



# Analysing single-cell data using R

Sini Junntila, PhD



# Outline

- Brief introduction to single-cell analysis tools
- Basic analysis workflow
- Reading in data and QC
- Normalization and highly variable features
- Clustering and visualization
- Integration
- Automatic annotation
- Differential expression analysis
- Trajectory analysis with Totem
- Session website: [https://github.com/elolab/Olissipo\\_hands-on\\_session2](https://github.com/elolab/Olissipo_hands-on_session2)

# Single-cell analysis resources in R

- Orchestrating Single-Cell Analysis with Bioconductor (OSCA)
  - <http://bioconductor.org/books/release/OSCA/>
  - Teaches users some common workflows for the analysis of single-cell RNA-seq data (scRNA-seq)
  - Shows you how to make use of cutting-edge Bioconductor tools to process, analyze, visualize, and explore scRNA-seq data
- Seurat
  - <https://satijalab.org/seurat/index.html>
  - Seurat is an R package designed for QC, analysis, and exploration of single-cell RNA-seq data
  - Seurat aims to enable users to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements, and to integrate diverse types of single-cell data

# Single-cell analysis resources outside R

- Scanpy
  - <https://scanpy.readthedocs.io/en/latest/index.html>
  - Scanpy is a scalable toolkit for analyzing single-cell gene expression data
  - It includes preprocessing, visualization, clustering, trajectory inference and differential expression testing
  - The Python-based implementation efficiently deals with datasets of more than one million cells
- Single-cell best practices
  - <https://www.sc-best-practices.org/preamble.html>
  - The goal of this book is to teach newcomers and advanced professionals alike, the best practices of single-cell sequencing analysis
  - This book will teach you the most common analysis steps ranging from preprocessing to visualization to statistical evaluation and beyond
- Cell Ranger
  - <https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger>
  - Cell Ranger is a set of analysis pipelines that process Chromium single cell data to align reads, generate feature-barcode matrices, perform clustering and other secondary analysis, and more

# Public scRNAseq data

- Gene Expression Omnibus (GEO)
  - <https://www.ncbi.nlm.nih.gov/geo/>
  - Contains processed data
  - FASTQ files usually at SRA
    - <https://www.ncbi.nlm.nih.gov/sra/>
- Single Cell Expression Atlas
  - <https://www.ebi.ac.uk/gxa/sc/home>
- Single Cell Portal
  - [https://singlecell.broadinstitute.org/single\\_cell](https://singlecell.broadinstitute.org/single_cell)

# Our data

- Lee et al: Sci Immunol. 2020 Jul 10;5(49):eabd1554.doi: 10.1126/sciimmunol.abd1554
- GSE149689
- 4 healthy controls, 11 COVID-19 patients, 5 flu patients
- Download the data
  - [https://bioinfoshare.utu.fi/DataTransfer/Olissipo\\_hands-on\\_session2/data](https://bioinfoshare.utu.fi/DataTransfer/Olissipo_hands-on_session2/data)
  - Size: 468M
  - Username: btk
  - Password: WzfwWGUMo7

