Introduction to single cell RNA-seq

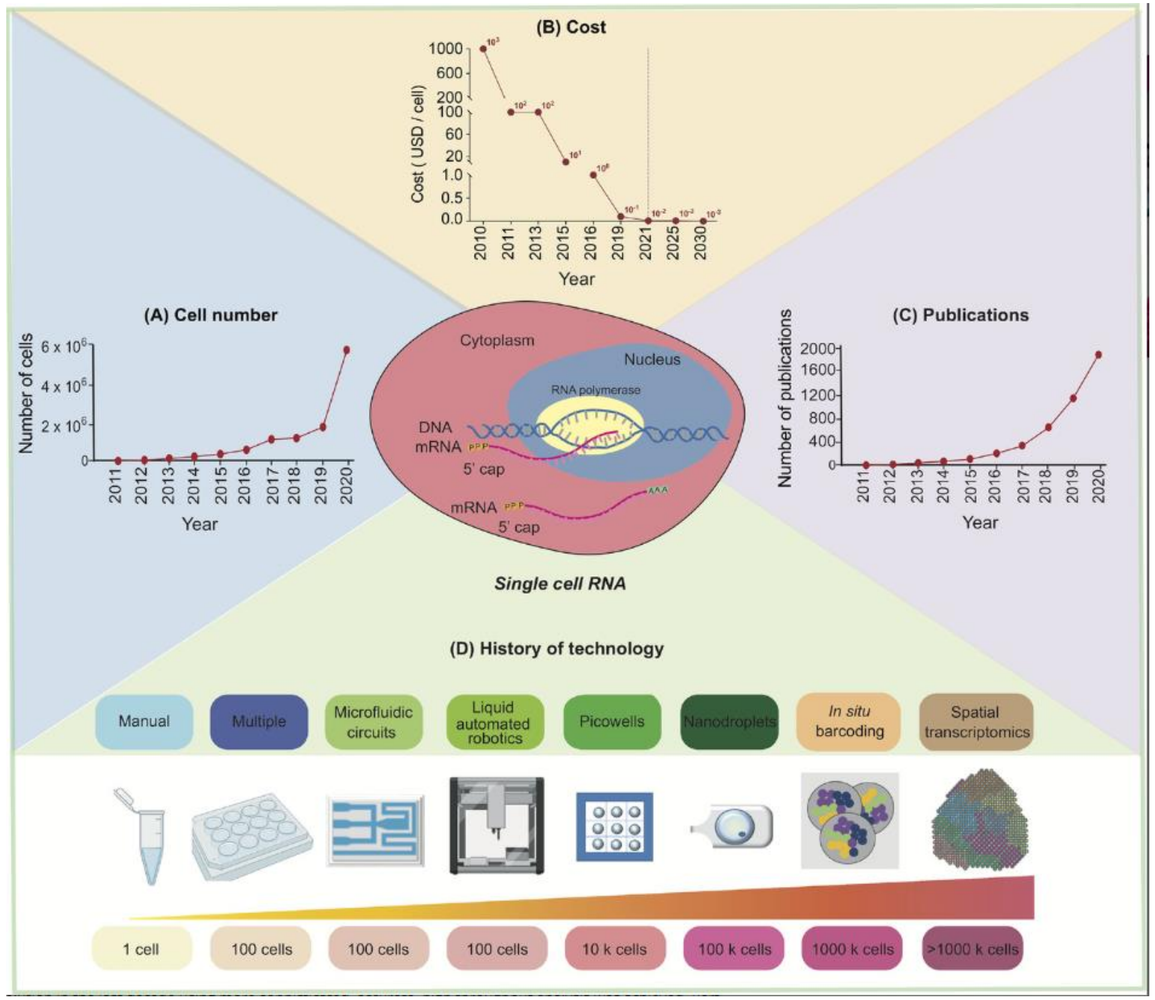
Xabier Bujanda Cundin
17/07/2023

Why single cell?

- **Cell**: the smallest unit that can live on its own and that makes up all living organisms and the tissues of the body
- Identification of cell types and state in a heterogeneous sample
- Assess differences between and within cell types

Challenges:

- Huge amount of data to handle
- Low read depth
- High variability between cells/samples
- Data processing, analysis, presentation and interpretation



Jovic, D, Liang, X, Zeng, H, Lin, L, Xu, F, Luo, Y Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med*. 2022; 12:e694. https://doi.org/10.1002/ctm2.694

scRNA-seq data can answer multiple biological questions

- Cell populations
- Expression differences cell clusters and conditions
- Time-series gene expression development

But…

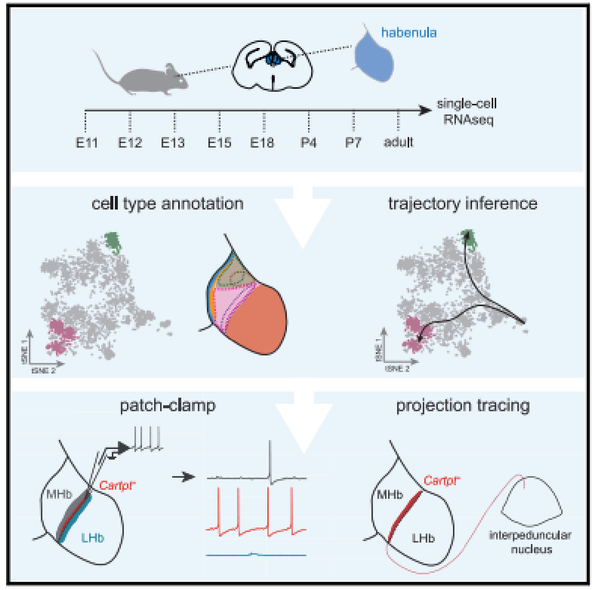
- Low number of transcripts can be detected
- Some cell types may not be compatible for some scRNA-seq techniques
- Low expressed transcript may not be detected

Article

Cell Reports

Molecular signatures and cellular diversity during mouse habenula development

Graphical abstract



Authors

Lieke L. van de Haar, Danai Riga, Juliska E. Boer, ..., Frank J. Meye, Onur Basak, R. Jeroen Pasterkamp

Correspondence

r.j.pasterkamp@umcutrecht.nl

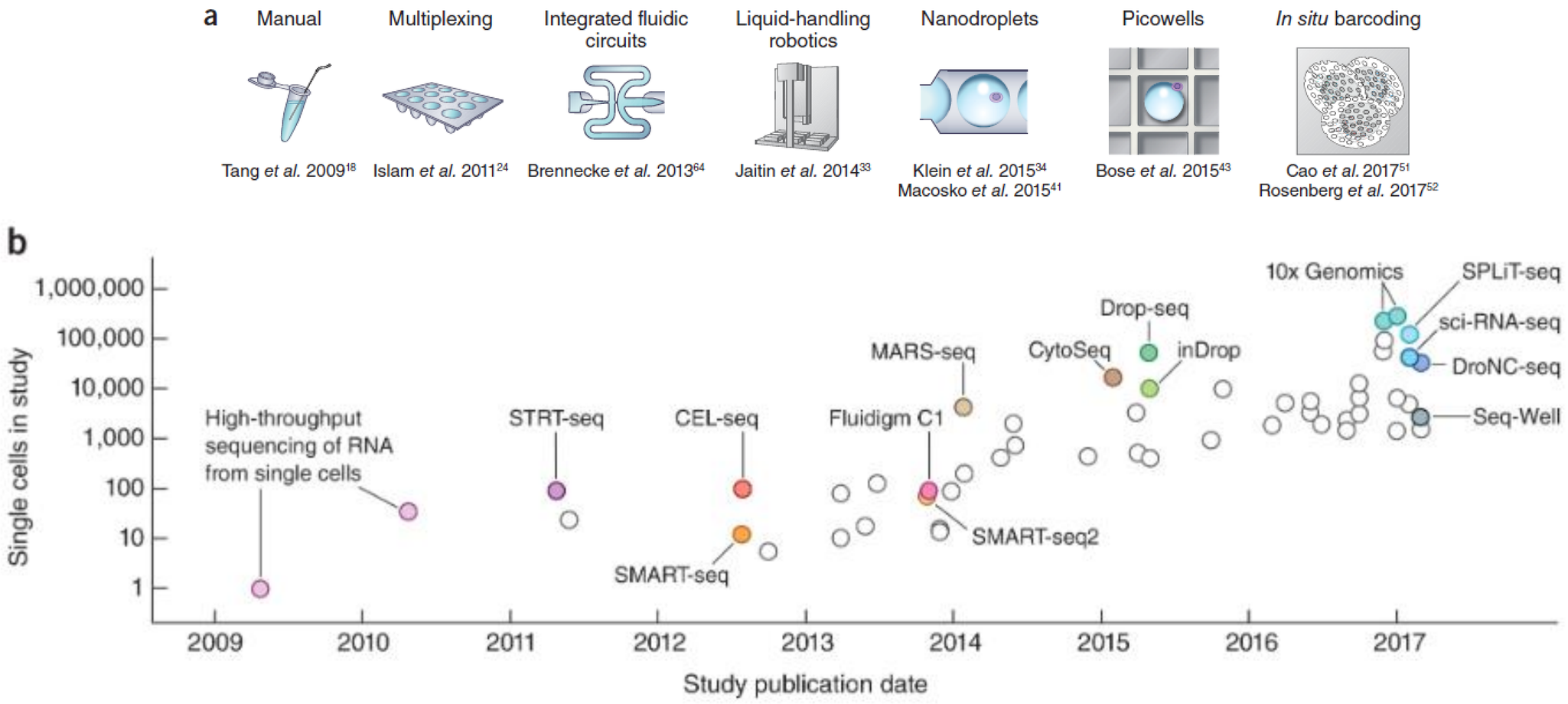
In brief

van de Haar et al. characterize cell populations of the developing mouse habenula using single-cell RNA sequencing. They reconstruct the developmental trajectories of molecularly distinct neuronal subsets and establish physiological and connectivity properties of *Cartpt* neurons. Their analysis indicates a possible link between developing habenula neurons and human psychiatric disorders.

