Identifying cell and gene candidates in severe COVID-19 patients using single-cell RNA-seq

Instructors: Paulo Czarnewski
Erik Fasterius

Students: Violeta Castelo Szekely
Giada Sandrini
Anchang Charles Gwellem
Anara Alshanbayeva

Identifying cell and gene candidates in severe COVID-19 patients

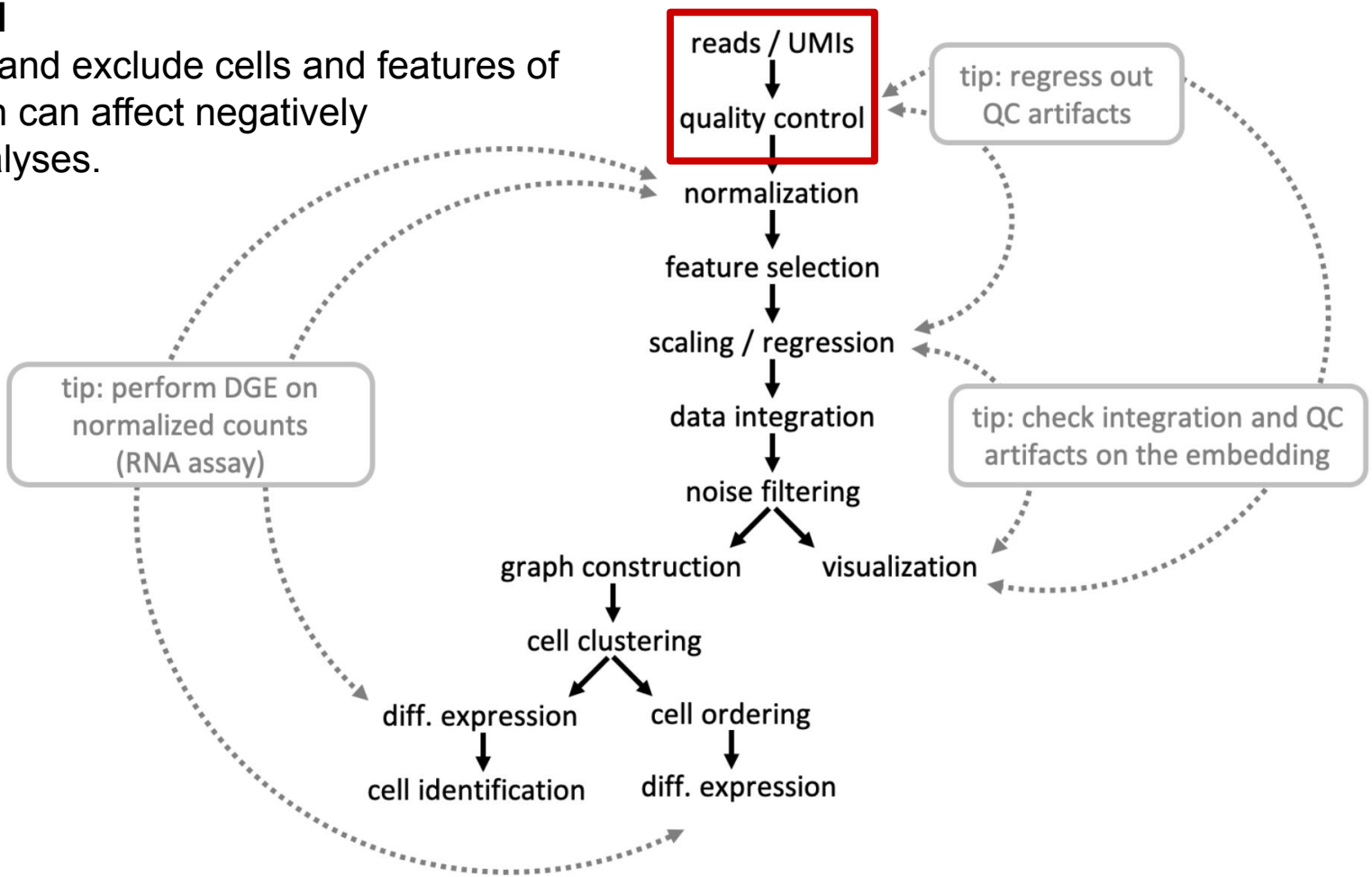
Background: COVID-19 is an infectious disease driven by the virus SARS-CoV-2, which primarily infects lung epithelial cells. However, elderly patients usually develop severe lung inflammation and lung dysfunction, ultimately leading to respiratory failure (Guan et al 2020). The onset of the disease is characterised by a cytokine storm comprising several inflammatory mediators (Pedersen et al 2020), specially in severe cases of the disease. Many cell types orchestrate the immune response to the virus, but their relative contribution at the single-cell resolution is still unclear. Herein, our main goal is to identify which cell types and gene pathways are altered in the blood of patients with severe COVID-19.

Main research question: Which cell types and genes are altered when comparing blood immune cells from healthy versus COVID-19 patients.

Importance: Identifying such genes will allow us to: 1) better understand why severe COVID-19 patients develop stronger immune responses; 2) find potential cells for blockage or immune enhancement therapy or; 3) identify pathways that could be targeted pharmacologically.

Quality control

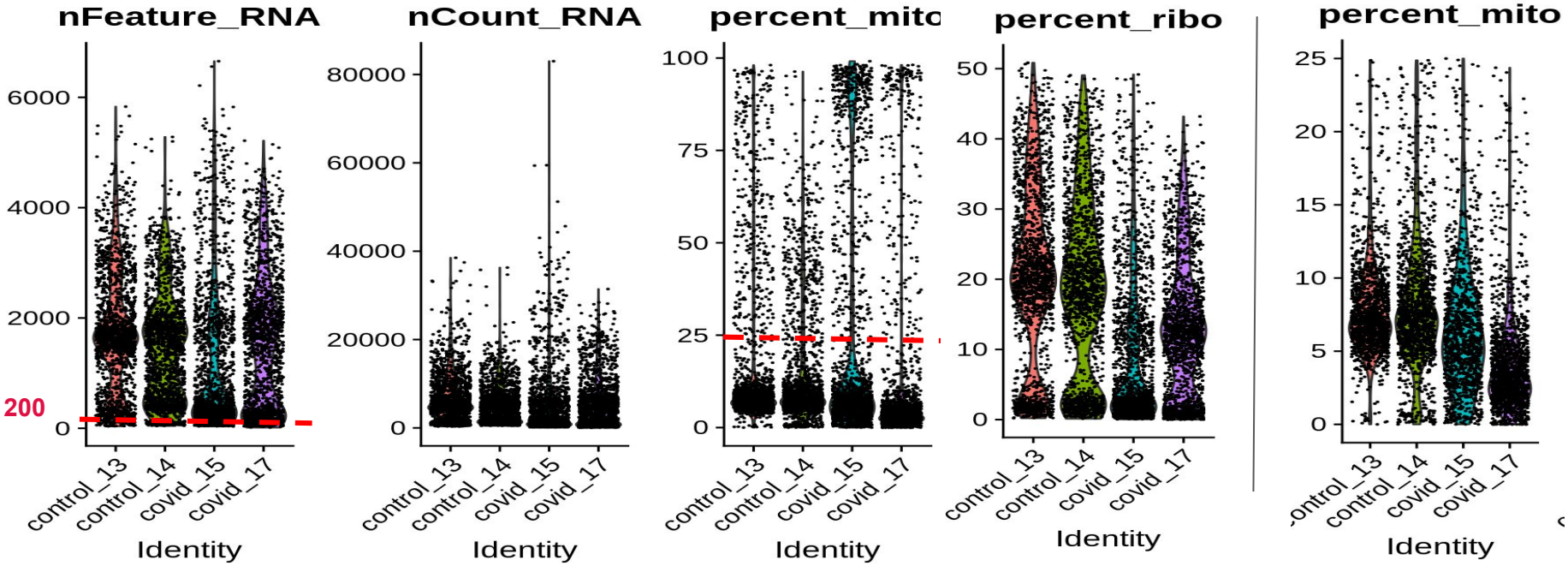
Done to identify and exclude cells and features of low quality which can affect negatively downstream analyses.