SCIENCE IMMUNOLOGY | RESEARCH ARTICLE

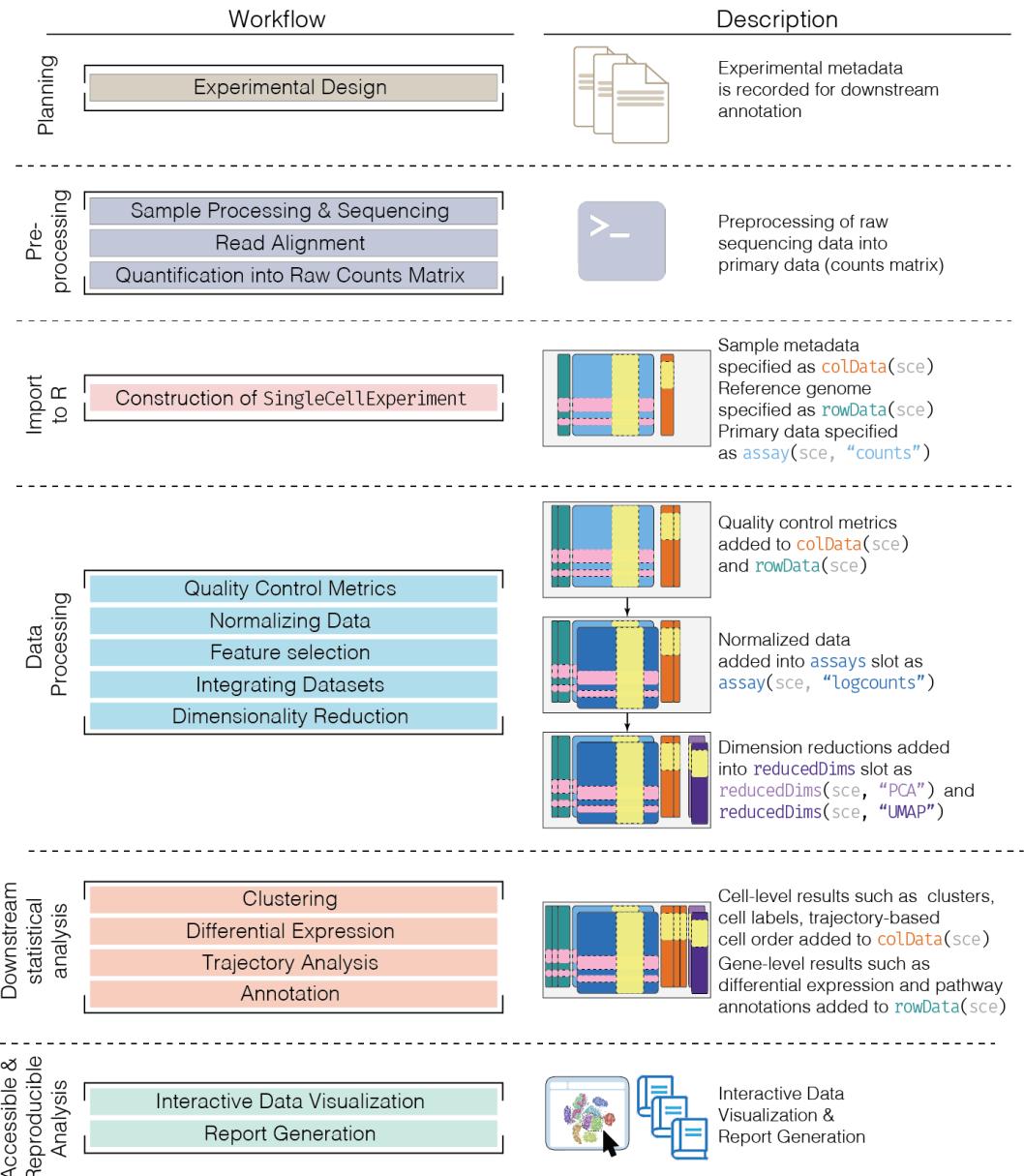
CORONAVIRUS

## Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19

Jeong Seok Lee<sup>1\*</sup>, Seongwan Park<sup>2\*</sup>, Hye Won Jeong<sup>3\*</sup>, Jin Young Ahn<sup>4\*</sup>, Seong Jin Choi<sup>1</sup>, Hoyoung Lee<sup>1</sup>, Baekgyu Choi<sup>2</sup>, Su Kyung Nam<sup>2</sup>, Moa Sa<sup>1,5</sup>, Ji-Soo Kwon<sup>1,6</sup>, Su Jin Jeong<sup>4</sup>, Heung Kyu Lee<sup>1,5</sup>, Sung Ho Park<sup>7</sup>, Su-Hyung Park<sup>1,5</sup>, Jun Yong Choi<sup>4†</sup>, Sung-Han Kim<sup>6†</sup>, Inkyung Jung<sup>2†</sup>, Eui-Cheol Shin<sup>1,5†</sup>

# Basic data analysis workflow

- Script:  
`Lee_data_analysis_with_seurat.R`



<http://bioconductor.org/books/3.16/OSCA.intro/analysis-overview.html>

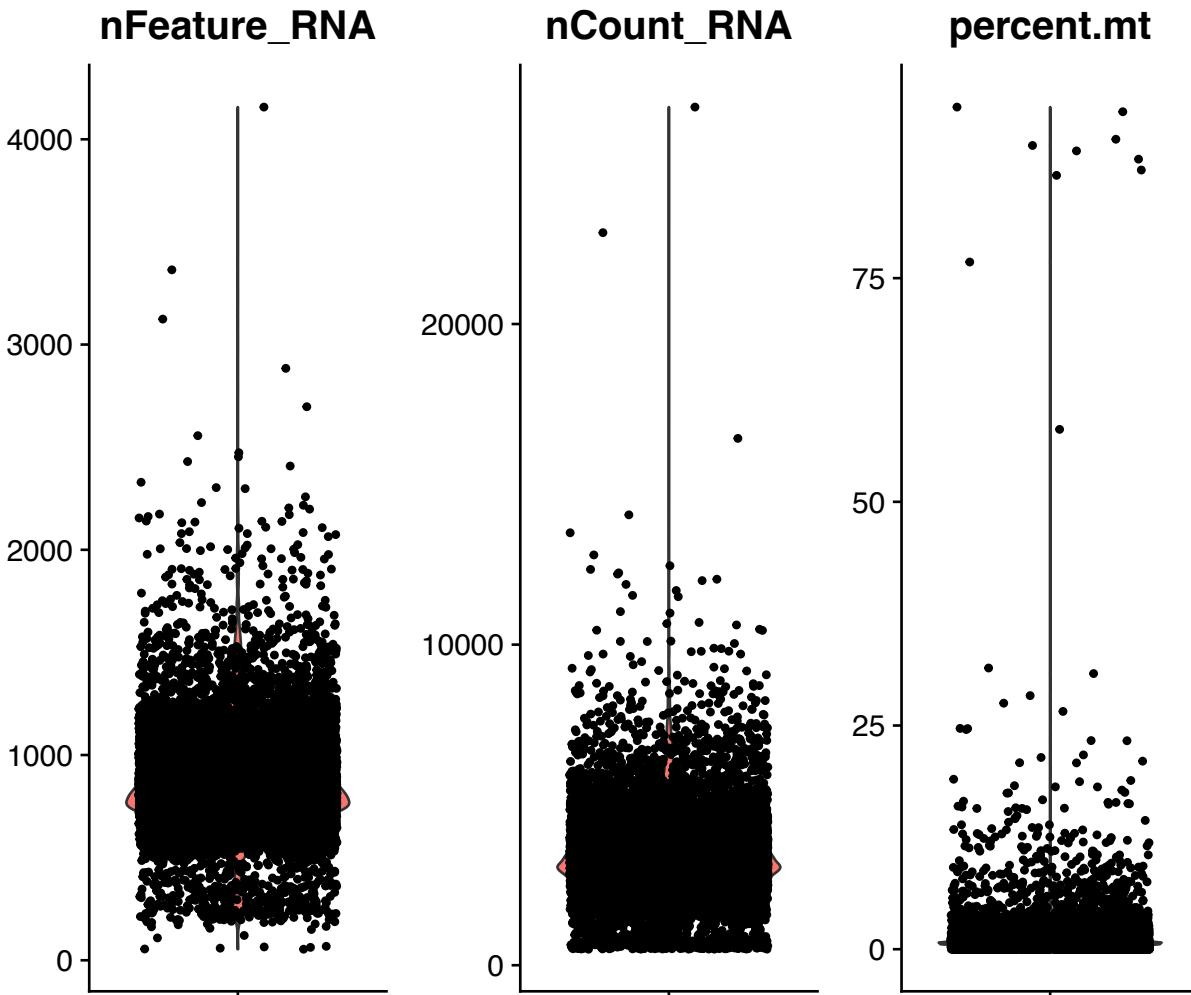
# RDatas

- If you have problems running parts of the code, you can download the ready RData objects
- [https://bioinfoshare.utu.fi/DataTransfer/Olissipo\\_hands-on\\_session2/RData](https://bioinfoshare.utu.fi/DataTransfer/Olissipo_hands-on_session2/RData)
  - Size: 1.1G
  - Username: btk
  - Password: oKxDvPn9xz

