Lecture Title and Date

25m9e - Single Cell Analysis 03/03/2025

Objectives of the Lecture

By the end of this lecture, students should be able to:

- Understand single-cell RNA sequencing (scRNA-seq) methods
- Understand computational methods for analyzing scRNA-seq data
- Understand doublet detection and removal algorithms
- Explain applications of dimensionality reduction algorithms in single-cell analysis
- Explain applications of cluster algorithms in single-cell analysis pipelines
- Understand imputation of undetected genes
- Understand the diffusion pseudotime algorithm for reconstructing cell lineage instead of discrete clustering.

Key Concepts and Definitions

- Single-cell RNA sequencing: Profiling RNA expression level from individual cells.
- Doublet: Oil droplets or microwells contain more than one cell, during cell isolation, in which case sequencing results no longer represent a single cell.
- UMAP & t-SNE: Nonlinear dimensionality reduction algorithms commonly used in scRNA-seq analysis for estimating a low-dimensional manifold that captures local structures instead of global structures.
- Louvain Clustering: A connectivity-based clustering algorithm by optimizing modularity of communities.
- Forest Fire Clustering: A stochastic clustering algorithm by randomly selecting a seed node and propagating the label based on threshold.

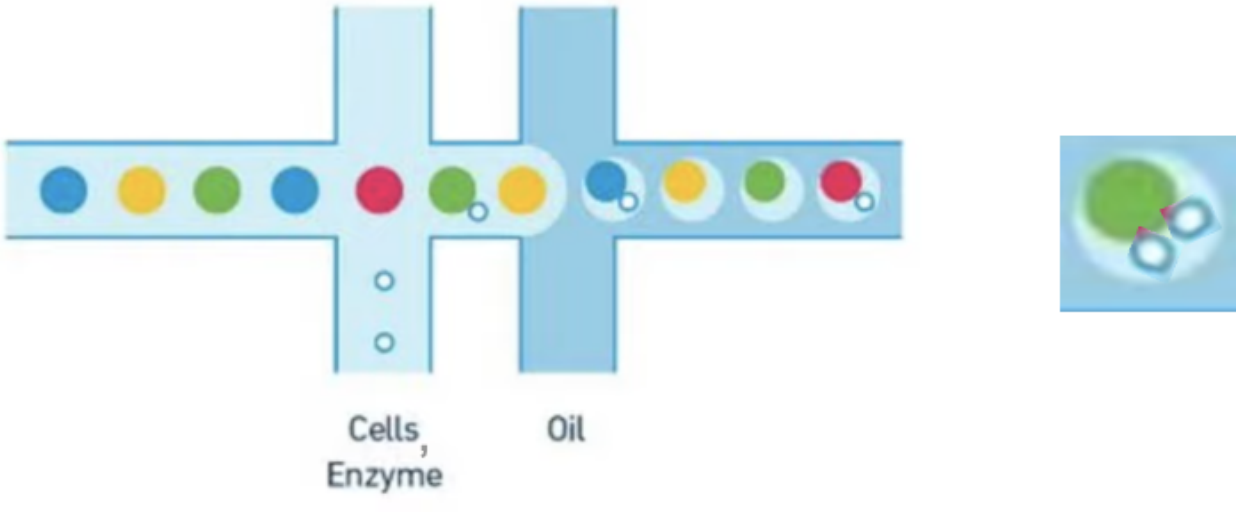
Main Content/Topics

Single-cell RNA sequencing collects gene expression profiles for individual cells instead of an average expression profile as in Bulk RNA sequencing. In general, the single-cell analysis workflow can be broken down into the following steps: quality control, normalization, feature selection, dimensionality reduction, cell-cell distances, and unsupervised clustering.

scRNA-seq experiment

In a scRNA-seq experiment, individual cells are isolated into microscopic droplets. Transcripts from each cell are tagged with barcodes and unique molecular identifiers(UMI), which are unique to the cell and RNA molecule, respectively. UMIs can identify duplicates in the original

cell and copies from PCR amplification for a more accurate expression level. The raw data of the scRNA-seq experiment is a cell-by-gene matrix of expression levels or counts.