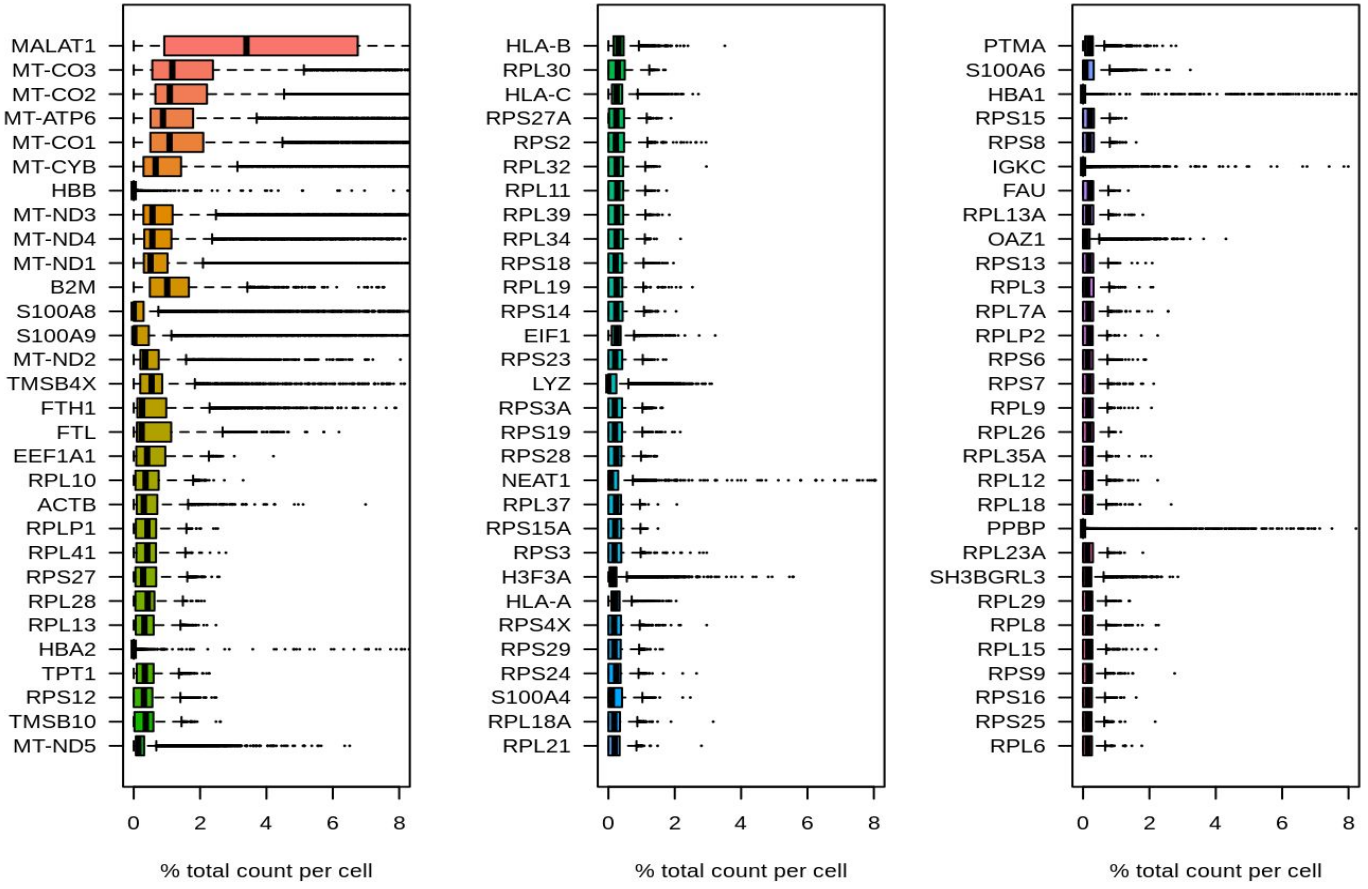
Most of our cells are of good quality taking into account the number of genes, transcripts, mitochondrial genes and ribosomal genes detected from the cells,