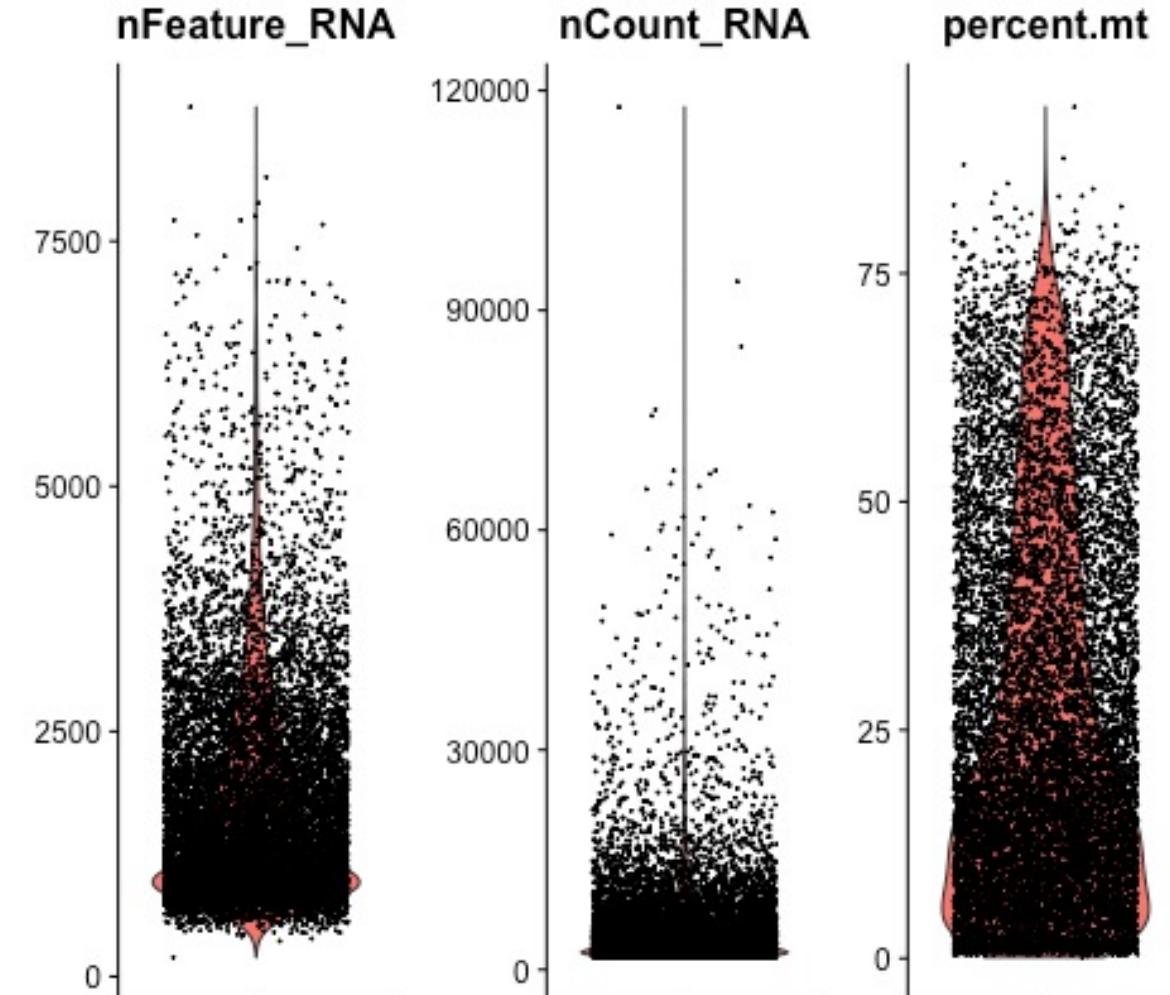
# Reading in data and QC

- Usually, raw count data is used as input
- QC is important to filter out low-quality cells
- Percentage of mitochondrial genes, number of genes identified and number of reads/UMI counts is usually checked
- High percentage of mitochondrial genes is indicative of low quality

## High-quality data



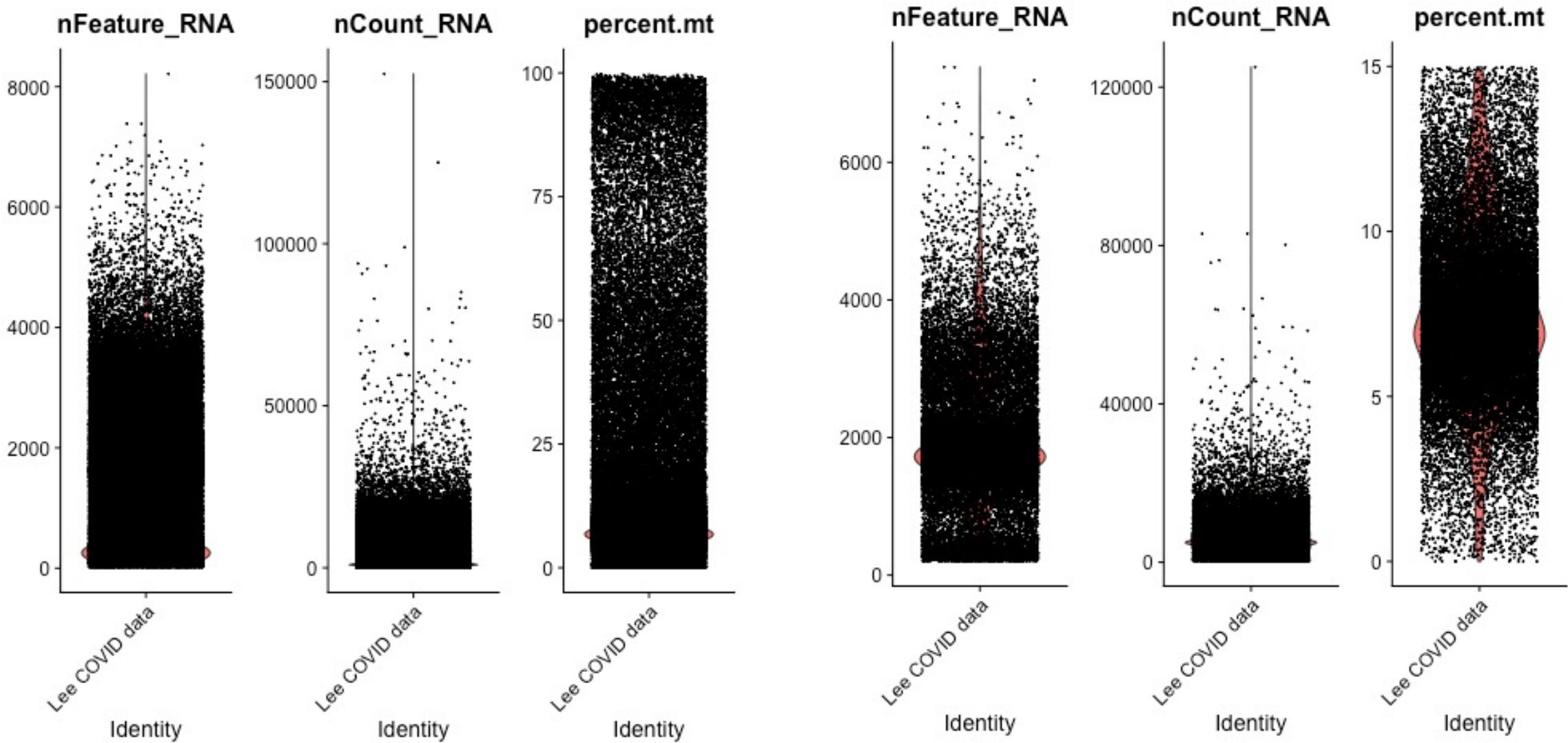
## Not-so-good-quality data



# Reading in data and QC

- Usually, raw count data is used as input
- QC is important to filter out low-quality cells
- Percentage of mitochondrial genes, number of genes identified and number of reads/UMI counts is usually checked
- High percentage of mitochondrial genes is indicative of low quality
- Run code until line 54
- In the code, the input \*.tsv.gz files are in folder called “data”
- All through the code, figures are saved to a folder called “Figures”
- What is your opinion on the data quality?
- How many cells were filtered out?

