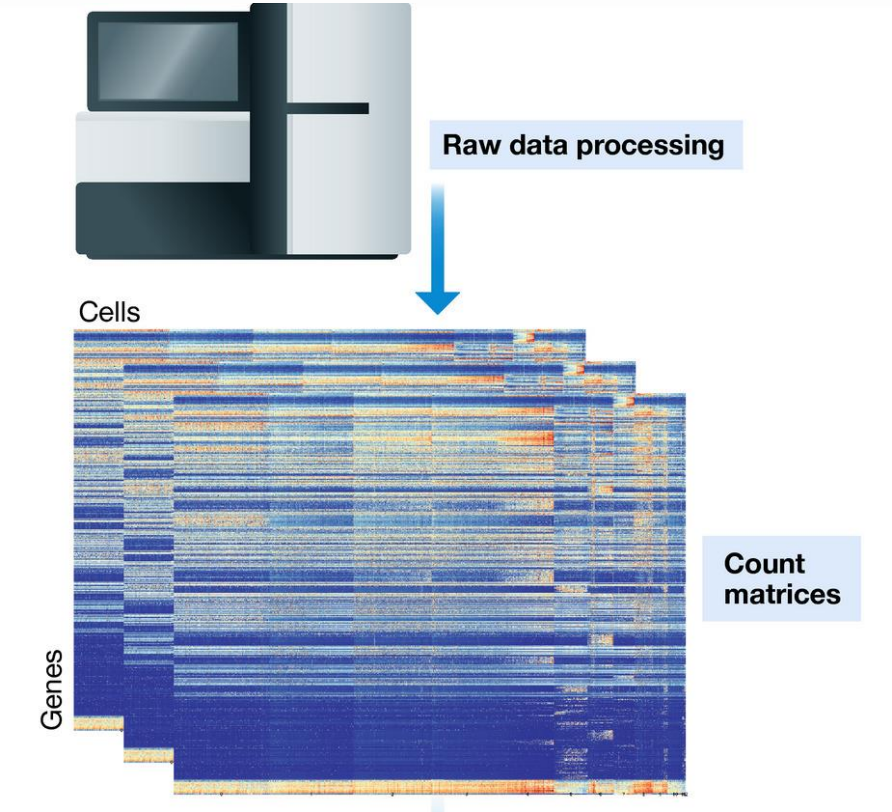

# From the Single Cell RNA-Seq Count Matrix to Biological Knowledge

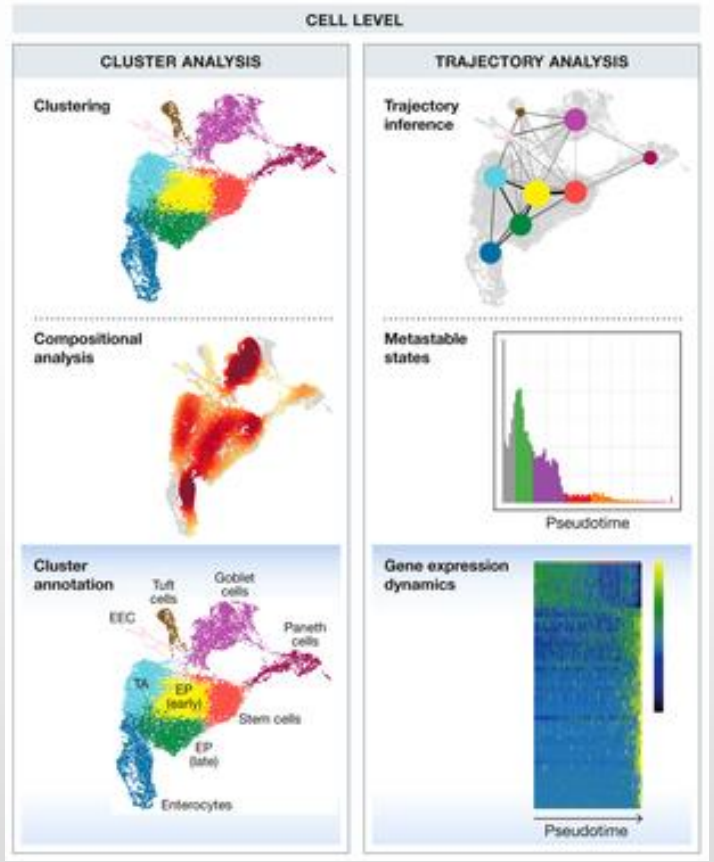
## With the help of Seurat R package

**Dena Leshkowitz**

**Bioinformatics Unit**

# From the Count Matrix to Biological Knowledge

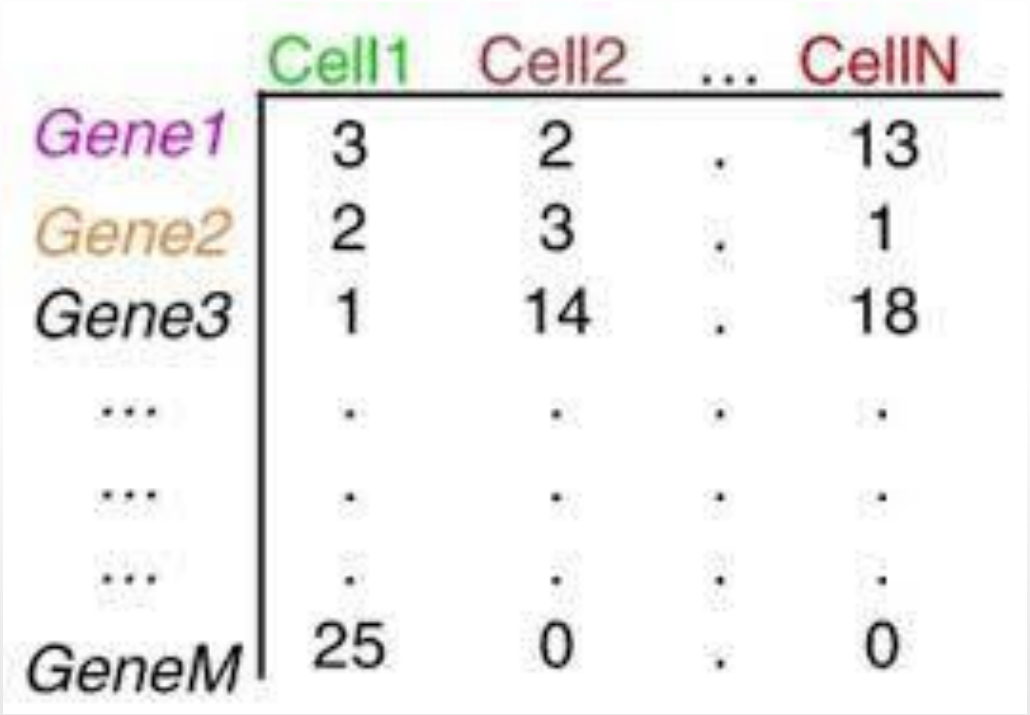



Mol Syst Biol, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

# scRNA-Seq Count Matrix

- The count matrix after running CellRanger consists of the cell barcodes we consider as “real cells”
- The counts in CellRanger output are the UMI counts per gene
- The count matrix is large and sparse

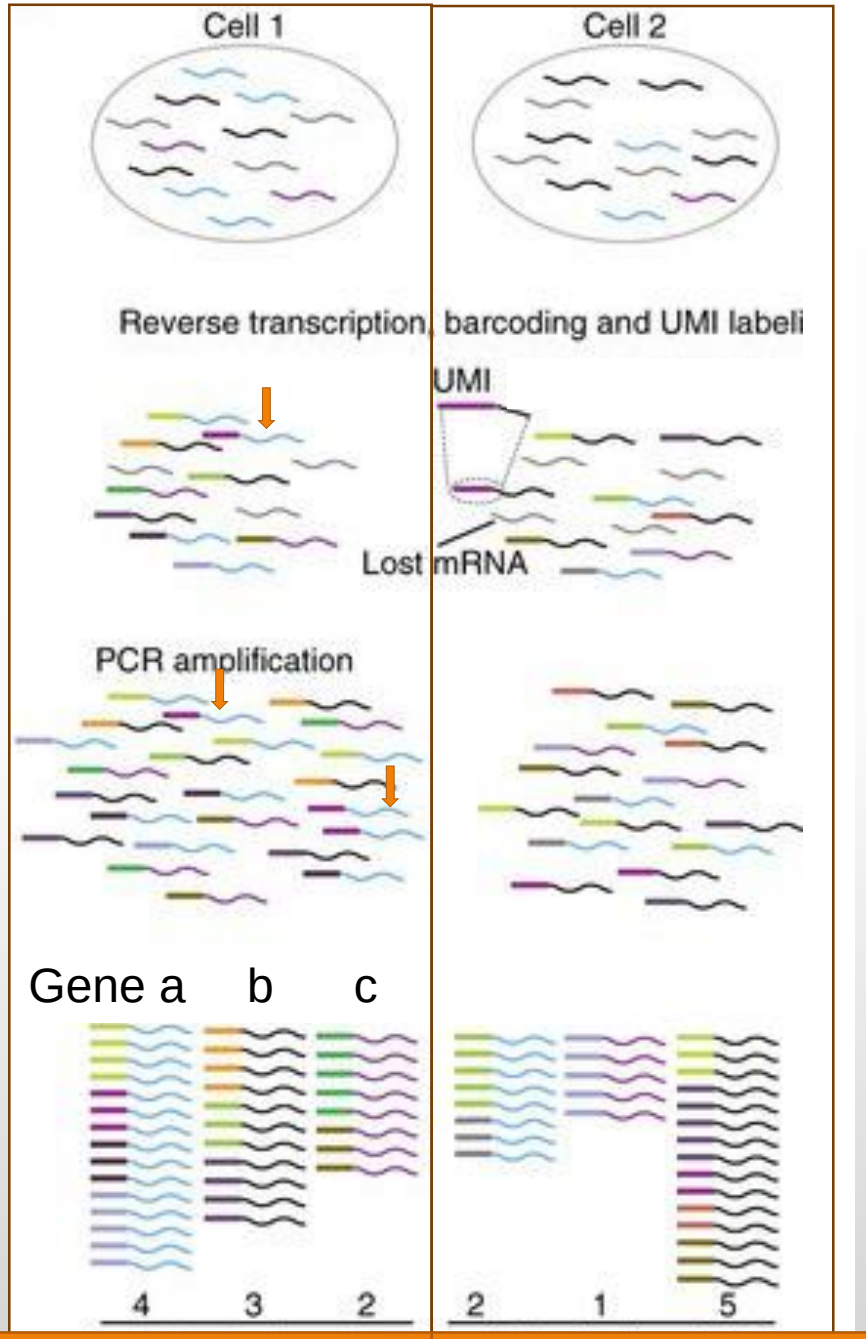


| | Cell1 | Cell2 | … | CellN |
|---|---|---|---|---|
| *Gene1* | 3 | 2 | . | 13 |
| *Gene2* | 2 | 3 | . | 1 |
| *Gene3* | 1 | 14 | . | 18 |
| … | . | . | . | . |
| … | . | . | . | . |
| … | . | . | . | . |
| *GeneM* | 25 | 0 | . | 0 |