Before filtering

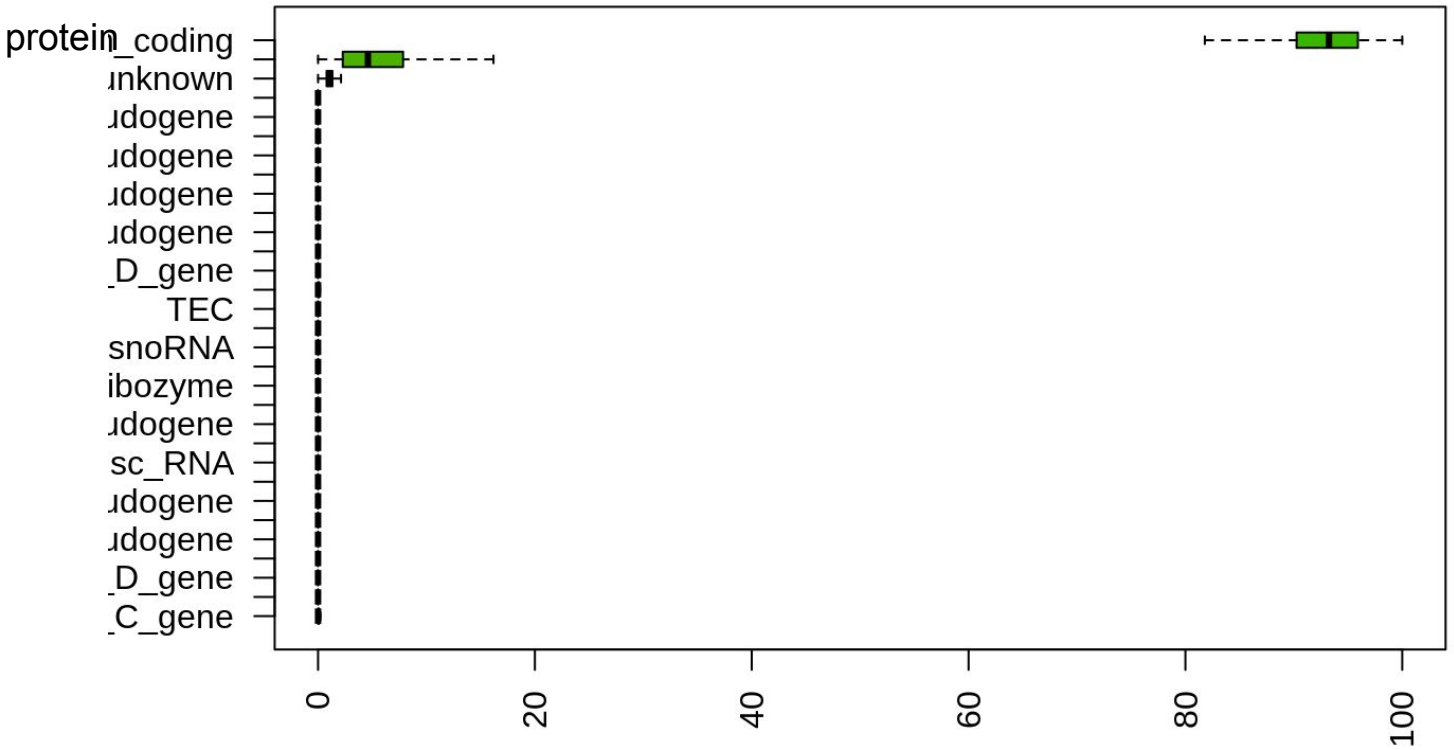
After filtering



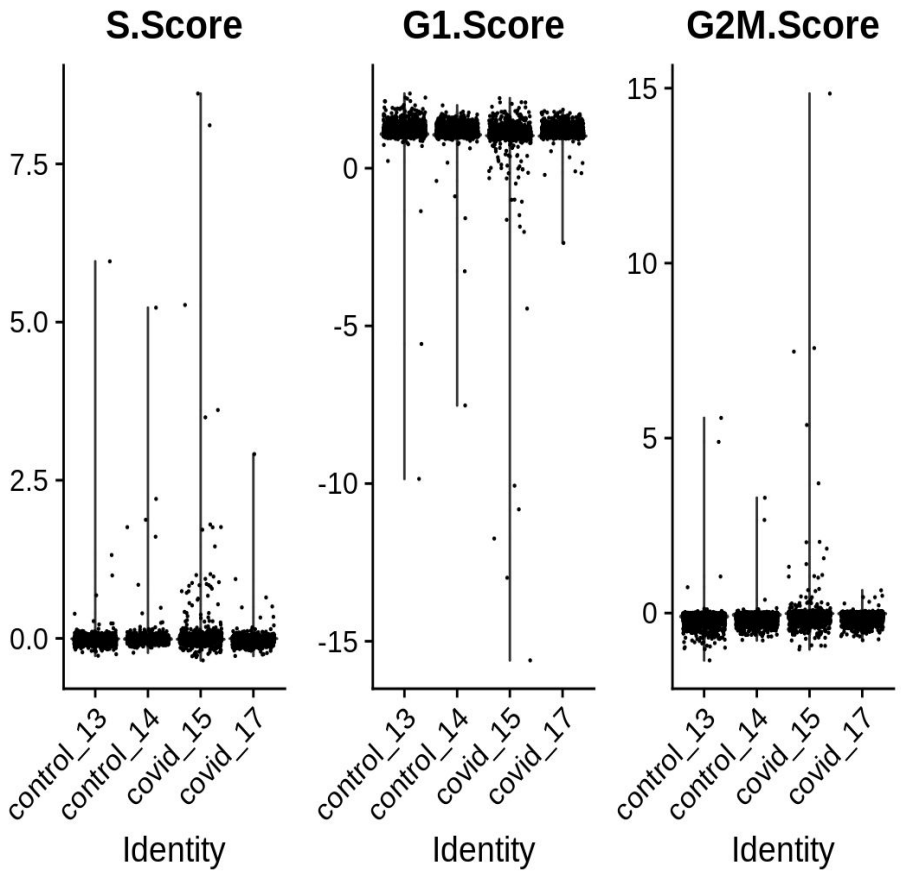
Plot of most highly expressed genes - house-keeping genes, mitochondrial, ribosomal



As expected most genes detected are protein coding genes - the rest are filtered out

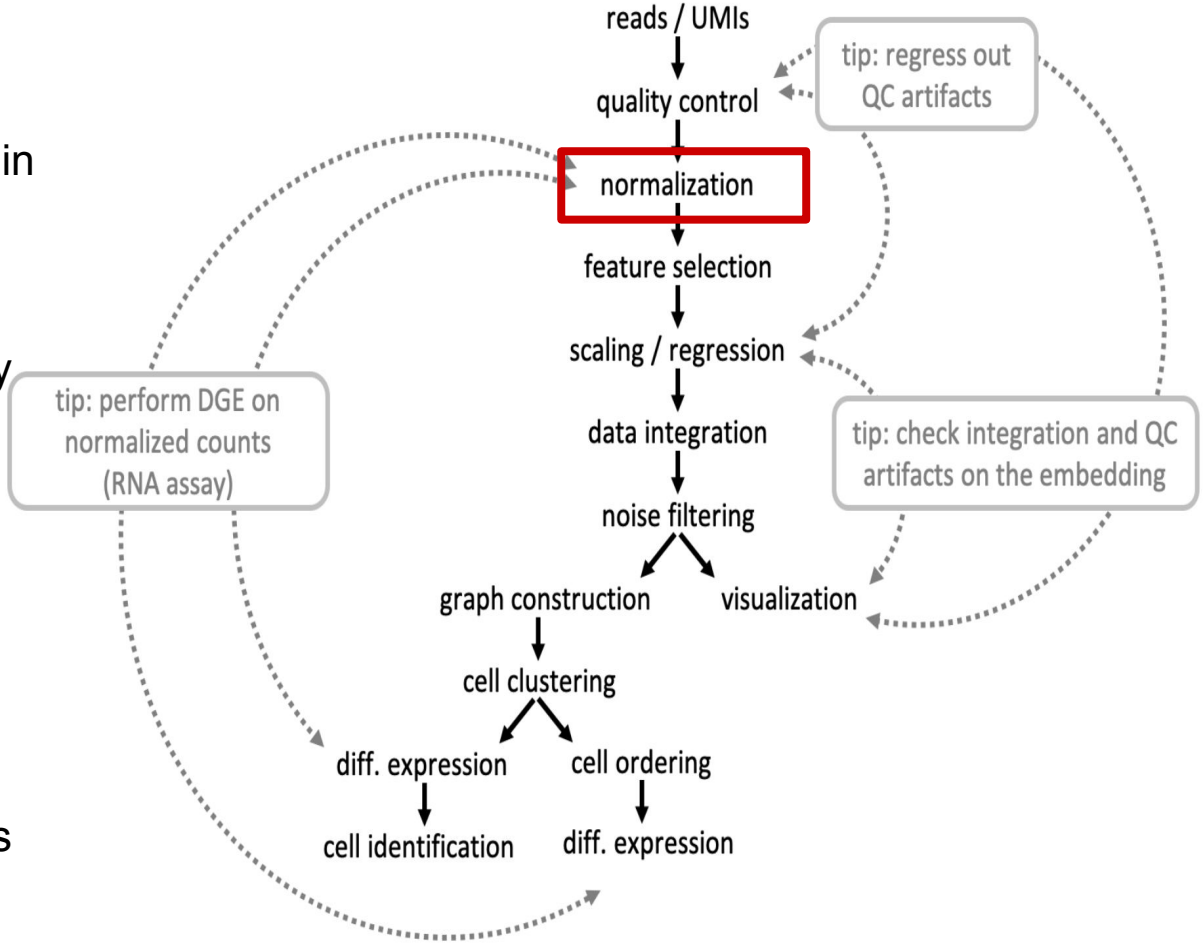


Most cells do not show variation in their cell cycle score



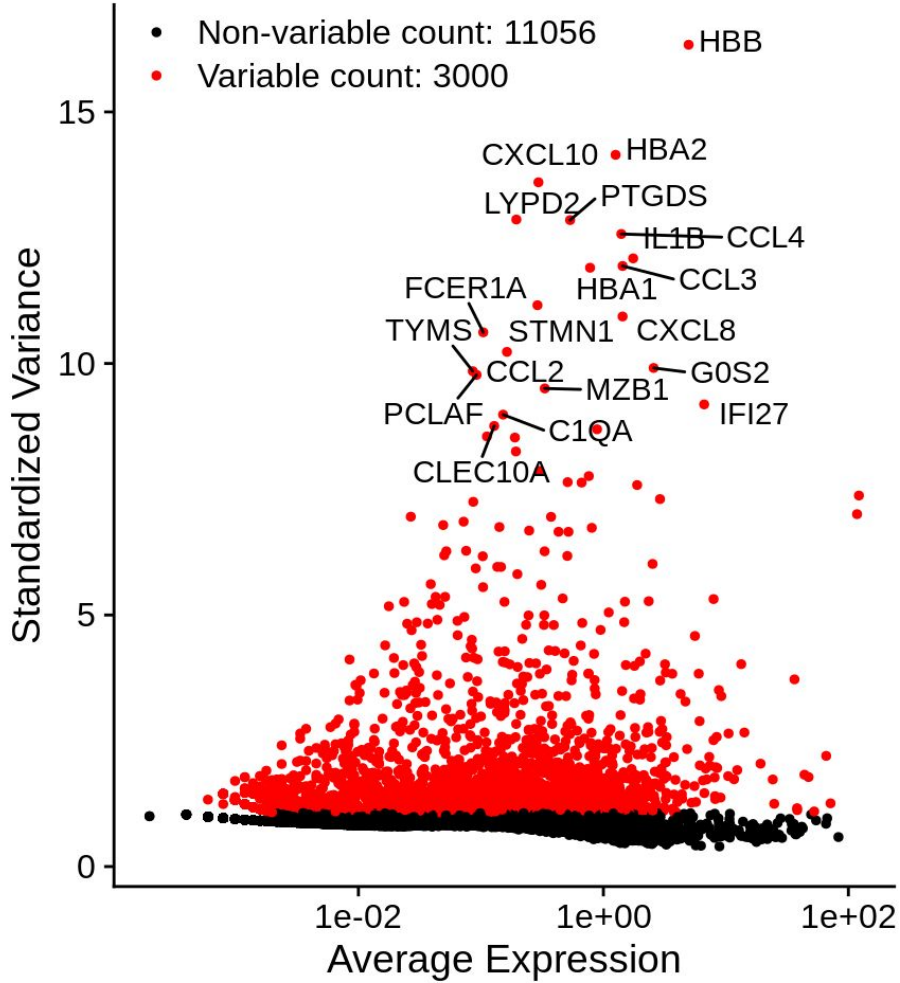
Normalisation is done to

- Compensate for differences in Sequencing depth
- Log-normalisation is done by dividing each gene count by the library size
- The result is multiplied by a constant number - 10000
- Library-size corrected values are log-transformed