Highlights

- scRNA-seq reveals cellular heterogeneity of the developing mouse habenula
- Trajectory inference reconstructs developmental paths of habenula cell types
- *Cartpt* marks physiologically distinct neurons that innervate the dorsal IPN
- Developing habenular cell types align to risk loci of human psychiatric disorders

Single cell RNA-seq technologies



Svensson, V., Vento-Tormo, R. & Teichmann, S. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* **13**, 599–604 (2018). https://doi.org/10.1038/nprot.2017.149

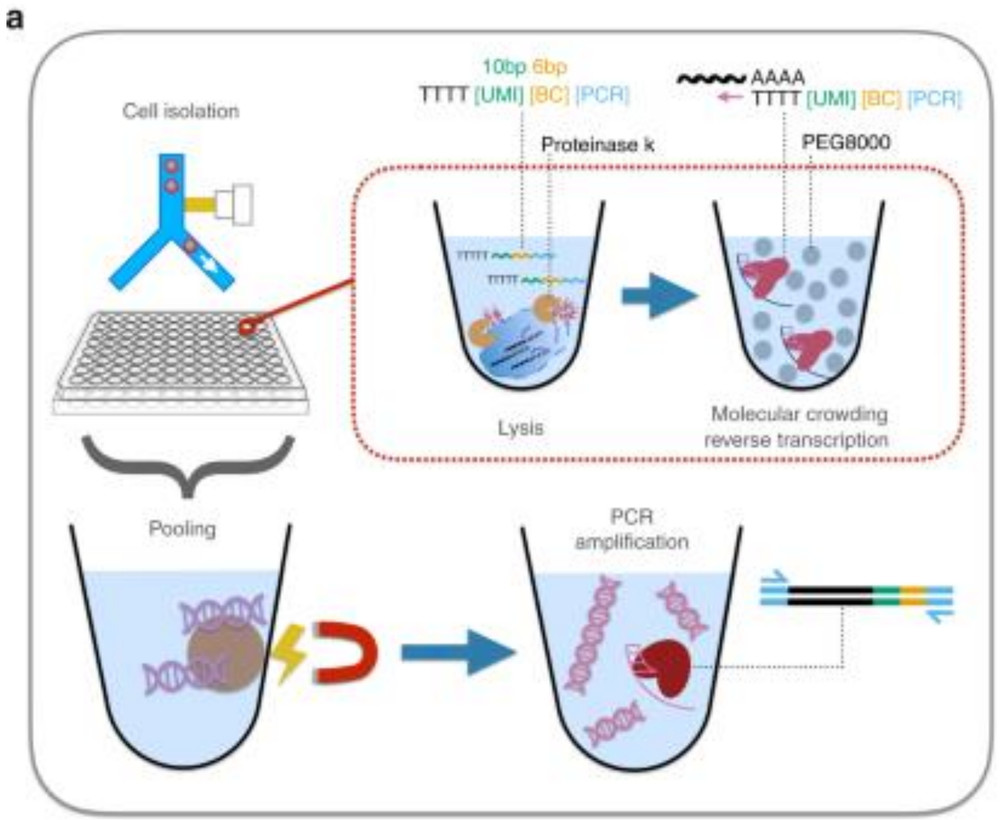
Single cell RNA-seq technologies

TABLE 1 Comparison of different single-cell RNA sequencing (scRNA-seq) technology and experimental protocols

Platforms	Isolation strategies	Tissue	Cell numbers	Targets	UMI	Amplification methods	Region	Published year
Smart-seq	FACS	Dissociated cell	Hundreds of cells	/	×	PCR	Full-length	2012
Smart-seq2	FACS	Dissociated cell	Hundreds of cells	/	×	PCR	Full-length	2013
Fluidigm C1	Micro-fluidic	Dissociated cell	Hundreds of cells	No poly(A) minus RNA detection	×	PCR	Full-length	2013
Drop-seq	Microdroplets	Dissociated cell	Large number of cell	No poly(A) minus RNA detection	√	PCR	3′ end	2015
10x Genomics	Microdroplets	Dissociated cell	Large number of cells	No poly(A) minus RNA detection	√	PCR	3′ end	2016
MATQ-seq	FACS	Dissociated cell	Hundreds of cells	No poly(A) minus RNA detection	√	PCR	Full-length	2017
Seq-Well	Micro-fluidic	Dissociated cell	Large number of cells	No poly(A) minus RNA detection	√	PCR	3′ end	2017
CEL-seq	FACS	Dissociated cell	Hundreds of cells	No poly(A) minus RNA detection	√	IVT	3′ end	2012
MARS-seq	FACS	Dissociated cell	Hundreds of cells	No poly(A) minus RNA detection	√	IVT	3′ end	2014
inDrop-seq	Microdroplets	Dissociated cell	Large number of cell	No poly(A) minus RNA detection	√	IVT	3′ end	2015
DNBelab C4	Microdroplets	Dissociated cell	Large number of cells	No poly(A) minus RNA detection	√	PCR	3′ end	2019

SMART-seq technology

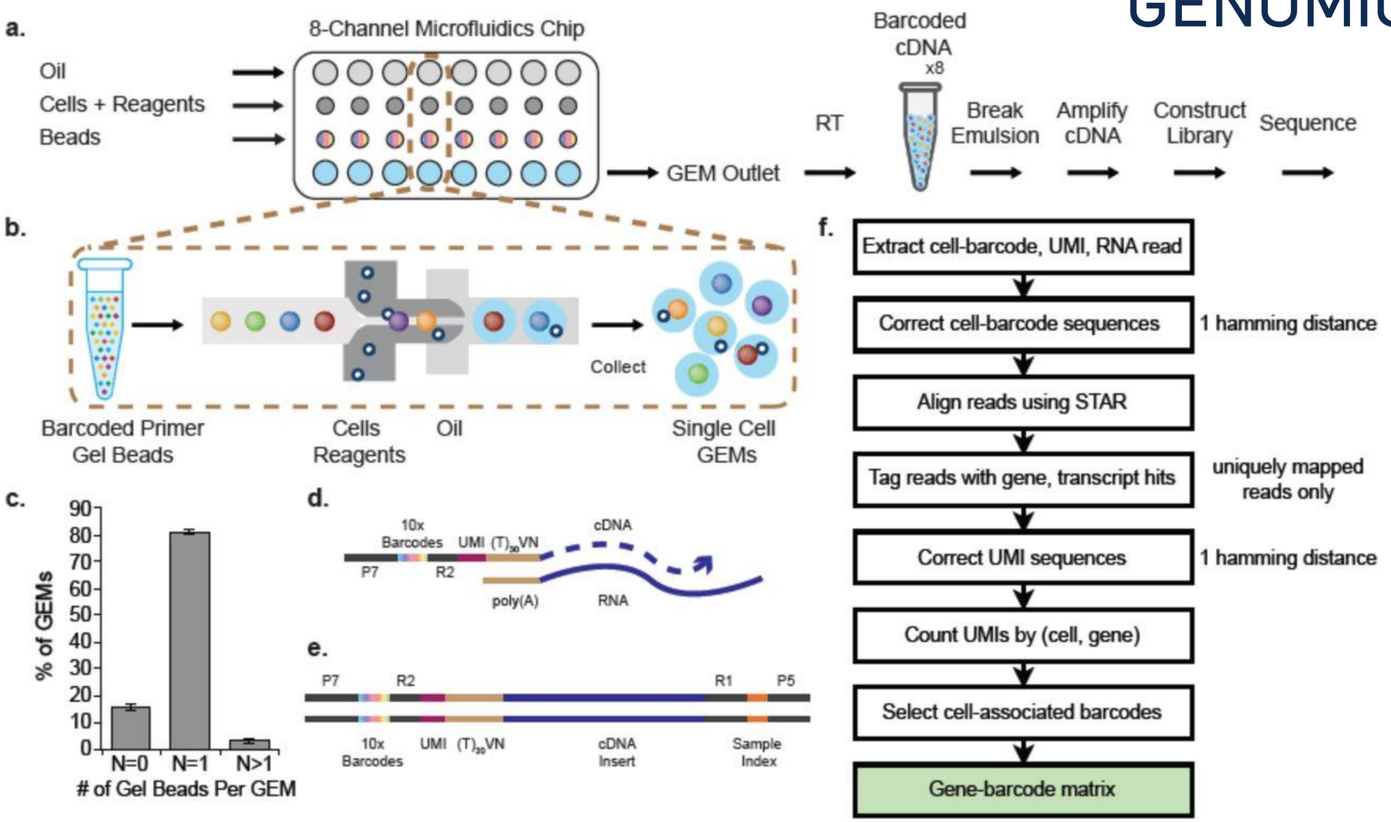
- Cell separation through FACS
- Library preparation by cell
- Poor depth
- Whole gene sequencing (isoforms identification)



Bagnoli, J.W., Ziegenhain, C., Janjic, A. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat Commun* **9**, 2937 (2018).

10X Genomics technology

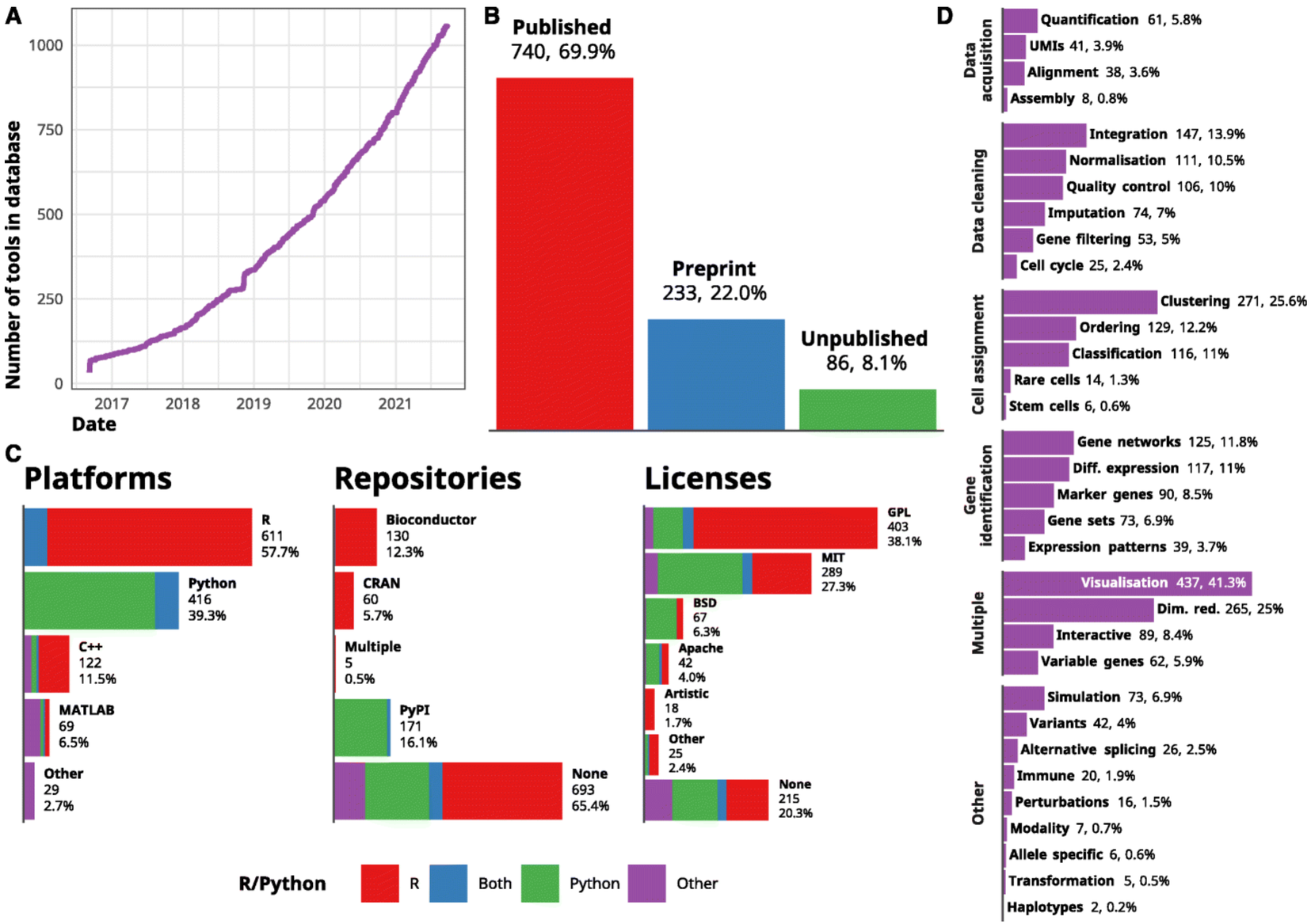
- High Sensitivity
- Lower technical noise
- High price
- Simpler data processing steps



Zheng, G., Terry, J., Belgrader, P. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017). https://doi.org/10.1038/ncomms14049

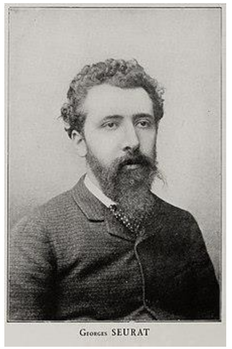
Platforms:

- Seurat (R)
- Scanpy (Python)
- Galaxy (webtool)
- Scater + scran (R, Bioconductor)
- Monocle3 and psupertime (R)

