scRNAseq analysis with Seurat

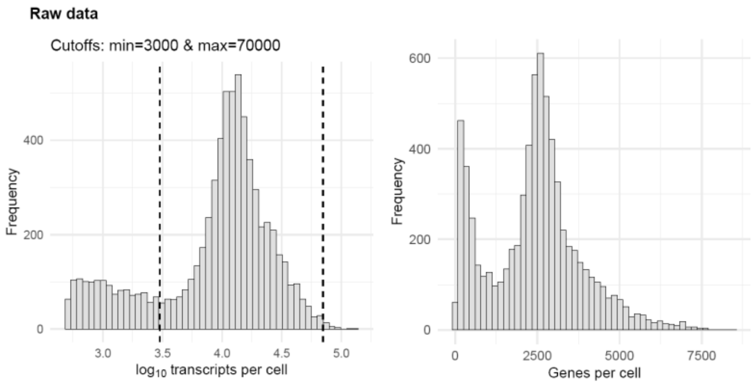
Read more on the analysis based on Seurat : https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html

The analysis folder contains the following folders and files:

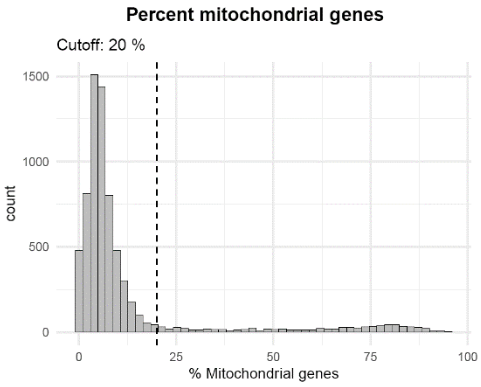
Data QC

Standard preprocessing of scRNAseq data starts with the selection and filtering of cells based on the quality of the data. Low quality cells are generally detected by low number of transcripts, low number of genes and relatively high number of mitochondrial genes. These cells are removed from the dataset using filtering thresholds for transcripts/cell and %mitochondrial genes, determined by the general QC metrics plotted in the graphs explained below.

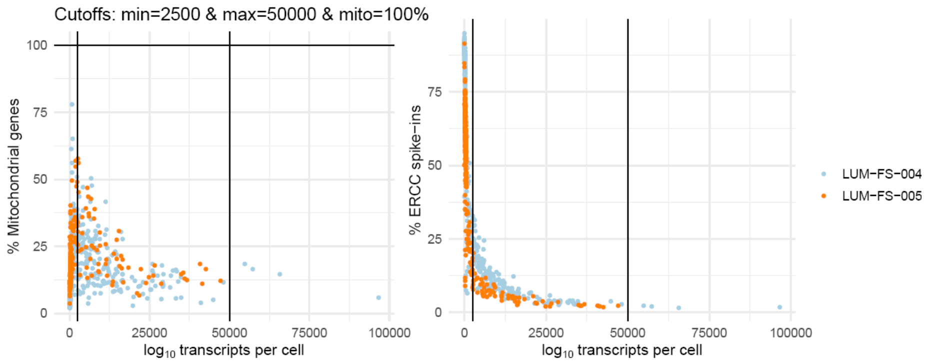
- The histograms of the raw data representing the number of transcripts (also called UMIs) and genes per cell. Dashed lines represent the thresholds used for filtering.



- The density plots represent the same data (number of transcripts and genes) but colored by SORTseq plate (the individual density plots of raw and filtered data per SORTseq plate are saved in separate folders).
- The histogram representing ratio mitochondrial genes/total genes per cell in PercentMito_Raw data. The dashed line shows the filtering threshold if needed to filter out potentially stressed cells with high mitochondrial gene content. *Note that the percent of mitochondrial genes per cell is dependent on the studied cell type.*



- Scatterplots of the number of transcripts per cells vs mitochondrial gene content and ERCC spike-in content, annotated with the filtering thresholds used for transcripts per cell (min and max) and % mitochondrial genes. Cells with low number of transcripts and high mitochondrial genes are removed from the analysis. ERCC spike-ins are technical controls during the library prep and similarly high % of ERCC spike-ins per reaction implies low quality cells.