# UMI Counts

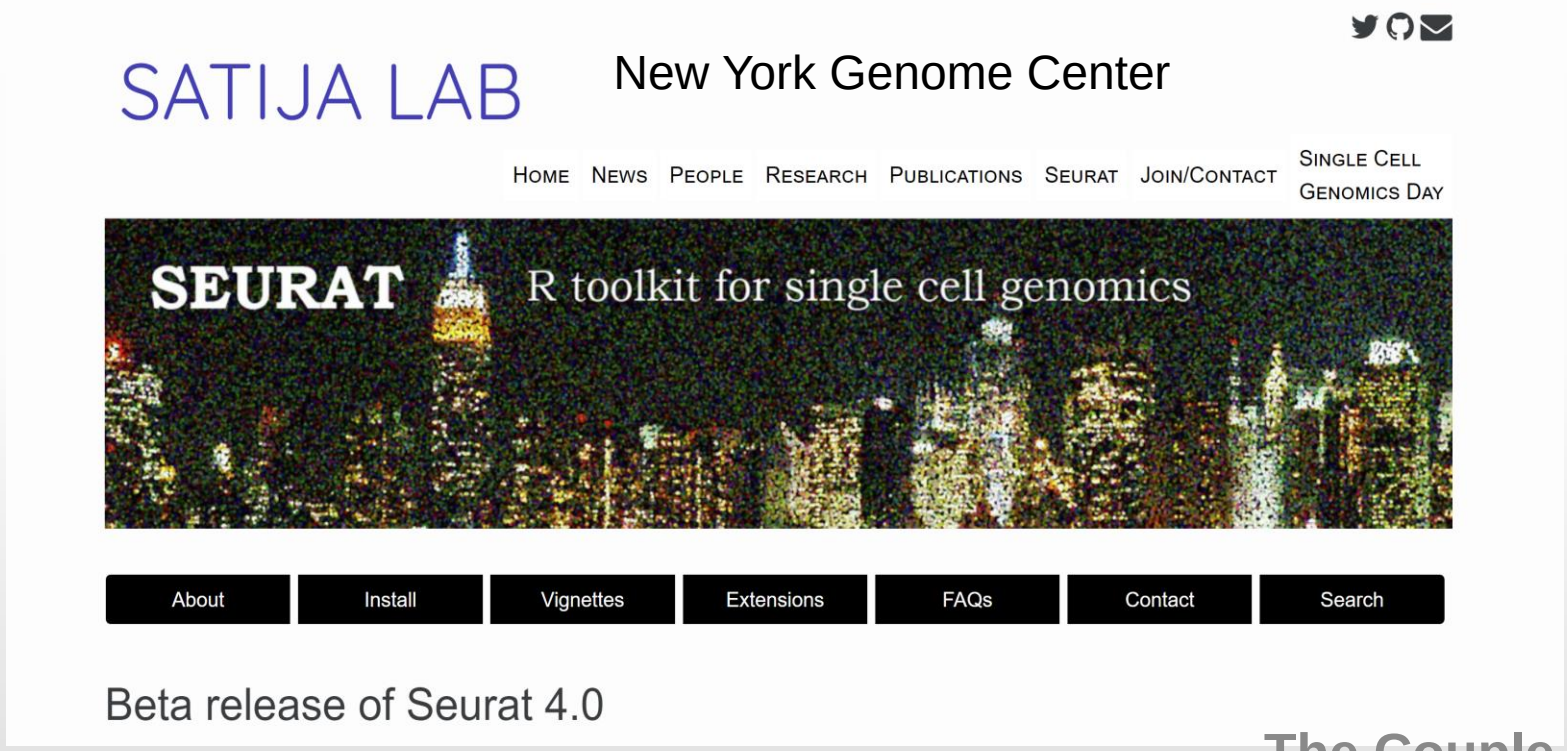
We count UMI (Unique Molecular Identifier) in order to remove PCR amplification

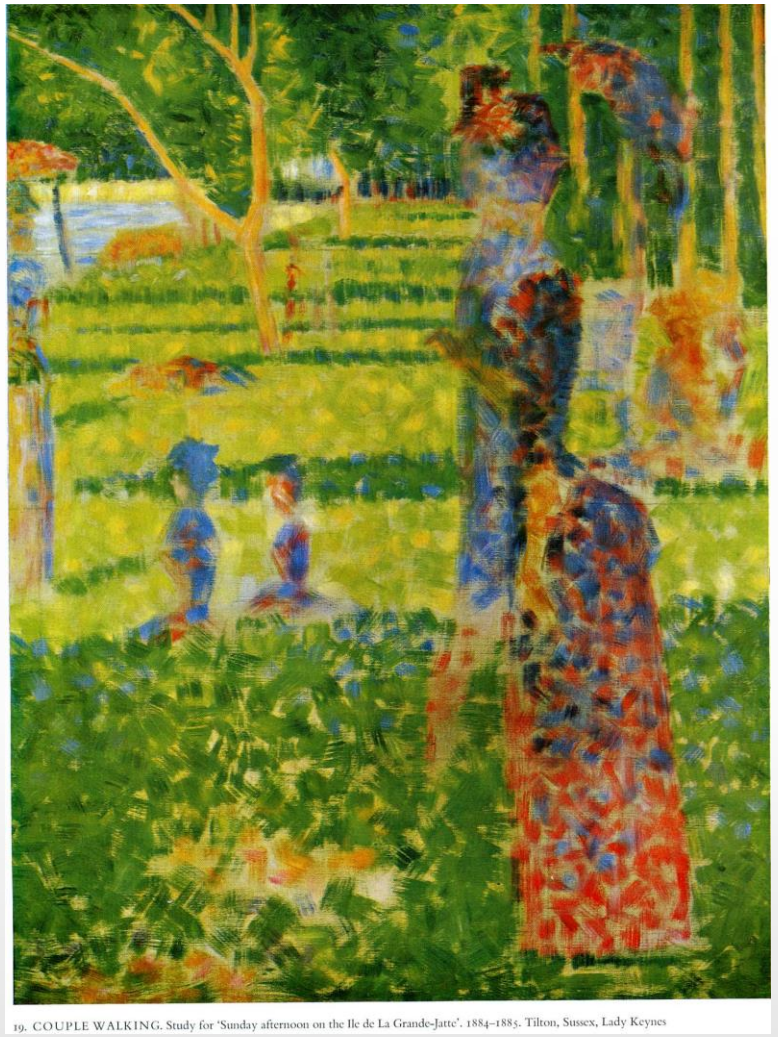
- Reads are considered duplicated, if they map to the same gene and have the same UMI
- Instead of counting reads-sequences we will count number of unique UMIs per gene per cell.




This figure is adapted from Islam et al (2014)

# Analysis with Seurat Package





19. COUPLE WALKING. Study for 'Sunday afternoon on the Ile de La Grande-Jatte'. 1884–1885. Tilton, Sussex, Lady Keynes

**The Couple**

Georges Seurat

Date: 1884; France

Style: Pointillism, Neo-Impressionism

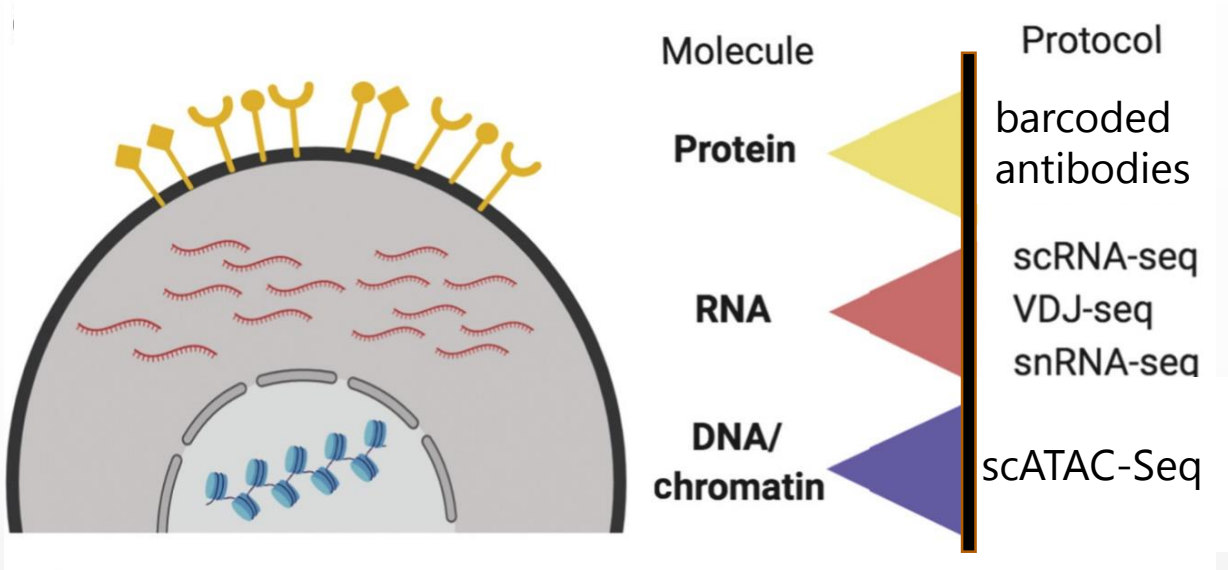
https://www.wikiart.org/en/georges-seurat/the-couple-1884

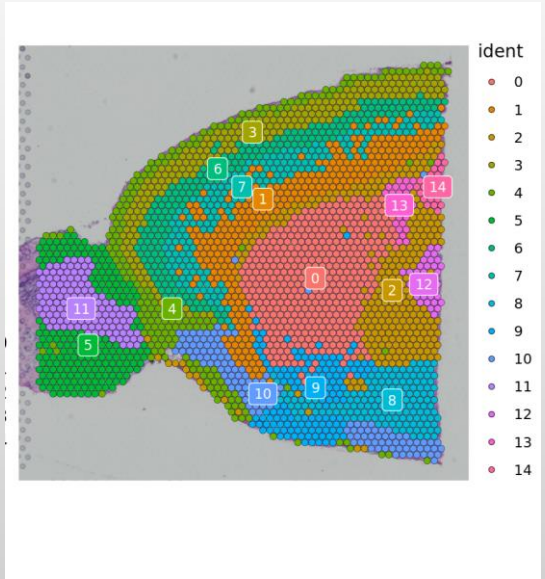
# Seurat package

- R language package designed for QC and exploration of single cell RNA-seq data
- Using Seurat we can control many of the various parameters in the analysis - advantage over CellRanger
- Widely used in the community & frequent updates – new computational approaches
- Supports **Integrative multimodal analysis.** The ability to analyse measurements of multiple data types that were collected simultaneously from the same cells including protein levels, chromatin state, spatial location and more.

