Younginn Park

scRNA-seq analysis and cellular populations

Motivation

High-throughput RNA sequencing and microarrays has allowed for rapid development of transcriptomics, surpassing methods that relied on northern blotting or qPCR. However, these studies still had limitations in terms of resolution as they took measurements across the whole tissue ("bulk"), which meant taking averages of expression levels and treating them as representative across cells in a population. The benefits of this approach include having a clean homogeneous picture of the expression levels, identifying representative markers of a tissue and being easier to handle in comparative analyses. However, this posed a challenge, especially in the clinical context, for several reasons:

1) **Heterogeneity of cells** - in multicellular organisms, different cells in a tissue can have varying roles during biological processes forming subpopulations with distinct expression profiles. This has further implications if there are imbalances in the representations of certain subpopulations, especially for rare subpopulations.
2) **Temporal processes** - during temporal processes like cell differentiation or cell proliferation, average expression levels can only show changes through time rather than the stages of the process. This can be a problem when cells of the same population undergo the same process, but in different time scales. This is often the case in more complex tissues where signals from neighboring cells can influence the course of the temporal processes.

Single cell transcriptomics has the potential to tackle these challenges, bringing insights into the heterogeneity of cells within tissues. Many diseases exhibit high levels of heterogeneity, driven by factors such as exposure to mutagens like UV radiation or tobacco, which can induce mutations. This leads to heterogeneity in tumor tissues, which has consequences on its proliferation and metastatic potential. Moreover, understanding the immunological response to treatments and drugs would help maximize efficacy and give accurate prognosis for such treatments.

Single cell RNA sequencing technology

In the workflow of single cell sequencing technologies, the process typically begins with a tissue sample as the input material. The subsequent steps can be broken down into 3 stages: **cell dissociation**, **cell isolation** and **library construction**.

First, single cell dissociation is the step where cells are separated to form a single cell suspension. Next, single cells are isolated using either plate-based methods, where cells are placed into individual wells on a plate, or droplet-based methods, which involve capturing cells in microfluidic droplets. However, errors may occur during isolation, leading to the capture of doublets or multiplets, non-viable cells, or no cells at all. Empty droplets are particularly common, as they are formed based on the low concentration flow of input cells.

Following isolation, library construction takes place within each droplet. Each droplet contains chemicals that facilitate the breakdown of cell membranes and the construction of libraries, including mRNA capture, reverse transcription to cDNA, and amplification. At this stage, mRNA can be labeled with a droplet-specific barcode and molecule-specific unique molecular identifier (UMI). Finally, the constructed libraries are pooled together (multiplexed) for sequencing. The sequencing process generates reads that can be further analyzed.

Bioinformatics analysis

The raw reads from sequencing are processed to get **count matrices** or read matrices (if UMIs were used). Pre-processing reads usually involves quality control, assigning reads to barcodes (“demultiplexing”), alignment to genome and quantification.

During quality control steps three covariates are mainly taken into consideration: number of counts per barcode (**count depth**), **number of genes per barcode** and **fraction of counts from mitochondrial genes per barcode**. Generally, quality control is based on looking for outlier peaks that are filtered out by thresholding as they usually correspond to dying cells, cells with broken membranes or doublets, e.g. barcodes with low count depth, few genes and high fraction of mitochondrial counts are indicative of cells with cytoplasmic mRNA leaked out through a broken membrane thus only mitochondrial mRNA is conserved. On the other hand, cells with high count and large number of genes may be doublets. Other alternative methods for detecting doublets/multiplets have been proposed in: DoubletDecon, DePasquale, Scrublet, Wolock, Doublet Finder. All these methods use more advanced techniques to differentiate out the viable cells, e.g. deconvolution of expression levels and checking if they match to more than one profile (DoubletDecon, DePasquale et al. 2019).

The best practice is to analyze these three covariates jointly. Analyzing any of these in isolation can lead to misidentification of signals. For example, a high fraction of mitochondrial counts alone may indicate cells related to respiratory processes. Most of the time, it’s better to be as permissive as possible to avoid filtering out viable cell populations. It’s also often the case that genes that are not expressed in more than a certain number of cells are also filtered out to reduce the size of the count matrix. Here, the threshold should be appropriate to the desired resolution needed in subsequent analyses.