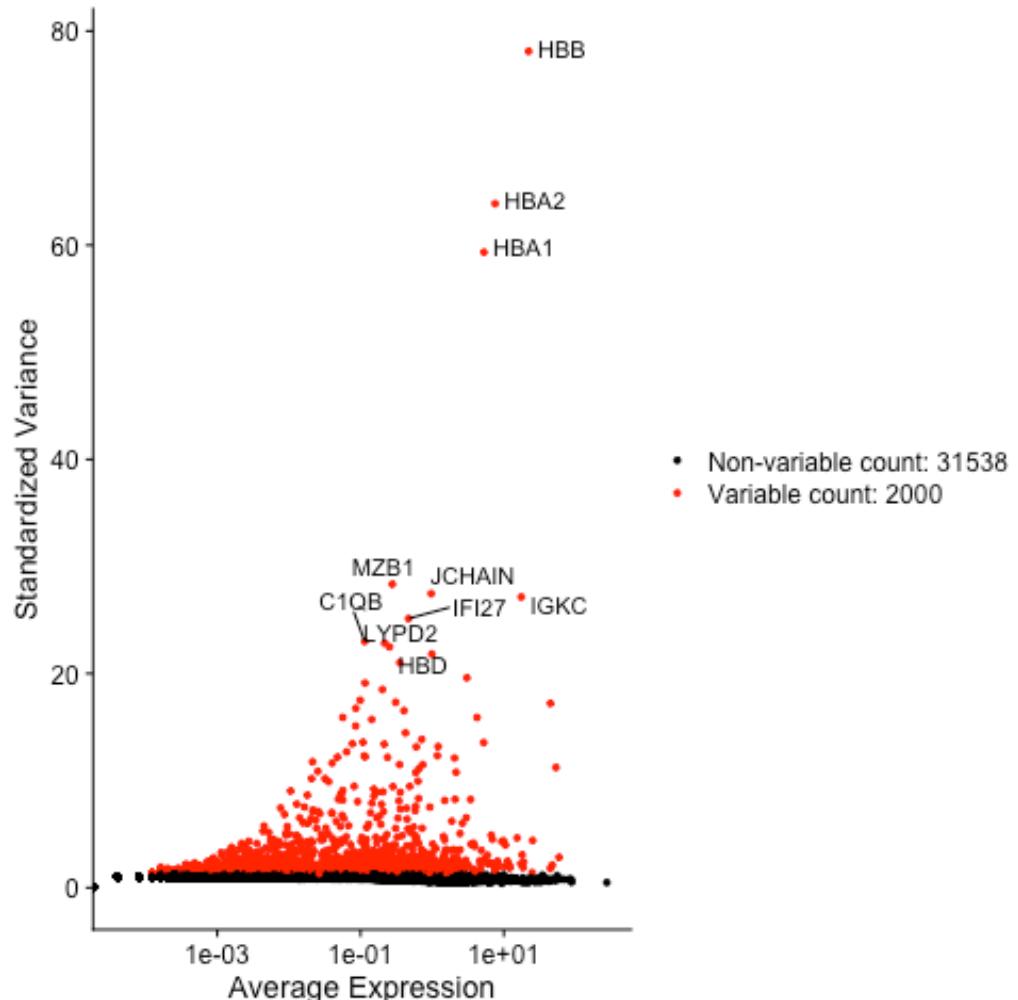
# Reading in data and QC



# Normalization and highly variable features

- Default normalization method in Seurat is "LogNormalize"
  - Normalizes the feature expression measurements for each cell by total expression, multiplies this by a scaling factor and performs log transformation
- SCTransform
  - a modeling framework for the normalization and variance stabilization of molecular count data from scRNA-seq experiment
- Identifying highly variable features
  - Find a subset of features that exhibit high cell-to-cell variation
- Run code until line 78
- We'll keep only four COVID patients and four healthy controls

