Nomenclature:

- UMI - unique molecular identifiers - short sequences in reversely transcribed cDNA, allowing for unique transcript identification. In this way, we can separate actual transcripts from PCR-amplified sequences.
- barcode - short sequences in reversely transcribed cDNA, allowing for separating samples (cells in scRNA-seq).
- count depth - number of counts per barcode
- size factor - number of reads in each library
- dropout - zero or near-zero counts
- CPM - counts per million (scaling method)

Key steps in scRNA-seq:

1. Quality control

The major challenge in scRNA-seq quality control is the correct identification of viable cells. It is based on the following metrics: count depth, number of genes per barcode, and the fraction of counts from mitochondrial genes. These metrics should not be analyzed separately, their interpretation should always be collective, coupled with the knowledge of the characteristics of biological material that we are investigating. For example, low count depth and high fraction of mitochondrial gene counts may correspond to dead cells in which DNA has leaked out. But, in some cells high fraction of mitochondrial counts is characteristic of their role (respiratory cells). It is also recommended to start with permissive QC and further perform more stringent QC when analyzing h eterogeneous samples [1].

Another common challenge is the presence of doublets/multiplets and empty droplets due to imperfections in the droplet-based sequencing technology. However, empty droplets may be useful in QC for calculating ambient gene expression. This term means having impurities for other barcodes due to the degradation of cells and leakage of their DNA. SoupX leverages this problem for QC correction. [1]

2. Normalization

The goal of data normalization is to remove technical variation to correctly draw further conclusions based on biological variation. Normalization consists of two steps: 1) Scaling - calculating size factors and scaling counts by them. 2) Transformation - addressing the compositional bias. [2-4]

2.1. Scaling

<u>Scaling by library size</u>

Scaling by size factors is used by methods such as CPM (counts-per-million) in which counts are divided by the library size factor and multiplied by 1.0 x 10^6. That method is derived from bulk RNA-seq. However, it is not a perfect method for single cells because it does not account for composition bias. Composition bias affects the direction of separation among clusters. [3,4].

<u>DESeq's size factor</u>

CPM and DESeq size factor are based on the assumption that the majority of genes are not differentially expressed. The DESeq size factor is based on dividing counts by geometric mean calculated across the samples. This method is not suitable for single cells, because of

the low amount of RNA from a single cell (efficacy of extracting RNA also plays a huge role here) we obtain a lot of zero and near-zero counts. These are so-called dropout events. [5]

Spike-in normalization

During the library preparation process spike-in RNA is added to the samples for later scaling to optimize for the equal coverage of spike-in across the barcodes. However, droplet-based technologies do not allow for easy spike-in incorporation. [5]

Normalization by deconvolution

To overcome the problem of dropouts, the counts across the cells are summed. Later the pooled cells are deconvolved. [5]

2.2. Transformation

The aim of transformation is to reduce the domination by highly expressed genes. It allows for data comparison across samples. The most common method is log-transformation. However, this method favors genes with low expression and dampens highly expressed genes. The proposed solution to this problem is using Pearson residuals.. But the advantage of log-transform is that this approach is utilized in later processes [2, 6]

3. Data correction

In the absence of spike-ins we can model UMI counts in the Poisson distribution which will correspond to the technical noise. Again, we are referring to the assumption that variation in most of the genes derives from technical variation rather than biological. [3]
Some biological variation may also arise from cell cycle and/or mitochondrial gene expression. We can correct this by applying linear regression against cell cycle. [1]

The crucial part of data correction is batch effect removal. Batch effect means technical noise that is caused by different environments the experiment was performed across samples. The most popular tool for batch effect correction is ComBat which uses linear models to capture differences in the mean-variance across batches [1, 3].

The next step is to choose variable genes for further downstream processes. The choice of the number of selected genes should depend on the level of heterogeneity of biological material. [3]

4. Dimensionality reduction

To efficiently perform downstream analyses we need to reduce the dimension of our data (which is barcodes x number of genes). Eigengene is a reduced representation of multiple dimensions (genes). The goal of PCA is to find axes in n-dimensional data to describe as much of variation as possible. The main idea behind scRNA-seq PCA is that we see biological variation as a correlation among genes which is present in top PCs. In contrast, technical affects genes independently.

5. Clustering

Graph-based clustering consists of nodes which are representation of cells connected with each other by edges which represent similarity between cells. Graph clustering in comparison to k-means clustering has lower time complexity (logn vs n^2). Moreover, graph clustering does not require a certain distribution of cells in clusters in contrast to other methods (k-means - spherical clusters, Gaussian mixture models - normal distribution). [3]

Hierarchical clustering is an unsuitable method for scRNA-seq data due to the need for the generation of a cell-to-cell distance matrix, which is too computationally intensive. [3]

Subclustering is clustering within a cluster. It is used for obtaining a higher resolution to better separate subtypes of cells. [3]

References:

[1] Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular systems biology, 15(6), e8746. https://doi.org/10.15252/msb.20188746
[2] https://hbctraining.github.io/scRNA-seq_online/lessons/06_SC_SCT_normalization.html
[3] https://bioconductor.org/books/3.17/OSCA.basic/normalization.html
[4] https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Sept23/UnivCambridge_ScRnaSeqIntro_Base/Slides/05_NormalisationSlides.pdf
[5] L. Lun, A.T., Bach, K. & Marioni, J.C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 17, 75 (2016). https://doi.org/10.1186/s13059-016-0947-7
[6] Hafemeister, C., Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 20, 296 (2019). https://doi.org/10.1186/s13059-019-1874-1