BMEG3105 Lecture 19
8/11/2023

Topic: Single-cell analysis and Protein-RNA/DNA
Lecturer: Prof Li Yu

Topic 1: About Last lecture

Epigenetics

The overall data analytic pipeline for epigenetics

Single-cell analysis: For Tumour micro-environment

Definition of single cell-analysis

How to do single cell-sequencing

Topic 2: Single-cell RNA-seq data analytics

Challenges in single-cell analytics:
- Noise: Refer to previous lectures(How to denoise?...)
- Doublet: Not perfect(In cell-isolation process), especially the data
  Needs to remove duplicates
- Dropout: About missing value in the data matrix analysis
- Batch effect: Artefacts from different experiments
  Wet lab: Different results by different students.
  Difference induced by different environment

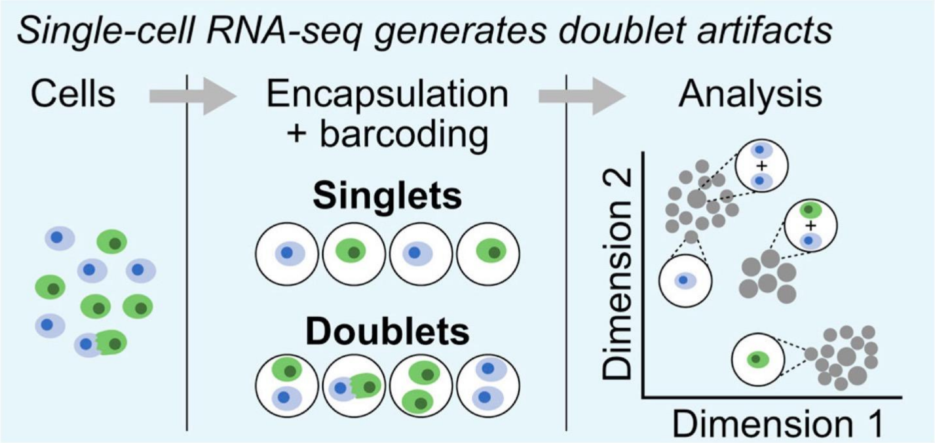
Gene expression matrix:
Need to normalise because there are too much genes
N x M no. of counts -> Normalised to counts per million to account for different library sizes in data
CPMi = Xi/(N/10^6) = Xi/N * 10^6
Xi is potentially small. Xi/N is very small. We need to multiply it by 10^6 to make it reasonably large.

Noise in the matrix

1. The number of genes expressed in the count matrix
2. The total counts per cell
3. The percentage of counts in mitochondrial genes

Quality control not enough: Resequencing

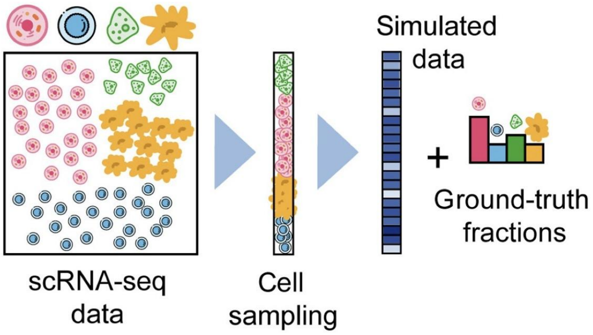
Doublet



Single-cell RNA-seq generates doublet artifacts
Cells
Encapsulation + barcoding
Analysis
Singlets
Doublets
Dimension 2
Dimension 1

We need to use certain techniques to remove doublets. How?

We are going to simulate the doublets



Four different types of cells: Categorized and the doublets will be removed.

Dropout

Expressed in the cell, but undetected in the mRNA profile.
E.g.

The original gene:

3
4
5
4
3
2

The mRNA profile:

3
4
0
4
3
2

Dropout occur due to low amounts of mRNA in individual cells.
Problem: Budget not enough. With more cells, there will be more dropouts.

We need to handle missing values by advanced statistical/ML methods.