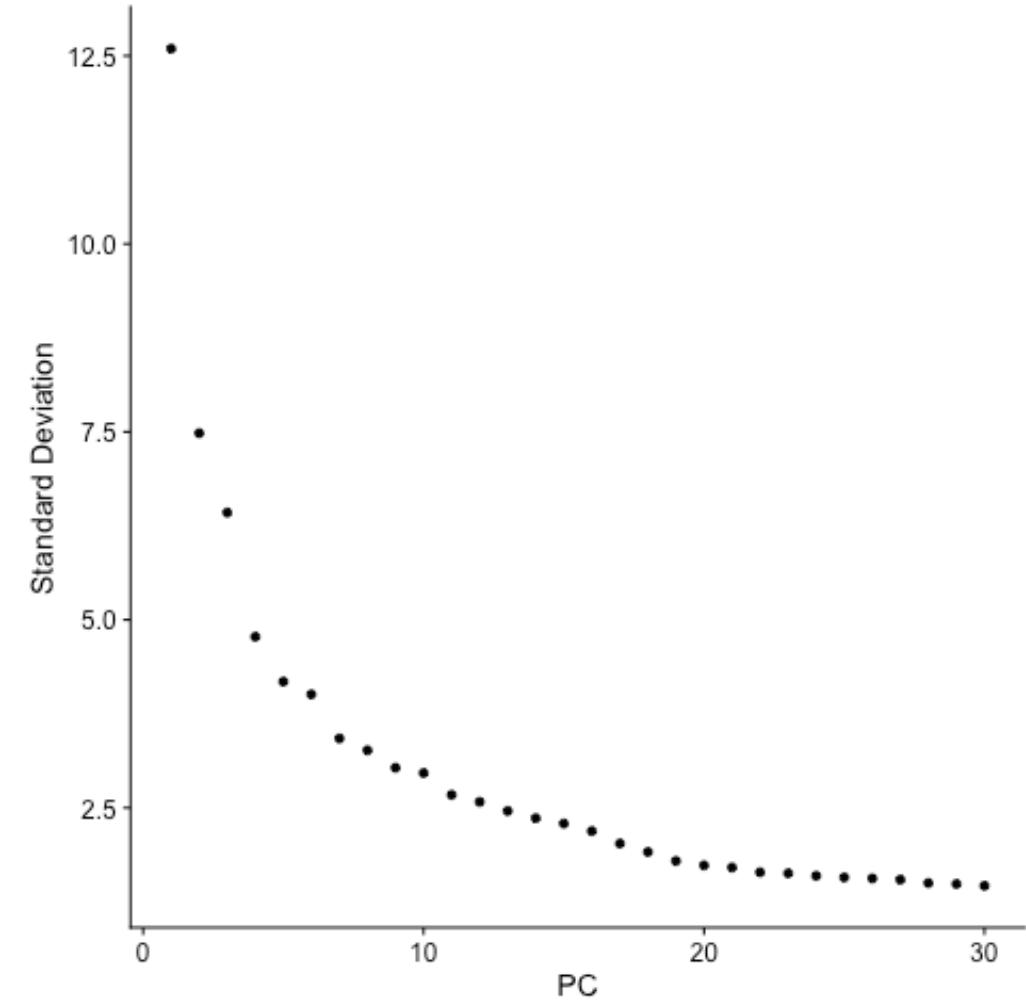
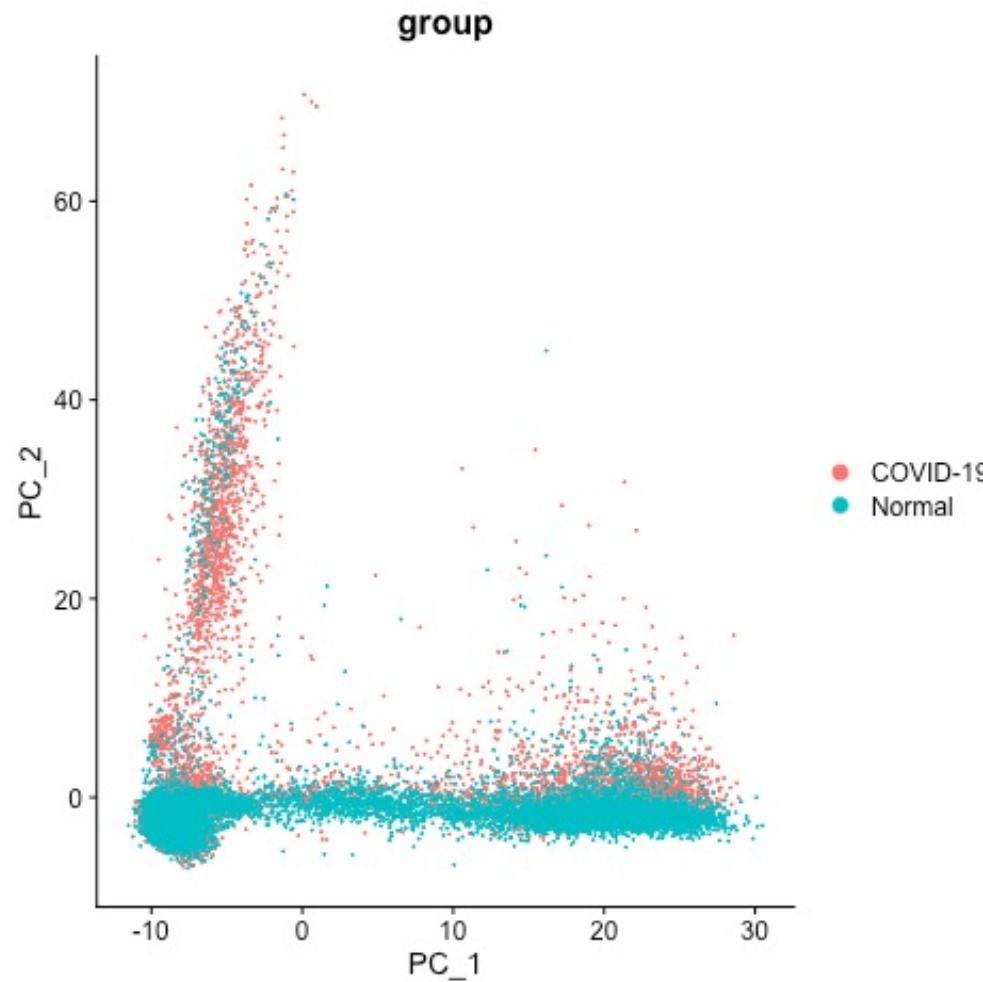
# Normalization and highly variable features



# Data scaling and PCA

- Scaling applies linear transformation prior to dimensional reduction
- By default, PCA is performed on the variable features
  - Other features can be defined using the features argument
- Top principal components represent a robust compression of the data
- Visualising the “dimensionality” of the data can help to decide how many PCs should be included
- Run code until line 95
- How many PCs would you use?

# Data scaling and PCA

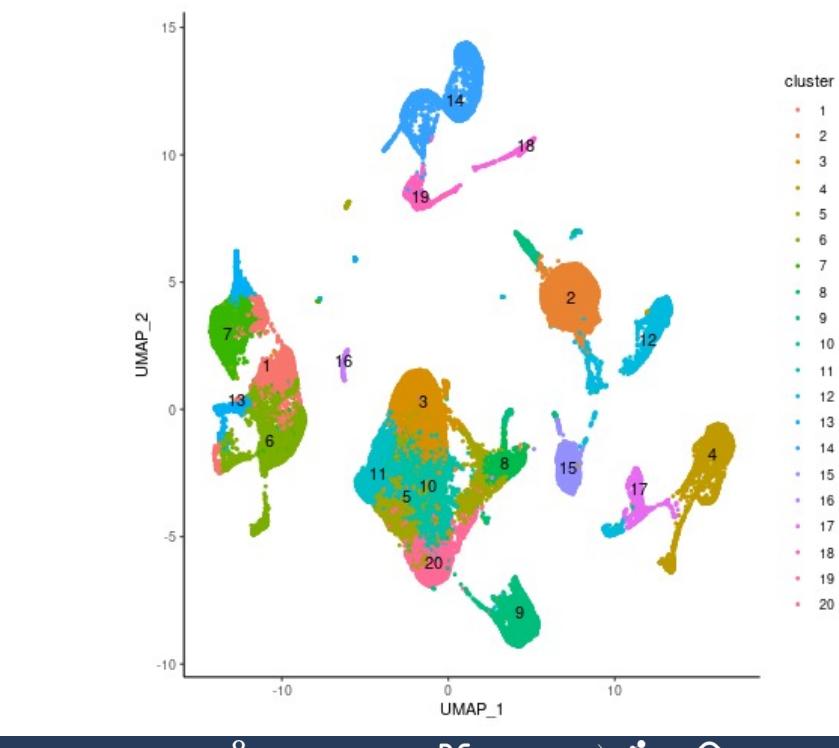
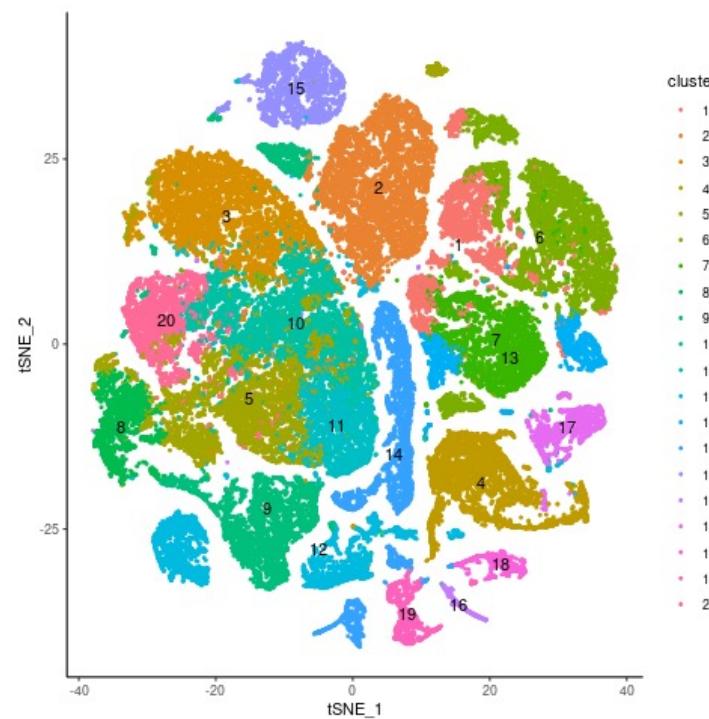