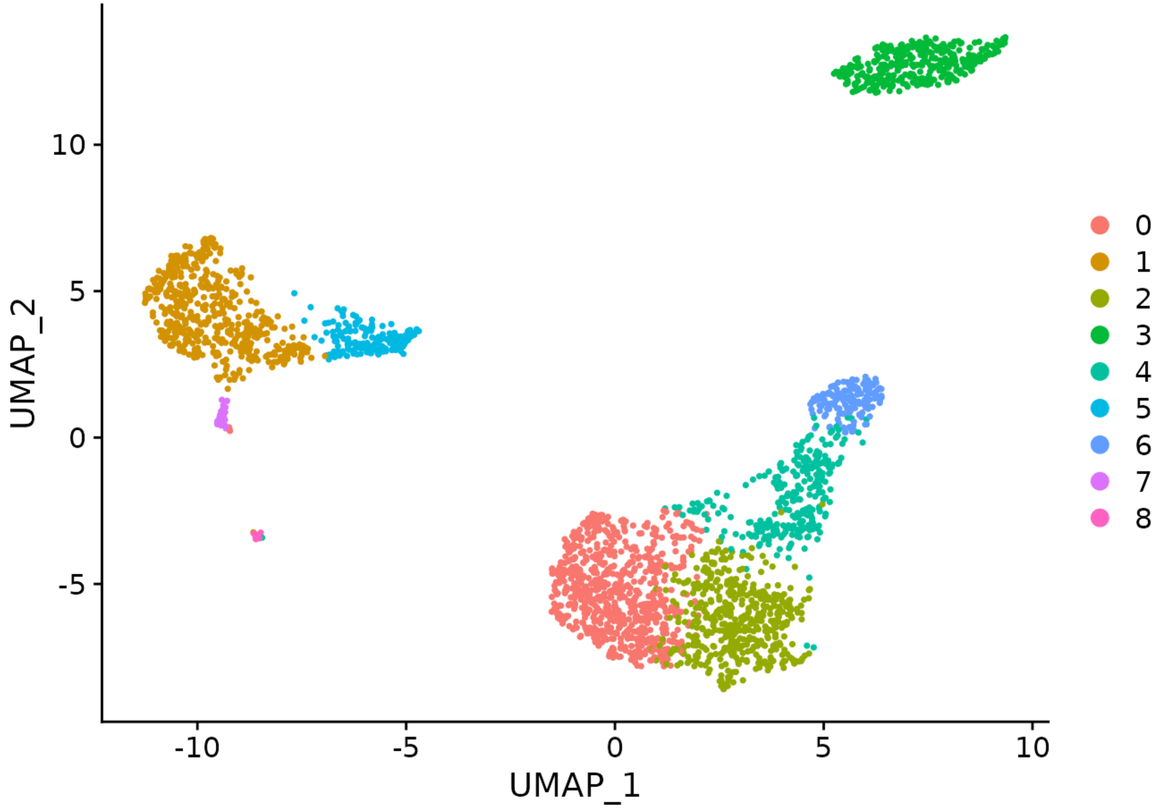


Zappia, L., Theis, F.J. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biol* **22**, 301 (2021). https://doi.org/10.1186/s13059-021-02519-4

3. Seurat



Georges SEURAT



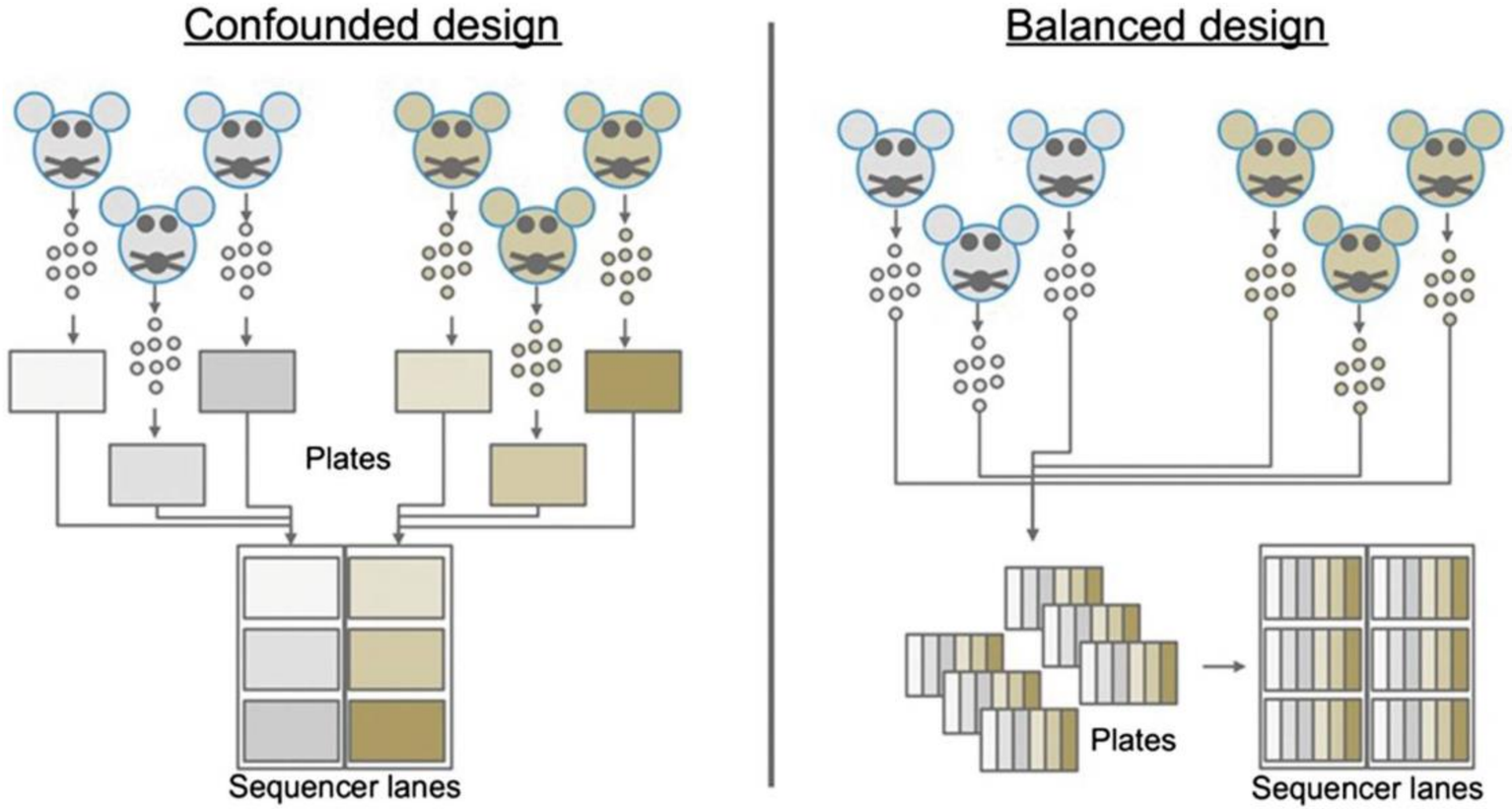
Why Seurat?

- Easy to use
- Very popular, supported and updated
- Deep documentation and good data model
- Several built-in functionalities
- Easy to extend with your own functionality
- Seurat objects extend to other packages

Adapted from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

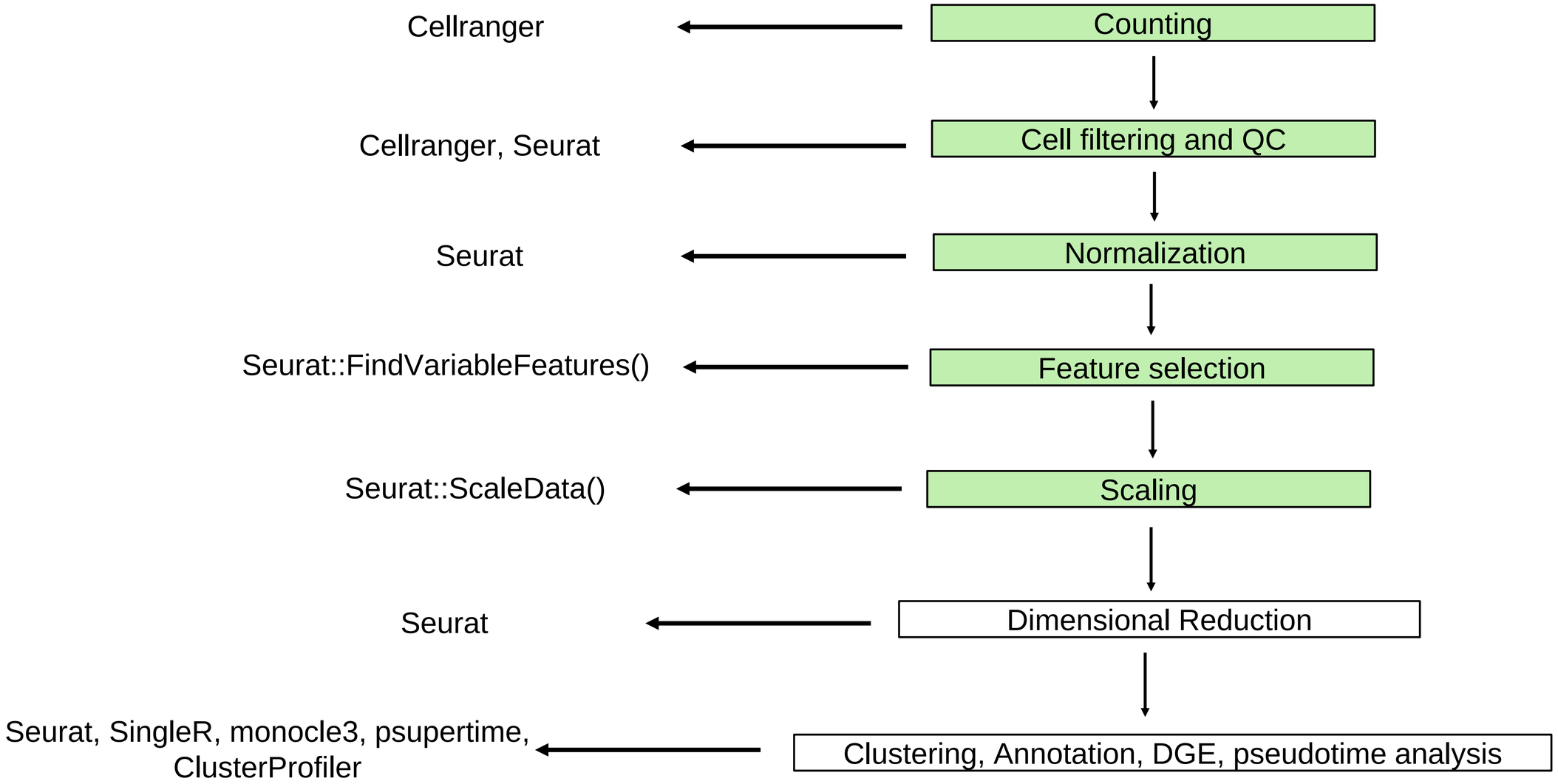
Experimental Design

- Replication and randomization
- Be aware of potential biases that can affect the results
- Record any factor for downstream correction



Baran-Gale J, Chandra T, Kirschner K. Experimental design for single-cell RNA sequencing. Brief Funct Genomics. 2018 Jul 1;17(4):233-239. doi: 10.1093/bfgp/elx035.