Doublet Detection & Removal

Doublet is a common artifact in scRNA-seq, where a droplet contains more than one cell. One way to detect doublets is to simulate doublets from the original data and perform a k-nearest neighbor (KNN) classifier to identify and remove doublets from the data.

Batch Effect Correction

The batch effect could occur due to non-biological factors in experiments. It has to be properly handled to reveal true biological signals. Batch effect correction is an iterative algorithm for removing batch effects using a linear model. The algorithm assigns cells into clusters and minimizes distances between different datasets for each cluster.

Dimensionality Reduction

In the gene expression matrix, each gene represents a dimension, resulting in a matrix with ~10k dimensions. Dimensionality reduction is necessary for visualization and many algorithms to work. Typically, PCA is first applied to reduce the gene expression matrix into a smaller number of dimensions (30-50). UMAP and t-SNE can then be used to estimate a low-dimensional embedding that can better capture global and local structures. Unlike PCA, UMAP and t-SNE both have hyperparameters to adjust. For example, `perplexity` for t-SNE or `n_neighbors` for UMAP controls how much global information is preserved.

Clustering

After performing UMAP, the cleaned-up low-dimensional projection should contain disjoint clusters of cells. One computational method for identifying clusters is the Louvain algorithm, which identifies clusters by maximizing modularity. Modularity is a metric measuring connectivity within communities. For each iteration in the Louvain algorithm, we scan all nodes and find if moving a node to neighboring communities could increase modularity. After no changes can be

made, each community is combined into one node as well as for edges. This process is then repeated until reaching a desired resolution. The Forest fire clustering is an alternative clustering algorithm with a faster runtime and smaller memory footprint than the Louvain algorithm. It randomly selects a seed node and propagates the label based on a threshold. The stochastic process can be repeated to serve as internal validation.

Imputation of undetected genes

Imputation can be performed after clustering to further clean up the data. Some genes might be missing from the dataset but are actually expressed. It is possible to find overall structure and estimate undetected genes based on similar cells.

Transferring Cell-type Annotation

The next step is to associate clusters with types of cells. Cell types can be associated with clusters by using marker genes, cluster-to-cluster correlations, or using machine learning algorithms. Alternatively, instead of labeling discrete cell types, the diffusion pseudo-time algorithm can reconstruct cell lineage from the probability of transition, which could be useful when studying cell development.

Discussion/Comments

- Single-cell RNA sequencing data is inherently noisy. Designing computational pipelines for analyzing such data requires great care.

List all suggested reading here and please answer:

Are the readings for the class useful? If so, are the specific subsections useful or would change. If not, are there other references you could suggest? Please suggest one.

- Andrews, T. S., Kiselev, V. Y., McCarthy, D., & Hemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nature Protocols*, *16*(1), 1–9. https://doi.org/10.1038/s41596-020-00409-w
    - Single Cell overview; goes over every step
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
    - Overview of Louvain clustering upgrade; read the methods section
- Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, *13*(10), 845–848. https://doi.org/10.1038/nmeth.3971
    - Optional: First page summarizes the Pseudotime algorithm
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

    - Optional: Description of cell type annotation. See fig. 1, which is explained in the beginning of Results section
- Andy Coenen, Adam Pearce. Understanding UMAP https://pair-code.github.io/understanding-umap
    - Optional: tSNE UMAP key concepts with great interactive visualization

All the suggested readings are helpful, and appropriate sections are already identified. It might be helpful to include a reference covering single-cell experimental procedures in more detail for those who are interested.

References ISL/ESL (if any)

- K-Nearest Neighbors (Doublet detection)
- Dimensionality reduction

Other Suggest references for many of the key concepts

N/A