Feature selection

- Done to identify features that vary between our samples
- To enable us work with genes informative enough to help us separate our cells
- Highly expressed genes will show higher variance, to normalise this variation, the log of their mean expression is taken

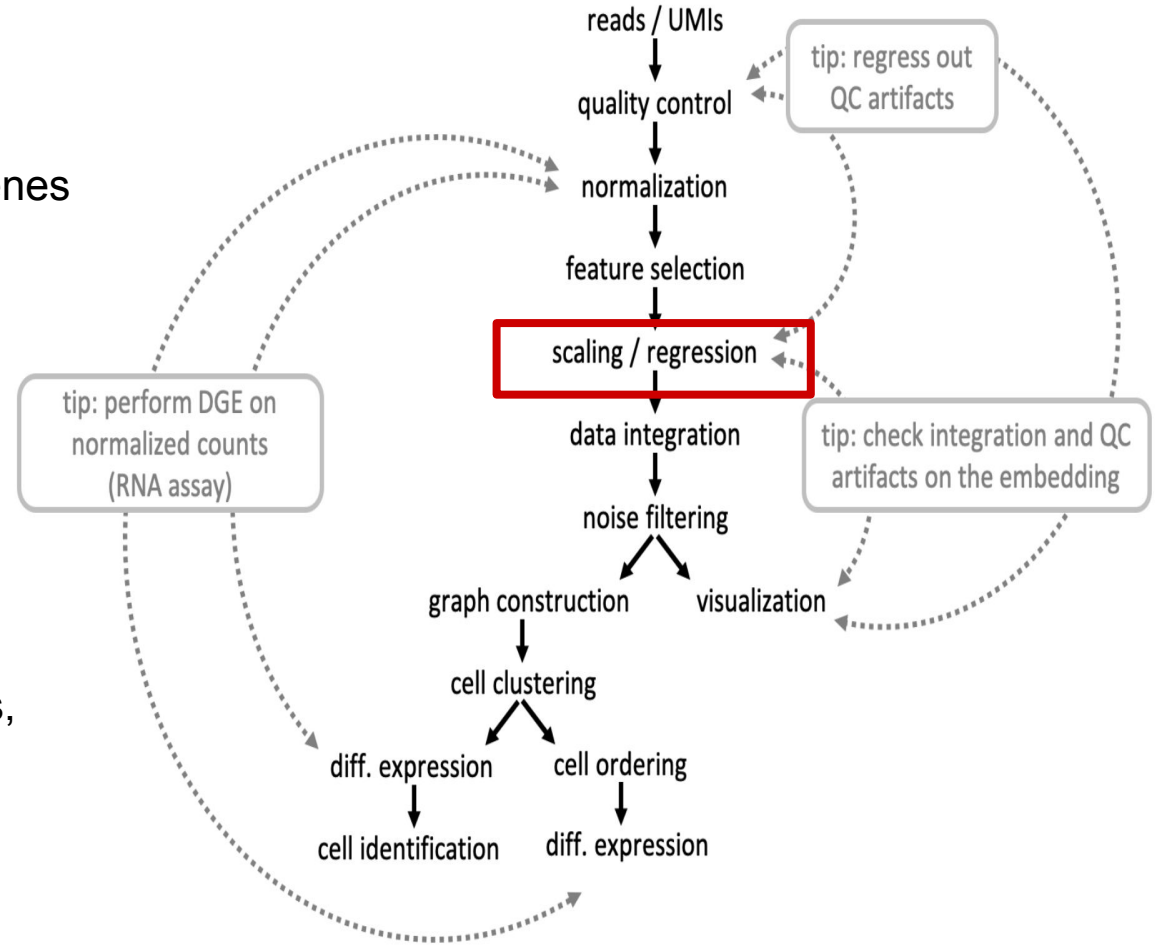


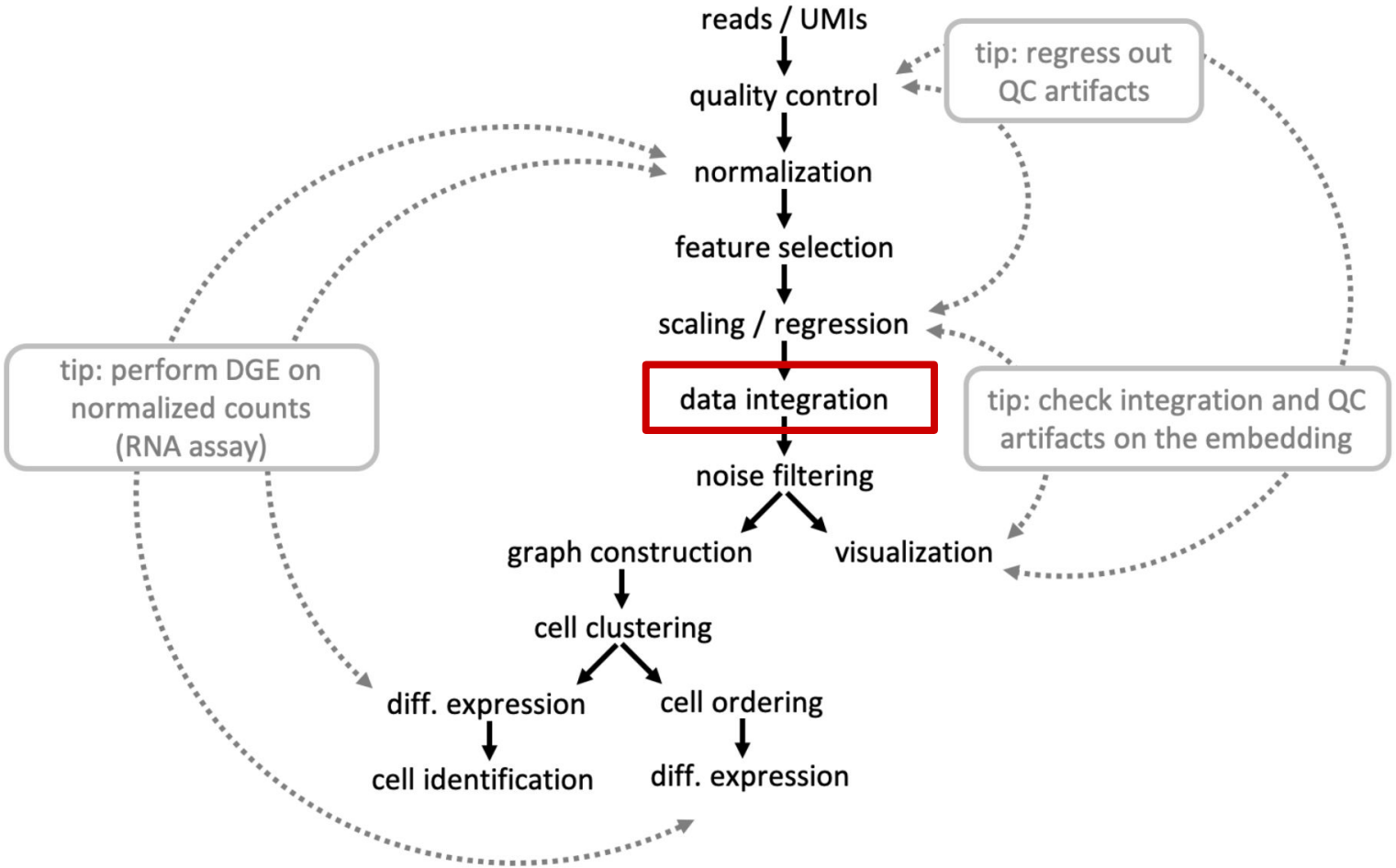
Scaling and linear centering

- Done to give equal weight to all genes irrespective of expression level

- This step can also be used to exclude the effect of certain variables

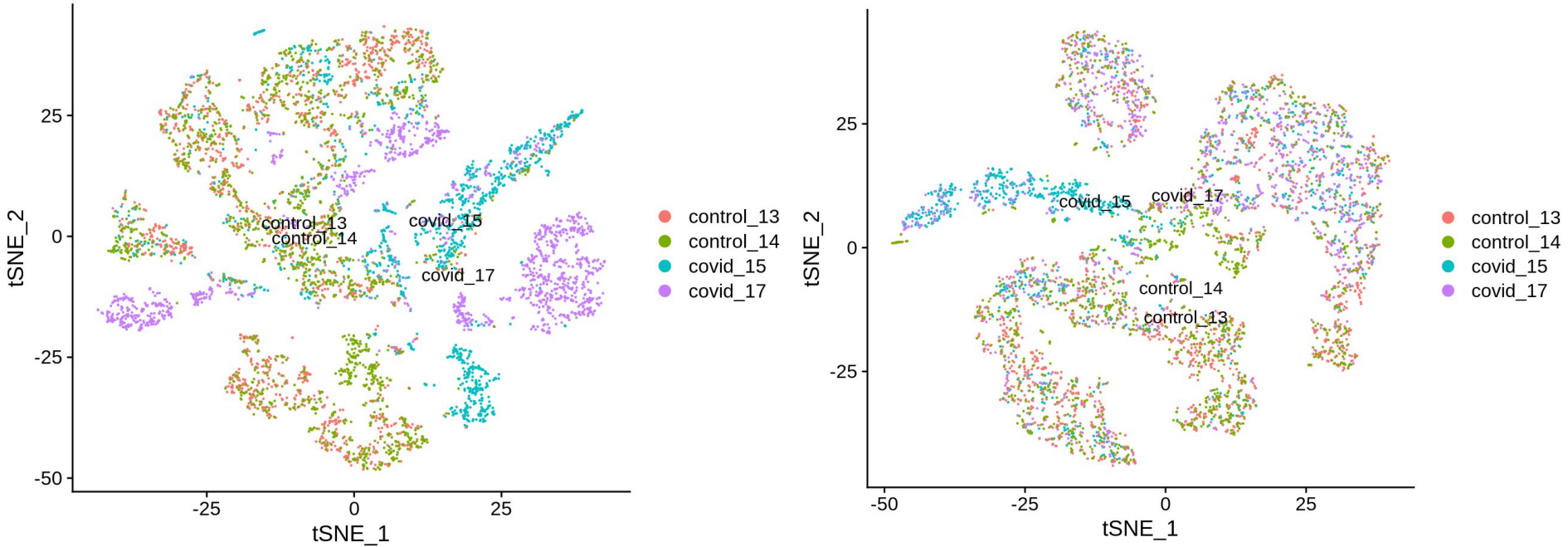
- We regressed out effects due to number of genes, transcripts, percentage of mitochondrial genes, S.Score

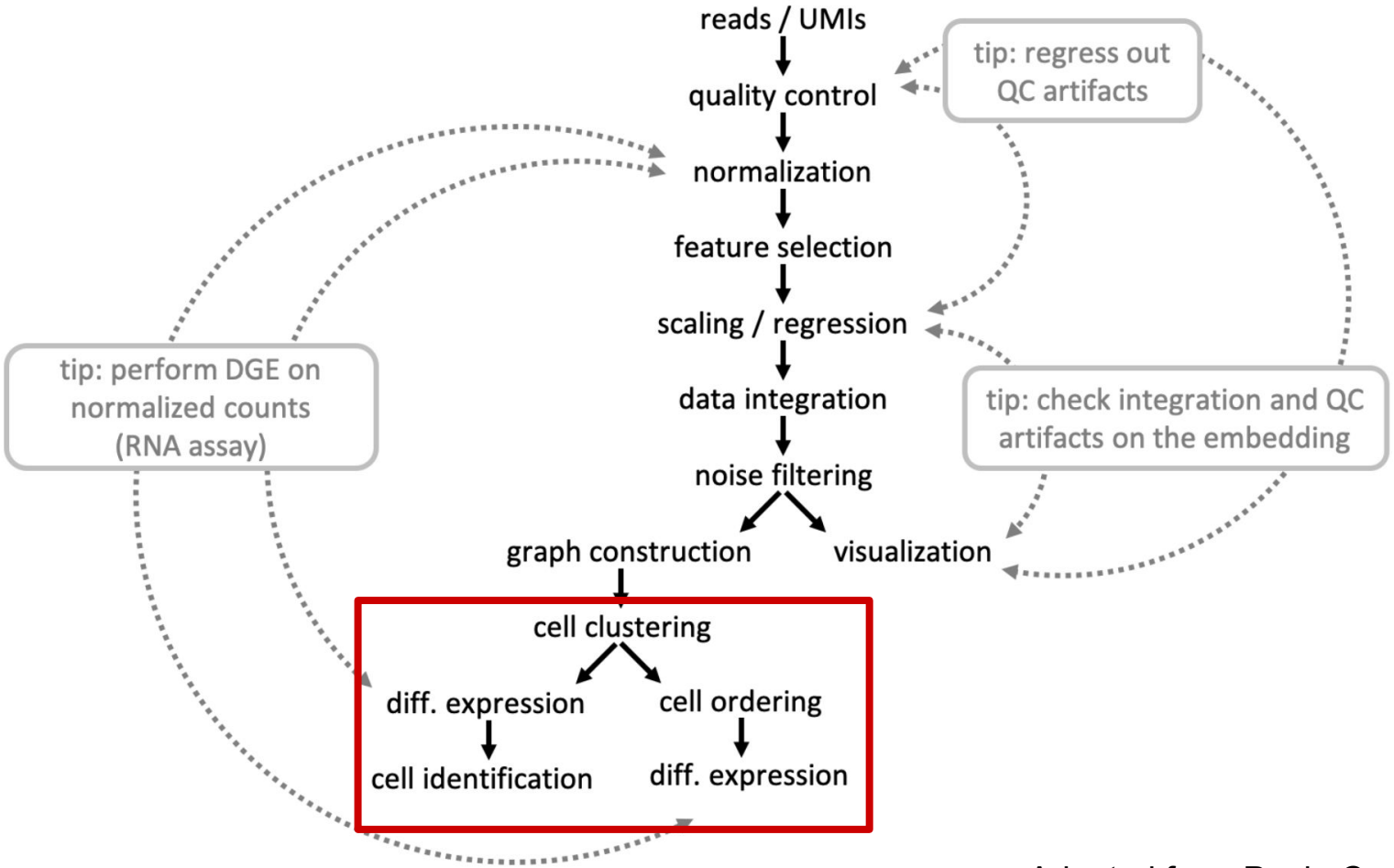



reads / UMIs
quality control
normalization
feature selection