Data Analysis Overview

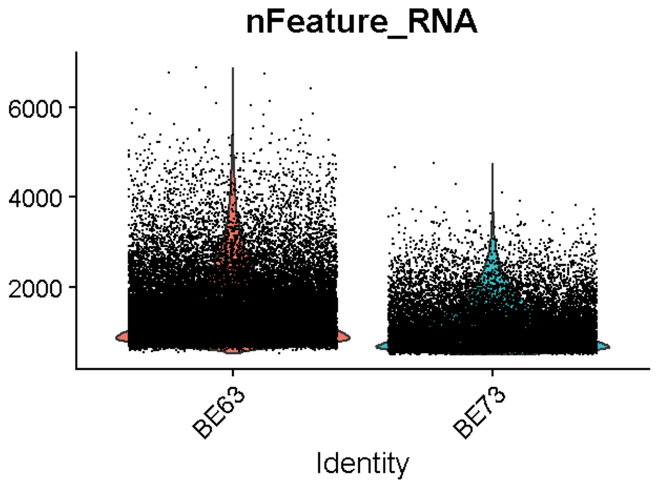


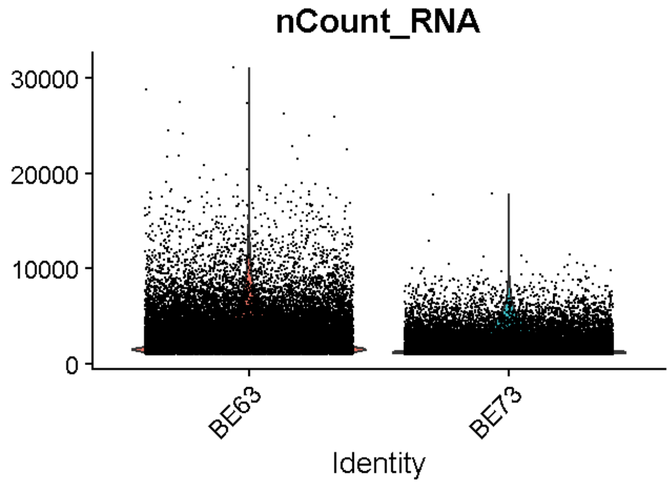
Counting

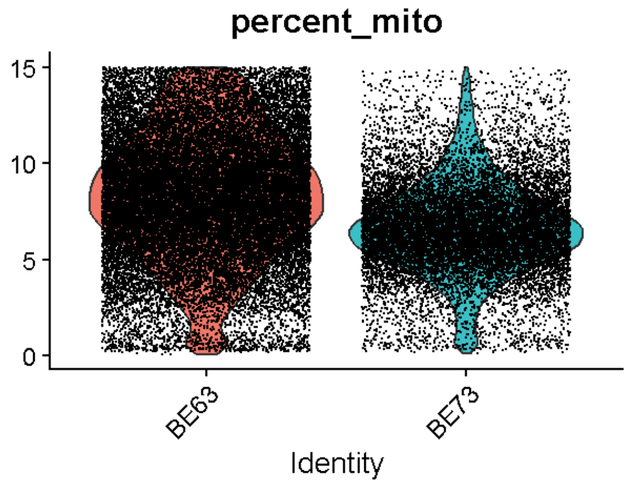
```
#Export path to make cellranger 7.1.0  avaliable
export PATH=$PATH:/home/tigem/software/cellranger-7.1.0

# Define the path to the main directory containing the subfolders with input files
main_dir="/home/tigem/x.bujandac/Polishchuk_scRNAseq/MAN2-HRP_TC_sc"

# Loop through each subfolder
for subfolder in "$main_dir"/*; do

        echo "Processing files in $subfolder..."

        # Run cellranger count on the files in the subfolder
        cellranger count --id=$(basename "$subfolder") \
        --transcriptome=/home/tigem/x.bujandac/Polishchuk_scRNAseq/GRCh38 \
        --fastqs=$subfolder \
        --sample=$(basename "$subfolder") \
        --localcores 8 \
        --localmem 64

        echo "Finished processing files in $subfolder."
        echo
done
```

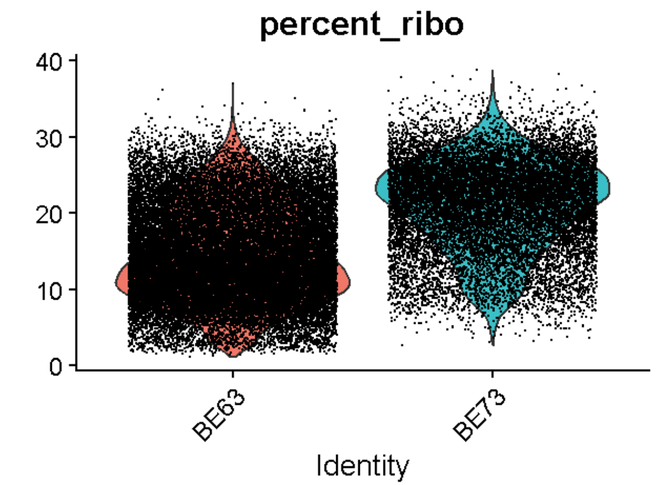
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

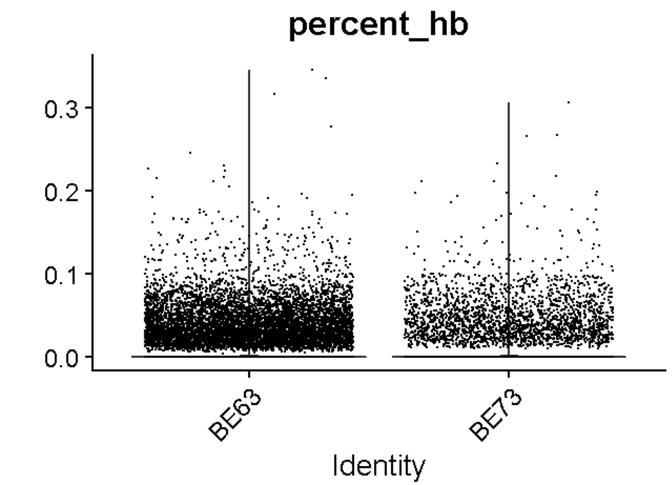
Quality Control











Differences among samples in:

- nFeatures and nCounts
- Eventual apoptotic cells
- Ceck for ribosomal genes and blood related genes

Quality Control

```r
# Create Seurat Object
counts <- CreateSeuratObject(counts = counts, meta.data = metadata, project = "scRNAseqBE")

#============= Quality Control ==============#

#calculate the % of mitocondrial genes for each cell
counts <- PercentageFeatureSet(counts, "^MT-", col.name = "percent_mito")

#calculate the % of ribosomal genes for each cell
counts <- PercentageFeatureSet(counts, "^RP[SL]", col.name = "percent_ribo")

# Percentage hemoglobin genes - includes all genes starting with HB except HBP.
counts <- PercentageFeatureSet(counts, "^HB[^(P)]", col.name = "percent_hb")

#Plot basic statistics
feats <- c("nFeature_RNA", "nCount_RNA", "percent_mito", "percent_ribo", "percent_hb")
VlnPlot(countsFiltered, group.by = "orig.ident", features = feats, pt.size = 0.1, ncol = 3) +
  NoLegend()
```

Data Normalization and Scaling

There are different ways to normalize data:

Logaritmic normalization:

$$x_i = \frac{The\ read\ count\ of\ gene\ X\ in\ cell\ i}{Total\ counts\ of\ cell\ i} \times 10^4 \ (Eq.1)$$

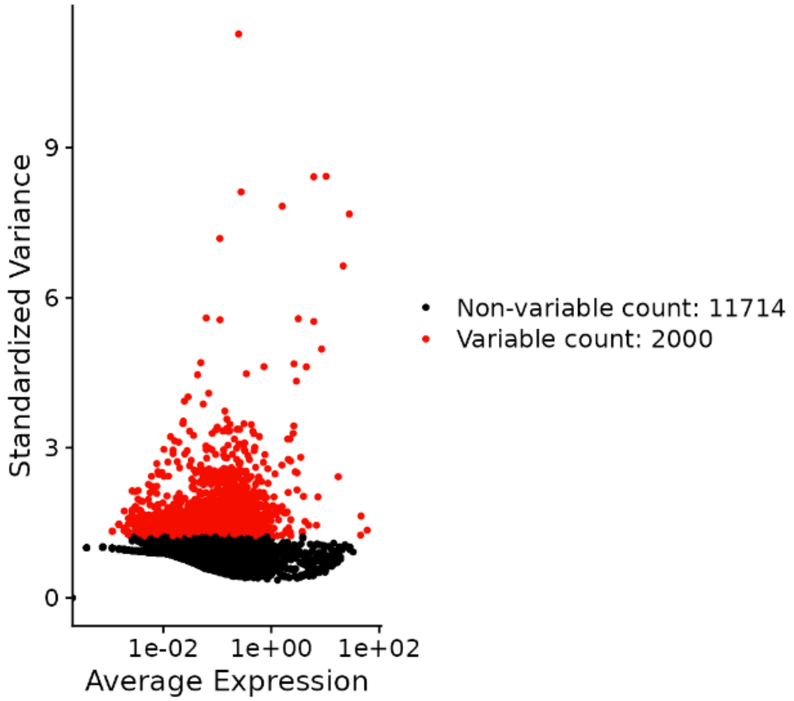
$$f(x_i) = ln(x_i + 1) \ (Eq.2)$$

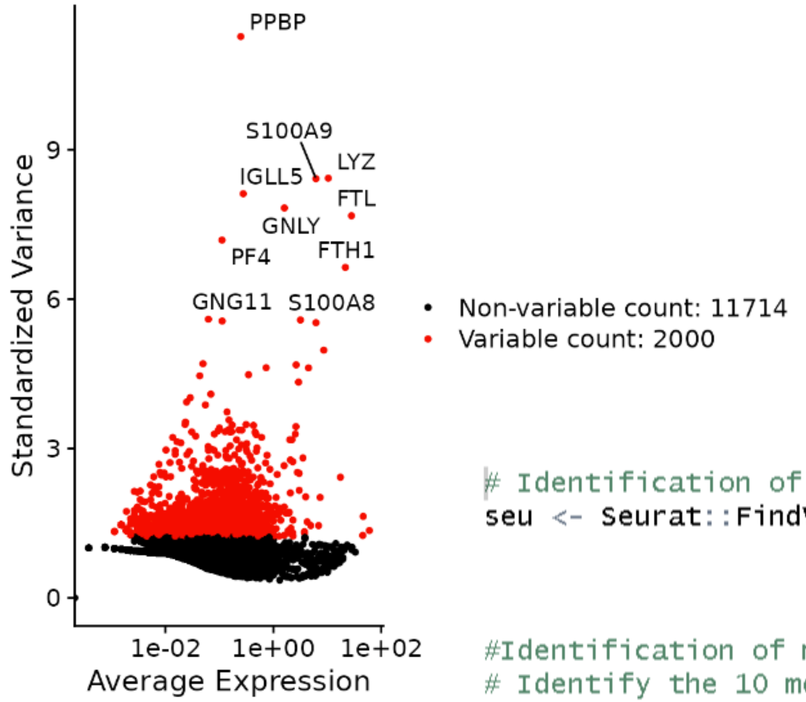
CPM, TPM, RPKM (old but gold)

Work well under the assumption that the amount of RNA is the same in all cells and a uniform scaling factor is applicable for all genes

Sctransform: Normalization and variation control in one step → Provides an excellent tool to not only normalize data, but also stabilize variance and regress unwanted variation