# Cell clustering and visualisation

- Clustering is an unsupervised learning procedure that is used to empirically define groups of cells with similar expression profiles
- This allows us to describe population heterogeneity in terms of discrete labels that are easily understood, rather than attempting to comprehend the high-dimensional manifold on which the cells truly reside
- After annotation based on marker genes, the clusters can be treated as proxies for more abstract biological concepts such as cell types or states

# Cell clustering and visualisation

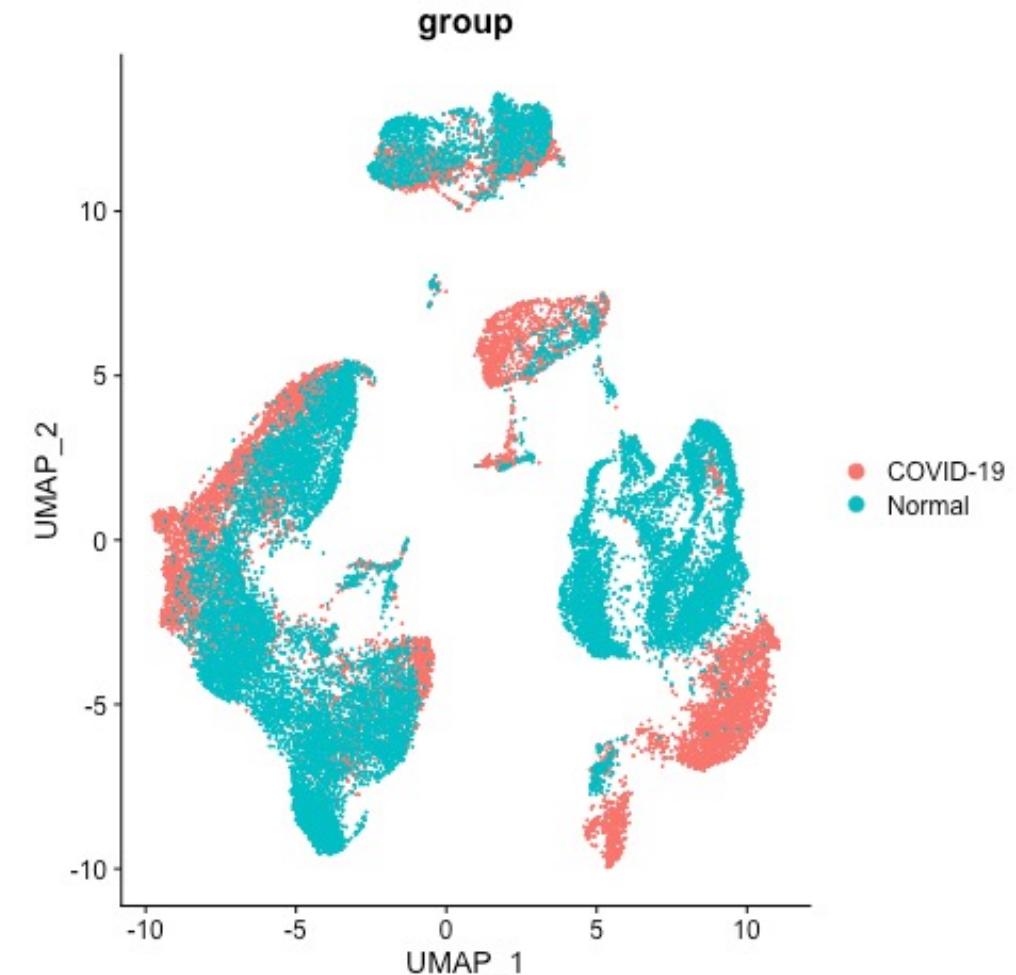
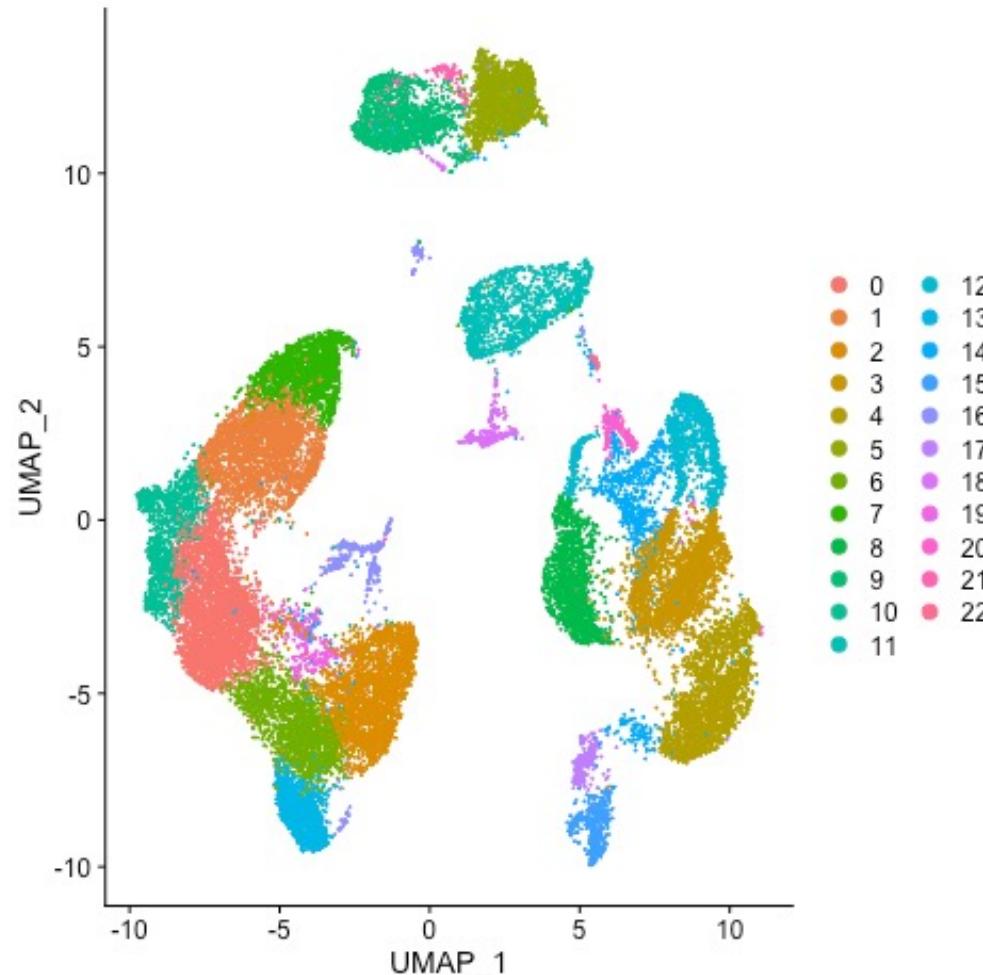
- Most popular visualisation methods are t-SNE and UMAP
- t-SNE retains only the local structure
- UMAP attempts to preserve more of the global structure than t-SNE