**Integrative multimodal analysis.**

*Modified - Trends in Immunology* 2019 401011-1021 DOI: (10.1016/j.it.2019.09.004)



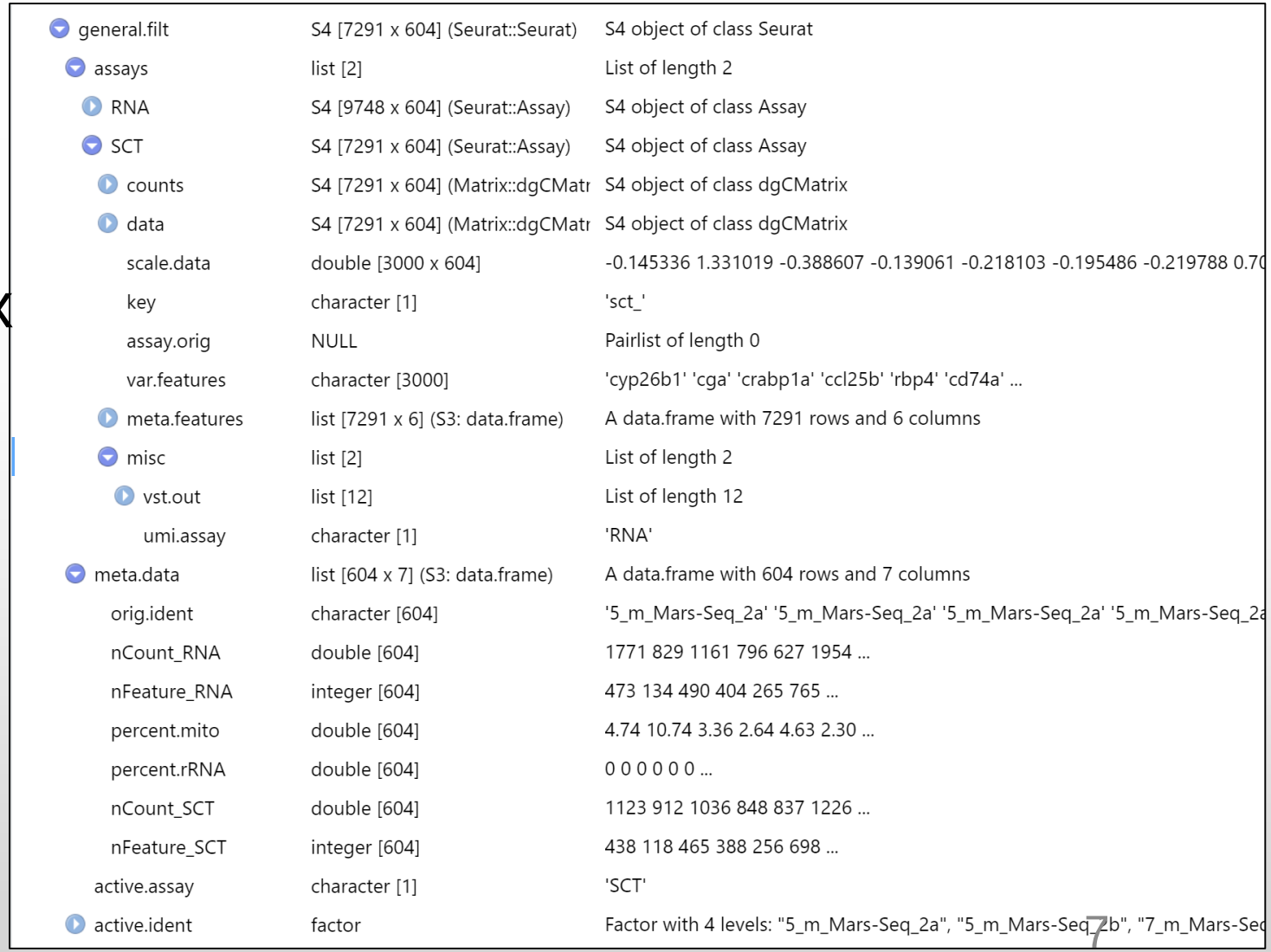




Spatial information Visium technology

# The Seurat Object

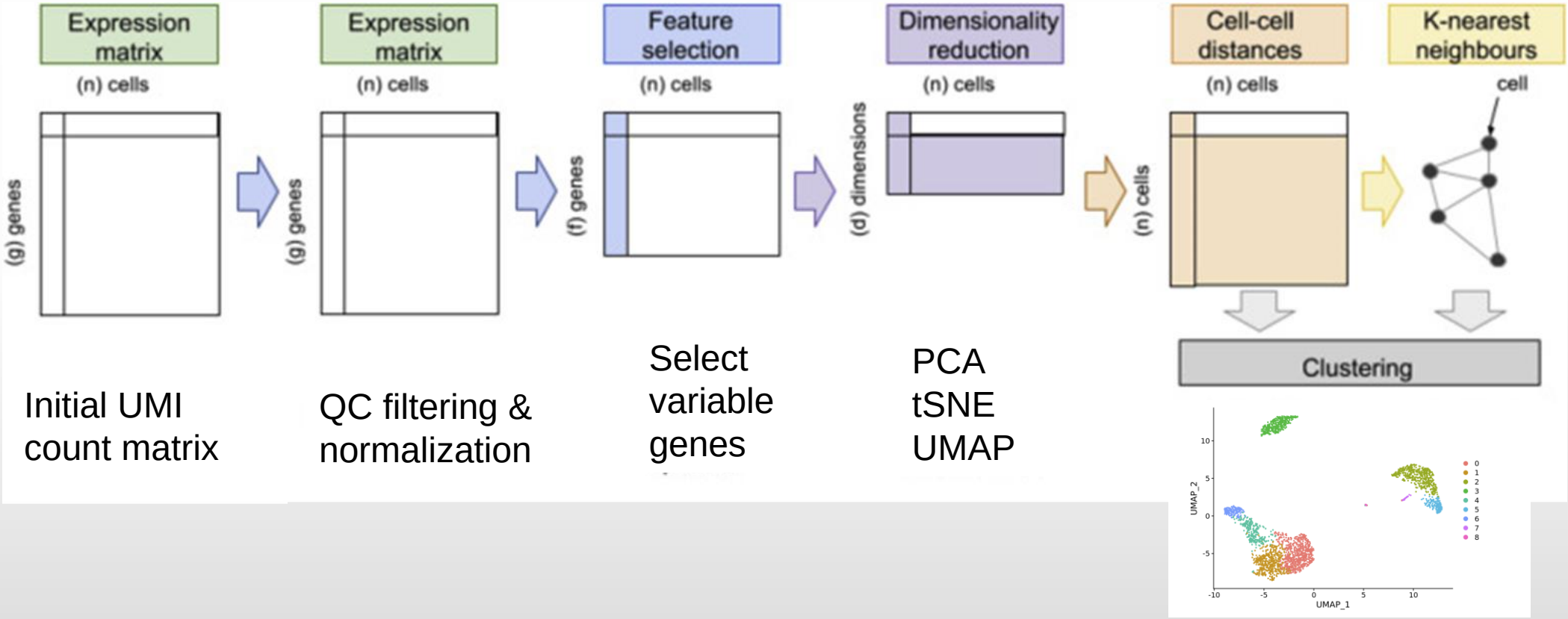
The object serves as a container for the:

- Data - like the count matrix
- Analysis - like PCA, or clustering results



| | | |
|---|---|---|
| general.filt | S4 [7291 x 604] (Seurat::Seurat) | S4 object of class Seurat |
| assays | list [2] | List of length 2 |
| RNA | S4 [9748 x 604] (Seurat::Assay) | S4 object of class Assay |
| SCT | S4 [7291 x 604] (Seurat::Assay) | S4 object of class Assay |
| counts | S4 [7291 x 604] (Matrix::dgCMatr | S4 object of class dgCMatrix |
| data | S4 [7291 x 604] (Matrix::dgCMatr | S4 object of class dgCMatrix |
| scale.data | double [3000 x 604] | -0.145336 1.331019 -0.388607 -0.139061 -0.218103 -0.195486 -0.219788 0.70 |
| key | character [1] | 'sct_' |
| assay.orig | NULL | Pairlist of length 0 |
| var.features | character [3000] | 'cyp26b1' 'cga' 'crabp1a' 'ccl25b' 'rbp4' 'cd74a' ... |
| meta.features | list [7291 x 6] (S3: data.frame) | A data.frame with 7291 rows and 6 columns |
| misc | list [2] | List of length 2 |
| vst.out | list [12] | List of length 12 |
| umi.assay | character [1] | 'RNA' |
| meta.data | list [604 x 7] (S3: data.frame) | A data.frame with 604 rows and 7 columns |
| orig.ident | character [604] | '5_m_Mars-Seq_2a' '5_m_Mars-Seq_2a' '5_m_Mars-Seq_2a' '5_m_Mars-Seq_2a |
| nCount_RNA | double [604] | 1771 829 1161 796 627 1954 ... |
| nFeature_RNA | integer [604] | 473 134 490 404 265 765 ... |
| percent.mito | double [604] | 4.74 10.74 3.36 2.64 4.63 2.30 ... |
| percent.rRNA | double [604] | 0 0 0 0 0 0 ... |
| nCount_SCT | double [604] | 1123 912 1036 848 837 1226 ... |
| nFeature_SCT | integer [604] | 438 118 465 388 256 698 ... |
| active.assay | character [1] | 'SCT' |
| active.ident | factor | Factor with 4 levels: "5_m_Mars-Seq_2a", "5_m_Mars-Seq_2b", "7_m_Mars-Sec |

# Analysis Workflow




Modified plot from-

Andrews et al. Molecular Aspects of Medicine
Volume 59, February 2018, Pages 114-122

# Cell Quality Control (QC)

Cell QC is commonly performed based on three criteria:

- The number of UMI counts per cell (transcription depth)
- The number of genes per cell
- The percent counts from mitochondrial genes

Aim: We would like to filter out the cells which are outliers, bad quality.

