Chair of Experimental Bioinformatics
TUM School of Life Sciences Weihenstephan

Lecture: Advanced Bioinformatics
SoSe 2022

Exercise Sheet 06 - Single-cell transcriptomics

Total: 15.0 points 21.06.22 - 28.06.22

1. Single-cell technology (1.5 points)

The cell capture rate is the probability that…

- [x] … a droplet has at least one bead
- [ ] … a droplet has at least one cell

The cell duplication rate is the rate …

- [ ] … of cells per droplet
- [ ] … of barcodes per droplet
- [ ] … at which captured single cells are associated with two barcodes
- [x] … at which captured single cells are associated with two or more different barcodes

UMIs

- [ ] … are used to identify individual cells
- [x] … are used to identify individual molecules

2. Barcodes (1.0 + 2.0 bonus points)

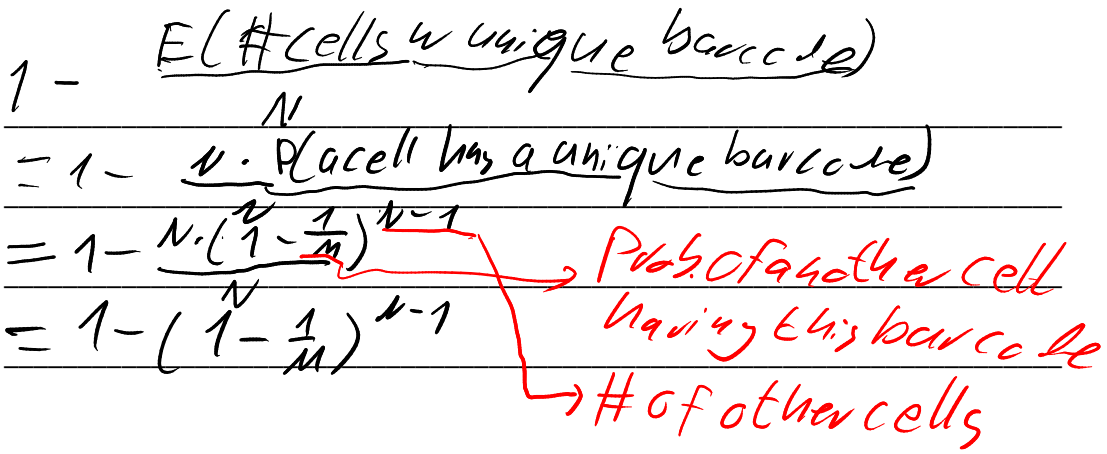
a) What is the minimum barcode length for a synthetic doublet rate of 5% when we assay 1 million cells. (1.0 points)

4^l = number of barcodes of length l
N = number of cells, M = number of barcodes
5% divergence rate -> M = 20*N

$L = \log_4(20*N)$, for 1 million cells:
$L = \log_4(20*1.000.000) = 12.127$

b) Derive the formula for barcode collisions $1 - 1\left(- \frac{1}{M}\right)^{N-1}$ shown in the lecture from first principles. (2.0 bonus points)

Hint: you need the probability for picking a barcode p = 1/M and the binomial coefficient to solve this.



3. Dimensionality reduction (2.5 points, 0.5 each)

The Kullback-Leibler divergence is

- x … used to measure the difference between two probability distributions
- … used to measure the difference between data points in low-dimensional space

The first step of t-SNE is to

- … select neighbours w.r.t. t-distribution over points in high-dimensional space
- x … select neighbours w.r.t. Gaussian distribution over points in high-dimensional space

The first step of UMAP is to

- … approximate a manifold in high-dimensional data using principal components
- x … approximate a manifold in high-dimensional data using simplicial complexes

t-SNE and UMAP are used

- x … for visualization
- … as pre-processing
- … for clustering
- … for cell type annotation

Which of the following are non-linear method(s)