Using the above filtering thresholds, low quality cells and cells with high number of transcripts (potentially doublets) are filtered out, resulting in the following QC plots:

- Histograms showing the number of transcripts and genes per cell after filtering
- The density plots representing the number of transcripts and genes per cell colored by SORTseq plate (individual density plots per SORTseq plate are saved in a separate folder).
- Violinplot showing the distribution of transcripts and genes per cell per SORTseq plate.
- Filteredcells_plate: text file summarizing the number of cells per plate that meet the filtering thresholds.

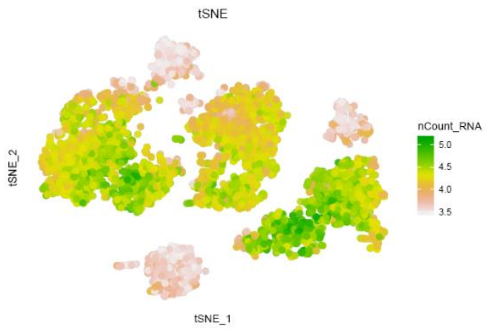
Processing

After filtering out low quality cells, the data is normalized and the most variable genes in the dataset are identified and used for dimensionality reduction and clustering.

- The top2000 highly variable genes are selected (shown in red in Scat_Vargenes, with the top 10 annotated). These genes are saved in the csv file.
- Principal component analysis is done on the 2000 variable genes to identify the dimensionality in the dataset. The chosen principal components for downstream analysis (tSNE and clustering) is shown in Elbow_PCs.

tSNE

Dimensionality reduction with tSNE allows to visualize all the cells in 2 dimensions. This is a non-linear dimension reduction; the structure of the data is preserved in such a way that cells clustered together are similar but the larger distances between the clusters is not linear.



- tSNE_QC – showing the nr of transcripts (nCount), nr of genes (nFeature) and % mitochondrial genes (percent.mt) per cell to evaluate if these factors drive clustering. The example tSNE plot on the right illustrates how cells should not cluster; separate clusters of cells with low transcripts/cell.
- tSNE_Libraries – colored by the SORTseq plate origin of each cell to evaluate plate/batch effects.
- tSNE_Artefacts – shows the expression of genes we often see in cells of either low complexity or low quality (genes which we believe to be mapping artefacts).

Clustering

Seurat uses a graph-based clustering approach based on the identified principal components (abovementioned in "Processing") to group similar cells together.

- tSNE_clusters: tSNEmap colored by the identified clusters
- Contribution_cluster&Plate: the contributions of cells from each SORTseq plate to clusters and vice versa to evaluate batch effects (*i.e.* when cells from SORTseq plates are unexpectedly clustering separately).

Differential expression folder (included in the clustering folder)

To identify the biology driving the identified cell clusters, genes differentially expressed in each cluster are calculated.

- Includes an excel file with the differentially expressed genes per cluster compared to all other clusters (each cluster in a separate tab). Differential expression testing is performed with the parameters listed in DiffExp parameters.txt (more details are included below).

Each excel file includes per gene:
P_val: the pvalue of the statistical test.
p_val_adj: pvalue with Bonferroni correction.
Avg_logFC: log fold change of the average expression in cells in cluster of interest vs rest of the cells (log scale). Positive values indicate that the gene is more highly expressed in the cluster of interest.
pct.1 and pct.2: the % of cells where the genes is detected (pct.1 is cluster of interest, pct.2 is the rest).
logAvExp: Average expression across all cells (log scale).

- The heatmap (HM_clusters_topX&padjY) summarizes the top 10 significant genes per cluster, ranked on fold change.
- The Volcanoplots (average expression across all cells vs average fold change) show the differential expression for each cluster, the top 10 is annotated for each comparison. Genes not significantly expressed (or not meeting the differential expression testing criteria – *i.e.* if a logFC or min.pct threshold is set - are depicted in grey).
- In the subfolders, the expression (transcripts and log transcripts) of the top10 genes per cluster is shown in the tSNE plot.

Parameters used for differential expression testing:

- only.pos: only return positive markers per cluster
- min.pct: testing genes detecting in a minimum fraction of cells in either of the two populations
- test: statistical test used (Wilcoxon Rank Sum test)
- n: top n genes shown per cluster in heatmap and in tSNE
- return.thresh: adjusted p value for significance
- logfc: testing genes which show at least X-fold difference (log-scale) between the two groups of cells (-Inf = no difference)