scaling / regression
data integration
noise filtering
graph construction
visualization
cell clustering
diff. expression
cell ordering
cell identification
diff. expression
tip: regress out QC artifacts
tip: perform DGE on normalized counts (RNA assay)
tip: check integration and QC artifacts on the embedding

Data integration

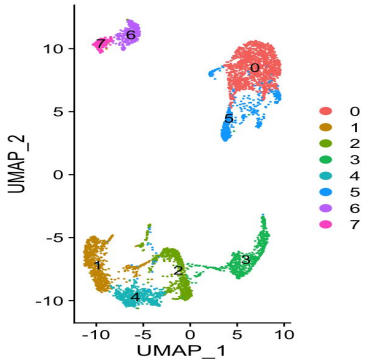
Remove batch effect





Adapted from Paulo Czarnewski

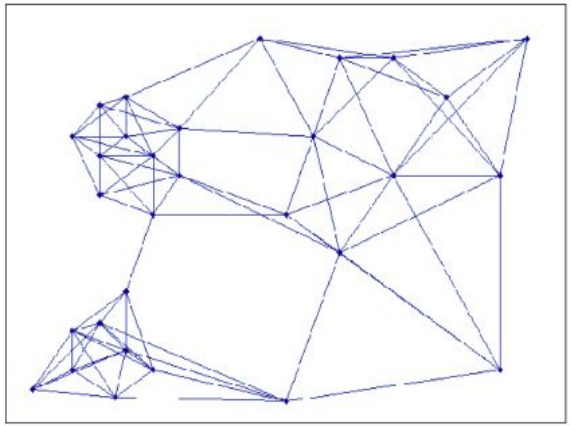
Cell Clustering

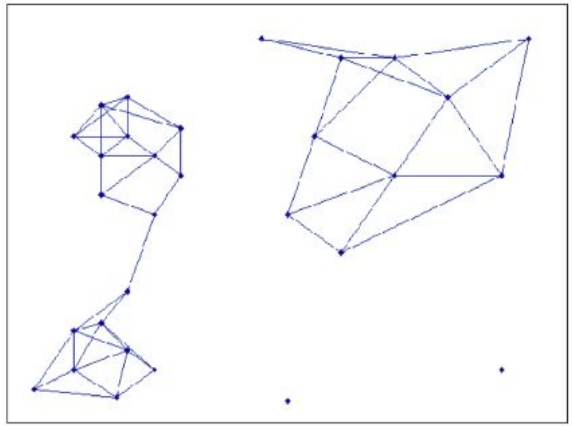


scRNA-seq – graph construction and clustering

The ***k*-Nearest Neighbor (*k*NN)** graph is a graph in which two vertices *p* and *q* are connected by an edge, if the distance between *p* and *q* is among the *k*-th smallest distances from *p* to other objects from *P*.

