scRNA-seq analysis - some keywords and definitions

Esthy Hung

July 16, 2019

1 Introduction

All of one individual's cells share the same DNA, but they perform very different functions. The differences lie in gene expression, the translation of DNA through RNA to functional proteins. The RNA present in a cell is therefore intimately connected to the cell's behaviour. Counts of RNA molecules determine what a cell is doing, what other cells it resembles, and what those cells might do in the future. These RNA measurements were first done in bulk by extracting RNA from a tissue, which gives an average of the expression levels across the sample. More recent techniques allow this to be done at a single cell level, producing counts of individual RNA molecules within individual cells.

Such single cell analysis enables precise, fine-tuned distinctions between cell types within the same tissue, with extensive applications across biology and medicine. Gene expression signatures can, for example, be used to identify pluripotent stem cells or malignant tumour cells and their precursors. Understanding the signs of cancer would lead to earlier, more accurate diagnosis and better treatment. However, despite significant progress in developing methods for single-cell RNA sequencing, neither the underlying experimental procedures nor the required inference methods are sufficiently well understood. I propose to work on both problems in parallel.

Gene expression: Translation of DNA through RNA to functional proteins

Count and read matrices:

Transcriptome: The set of all RNA molecules in one cell or a population of cells

Cell cycle:

Marker genes: Genes up-regulated in cluster compared to all other clusters.

Cell scoring:

Mean expression:

Pseudotime analysis:

Metastable states:

2 scRNA-seq analysis:

1. Pre-processing
    - **Quality control** - How we ensure that all cellular barcode data correspond to viable cells, and that data quality is sufficient for downstream analysis. All three co-variates should be considered jointly when univariate thresholding decisions are made.Three QC Covariates:
        - Count depth: no. counts per barcode
        - No. genes per barcode
        - Fraction of counts from mitochondrial genes per barcode
    - **Normalisation** - Process of scaling count data to obtain correct relative gene expression abundances between cells to enable comparison between samples.
    - **Data correction** - targets further technical and biological covariates like batch, dropout, or cell cycle effect. We consider correction of these effects separately.
    - **Feature selection** - Process of filtering out and keeping genes that are 'informative' of variability in the data. Highly variable genes (HGVs) are usually used. Genes with highest variance-to-mean ratio are selected as HGVs in each bin. This is preferably done after technical data correction.

- **Dimensionality reduction** - The biological manifold on which cellular expressions lie can be sufficiently described by far fewer dimensions than the number of genes; dimensionality reduction aims to find these dimensions.

2. **Downstream Analysis:** - Obtaining descriptions of the underlying biological system described by the data by fitting interpretable models. Cell- and gene-levels are considered.

    - **Cluster Analysis:** The attempt to explain the heterogeneity in the data based on a categorisation of cells into groups.
    - **Compositional Analysis:** At cell level, we can analyse clustered data in terms of compositional structure - looking at the proportion of cells that fall into each cell-identity cluster. These proportions can change in response to disease. Statistical tests over changes in the proportion of a cell-identity cluster between samples are dependent on one another.
    - **Trajectory analysis** We regard data as a snapshot of dynamic process and investigate the underlying process. *Trajectory inference:* Required in order to capture transitions between cell identities, branching differentiation processes, or gradual, unsynchronised changes in biological function, using dynamic models of gene expression.

3 Trajectory inference:

When we have multiple trajectories in data. We find trajectory struture in data first before pseudotime can be inferred along each trajectory.

Pseudotime variable: describes the ordering of cells along paths

PAGA: Partition-based Graph Abstraction.

- Recommended TI method for trajectories expected that are more complex .
- Used to reconcile clustering and pseudotemporal ordering.
- Can apply to whole dataset - no assumptions of continuous trajectories connecting all clusters

Slingshot: - Recommended fo simple trajectories ranging from linear to bi- and multi-furcating models

- Use normalised, log-transformed, batch-corrected data to minimise technical variation in inferred trajectories/
- TI with fixed endpoints vastly improves trajectory.