BMEG3105: Data Analytics for Personalized Genomics and Precision Medicine

Data analytics for personalized genomics and precision medicine Lecturer: Yu LI (李煜) from CSE

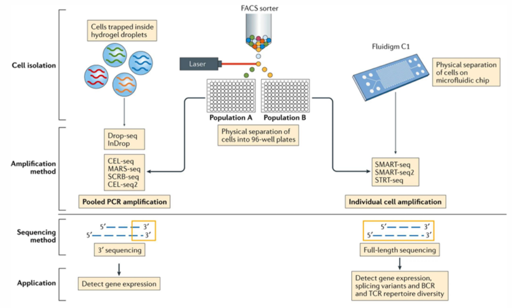
LECTURE 18: Visualization and Protein RNA/DNA

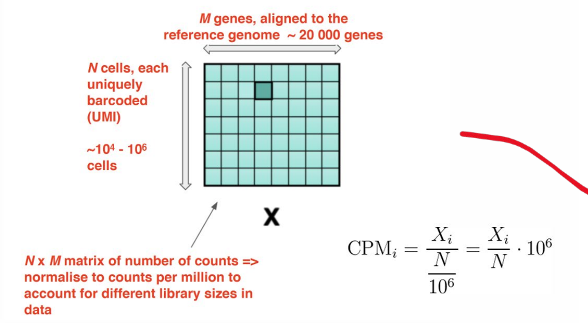
Scriber: Rana Sabri (1155228843)

What is single cell sequencing?

Single-cell sequencing refers to methods that isolate and sequence the DNA, RNA, or other molecular content of *individual cells* rather than a mixture of many.

How to do single-cell sequencing?





M genes, aligned to the reference genome ~ 20 000 genes
N cells, each uniquely barcoded (UMI)
~10⁴ - 10⁶ cells
X
N x *M* matrix of number of counts => normalise to counts per million to account for different library sizes in data

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

Since the count of each cell is very small we times it by a million.

CHALLENGES OF SINGLE CELL DATA ANALYSIS

❖Noise = Random fluctuations or errors in the data caused by technical limitations.

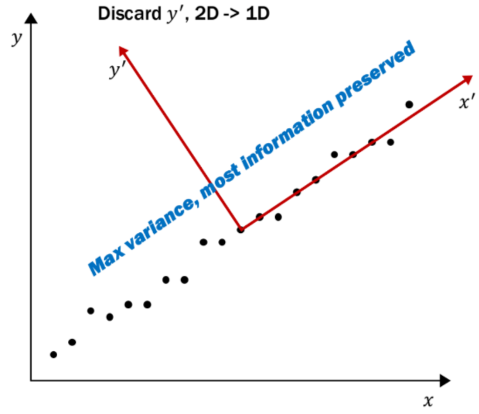
❖Doublet= Occurs when two cells are mistakenly captured and sequenced as one.

❖Dropout=Failure to detect gene expression even when the gene is active in the cell.

❖Batch effect= Systematic differences in data caused by processing samples in separate batches.

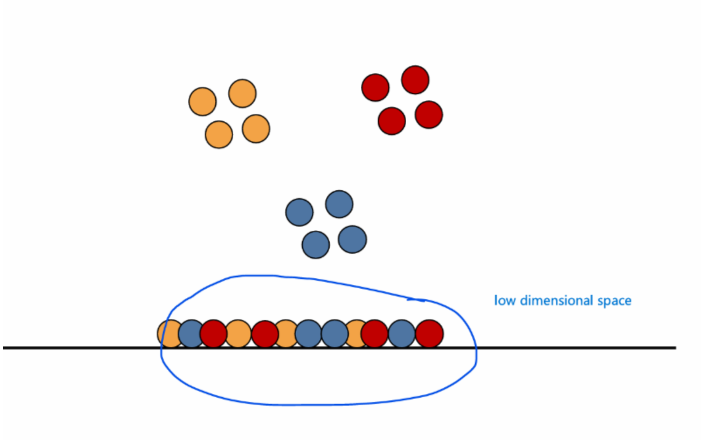
Dimension reduction---PCA

As we have learned in previous lectures PCA is a method we use to reduce dimensionality of a dataset while preserving as much of its variability (information) as possible.