The **Shared Nearest Neighbor (SNN)** graph has weights that defines proximity, or similarity between two edges in terms of the number of neighbors (i.e., directly connected vertices) they have in common.

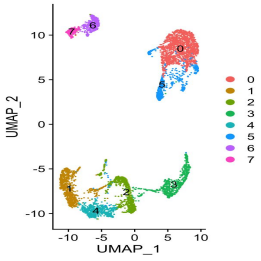




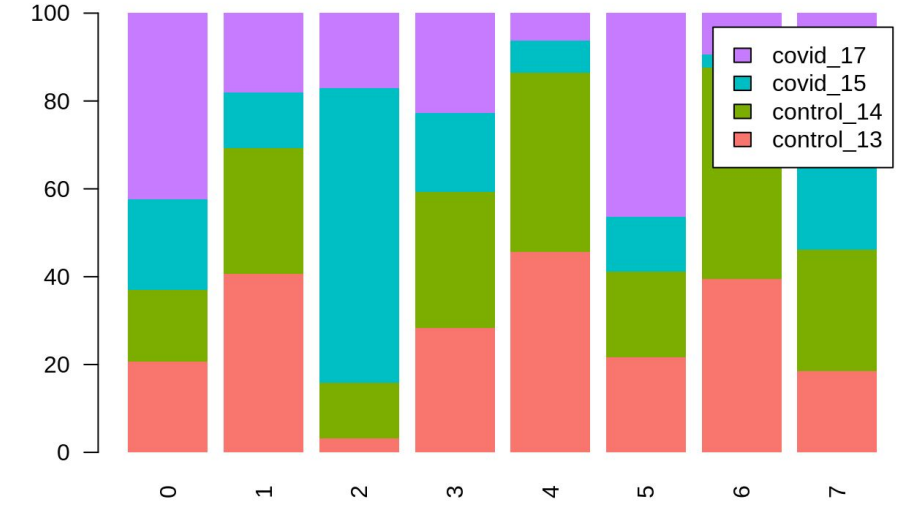
Adapted from Paulo Czarnewski

Cell Clustering

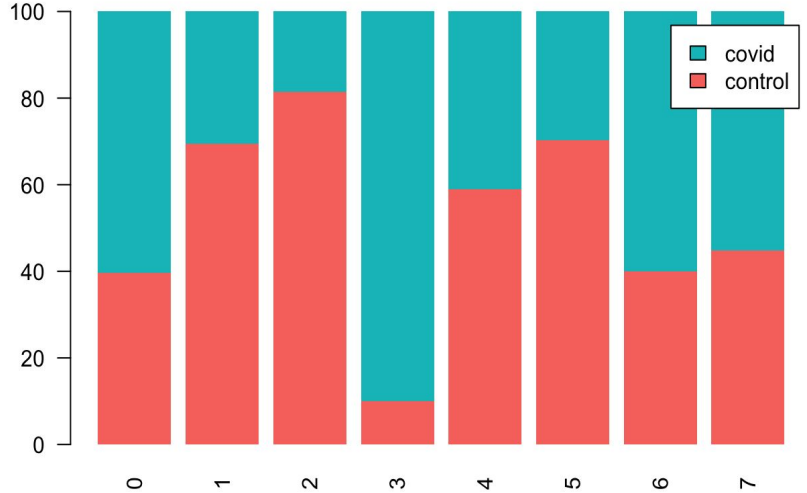
Clusters divided by **sample composition (batch) (a)** or by condition (covid/control) (b)



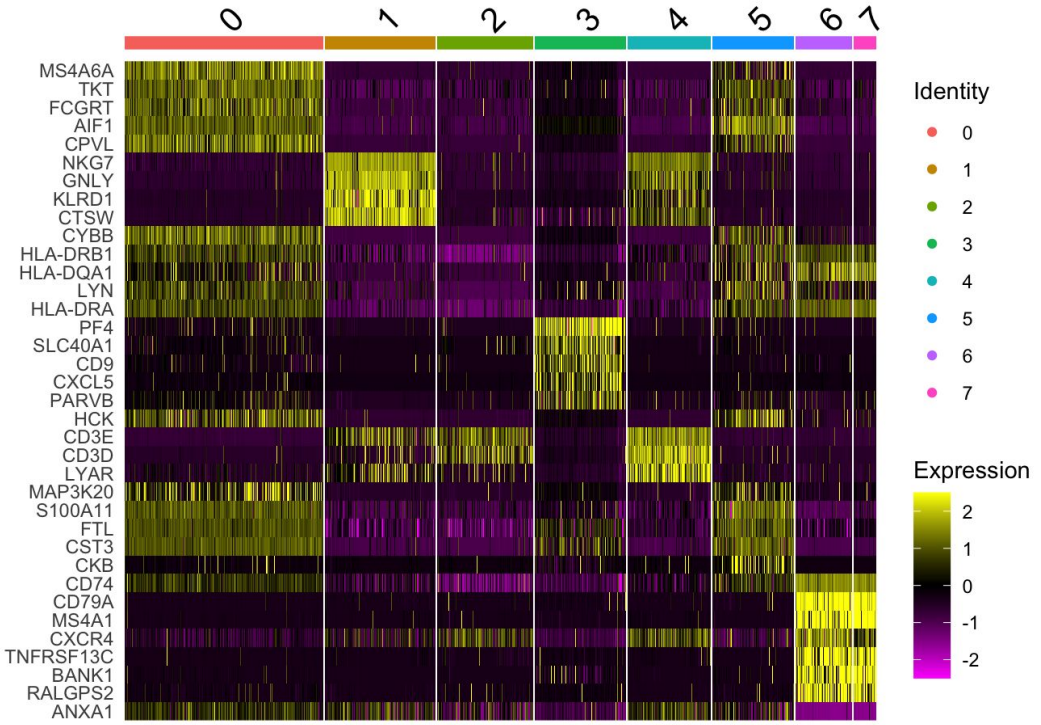
(a)

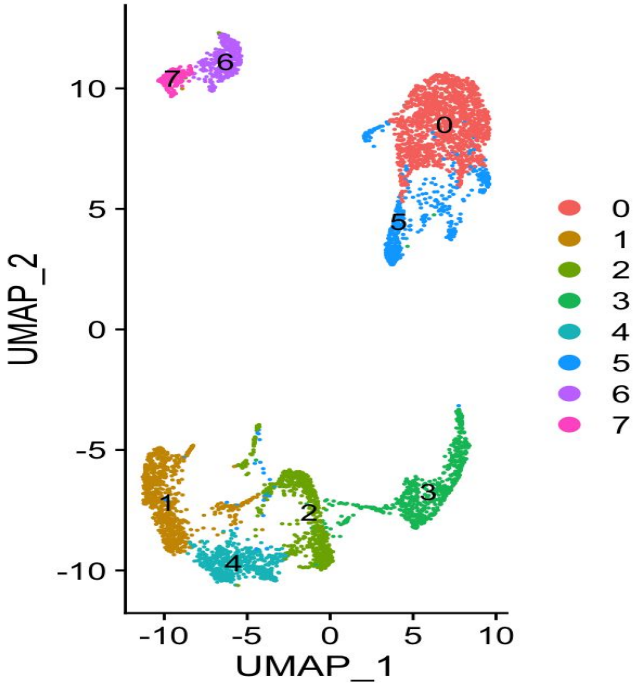


(b)