Feature selection



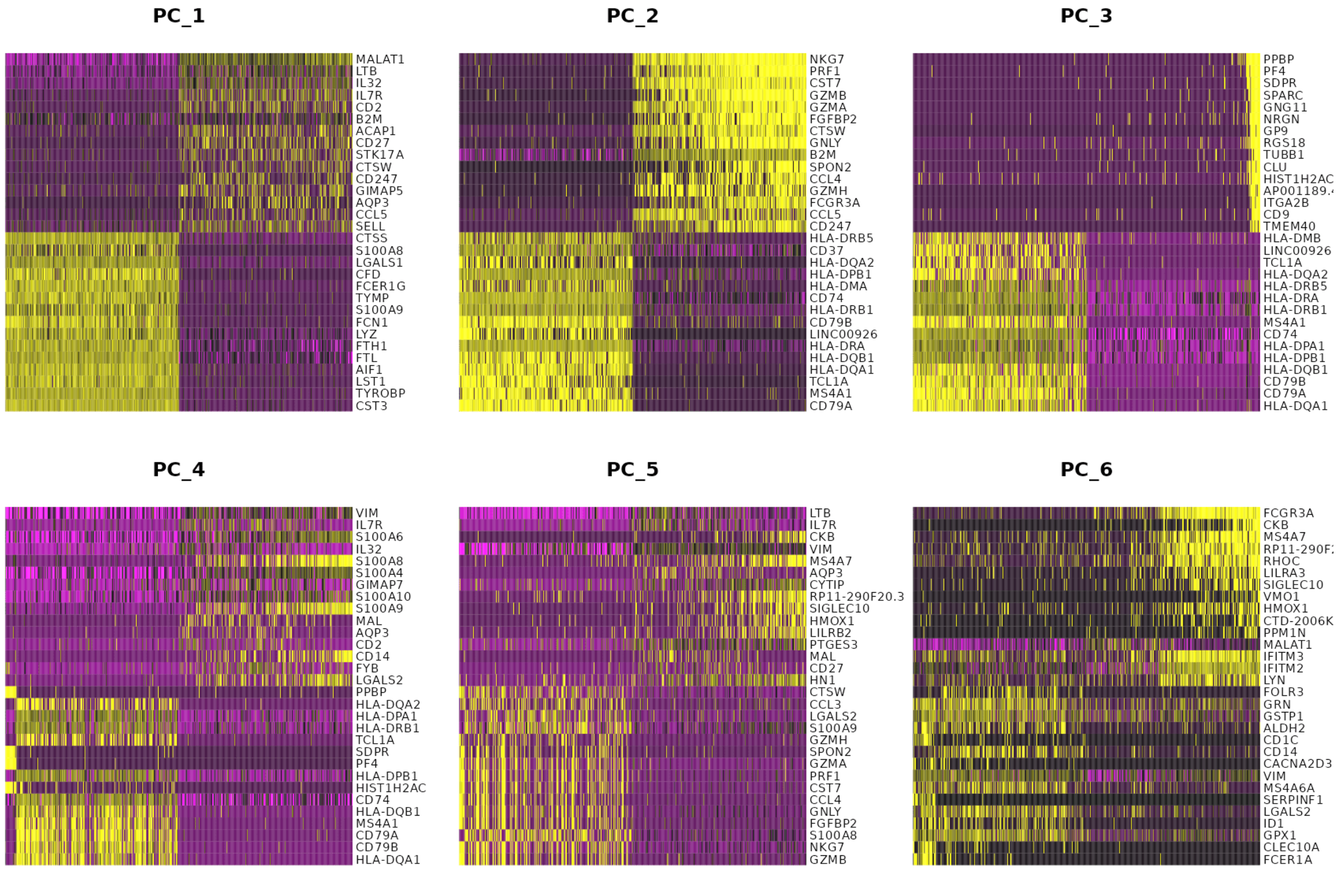


- Identify several genes that exhibit high variability between cells
- Simple mathematical model to detect the most variable genes
- These most variable genes represent the features to prioritize for the downstream analysis

```
# Identification of highly variable features (feature selection)
seu <- Seurat::FindVariableFeatures(seu,
                                    selection.method = "vst",
                                    nfeatures = 2000)

#Identification of most expressed genes
# Identify the 10 most highly variable genes
top10 <- head(Seurat::VariableFeatures(seu), 10)
top10

#Plot
vf_plot <- Seurat::VariableFeaturePlot(seu)
Seurat::LabelPoints(plot = vf_plot,
                    points = top10, repel = TRUE)
```

Adapted from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

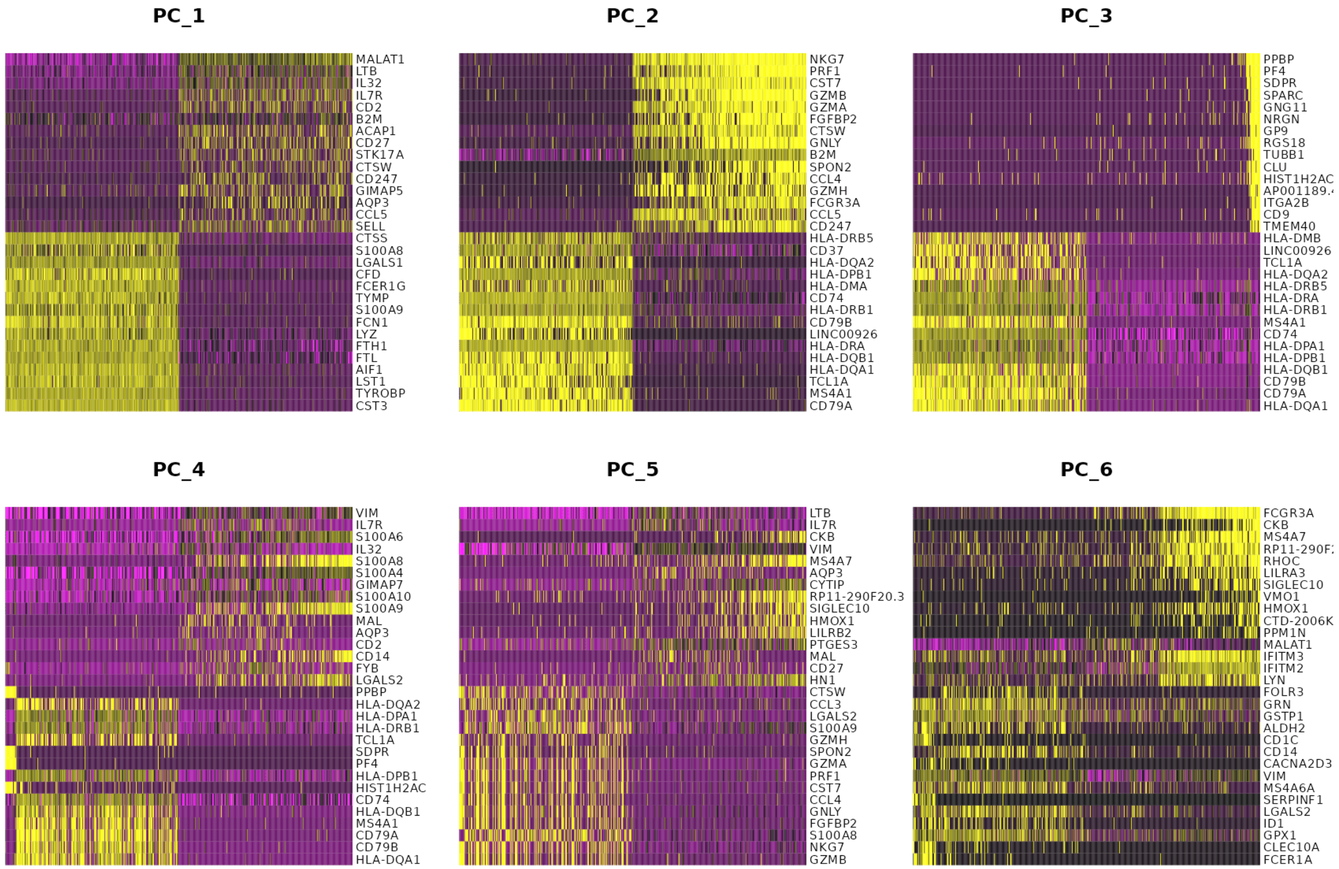
Linear Dimensional Reduction

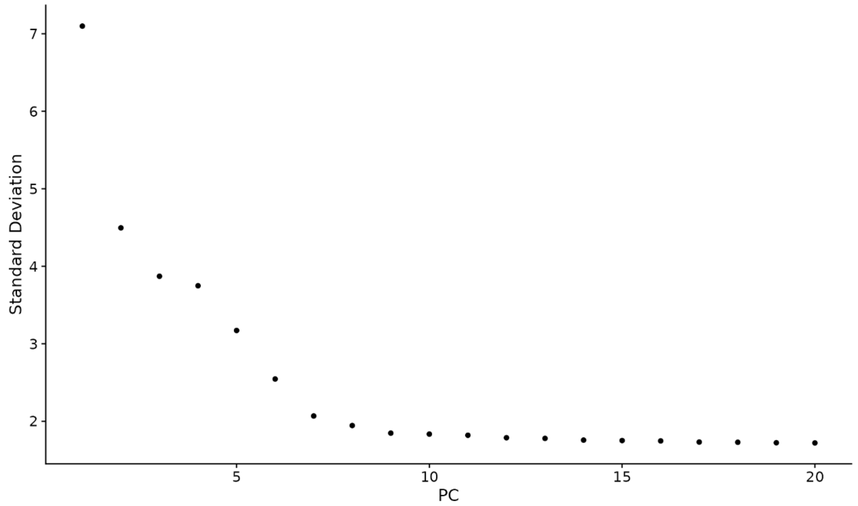


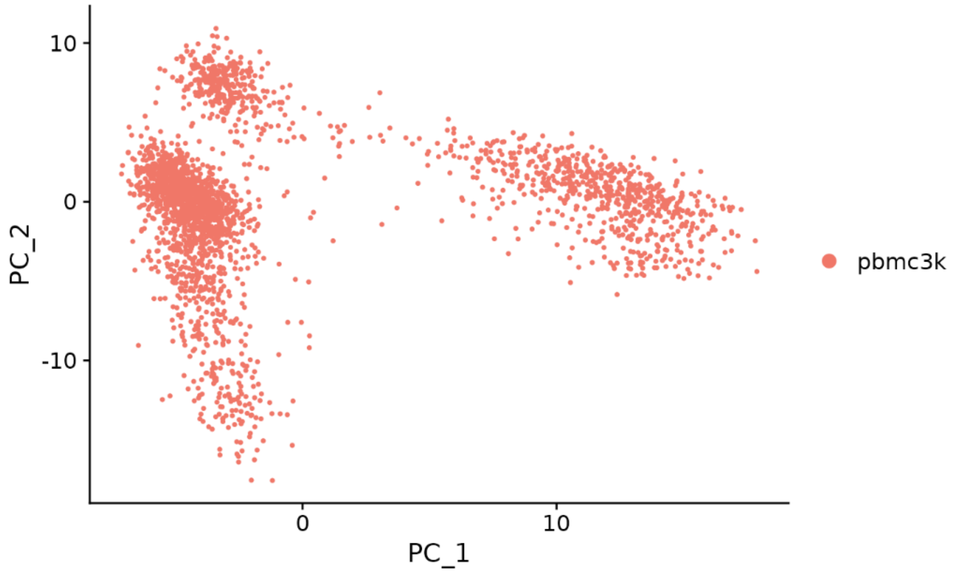
PCA is a statistical technique for reducing the dimensionality of the dataset by linearly transforming the data into a new coordinate system where most of the variation in the data can be described

```
#========== Dimensional Reduction =========#
# PCA
countsFiltered <- RunPCA(countsFiltered, features = VariableFeatures(object = countsFiltered))
```

