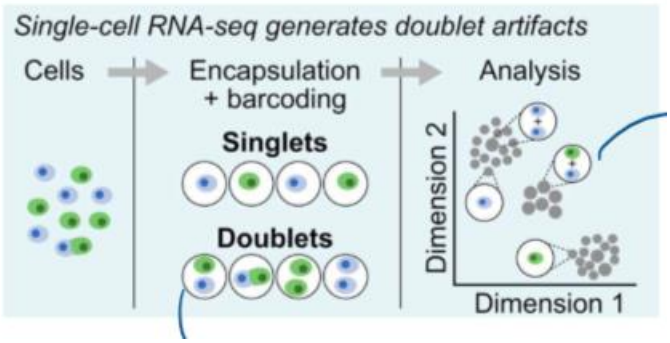
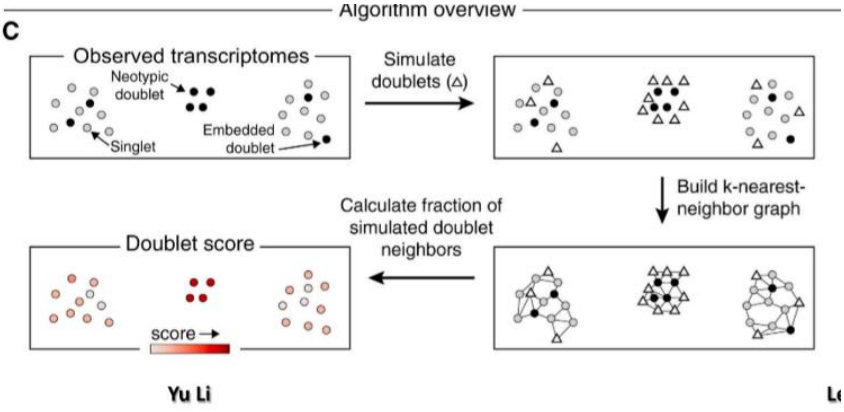
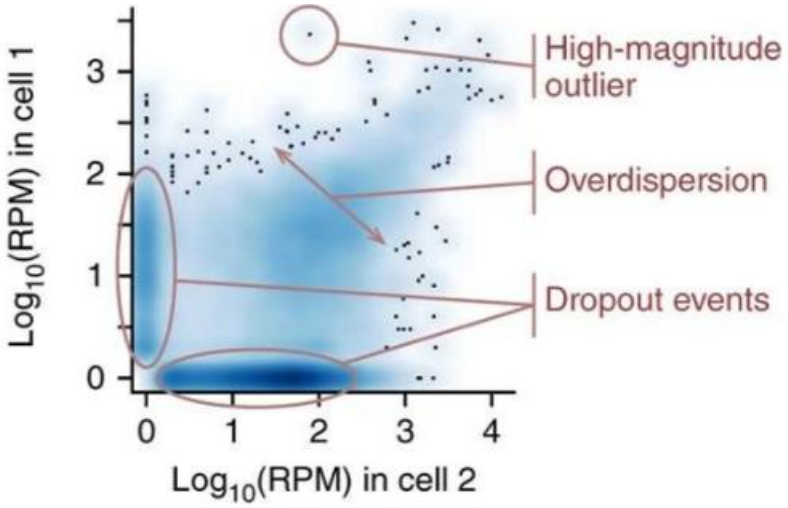
Lecture 19: Single-cell analysis & protein DNA/RNA

1. **Single cell RNA sequencing data analytic**
    a. Pipeline includes pre-processing, perform analysis, and finally downstream analysis.
        i. The preprocessing step can be quality control, normalization of data, feature selection, dimension reduction, and visualization.
        ii. The downstream analysis includes clustering and other complex algorithms.
    b. Challenges in preprocessing process
        i. Noise
            1. gene expression matrix of snRNA might have close to zero data. If we normalize them, it can be even smaller.
            2. Thus, we need to normalize them to count per million to preserve value after normalization.
            3. $$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$
            4. Another noise can be the noise from mitochondrial gene which is also present in total gene count.
        ii. Doublet
            1. Doublet is an error during the single isolation process when the diverse types of cells were isolated and sequenced together resulting in an error of gene expression.
            2. 