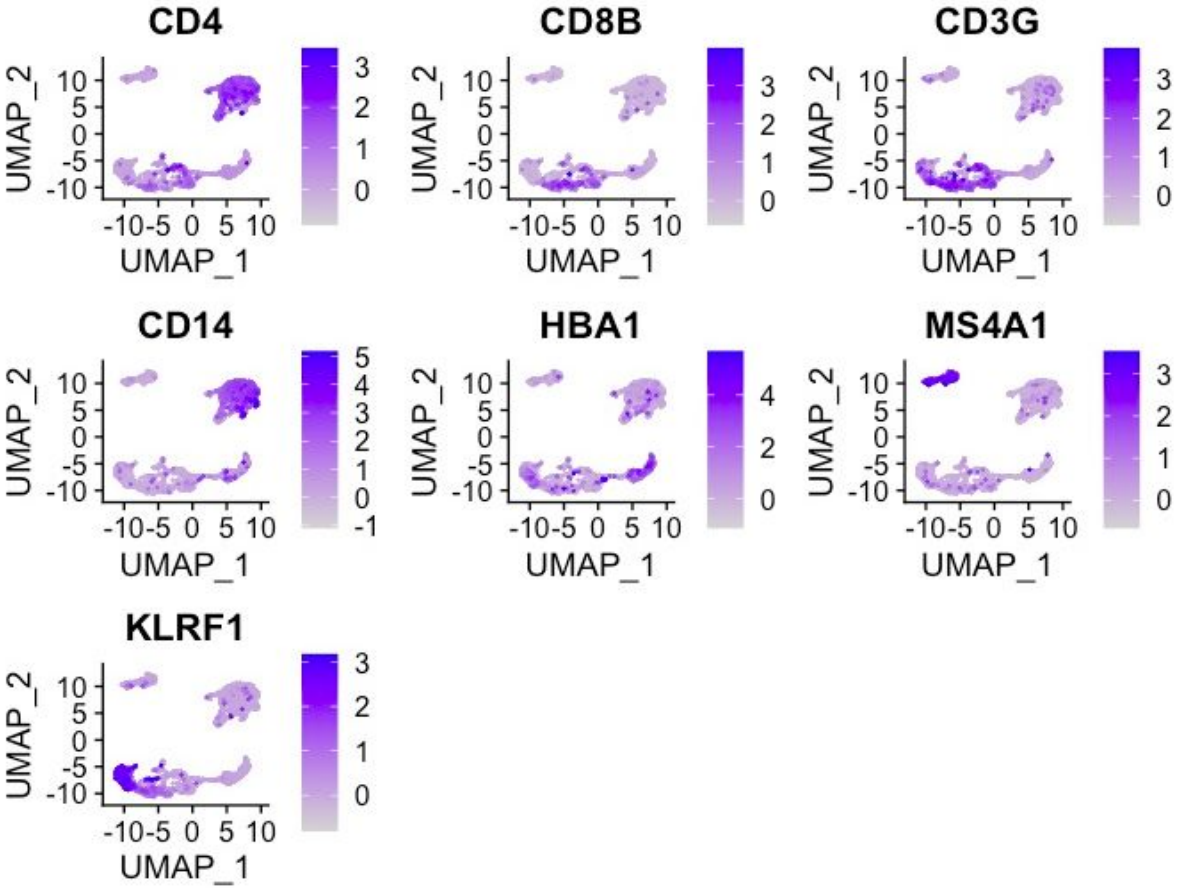


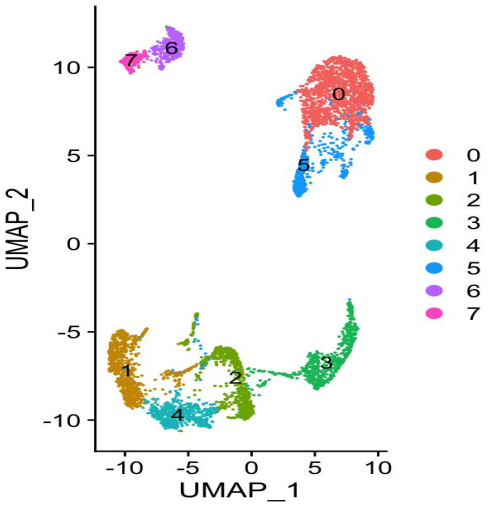
Cell Clustering



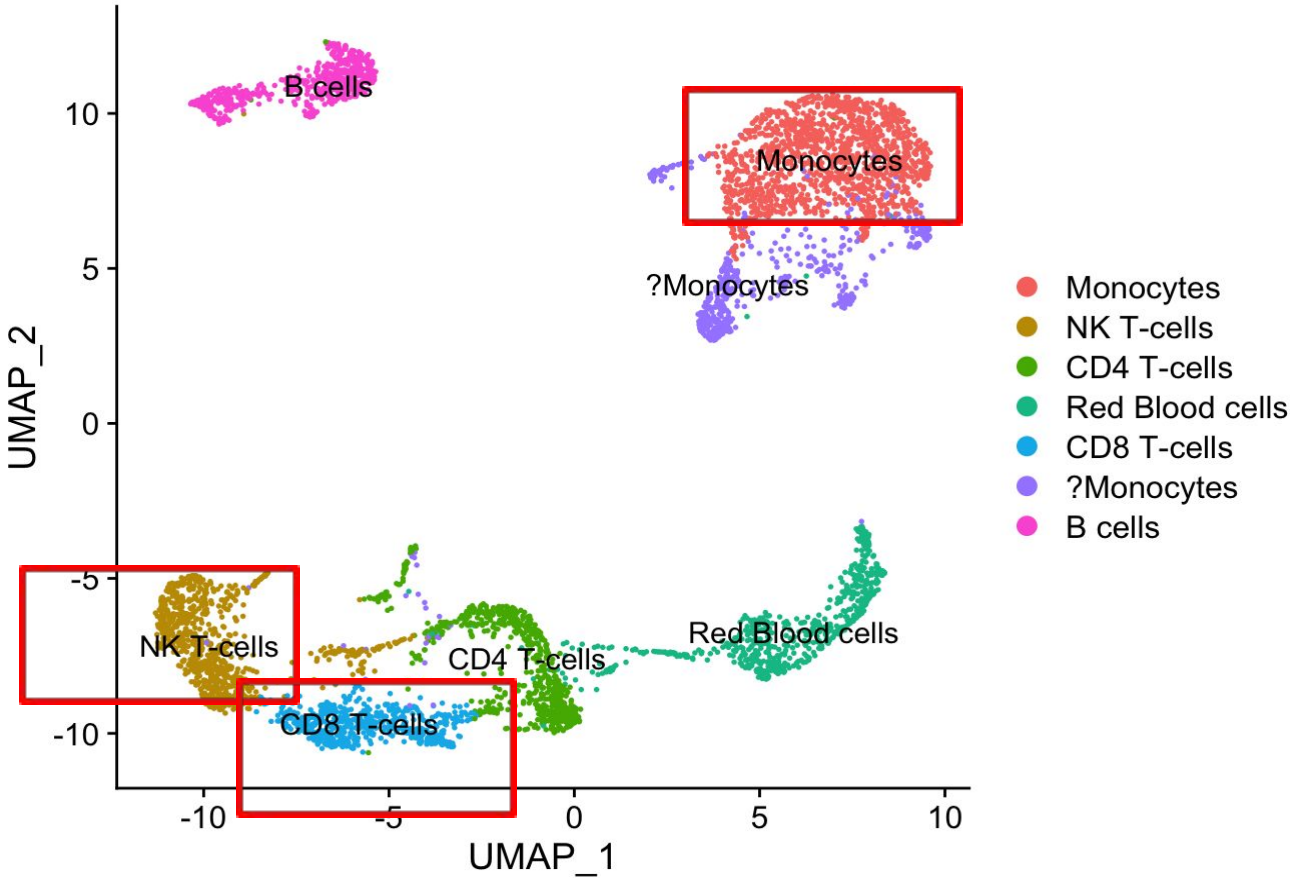


Differential expression analysis Type-1: Identifying cell populations (DEA among clusters)