Linear Dimensional Reduction

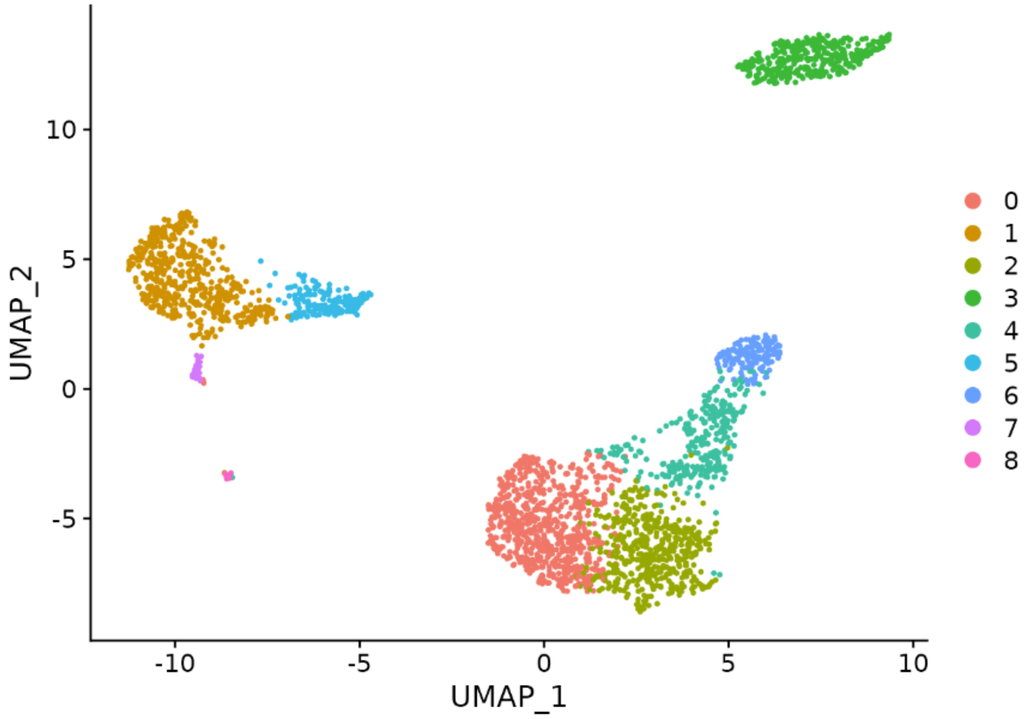







```
#========== Dimensional Reduction =========#
# PCA
countsFiltered <- RunPCA(countsFiltered, features = VariableFeatures(object = countsFiltered))
```

Non - Linear Dimensional Reduction




```
#============= Cluster the cells ==============#
countsFiltered <- FindNeighbors(countsFiltered, dims = 1:10)
countsFiltered <- FindClusters(countsFiltered, resolution = 0.5)
head(Idents(countsFiltered), 5)

#============= Non-linear dimensional reduction ==============#
#Run UMAP
countsFiltered <- RunUMAP(countsFiltered, dims = 1:6)

# individual clusters
DimPlot(countsFiltered, reduction = "umap",label = TRUE, repel = TRUE)
```

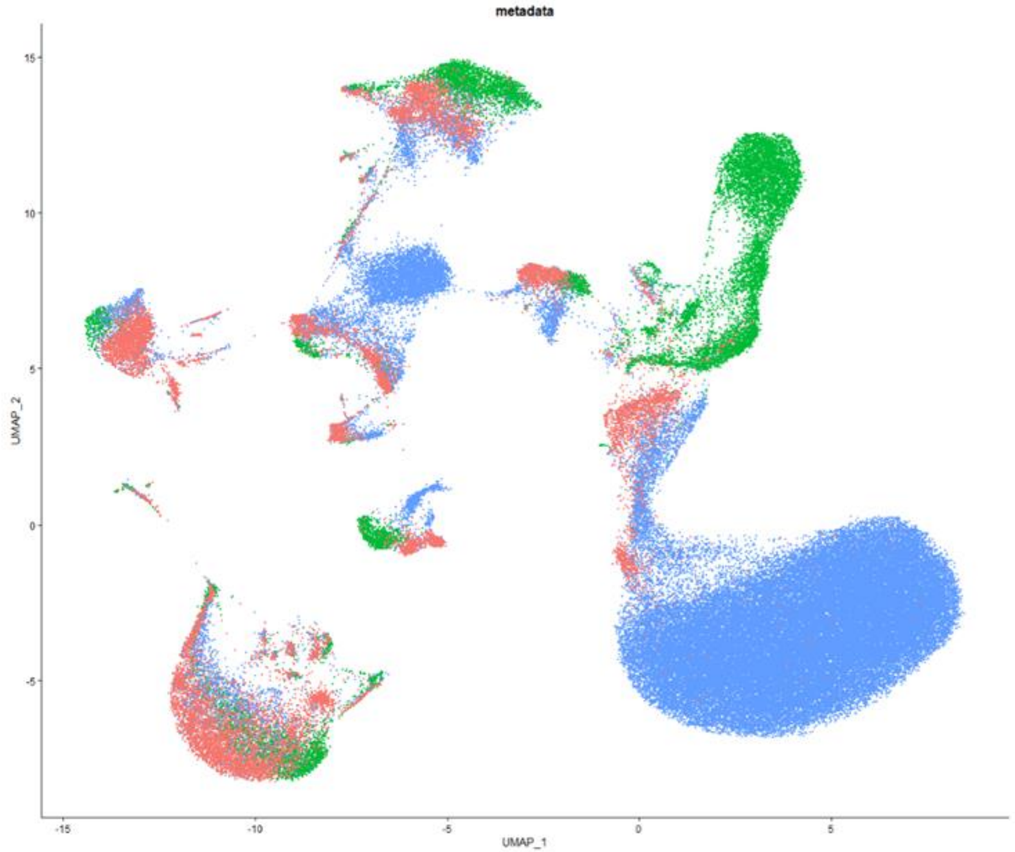
First we need to cluster cells!

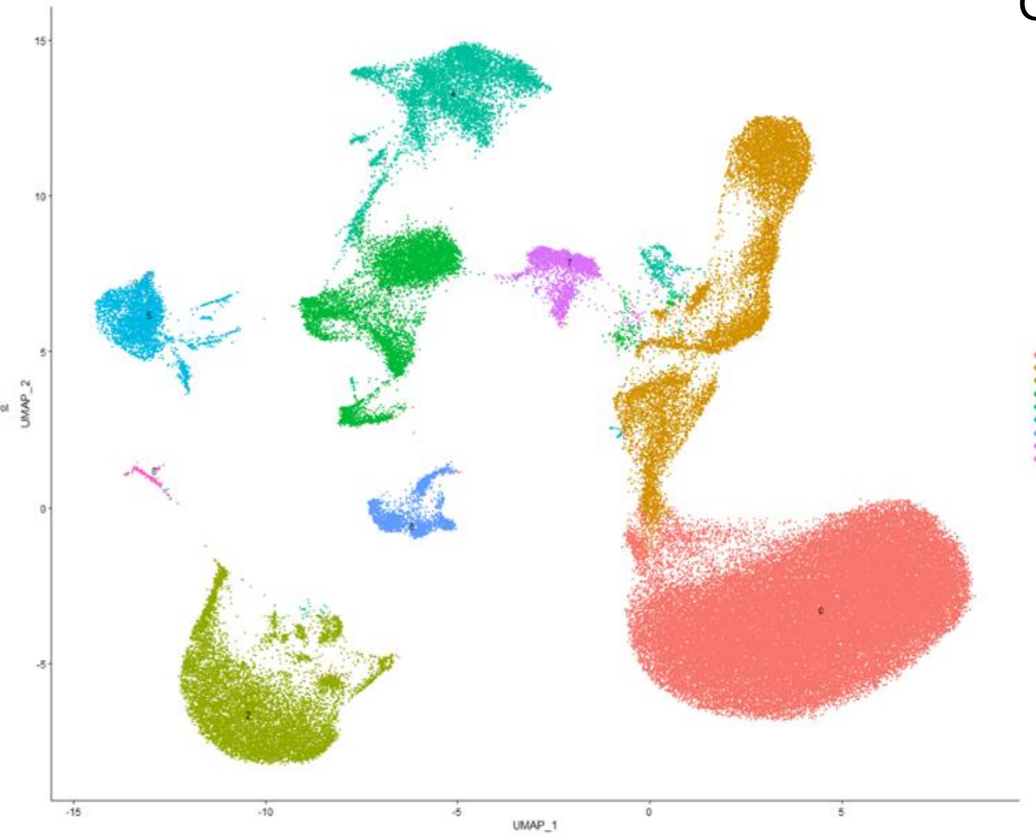
Clustering groups a set of objects in a way that objects in the same group (**cluster**) are more similar to each other than to those in other groups

Adapted from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Non - Linear Dimensional Reduction

