The process of analysis of single-cell (sc) RNA sequences consists of multiple steps and considers various issues, which appear due to the biology of the cells and imperfections of techniques. In this note, we will briefly cover the main steps of the analysis and take a more focused look at some of the methods. However, one of the main challenges in the analysis of scRNAseq is standardization. Since the method itself is new, the number of analysis tools is skyrocketing (=> hard to choose the tool), and the sizes of the datasets are exploding. All these factors make it harder to standardize the data and approach.

In general, there are 3 main analysis steps (pre-processing, cell-level, and gene-level downstream analysis), however, with multiple substeps and branches.

Pre-processing:

Raw data from a sequencer looks like read or count matrices. Next it is required to perform QC of reads, to assign reads to their molecular barcodes (basically barcode == cell), to assign mRNA molecules to their origin (demultiplexing), to perform genome alignment, and quantification.

QC:

Aim of the QC is to ensure that all cellular barcodes correspond to viable cells. Pretty often errors in single-cell isolation can happen, such as: doublets or multiplets, nonviable captured cells or no capture. Empty droplets are often the case in droplet method, as it relies on low concentration.

There are 3 QC covariates: 1. count depth: number of counts per barcode 2. the number of genes per barcode 3. fraction of counts from mitochondrial genes per barcode. The distribution of these QC covariates are examined for outlier peaks. QC at the level of transcripts also should be performed: removing genes that are not expressed in more than a few cells => not informative of the cellular heterogeneity. Usually it is better to start with permissive QC thresholds, investigate their effects, and only then to perform more strict analysis.

Normalization should also be performed in order to smooth the sampling effects. There are linear global scaling normalization methods, and non-linear normalization methods - for more complex unwanted variation. Log-transformation of normalized data is the next step.

Regressing out biological, and technical effects:

In terms of biological effects - 1. removing effects from mitochondrial gene expression (an indication of cell stress) 2. cell cycle effect on the transcriptome should be removed by linear regression against a cell cycle score (a list of marker genes is required in order to compute the cell cycle score). However, sometimes removing the biological effects is misleading, thats why context of data is always important: for example, proliferating cell populations can be identified based on cell cycle scores.

Technical effects include batch effect and count depth. Batc effect occurs when cells are handles in distinct groups. Count depth can be both biological and technical artefacts. Biology-wise cells can differ in size => also in mRNA counts, technical-wise - poor sampling.

Feature selection, dimensionality reduction and visualization:

As an output we will have ~ 25 000 genes, however, only a part of them will be informative. QC will filter out a lot, however, we will end up with ~15 000 dimensions of the dataset, which is a computational burden. Moreover, thereis need in noise reduction and visualization. Therefore, reduction of dimensions (RD) is needed. There are several steps of RD. The first one - feature selection: keeping only the genes that are informative of the variability in the data. Genes are binned by their mean expression -> genes with highest variance/mean ratio are selected as HVGs (highly variable genes) in each bin. 1000 - 5000 HVGs are usually selected for downstream analysis. The second step is the actual RD. There are different algorithms of RD, however, the global idea is the same: to put the expression matrix into a low-dimensional space, and to capture the underlying structure in the data in as few dimensions as possible.

2 main objectives of RD methods:

1. visualization
    a. attempt to describe the dataset in 2D or 3D;
2. summarization
    a. reduces data to its essential components by finding the inherent dimensionality of the data;

There are two popular RD techniques: PCA and discussion maps. Here I will explain in simple words how PCA works.

PCA (Principle Component Analysis) - is a linear approach, which generates RD by maximizing the captured residual variance in each further dimension, or in other words - maximizing the sum of squared distances from the projected data points to the origin. PCA does not capture the structure of the data.

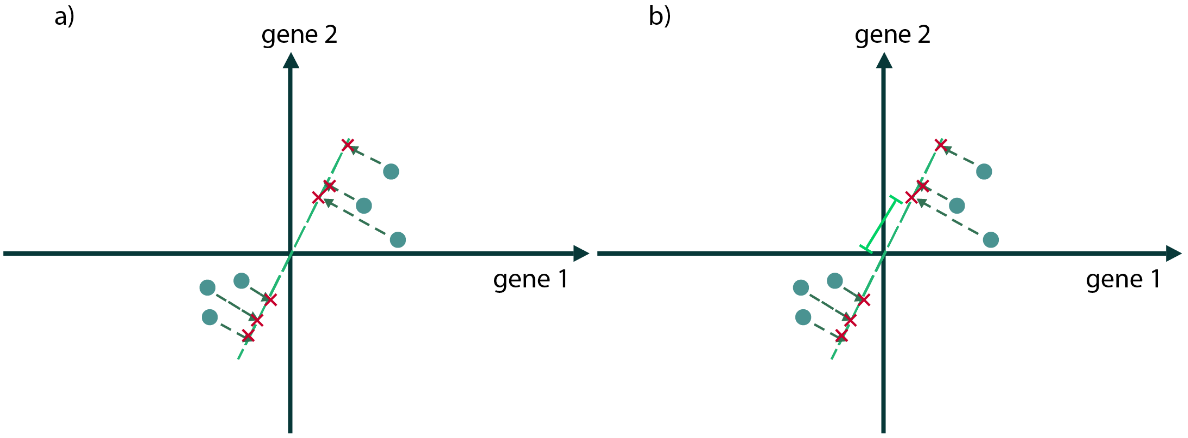
PCA functionality:

- Reduction of dimension (usually to 2D);
- Tells which variable is the most valuable for clustering data;
- Tells how accurate the 2D graph is;

Steps of PCA (simplified):

1. Average measurements for each gene are calculated. With the average values the center of the data can be calculated;
2. The data is shifted so that the center is on the origin of the graph => now a line should be fitted on it.
3. How the PCA decided if the line fits or not: PCA projects the data points onto the line. And then it measures distances from the data to the line and try to minimize the distances.
   Two options are possible (in fact same option but from different sides):
   a. It can measure the distances from the data to the line and try to find the line that minimizes those distances;
   b. It can find the line that maximizes the distances from the projected points to the origin. So, PCA finds the best fitting line by maximizing the sum of squared distances from the projected data points to the origin.

   The rotation of the line and calculation of the max sum of squared distances continues until the maximum one is found. It will be eigenvalue of PC1, meanwhile the line itself is PC1. The slope of the PC1 defines the linear combination of gene 1 and gene2 (eg the ratio between gene1 and gene2).



4. In 2D graph the PC2 is simply a perpendicular to PC1, and no additional adjusmetns are needed.

When PCA is performed together with SVD, then the PC1 is scaled to one.

So we discussed a little bit the aim of summarization in RD. Now let's talk about visualization. It is usually performed via non-linear methods (tSNE, UMAP, SPRING).

tSNE finds the way to project data into a low dimensional space so that the clustering in the high dimensional space is preserved. Sounds similar to PCA, however, there is a crucial difference between those two. PCA offers global preserving of the relationships between data clusters. Whereas, tSNE preserves more local structure.

Downstream analysis (DA).

It aims to extract biological insights and describe the underlying biological system. DA can be divided into cell-level and gene-level approaches. It focuses on the description of 2 structures:

1. clusters - try to explain the heterogeneity in the data based on a categorization of cells into groups
2. trajectories - snapshot of a dynamic process

Cluster analysis.

Cells organized into clusters - the first intermediate result of any single-cell analysis. Clusters are obtained based on the similarity of their gene expression profiles (determined via distance metrics). There are 2 approaches of generating cell clusters:

1. clustering algorithms
    a. unsupervised ML
    b. cells are assigned to clusters by minimizing intracluster distances or finding dense regions
    c. K-means clustering is a popular approach
2. community detection
    a. graph-partitioning algorithms
    b. cells are nodes, each cell is connected to its K most similar cells
    c. K-Nearest neighbor approach (KNN graph)

Cluster Annotation

Cell identity VS Cell type

Users can assume that clusters detected in SC data represent cell types. However, it is not always clear what constitutes a cell type: in some researches T-cell can be a label for a cell type, but in another - distinguishing between CD4+ and CD8+ T-cells is needed. For this reason it is better to refer to clusters as “cell identity”. Identifying and annotating clusters relies on using external sources of information: mouse brain atlas, human cell atlas.

Trajectory inference.

Clustering is not sufficient to describe cellular diversity. As the processes that drive the development of the observed heterogeneity are continuous => need to capture transitions between cell identities by applying dynamic models of gene expression - trajectory inference (TI).
TI interpret single-cell data as a snapshot of a continuous process. The process is reconstructed by finding paths through cellular space that minimize transcriptional changes between neighbouring cells. The ordering of the cells along these paths is described by a pseudotime variable.

Gene expression dynamics.

The received trajectory needs to be analyzed on the gene level, so it has support that it is not the result of fitting transcriptional noise. Approaches to the detection of trajectory-associated genes detect those that vary across a trajectory by regressing gene expression against pseudotime variable.

Cell-level analysis unification.

Clustering and trajectory inference represent 2 distinct views of single-cell data. They can be unified into one:
single-cell clusters = nodes, trajectories between clusters = edges
So it would represent both the static and dynamic nature of the data.