- x t-SNE
- PCA
- SVD
- x UMAP

4. Seurat (10.0 points)

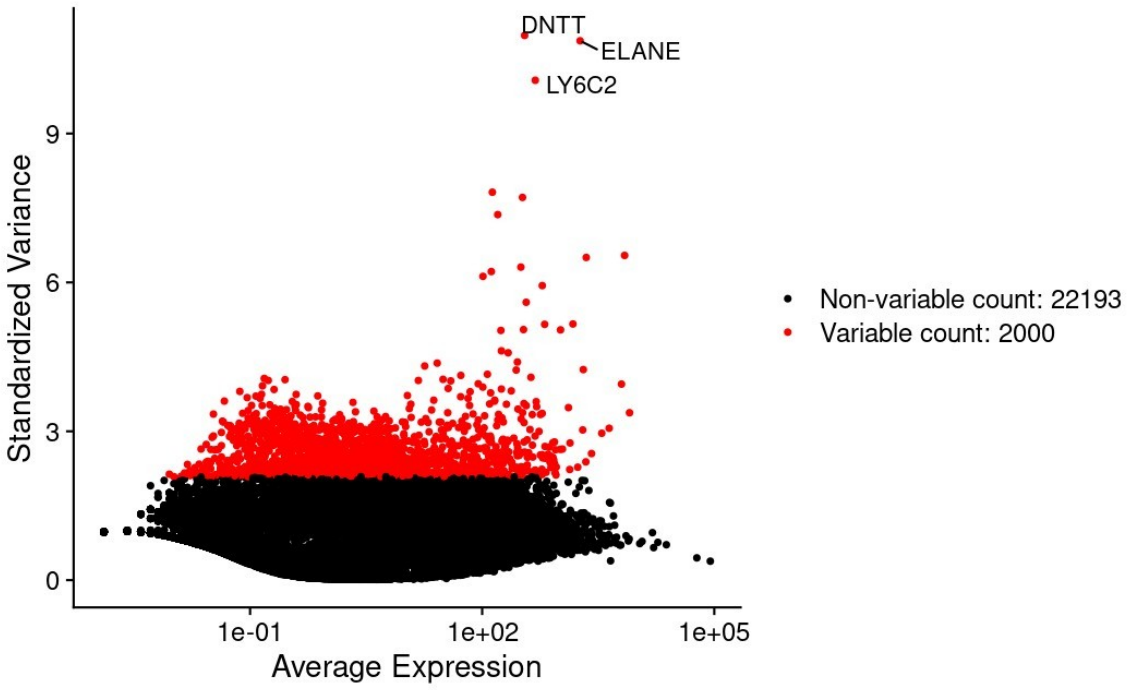
Seurat is a user-friendly R package for analyzing single-cell data. It's documentation comes with easy-to-follow tutorials: https://satijalab.org/seurat/index.html
See this baseline tutorial for more insights, that you will need in this task:
https://satijalab.org/seurat/articles/pbmc3k_tutorial.html
Install the package and answer the following questions (R markdown files usually work best for workflows like this):
Note: Use the default parameters for all functions, if not stated otherwise.

a) Follow the Cell-Cycle scoring and regression vignette to load and pre-process the data. (https://satijalab.org/seurat/articles/cell_cycle_vignette.html). One of the pre-processing steps contains the detection of variable features; explain what this step is doing and name the top 3 variable features. Also add a VariableFeaturesPlot where the three features are labeled. (2.0 points)
Note: do not perform the regression yet, you will need the intermediate results of this section later on.

It creates a mean variability plot and identifies outliers on it

DNTT, ELANE, LY6C2

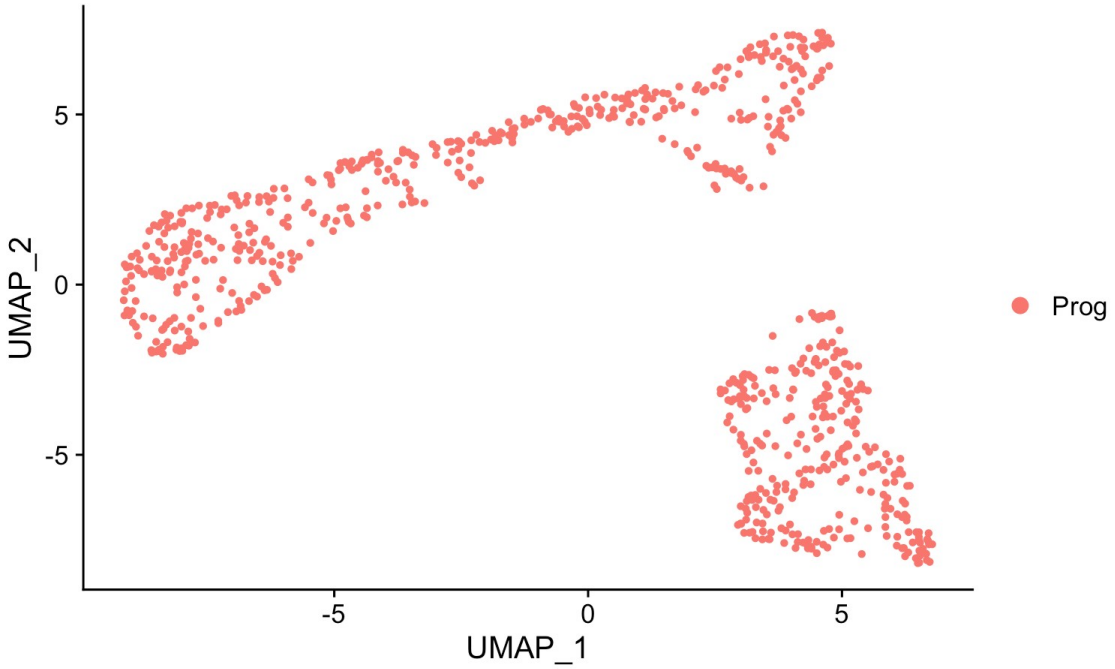
Variable Features Plot:



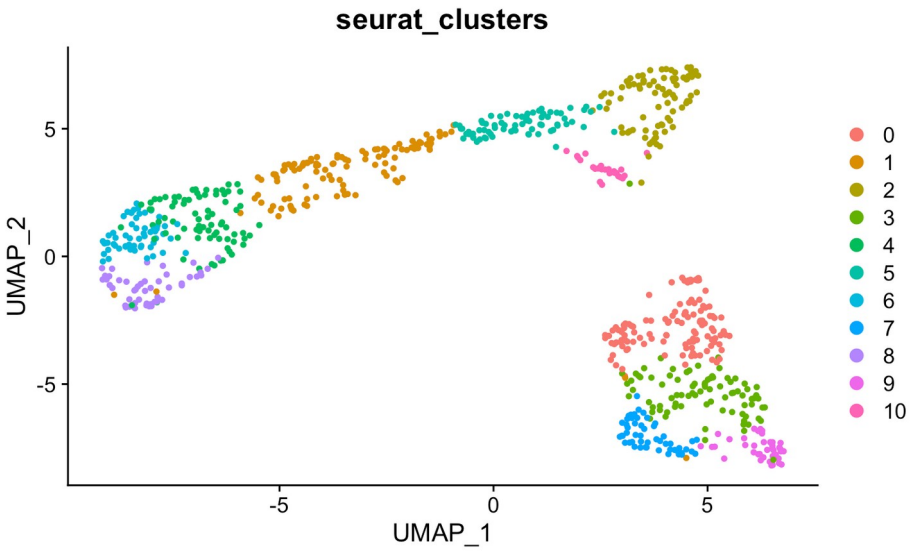
b) Perform the out-regression of cell cycle scores as done in the vignette. Why do we want to regress out cell cycle effects? Do you think this is necessarily a good idea? (2.5 points)

The raw datasets normally contain variation between single measurements that are not interesting for certain analyses.
These might be batch effects, some other technical noise or cell cycle effects. By removing this variation we can improve the downstream analysis and concentrate on the more important variation in the data.
If you need only cells of a certain cell cycle state for your analysis, you can also easily remove the others after the cell cycle detection.
However, one should not be too much relaxed about regressing certain effects out of the single cell data, since the removed variation might still contain important information or be related to other biological factors.

c) Extend your script to generate a UMAP plot based on the dataset that resulted in section a). Submit the plot with your solution (1.0 points).



d) Clustering is a vital part in scRNA-seq analysis. Perform Clustering (FindNeighbors & FindClusters) and color the UMAP from section c) according to the new clusters. Submit the plot with your solution. The second function has a parameter called 'resolution'; explain what it does. Use a resolution of 1 for the final UMAP. (2.0 points)



The resolution parameter changes the number of clusters/communities. Increasing it above 1.0 will result in a higher resolution, i.d. more clusters, while decreasing it results in a more coarse clustering with a few big communities.

e) Find the differentially expressed markers between all clusters (FindAllClusters). Can you find any of the cell-cycle marker genes in the results? Find a cluster, in which marker genes for the G2M cycle are differentially expressed, name the cluster, the genes (2 are enough) and the corresponding log2 Fold-change. (2.5 points)
Hint: You find the maker genes in the lists that are created in the first code-block of the cell-cycle scoring vignette (g2m.genes)

For example in cluster 8:

gene	**log2 Fold-change**
BIRC5	1.2002516
AURKA	1.0628728