We can take advantage of these methods to identify potential bath effects



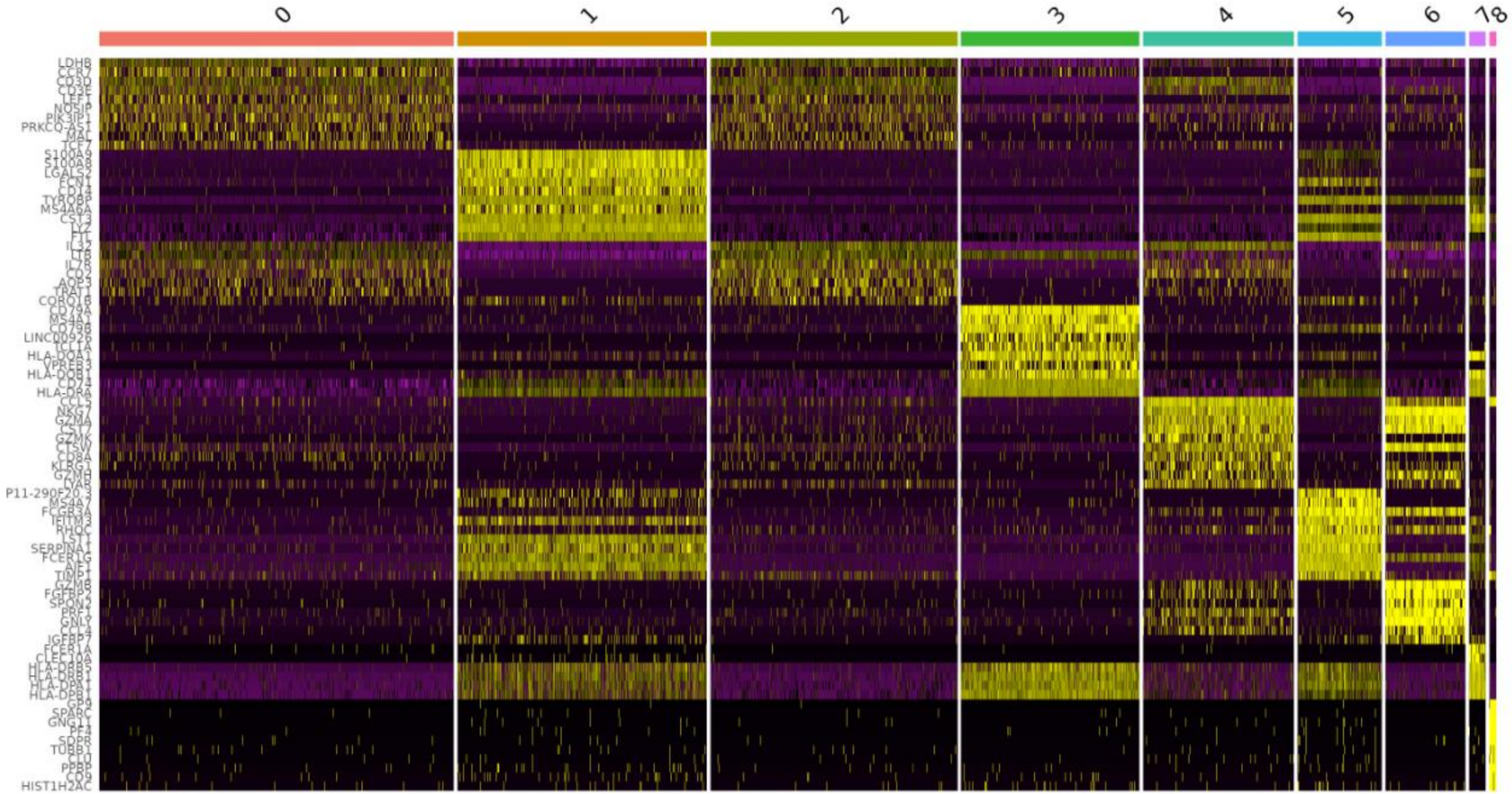


Clusters are driven by samples

↓

Batch correction

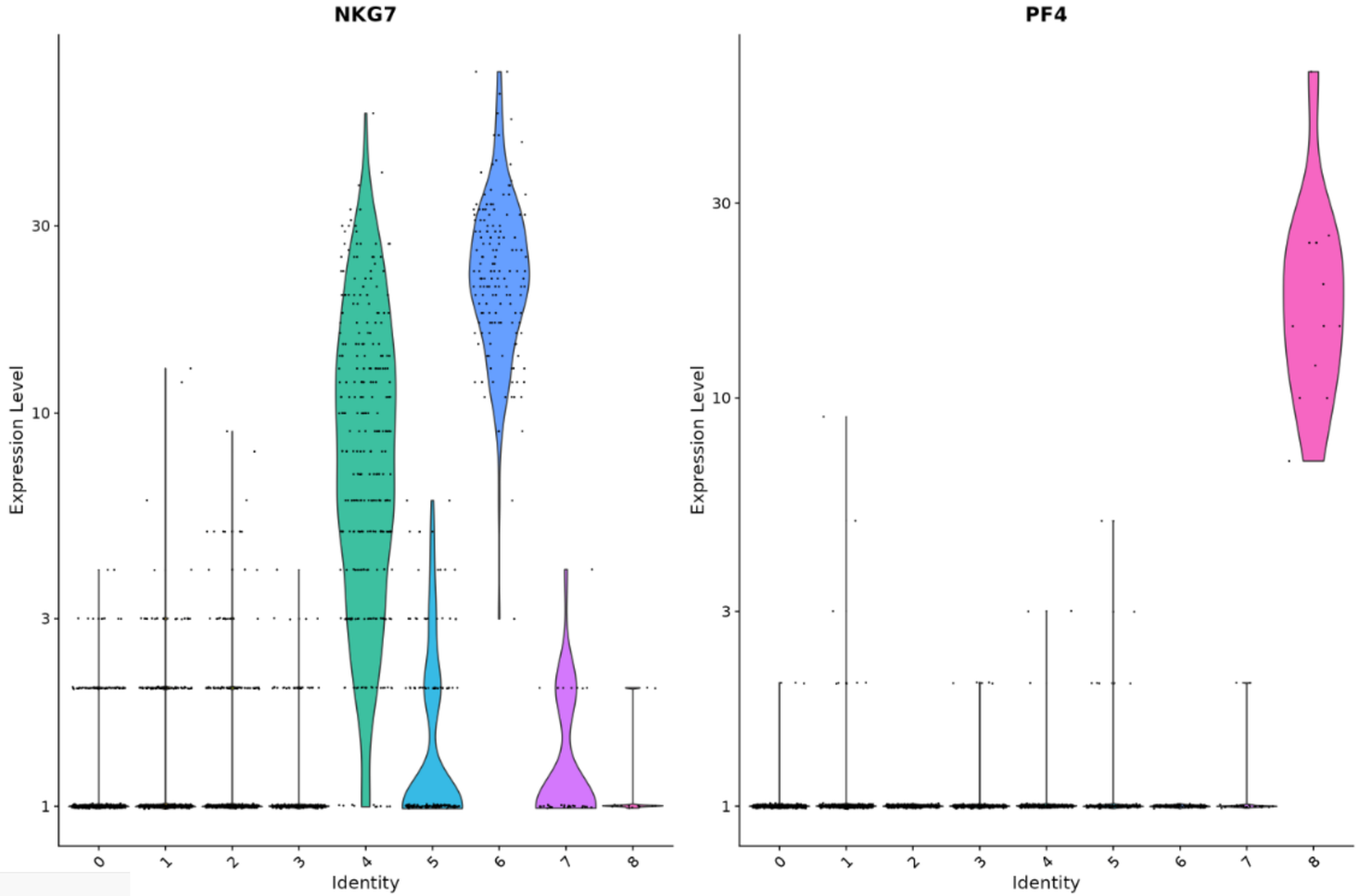
Finding Differentially Expressed Features




```
# find all markers of cluster 8
cluster8.markers <- FindMarkers(countsFiltered, ident.1 = 8, min.pct = 0.25)
```

Adapted from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

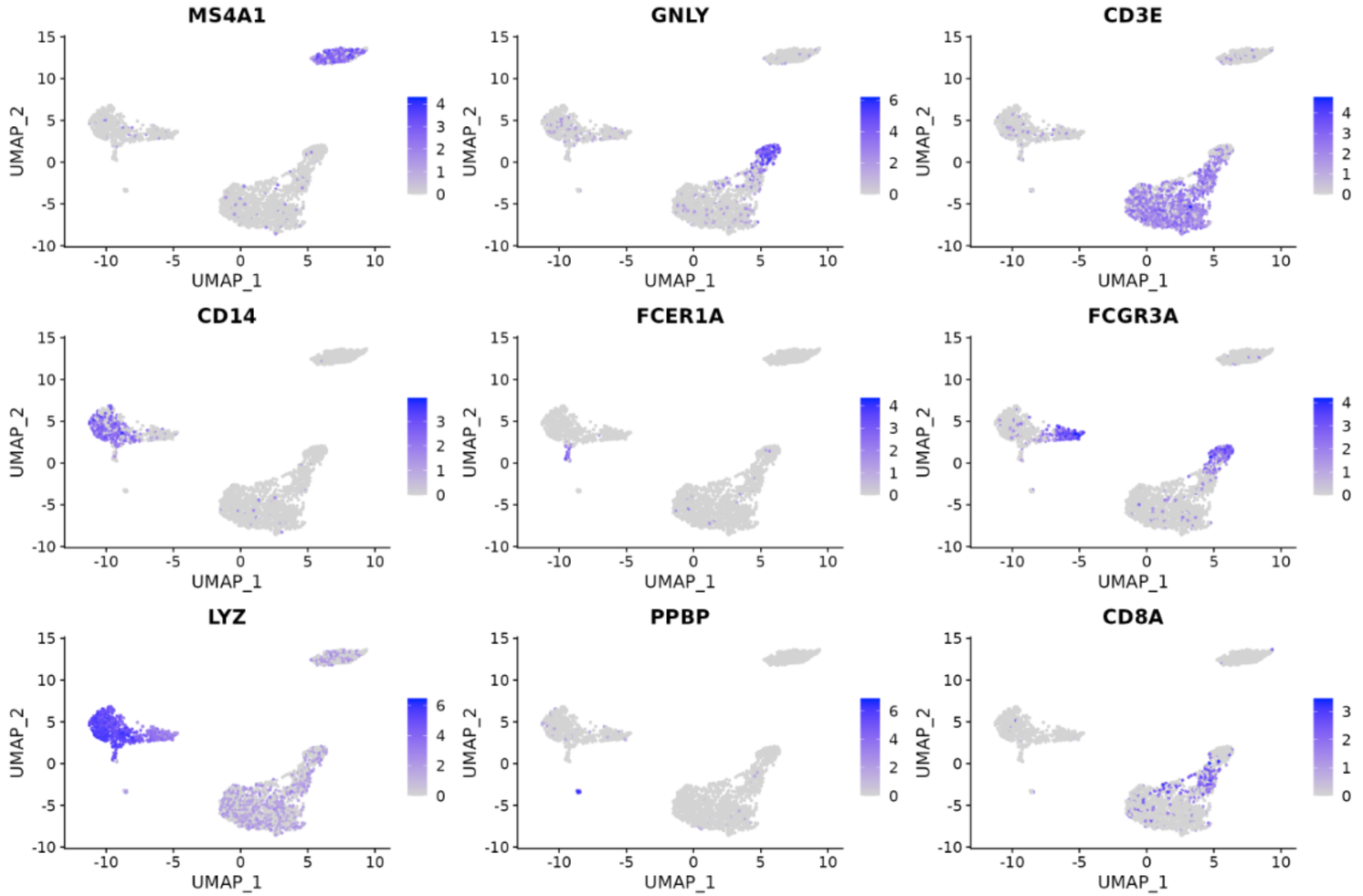
Finding Differentially Expressed Features




```
# you can plot raw counts as well
VlnPlot(pbmc, features = c("NKG7", "PF4"), slot = "counts", log = TRUE)
```

Adapted from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

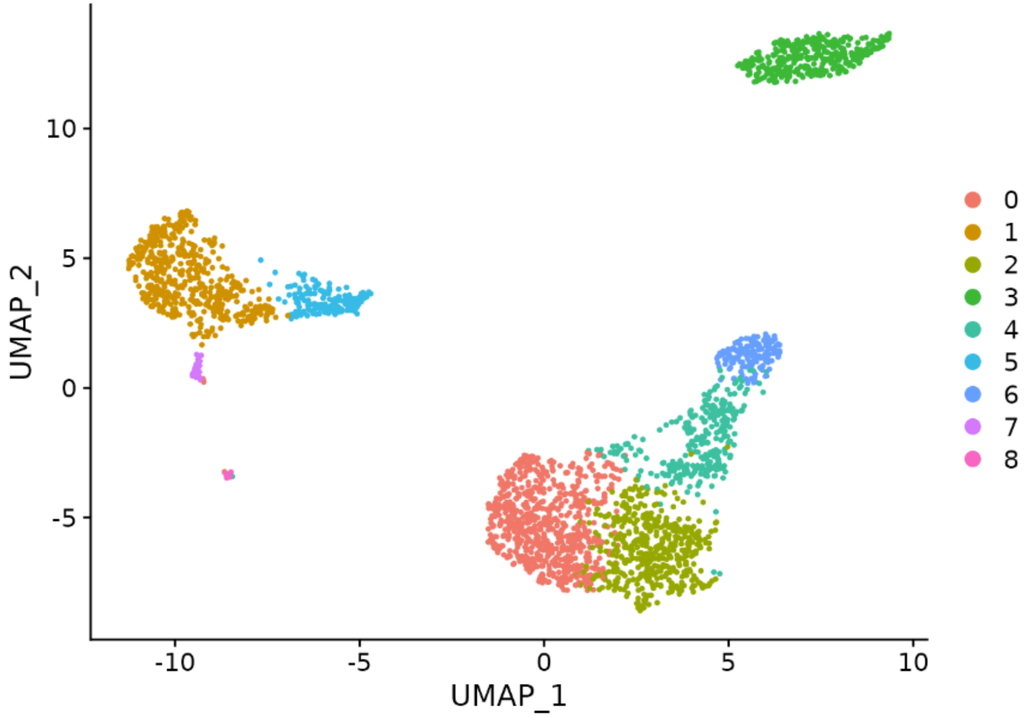
Finding Differentially Expressed Features




```
FeaturePlot(pbmc, features = c("MS4A1", "GNLY", "CD3E", "CD14", "FCER1A", "FCGR3A", "LYZ", "PPBP",
    "CD8A"))
```

