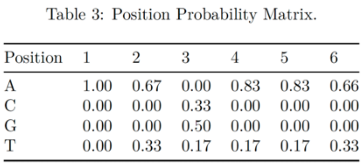


Table 3: Position Probability Matrix.

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

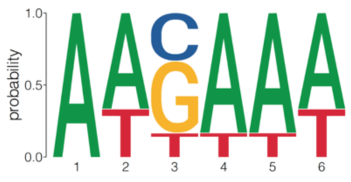
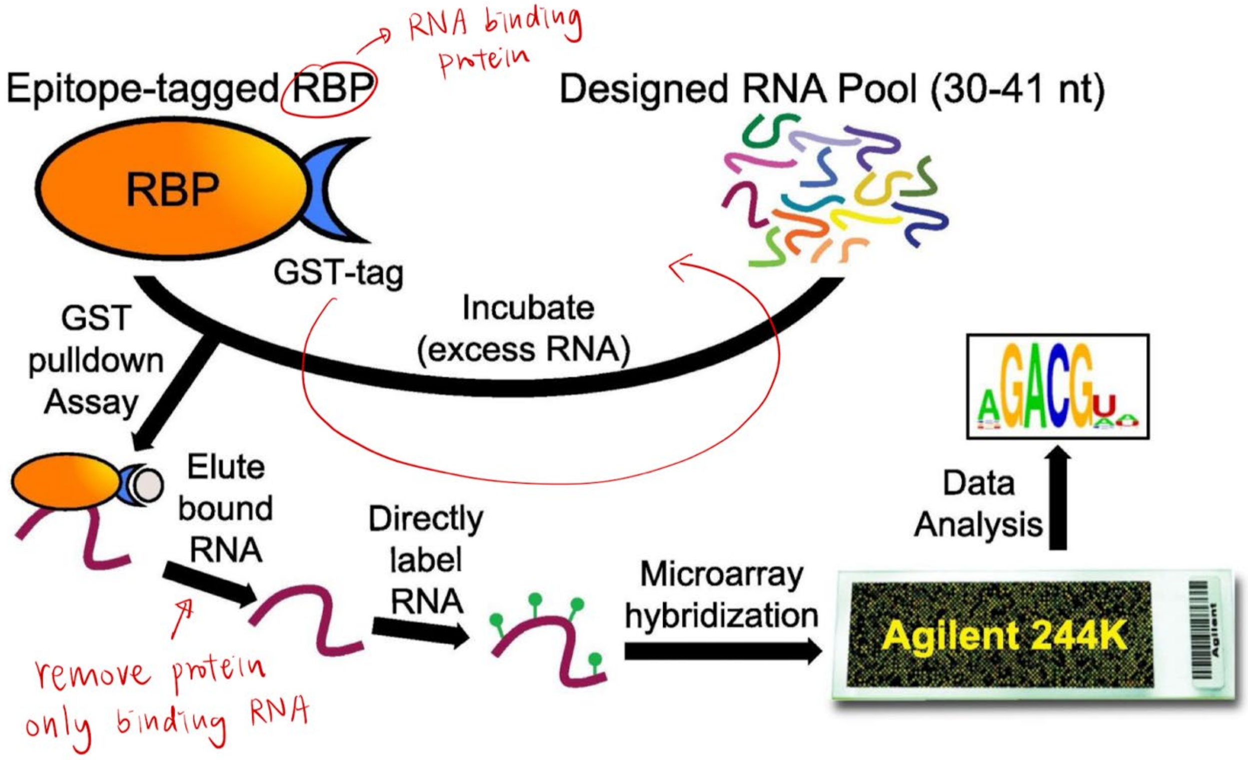


Figure 1: Sequence logo of a Position Probability Matrix


Epitope-tagged RBP
RNA binding protein
Designed RNA Pool (30-41 nt)
RBP
GST-tag
Incubate
(excess RNA)
GST pulldown Assay
Elute bound RNA
Directly label RNA
Microarray hybridization
Agilent 244K
Data Analysis
AGACGU
remove protein only binding RNA