            3. It can be eliminated by introducing simulated doublet gene expression that can identify doublet error from a normal gene expression.
            4. 

iii. Dropout
   1. Dropout is an error when a gene expression is undetected because of the low amount of sample mRNA in individual cell.
   2. It can be alleviated by an advanced statistical or machine learning method.
   3. It can also be prevented by introducing more sequence read which require more financial resources.
   4. From the figure cell 1 and cell 2 are the same cell type but the gene expression is represented in one cell not the other.



   5.

iv. Batch effect
   1. Batch effect is a non-biological factor in experiments that cause an error.
   2. It can be resolved by normalization, alignment, remove gene correlated with batch, regression of residuals with technical covariates, or latent space representation.

2. Visualization gene expression data in 2D

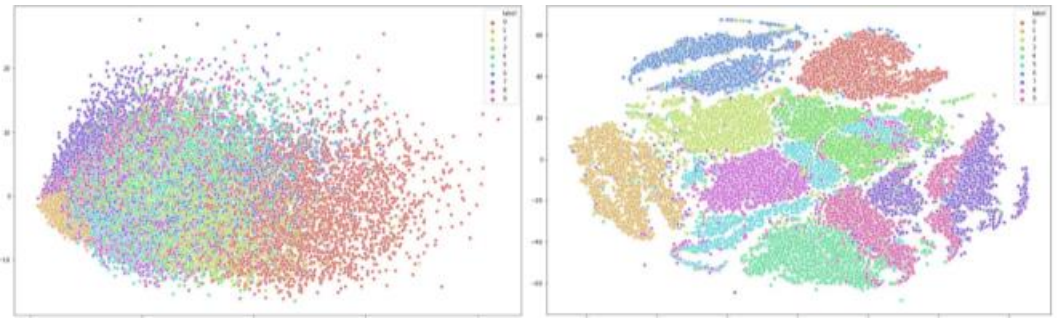
a. We can visualize the data when the dimension is reduced to two dimensions or less.

b. One of the methods would be the PCA.
   i. It will project the data to which has a higher variance and imperfectly capture data in the other axis.
   ii. It has a problem when the data variance is low or evenly distributed.
   iii. The problem with PCA is it does not preserve the cluster which might have information loss.

c. T-SNE (T-distributed stochastic neighbor embedding)
   i. It is a non-linear stochastic dimension reduction technique for mapping high dimensional data into low dimensional space.
   ii. Each round of t-SNE will result in a different result.
   iii. Similar objects are pulled together while dissimilar are pushed apart.
      1. Random initialization
      2. Update position base on similarities

3. Iterate the process until there are no more updates.

iv. T-SNE result will preserve the cluster while scarifying the physical meaning of cluster distance.

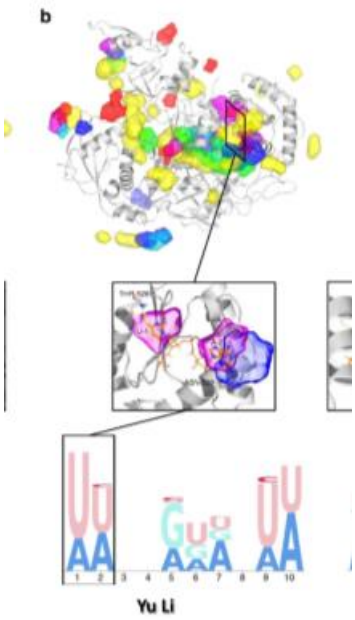
v. 

vi. Disadvantages of t-SNE

1. Long running time
2. Non-deterministic: each round has a different result.
3. Noisy pattern
4. Distance is not preserved.

3. Protein-RNA/DNA interaction

a. The protein binding analysis shows which nucleotide has more chance to bind with protein.



b.

c. It can be processed by multiple sequence alignment and count the read and put them into visualization graph.

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

d. Visualization Yu Li Lecture 19-47