However, PCA also come with its challenges, because its main goals are to reduce dimensionality we lose a lot of data, and the original clusters are not preserved.

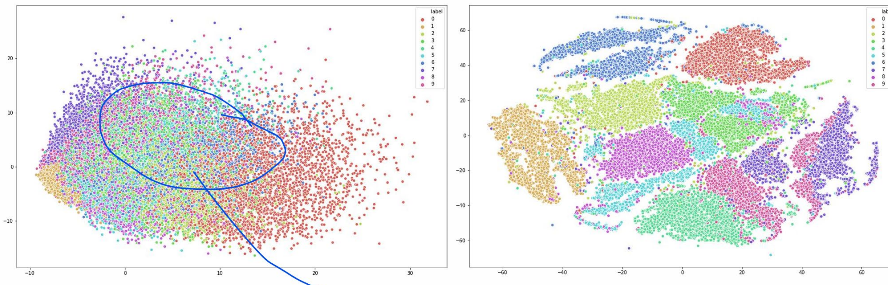
That’s why there is another method called Tsne (t-distributed stochastic neighbor embedding)



STEPS FOR Tsne

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

PCA vs tSNE



The first 2D from PCA

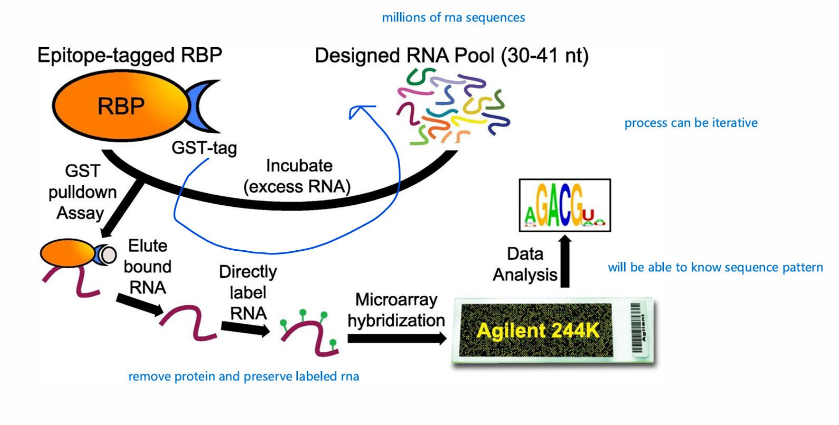
these clusters arent understandable

t-SNE

PCA is linear whereas tSNE is non linear method, and you can never reverse process and retrieve original information with tSNE.

Protein-RNA/DNA Interaction

Protein binding has preference, the problem is how do we get the binding motifs and how do we visualize them.



millions of rna sequences

process can be iterative

will be able to know sequence pattern

remove protein and preserve labeled rna

What is motif? → motif is a kind of sequence no matter protein, dna or rna, its repetitive

From aligned sequences to motif

MAKE SURE TO ALIGN SEQUENCES BEFOREHAND

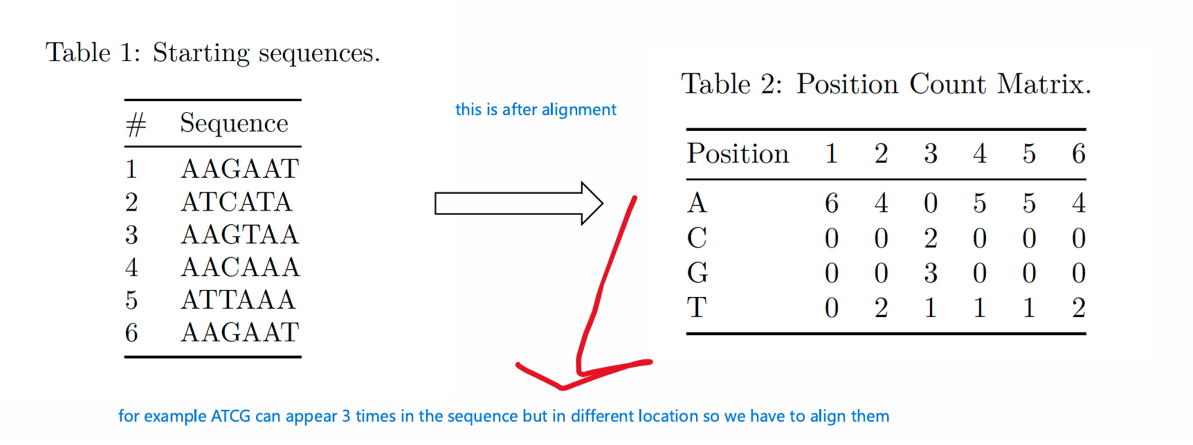


Table 1: Starting sequences.