# Cell Quality Control

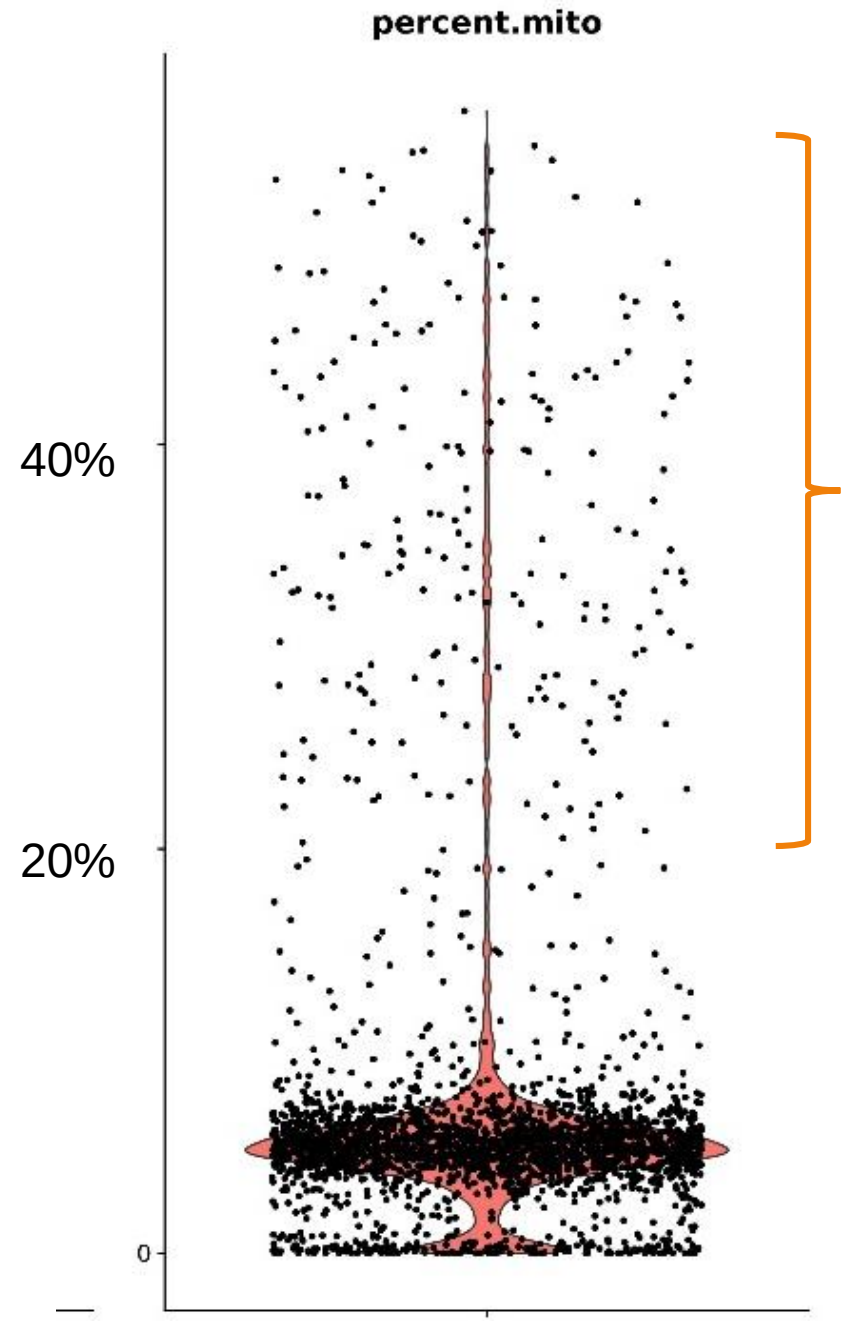
- Removing damaged cells
  - We want to remove damaged cells whose cytoplasmic mRNA has leaked out through a broken membrane, these cells can still maintain the mRNA located in the mitochondria.
  - These cells will have:
    - High percent of mitochondrial gene counts (out of total UMI counts)
    - Low UMI count depth
    - Few detected genes
- Question: Is their an additional reason to filter out cells with low gene or UMI counts?

# Question

Is their an additional reason to filter out cells with low gene or UMI counts?

# Violin plot

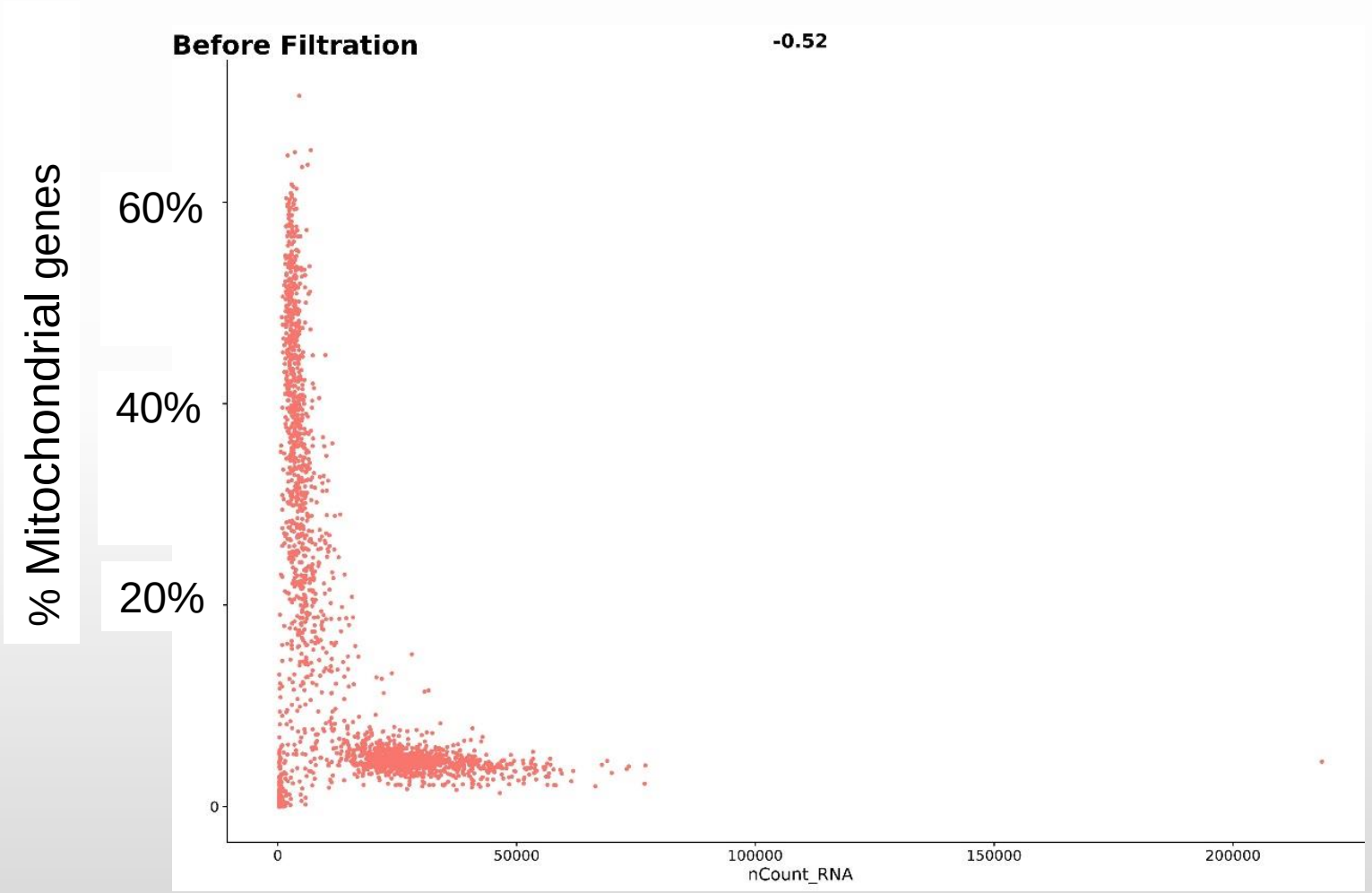
- We use violin plots to view the distribution of counts
- A violin plot is similar to a box plot, with the addition of a rotated kernel density plot on each side
- Each point represents a cell




Remove cells that have above 20% mitochondrial genes

# Mitochondrial genes

The % of mitochondrial genes is anti-correlated with the expression of cellular genes

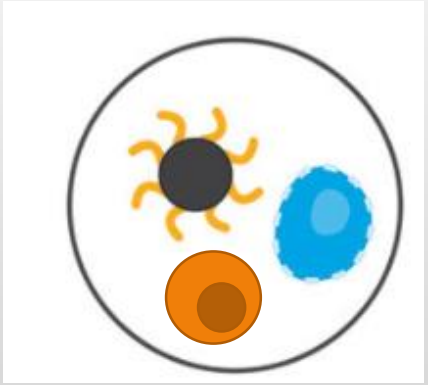



Number of genes detected per cell

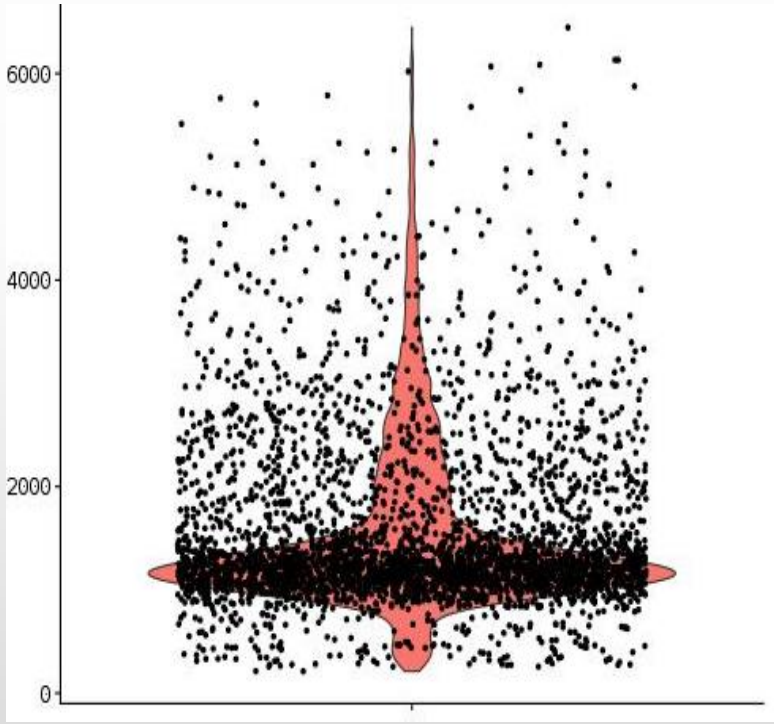
# Cell Quality Control

Removing doublets: barcodes representing doublet cells will have:

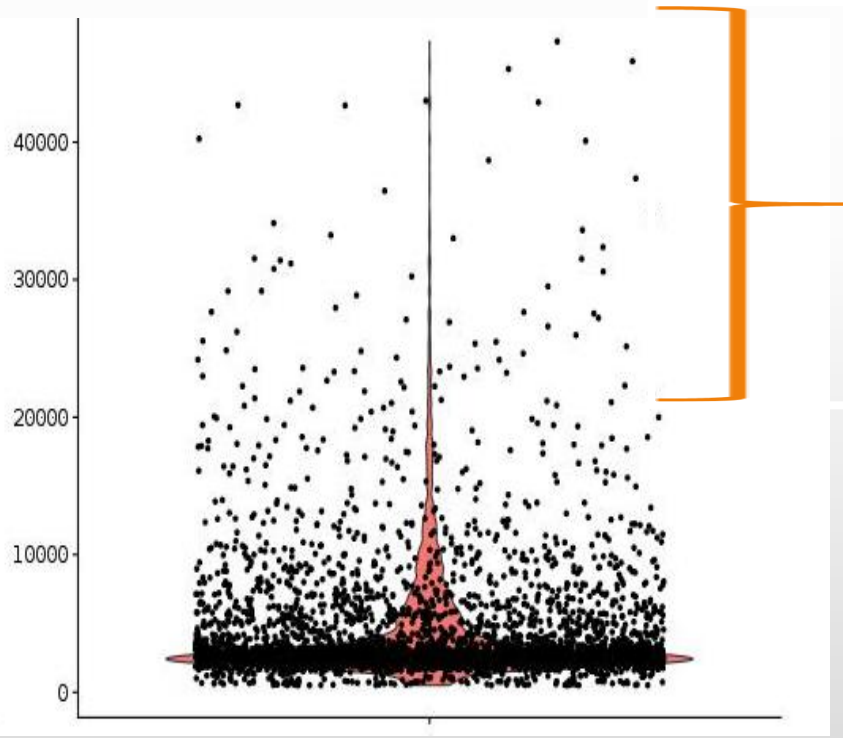
- High UMI counts
- Large number of detected genes