Normalization, data correction and dimensionality reduction

Normalization step helps compare gene count depths between cells more reliably by removing unwanted variability due to the effects of count sampling. Most common normalization technique is called “counts per million” (CPM). It involves scaling gene counts by the factor proportional to the count depth per cell. This method assumes that any differences between gene expression levels arise only due to sampling of cells.

It’s also possible to normalize using downsampling, which involves sampling reads only to a predefined number of counts or fewer, attempting to simulate a case where all cells have been sequenced to the same depth, making comparisons between expression levels more reliable. Other methods for normalization exist that for example take into account the share a gene has in the counts and then scale the expressions accordingly.

Data corrections may target technical and biological covariates like batch effects (technical variability), dropout events (low amounts of mRNA in cells) or cell cycle effects. Linear regression against a cell cycle score (implemented in Seurat and Scanpy) or LVMs may be used for regressing out cell cycle effects. Other biological covariates such as mitochondrial gene expression, which indicates cell stress, can be eliminated this way as well. Tools like ComBat allow for removing batch effects, which are variations between cells that arise due to non-biological differences, like variations in sample preparation procedures.

Dimensionality reduction can further help during feature selection, since not all genes have equal contribution to the variability of the cells. Representing expression matrices in low dimensions can be useful to describe the structure of the data with fewer dimensions than the number of genes while identifying directions of high variability in the gene space. It helps visualize the data which in turn will prove helpful in downstream analyses, like cluster analysis. Most commonly used dimensionality reduction methods include PCA (linear), t-SNE and UMAP.

Cluster analysis

Cluster analysis is a valuable tool for identifying cell subpopulations within single-cell transcriptomic data. To achieve this, dimensionality-reduced embeddings are commonly employed, alongside distance-based clustering methods. Most notable methods include algorithms like k-means and k-nearest neighbors. **K-means** identifies k clusters by finding k centroids and assigning cells to the nearest one, typically by using Euclidean distance as the default metric, though other measures like cosine similarity or correlation-based distances can also be used. On the other hand, **KNN** works by connecting cells to k-nearest cells by metrics like the Euclidean distance in the dimension-reduced space, offering faster performance by focusing solely on neighboring cells and reducing the search space. This approach can be further optimized using techniques such as the Louvain method.

Once clusters are defined, marker genes help characterize each cluster and associate them with potential biological labels. However, determining the exact biological context of each cluster remains challenging, leading to the use of terms like "cell identity" rather than "cell type" to describe them. Accurate annotation of these groups relies heavily on external gene marker databases, which are increasingly available thanks to initiatives like the Human Cell Atlas.

Trajectory analysis

Given that most processes within living tissues unfold continuously, each dataset of expression levels can be viewed as a snapshot of these dynamic processes across cells. This perspective enables a unique form of analysis aimed at ordering the varying expression levels along a pseudotime continuum, reflecting different stages of these processes and capturing transitions between cell identities. At a high level, this analysis seeks to identify paths between cells that minimize expression level differences, ultimately revealing patterns in cell progression similar to a developmental timeline, often referred to as *pseudotime*. By regressing changing gene expression levels across pseudotime, it is possible to discern smooth transitions at each step and annotate them with biological labels and provide insight into the underlying genes and the progression of temporal processes. Tools such as Monocle (Trapnell et al., 2014) and Wanderlust (Bendall et al., 2014) facilitate trajectory inference within programming environments.

References

Malte D Luecken & Fabian J Theis (2019), *Current best practices in single‐cell RNA‐seq analysis: a tutorial*, Molecular Systems Biology 15: e8746, doi.org/10.15252/msb.20188746

Evan Z. Macosko et al. (2015), *Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets,* Cell. 2015 May 21; 161(5): 1202–1214. doi:10.1016/j.cell.2015.05.002

Andrew Butler et al. (2018), *Integrating single-cell transcriptomic data across different conditions, technologies, and species*, Nature Biotechnology vol. 36 nr 5 May 2018

Trapnell, C., Cacchiarelli, D., Grimsby, J. et al. *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells*. Nat Biotechnol 32, 381–386 (2014). https://doi.org/10.1038/nbt.2859