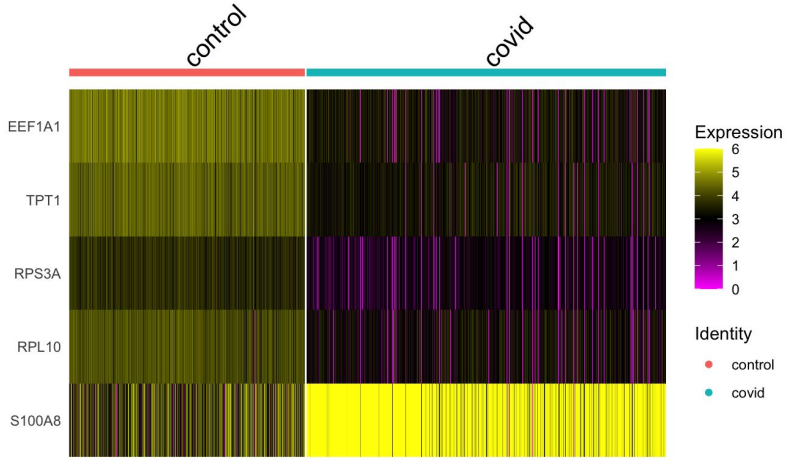


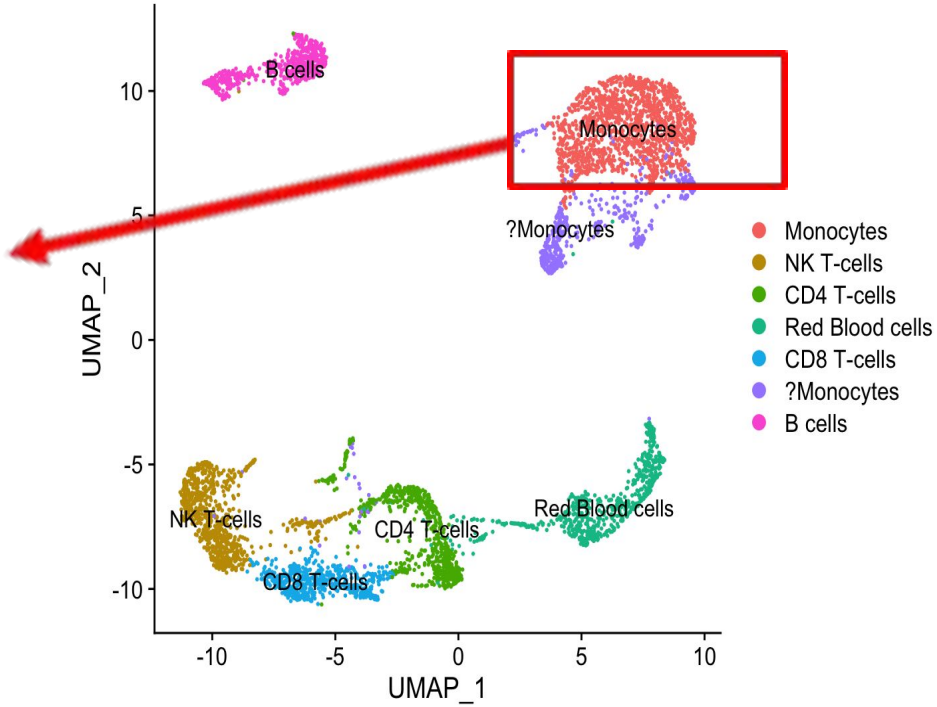


Differential Expression Analysis Type-2: identification between conditions (DEA within Clusters)



Cluster Specific Differential Expression Analysis





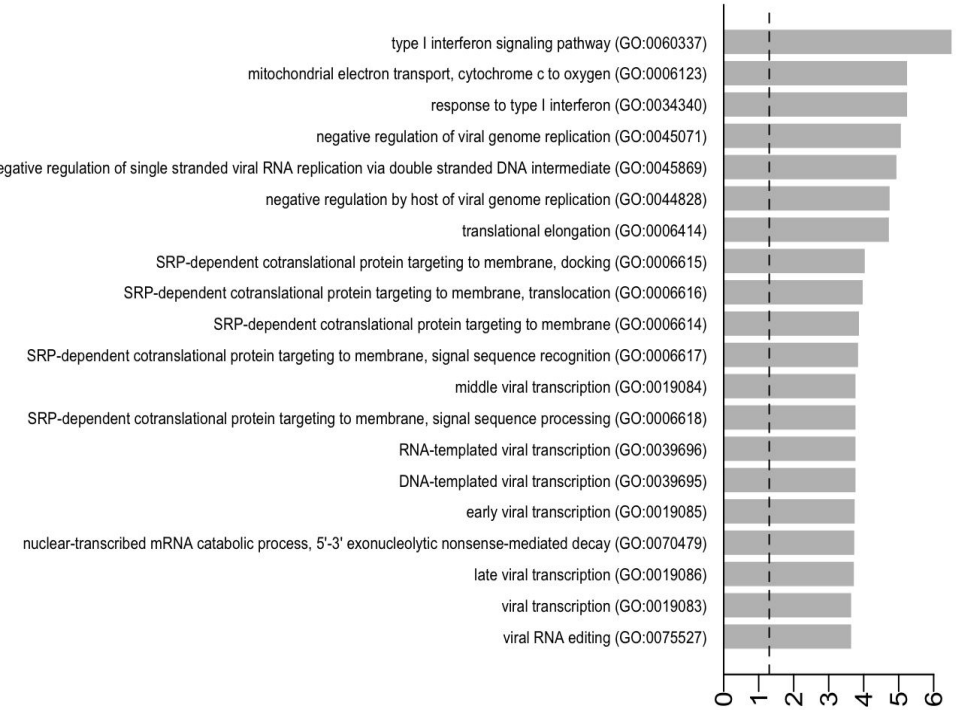
Exemplary databases:

- GO_Biological_Process_2017b
- KEGG_2019_Human
- WikiPathways_2019_Human

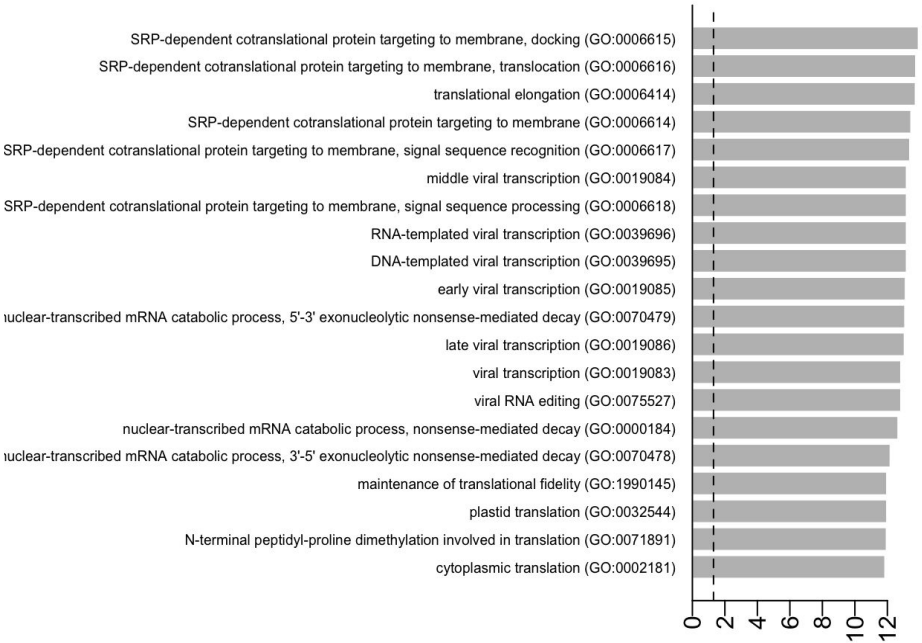
Differential Expression Analysis (Covid vs Control)

Biological Processes (corresponding to upregulated genes) in NK T-cells (a) and CD4 T-cells (b)

(a)

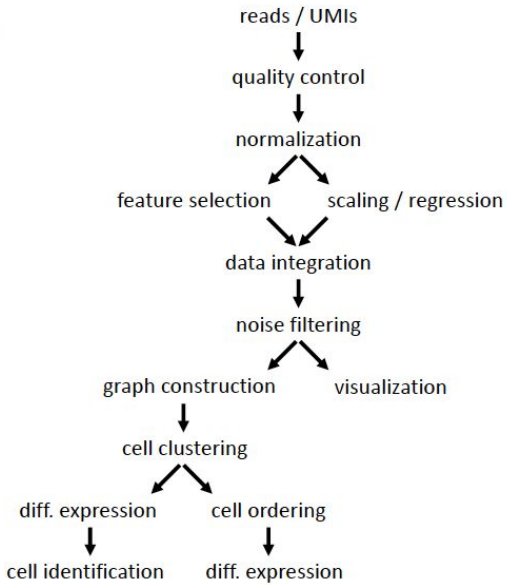


(b)



Take-home messages:

scRNA-seq analysis workflow



Research Question:

Which cell types and genes are altered when comparing blood immune cells from healthy versus COVID-19 patients

Importance

1) better understand why severe COVID-19 patients develop stronger immune responses (identified DE-genes)

2) find potential cells for blockage or immune enhancement therapy (NK T-cells, NK cells)

3) identify pathways that could be targeted pharmacologically