Number of genes

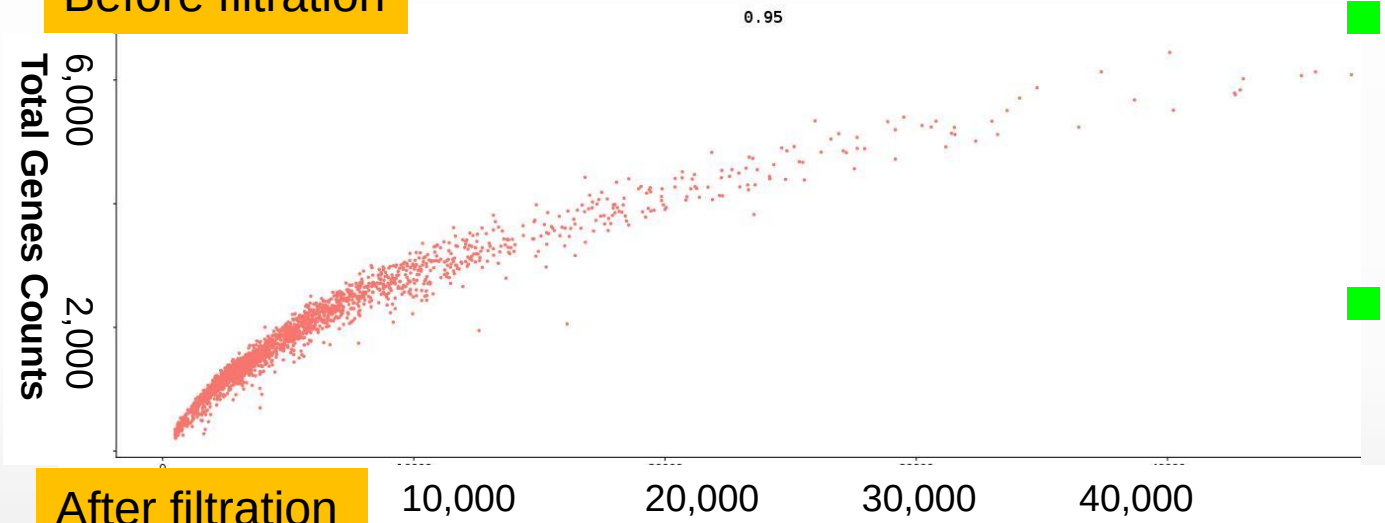



Number of UMIs

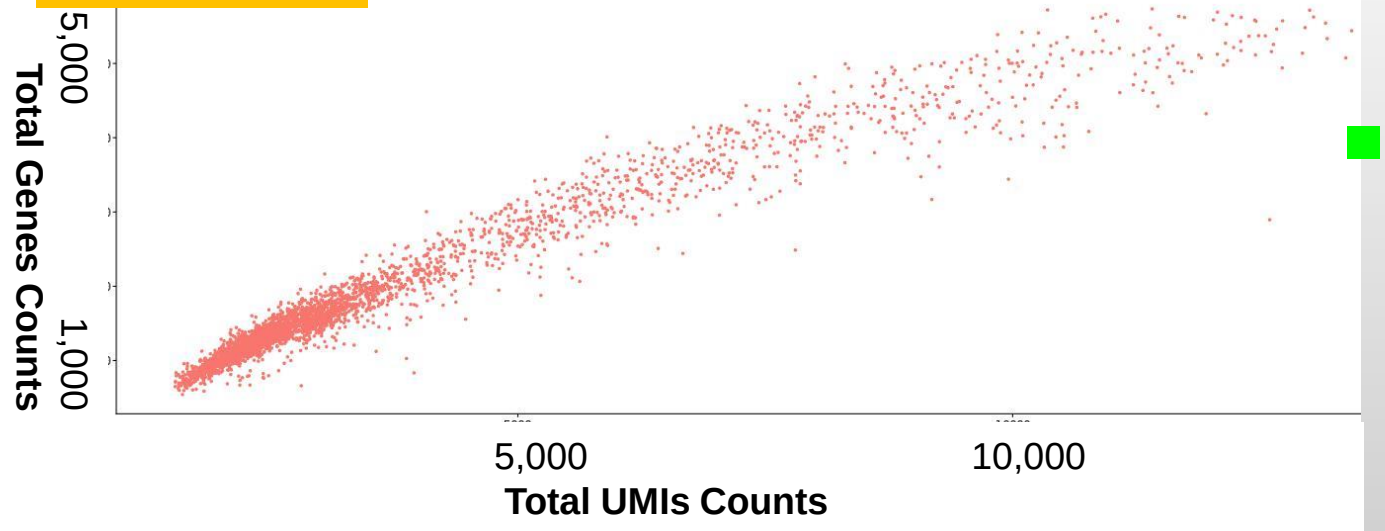


# UMI counts and Gene counts

Before filtration




After filtration




- Total number of UMI counts & genes counts are highly correlated
- The counts can vary significantly between cells, spanning more than one order of magnitude
- Question: What is required in order to perform comparisons of gene expression between the cells?

# Question

What is required in order to perform comparisons of gene expression between the cells?

# Normalization

Our goal - is to remove the influence of technical effects in the underlying counts, while preserving true biological variation.

- The use of UMIs in scRNA-seq removes technical variation associated with PCR
- Yet, there are many other sources for technical variation:
  - Cell lysis efficiency
  - Reverse transcription efficiency
  - Stochastic molecular sampling during sequencing

# Normalization

Seurat normalizes the UMI counts measurements for each cell and gene, by:

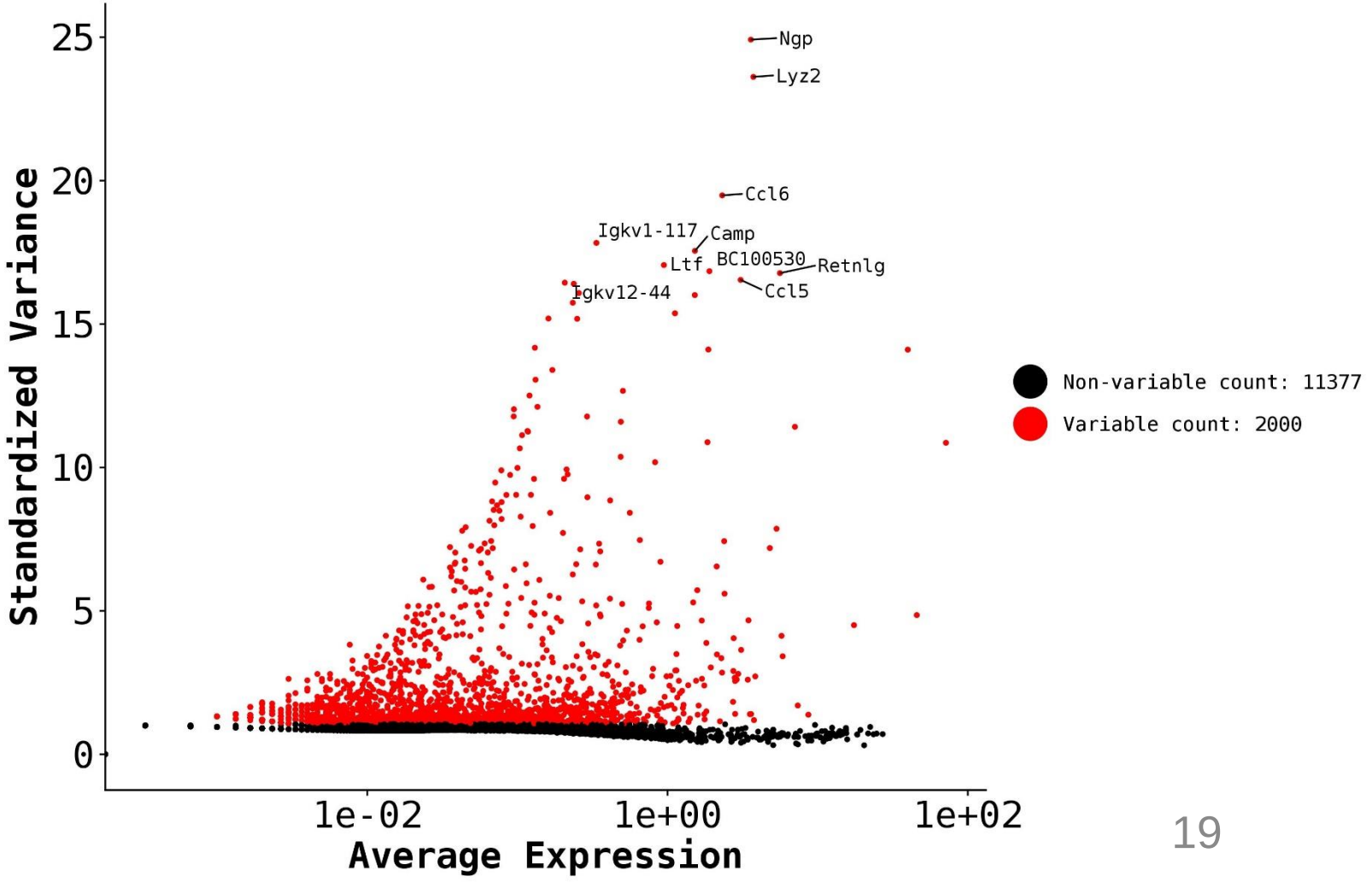
- Dividing by the total counts for that cell
- Multiply by a scaling factor (10,000 by default)
- Log transformation

Normalized values are stored in SueratObject[["RNA"]]@data

Alternatively there is new procedure - sctransform

# Selection of Genes with High Variability

- Not all the genes are used for downstream analysis
- We calculate a subset of genes  that exhibit high cell-to-cell variation
- Selection is done per bin of gene expression
- We select around 2000 genes