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

this is after alignment

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

for example ATCG can appear 3 times in the sequence but in different location so we have to align them

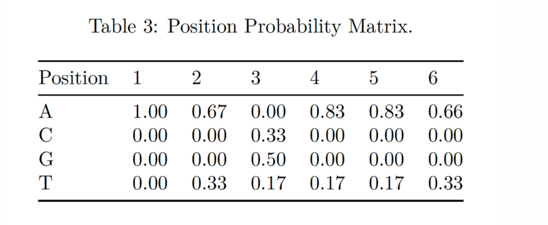
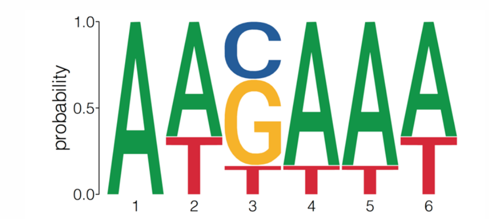


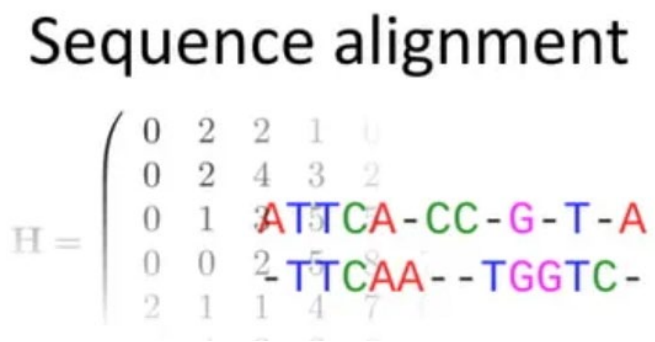
Table 3: Position Probability Matrix.

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



After position count we can come up with position probability matrix.

Sequence alignment is very important, if we didn't do maybe each position could equal 0.25 probability.



Sequence alignment

ATTCA-CC-G-T-A

-TTCAA--TGGTC-

❖AI + Health data

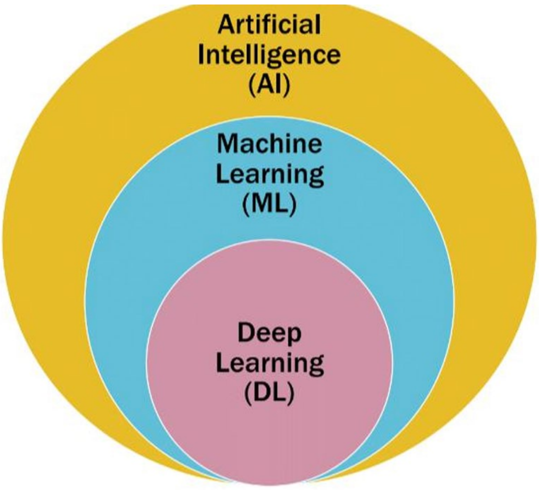
Why do we care about health data?

- **Personalized Care:** Health data helps doctors tailor treatments to individual needs, improving outcomes and reducing side effects.
- **Disease Prevention:** Tracking patterns in health data allows early detection of outbreaks and chronic conditions.

Basically, without the data, doctors cannot diagnose precisely.

AI vs ML vs DL

- **AI (Artificial Intelligence)** is the broad field of creating machines that can mimic human intelligence and behaviour.
- **ML (Machine Learning)** is a subset of AI where systems learn patterns from data to make predictions or decisions without being explicitly programmed.
- **DL (Deep Learning)** is a specialized branch of ML that uses multi-layered neural networks to model complex patterns in large datasets.


Artificial Intelligence (AI)
Machine Learning (ML)
Deep Learning (DL)