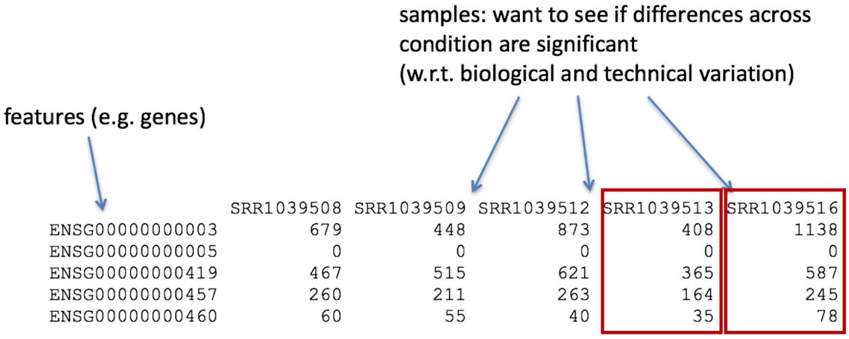
Batch effect

Non-biological factors in an experiment cause changed in the data produced by experiment.

Approaches to batch correction

1. Normalisation

Rich lab vs Poor lab: Rich lab can get more samples(Right).



samples: want to see if differences across condition are significant (w.r.t. biological and technical variation)

features (e.g. genes)

```
                SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
ENSG00000000003        679        448        873        408       1138
ENSG00000000005          0          0          0          0          0
ENSG00000000419        467        515        621        365        587
ENSG00000000457        260        211        263        164        245
ENSG00000000460         60         55         40         35         78
```

2. Alignment

3. Removing genes correlated with batch

4. Regression of residuals with technical covariants
Stat method: Not to be discussed

5. Latent space representations
Machine learning method

Conclusion
What Challenges? Noise Doublet Dropout Batch
Why do we have these challenges?
The intuition behind the solutions

Question: Which is not a batch effect?
Answer: Difference from different conditions(Normal vs disease)
Biological effect: Not a batch effect
Correct batch effect results:
1.Dif Machines 2.Dif sequencing depths 3. Dif sequencing locations

Topic 3: Visualisation

We want to preserve the clusters when visualising the data in 2D.

Technique: PCA(For Dimension Reduction)
By projecting the data to the direction with the highest variance, we preserve as much information as possible.
Not perfect
Losing info along y-direction
Original clusters are not preserved
Higher dimensions: More problematic.

How to preserve the clusters?
T-SNE: t-distributed stochastic neighbour embedding

- A non-linear dimensionality reduction technique well-suited for embedding high-dimensional data for visualisation in a low-dimensional space.
- Similar objects are modelled by nearby points with high probability
- An iterative process

Process

1. Random Initialization
2. For each point, update the position a little bit.
3. Repeat until no more updates.

* no need to know the maths behind.

The clusters pushed away each other.

Disadvantages

1.Long run time (Iterative)

2.Non-deterministic: Different runs may have different results

3.Noisy patterns

4.The original distance may not be precisely preserved

5.UMAP as alternative(not to be introduced)

We can use T-SNE/UMAP in python.

Question: Which is false?

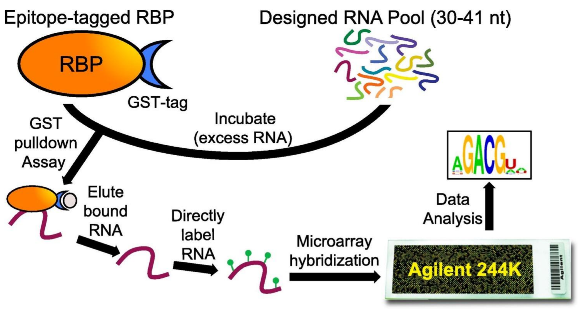
A. It cannot guarantee to give the same result
B. Physical meaning for x-axis and y-axis?
C. Random initialization affects final results
D. Can use to visualise the results from PCA

Ans: B

Topic 4: Protein-RNA/DNA interaction

Protein binding has preference

How to get the binding motif by experiments?



We learnt it in RNA-seq analysis

From aligned sequences to motif

We aligned the sequences first, than we perform normalisation and convert it into motif.
* Need to maximise the similarity to identify the pattern.

The End