# Cell clustering and visualisation

- Clustering is an unsupervised learning procedure that is used to empirically define groups of cells with similar expression profiles
- This allows us to describe population heterogeneity in terms of discrete labels that are easily understood, rather than attempting to comprehend the high-dimensional manifold on which the cells truly reside
- After annotation based on marker genes, the clusters can be treated as proxies for more abstract biological concepts such as cell types or states
- Run code until 124
- Are there batch effects in these data? Do you think integration is necessary?

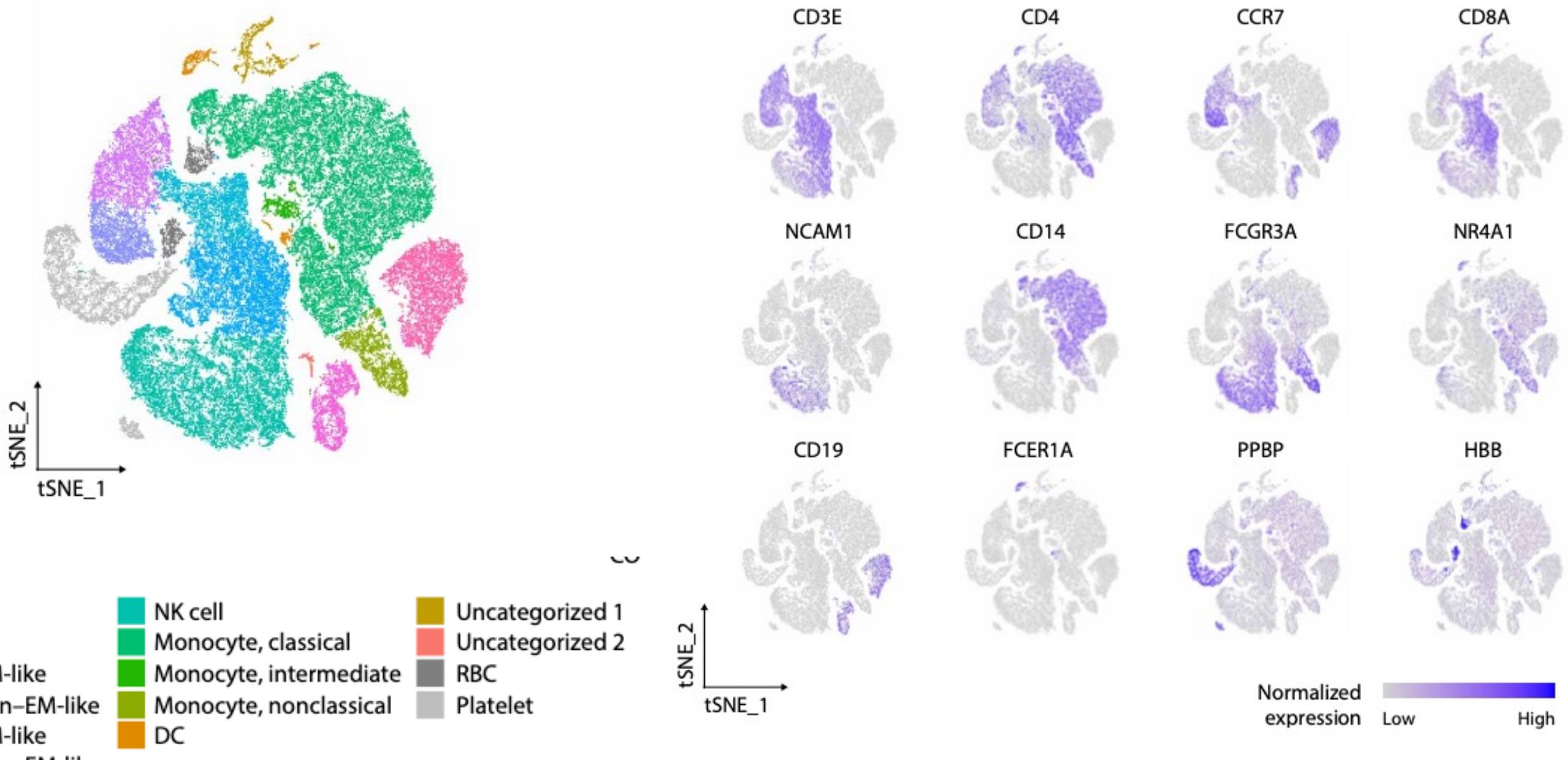
# Cell clustering and visualisation



# Integrating the data

- When the data contains batch effects (individuals, treatments, library prep...) integration might be needed
- Computational removal of batch-to-batch variation allows us to combine data across multiple batches for a consolidated downstream analysis
- Two integration methods exist in Seurat
  - CCA (Canonical Correlation Analysis)
  - RPCA (Reciprocal PCA)
- Other integration methods are available
  - Harmony, Scanorama, MultiMAP...
- Run code until line 180
- Do you think the annotation corresponds well with the marker genes?