Adapted from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Assigning Cell Identities to Clusters



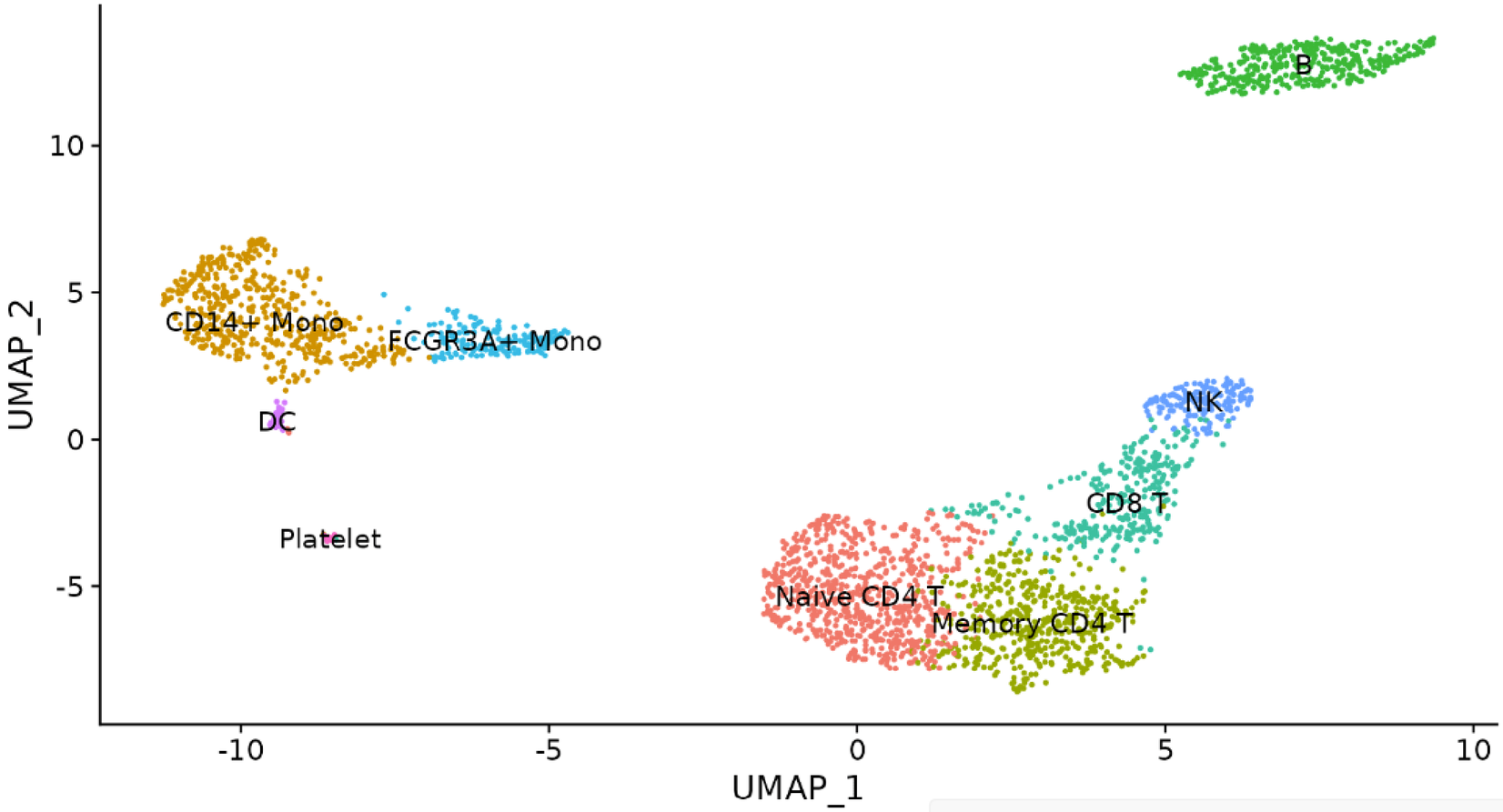
How to assign cell identities:

- Manually
- SingleR
- Using Machine Learning techniques

```
new.cluster.ids <- c("Naive CD4 T", "CD14+ Mono", "Memory CD4 T", "B", "CD8 T", "FCGR3A+ Mono",
    "NK", "DC", "Platelet")
names(new.cluster.ids) <- levels(pbmc)
pbmc <- RenameIdents(pbmc, new.cluster.ids)
DimPlot(pbmc, reduction = "umap", label = TRUE, pt.size = 0.5) + NoLegend()
```

Adapted from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Assigning Cell Identities to Clusters



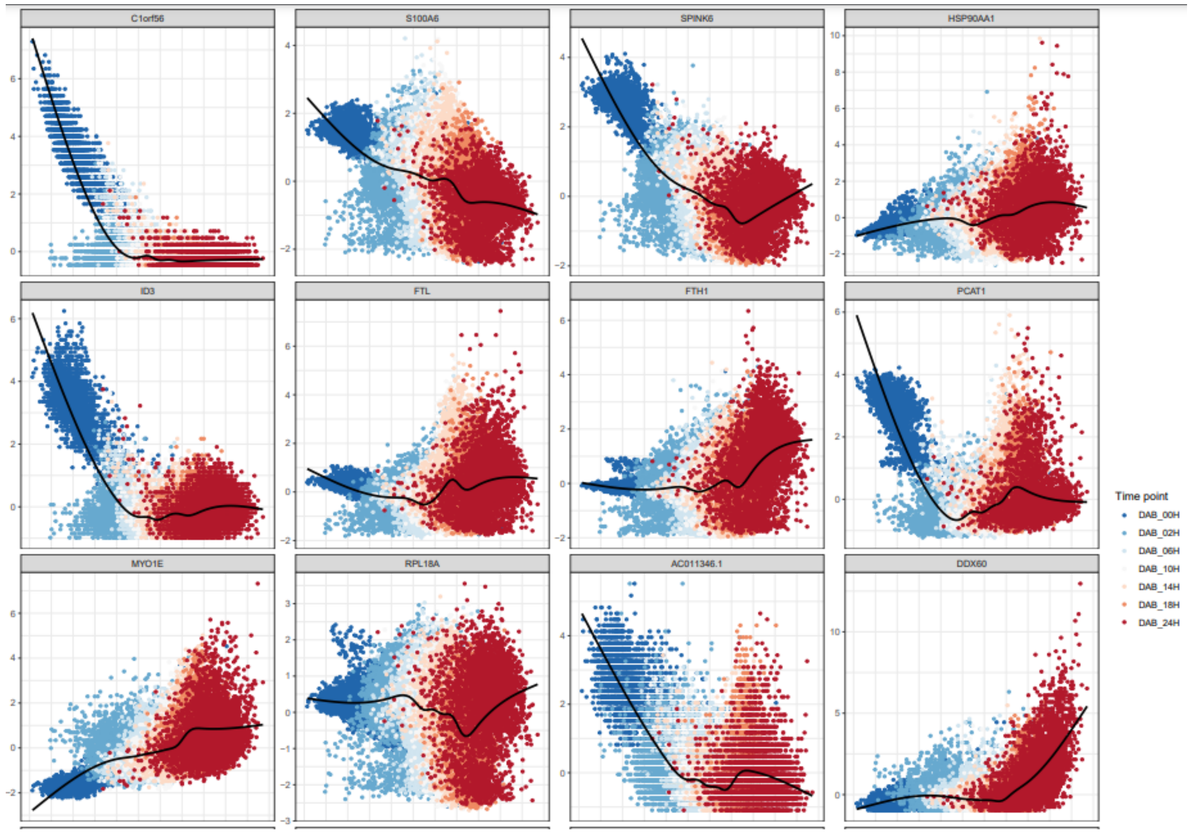
How to assign cell identities:

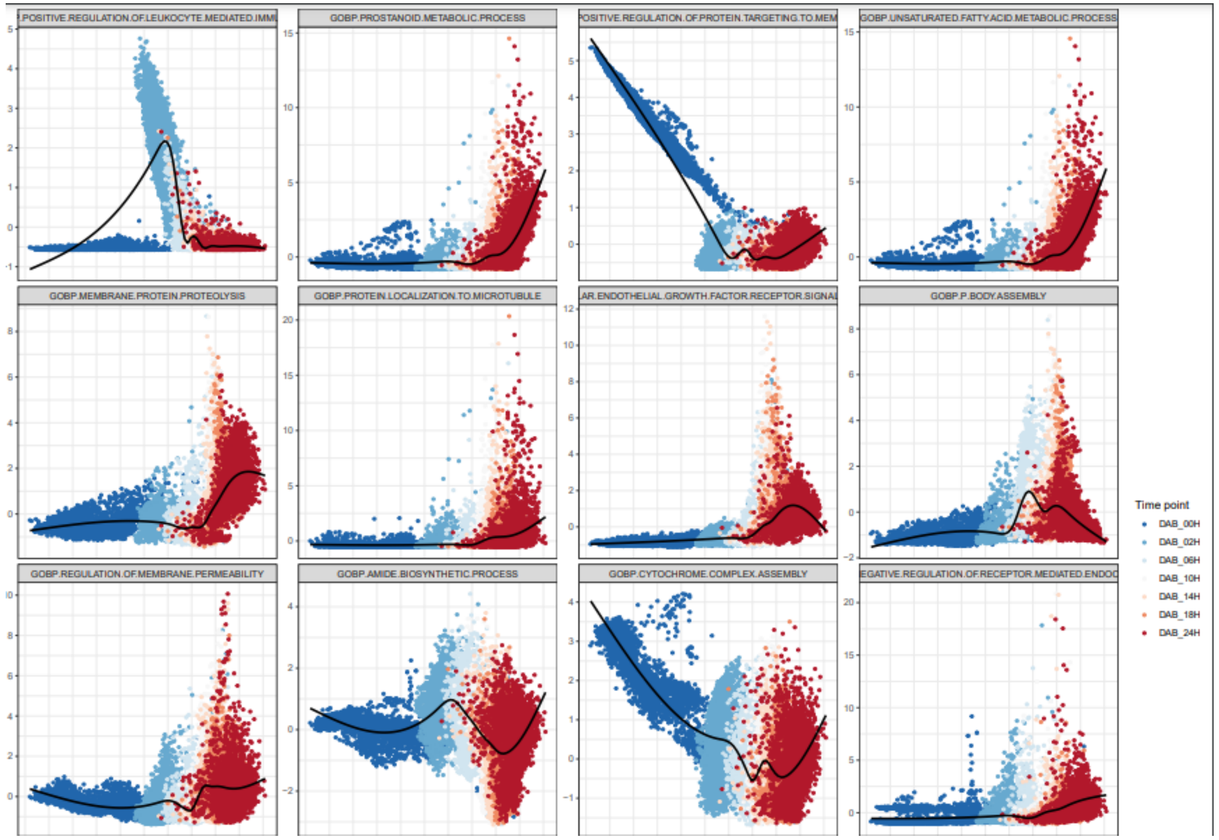
- Manually
- SingleR
- Using Machine Learning techniques

```
new.cluster.ids <- c("Naive CD4 T", "CD14+ Mono", "Memory CD4 T", "B", "CD8 T", "FCGR3A+ Mono",
    "NK", "DC", "Platelet")
names(new.cluster.ids) <- levels(pbmc)
pbmc <- RenameIdents(pbmc, new.cluster.ids)
DimPlot(pbmc, reduction = "umap", label = TRUE, pt.size = 0.5) + NoLegend()
```

Adapted from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Pseudotime Analysis





Melania Franchini and others, Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data, *NAR Genomics and Bioinformatics*, Volume 5, Issue 1, March 2023,

Introduction to single cell RNA-seq

Xabier Bujanda Cundin
17/07/2023