# Scaling the Data

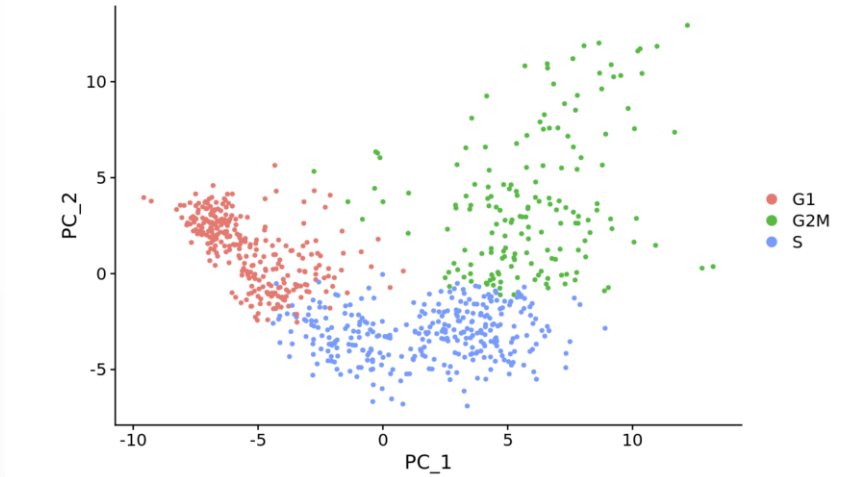
- We scale the data by linear transformation
  - Shifts the expression of each gene, so that the mean expression across cells is 0
  - Scales the expression of each gene, so that the variance across cells is 1
- The results of this are stored in SueratObject[["RNA"]]@scale.data
- By preforming scaling we prevent highly expressed genes from dominating the downstream analysis (highly expressed genes might also have the highest variability)

# Regress-out

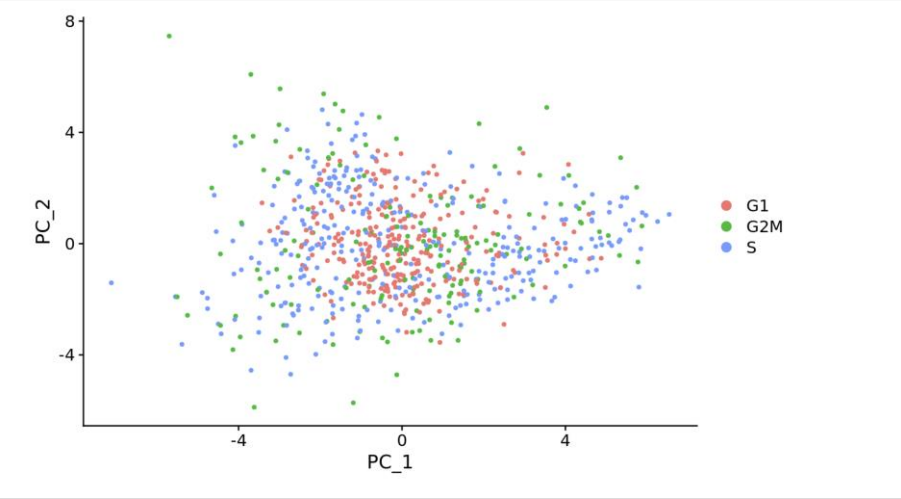
Seurat attempts to subtract or 'regress out' heterogeneity derived from either biological or technical sources, such as:

- Cell cycle scores
  - We assign each cell a score, based on its expression of G2/M and S phase gene markers
- Total UMI count
- Mitochondrial gene expression (also an indication of cell stress)
- Remark- we need to consider our goals when selecting the sources to regress-out

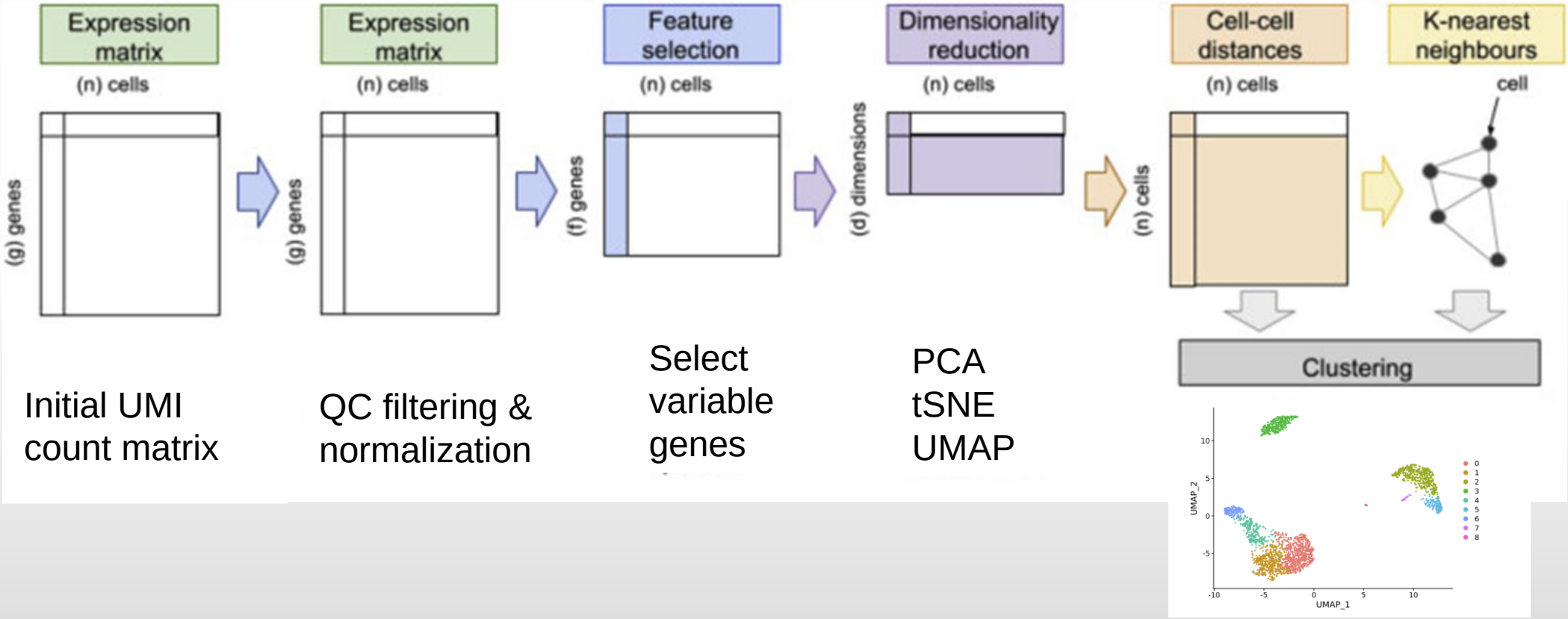
Before – only cell cycle genes




After regression – only cell cycle genes



# Analysis Workflow



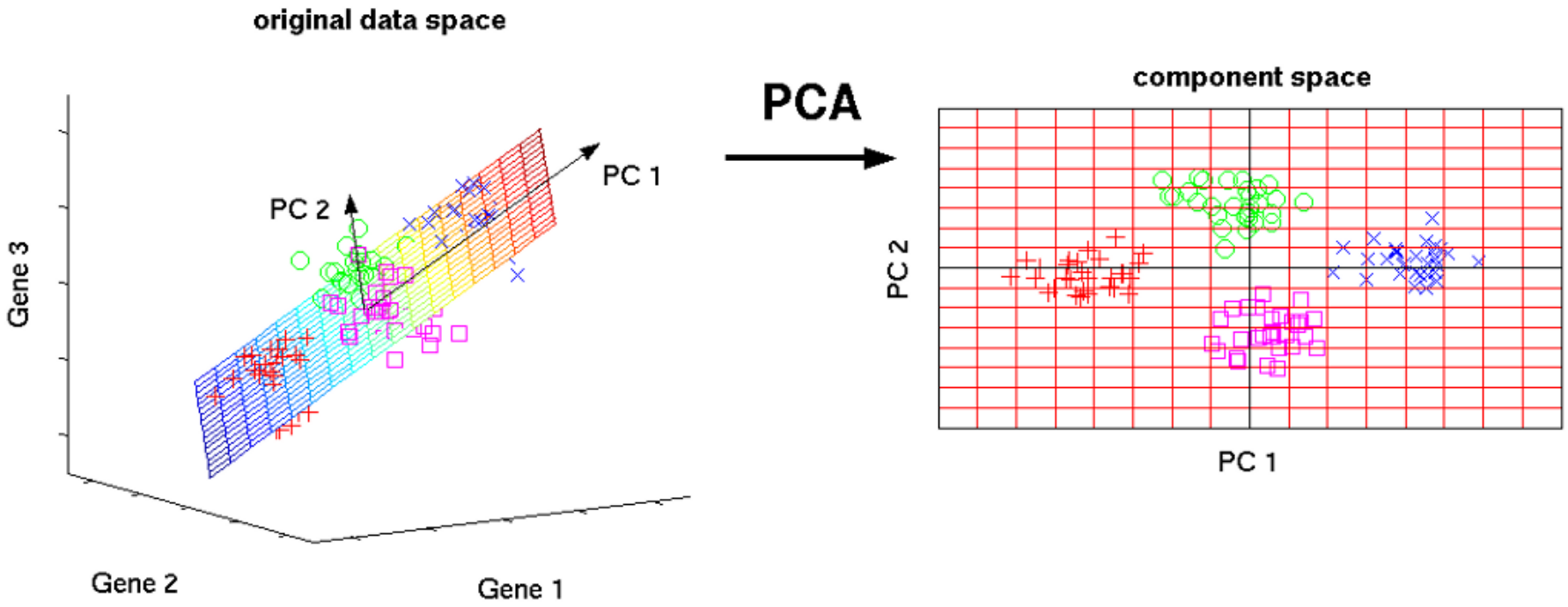

Modified plot from-

Andrews et al. Molecular Aspects of Medicine
Volume 59, February 2018, Pages 114-122

# PCA – Principle Component Analysis

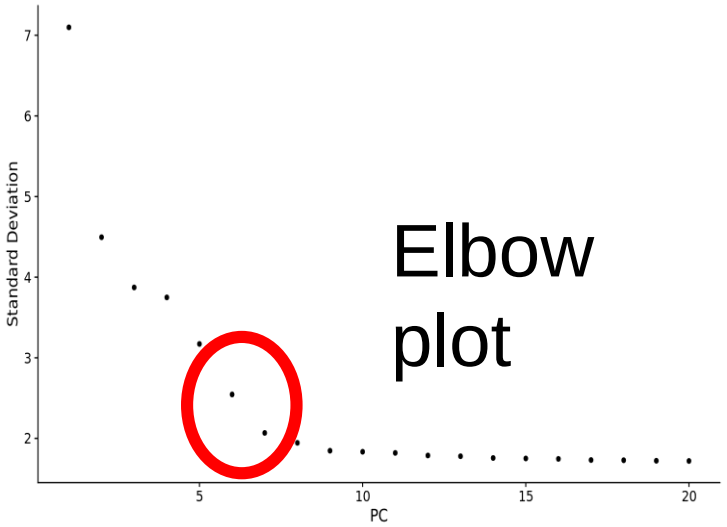
Perform linear dimensional reduction (PCA) on the scaled data

Finds a linear projection of high dimensional data so that the variance is maximized (and reconstruction error is minimized)