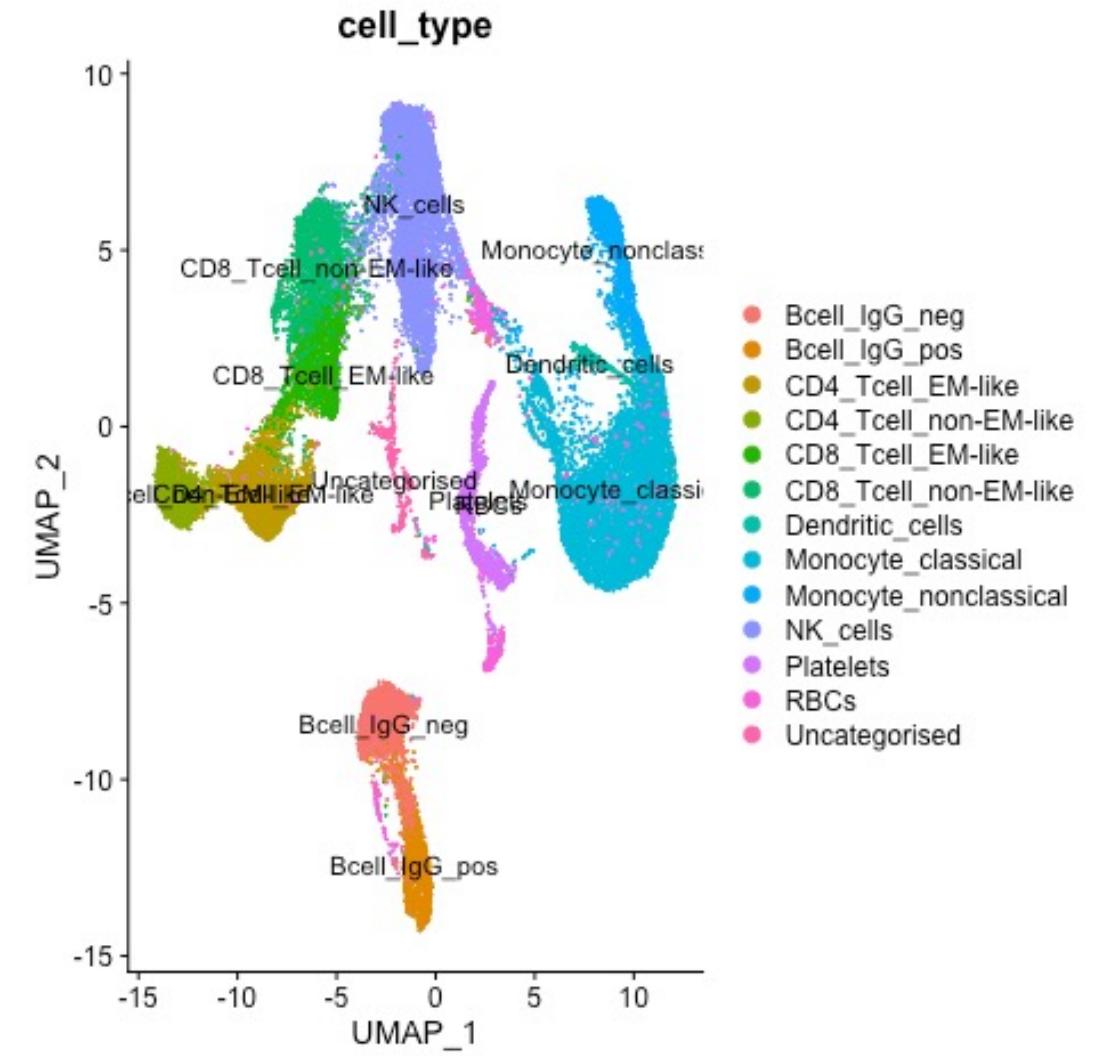
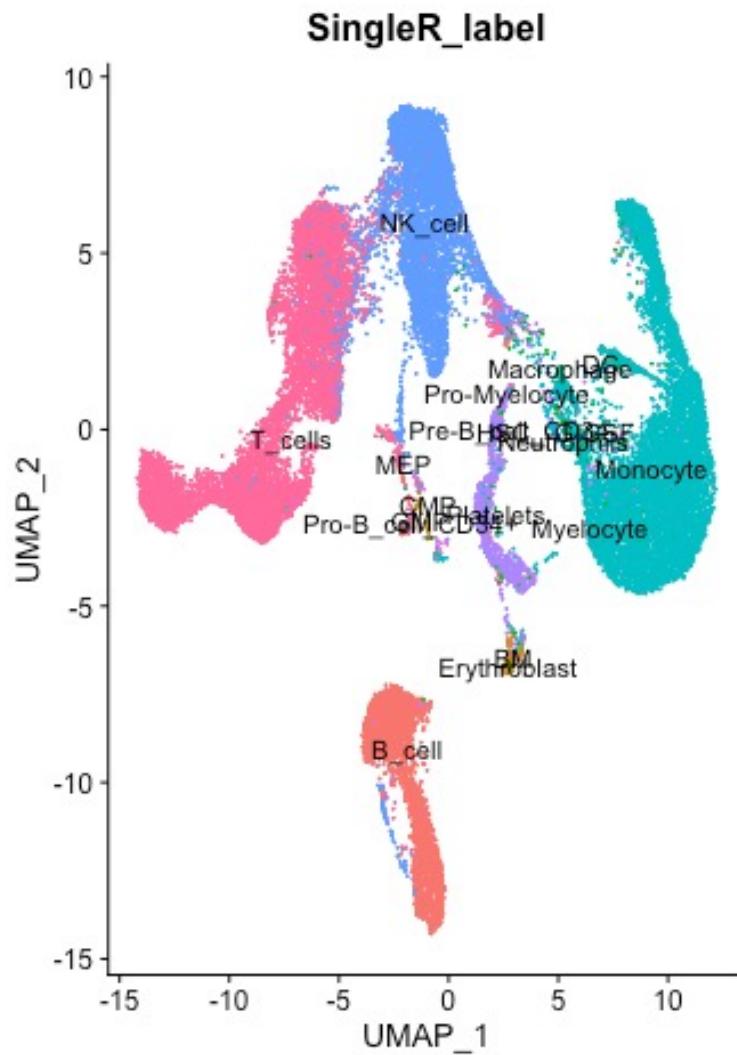
# Marker gene expression in the original publication



# Automatic annotation

- Usually, we don't know the correct cell type for the cells
- Manual annotation is time-consuming and requires expert knowledge
- Automatic annotation can be a fast and easy first step in the annotation process
- Other annotation tools
  - Azimuth (<https://azimuth.hubmapconsortium.org>)
  - CellTypist (<https://www.celltypist.org>)
- Run code until line 219
- We'll use SingleR and the Human Primary Cell Atlas for the automatic annotation
- How well did the automatic annotation work?
- What are the main differences to the manual annotation?

# Automatic annotation



# Differential expression analysis

- Finding marker genes for clusters
  - Compare the gene expression in one cluster to one other cluster or all other clusters
- Identifying differentially expressed genes between conditions
  - Compare the gene expression of samples in one condition to samples in another condition within one cluster
  - Different statistical approaches
    - Naïve methods that use cells as replicates
    - Pseudobulk methods
      - ROTs, limma, DESeq2
    - Mixed models that model the samples as random effect
      - Muscat, MAST\_RE

# Differential expression analysis

- Run code until line 317
- DE analysis run for dendritic cells and monocytes classical
- How do the DE results differ between the two methods?