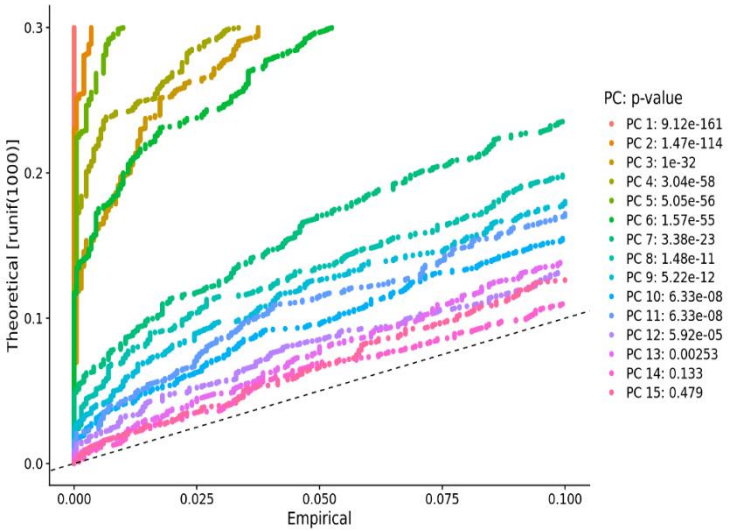



http://www.nlpca.org/pca_principal_component_analysis.html
https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigen

# Choose the Number of PCs



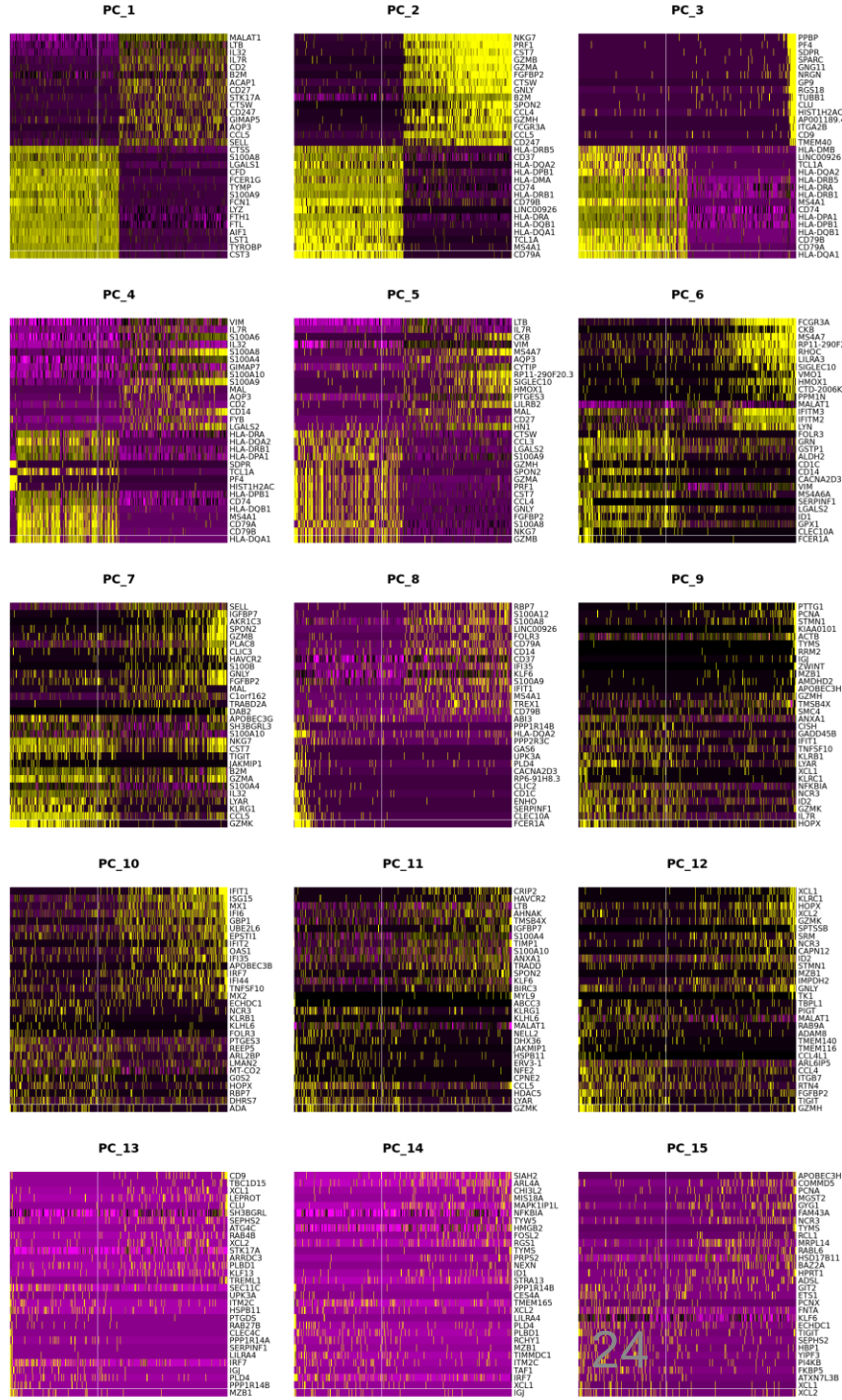

Elbow plot



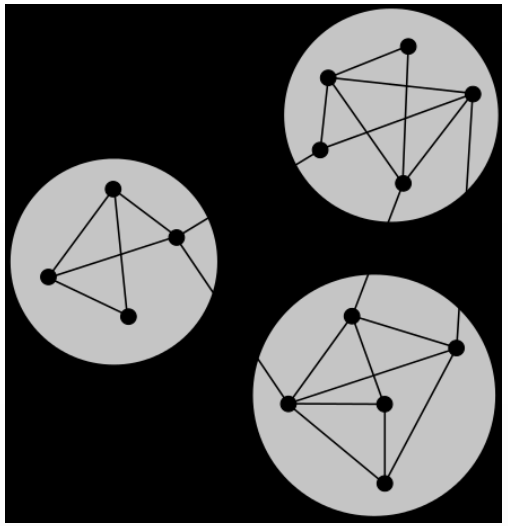

Jackstraw plot

## PC heatmaps

- Genes (rows)
- Cells (columns)
- Both cells and features are ordered according to their PCA scores
- Only the 'extreme' cells on both ends of the spectrum are plotted
- We can consider using less than 15 PCs



https://en.wikipedia.org/wiki/Community_structure



# Clustering

Clustering is an unsupervised learning procedure that is used in scRNA-seq data analysis to define groups of cells with similar expression profiles

- Similarity of cell expression profile is calculated in the PC space
- Very brief - Seurat uses a graph-based clustering approach, which embeds cells in a graph structure, using a K-nearest neighbor (KNN) graph, with edges drawn between cells with similar gene expression patterns. Then, we partition this graph into highly interconnected ‘communities’.

# Clustering

- We can control the number of clusters by two parameters:
  - The number PCs
    - We select the top PCs (since the PCs are sorted by the amount of variance they explain)
    - Selecting too many PCs can introduce noise
  - Resolution
    - a parameter which sets the 'granularity' of the clusters. Increased values leading to a greater number of clusters.
  - We do not know a priori what parameters to select

- Question- Given a dataset is there a true number of clusters?

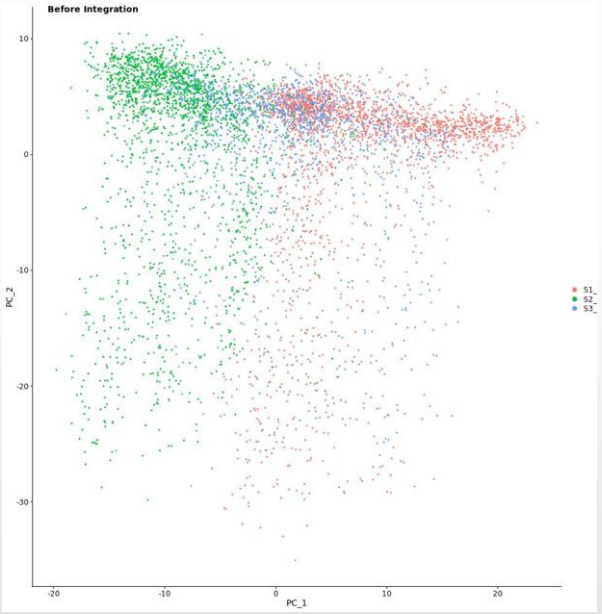
# Question

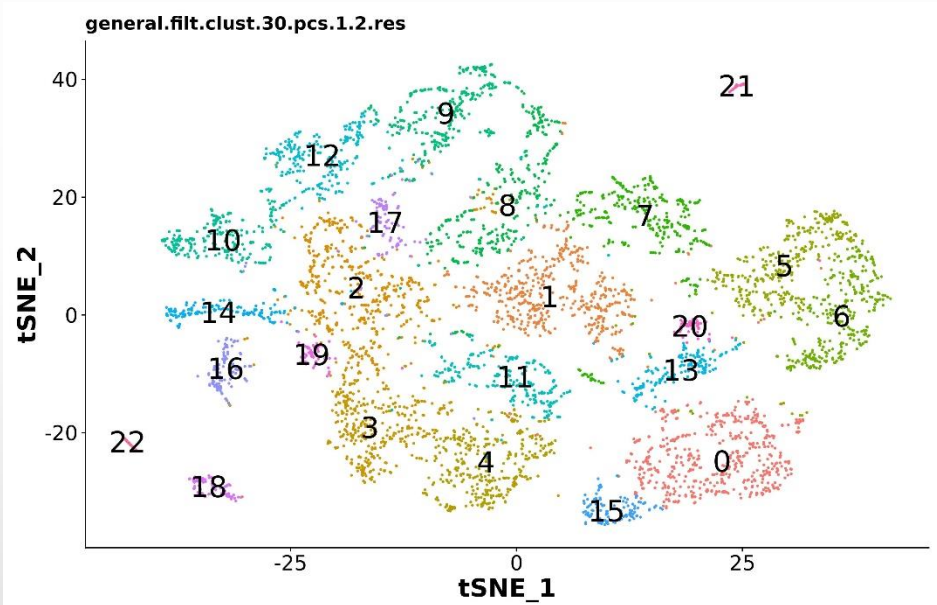
Given a dataset is there a true - definite number of clusters?

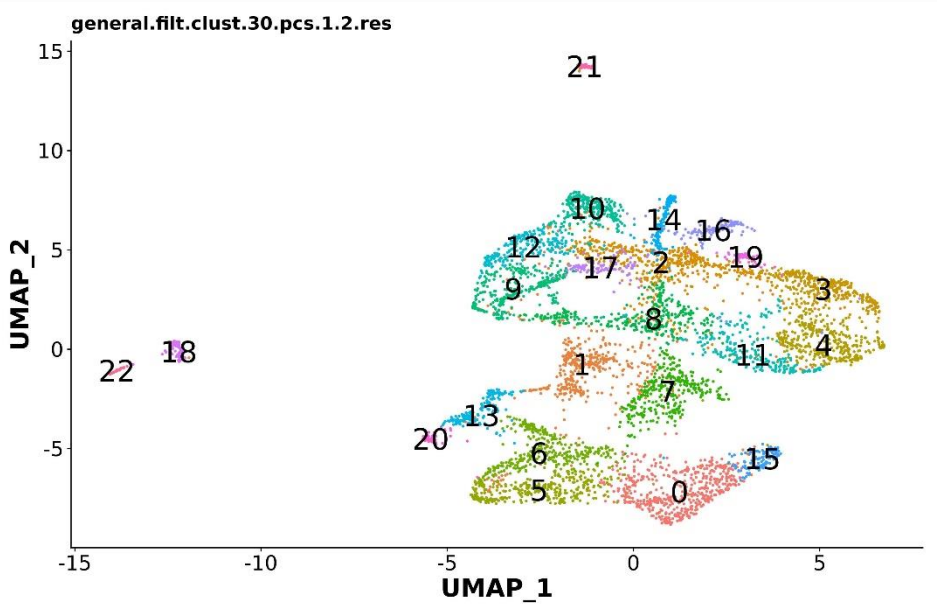
# Data Visualization

In order to view the data we further reduce the PCs space to 2-3 dimensions

PCA (2 PCs) tSNE (30 PCs) UMAP (30 PCs)




- t-SNE preserves local structure in the data
- UMAP preserves both local and global structure

# Finding Differentially Expressed Genes (DEG)

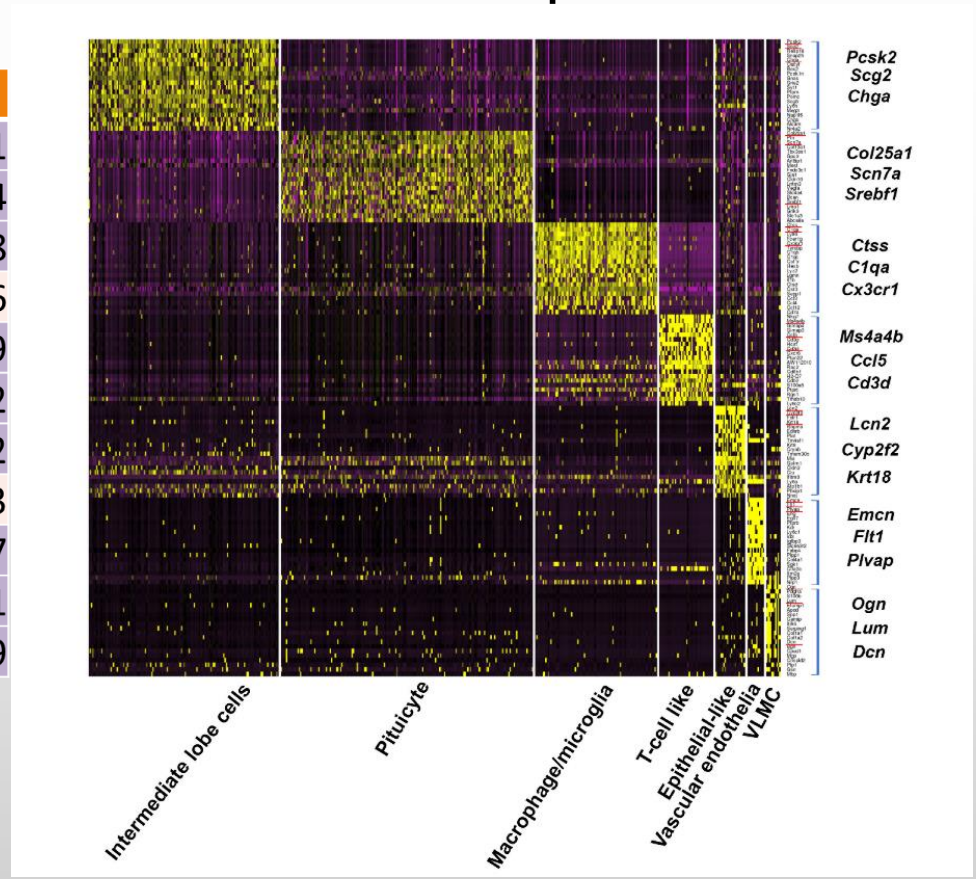
- Find cluster marker genes
- Identify positive (and negative) DEG of a single cluster compared to all other cells (Wilcoxon Rank Sum - default)

Table

| gene | cluster | avg_logFC | p_val | p_val_adj | pct.1 | pct.2 |
|---|---|---|---|---|---|---|
| Wfdc17 | 0 | 3.985889 | 0 | 0 | 0.975 | 0.091 |
| Ifitm1 | 0 | 3.363648 | 0 | 0 | 0.971 | 0.074 |
| Lrg1 | 0 | 2.987844 | 0 | 0 | 0.984 | 0.08 |
| Igfbp4 | 1 | 1.274831 | 8.86E-98 | 1.19E-93 | 0.661 | 0.166 |
| Cd8b1 | 2 | 2.439779 | 0 | 0 | 0.931 | 0.029 |
| Cd8a | 2 | 1.700122 | 0 | 0 | 0.672 | 0.012 |
| Ctsw | 2 | 1.04197 | 2.42E-154 | 3.24E-150 | 0.64 | 0.082 |
| Igfbp4 | 3 | 1.263992 | 7.48E-81 | 1.00E-76 | 0.621 | 0.173 |
| H2-Eb1 | 4 | 1.604532 | 2.75E-162 | 3.68E-158 | 1 | 0.207 |
| H2-DMb2 | 4 | 1.494226 | 4.19E-162 | 5.60E-158 | 0.979 | 0.191 |
| H2-Aa | 4 | 1.688614 | 2.70E-159 | 3.62E-155 | 1 | 0.229 |

pct.1 = fraction of cells in the cluster that express the gene
pct.2 = fraction of cells in all other cells that express the gene

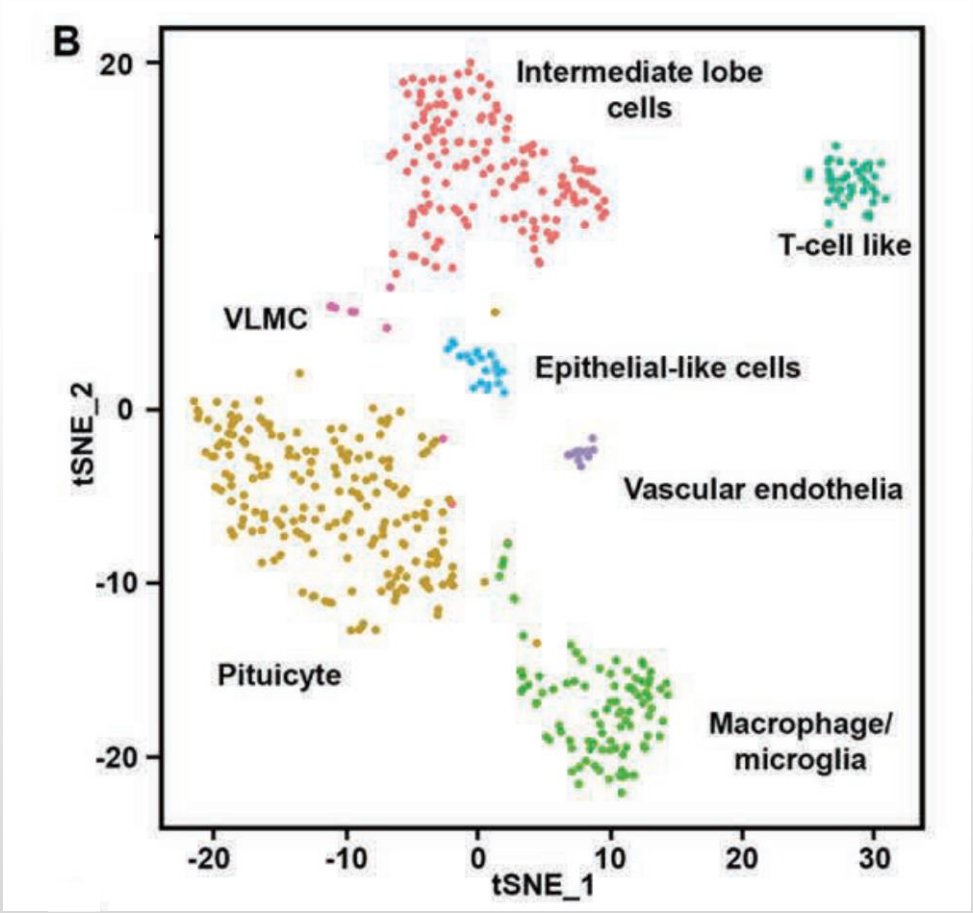
Heatmap



# Cluster Annotation

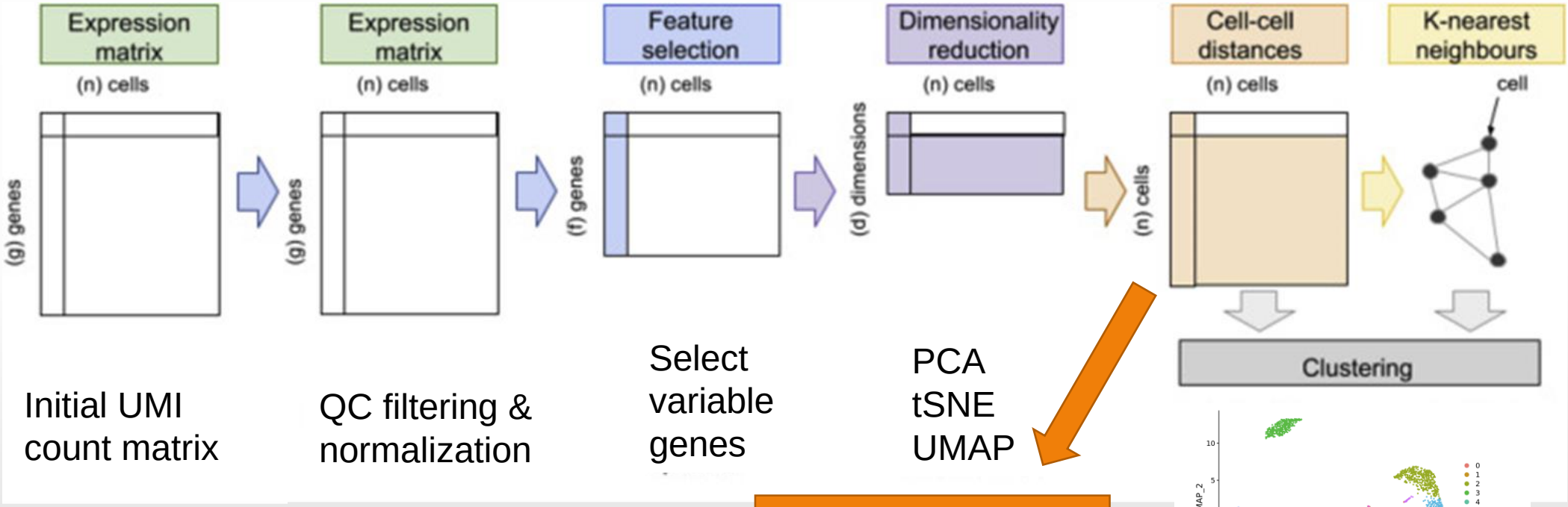
- Clusters can be annotated by enrichment tests comparing cluster marker genes to marker genes from an annotated reference database (hypergeometric test)
- Source for reference - Panglaodb https://panglaodb.se/, contains a comprehensive collection of single cell data

| | ***Mus musculus*** | ***Homo sapiens*** |
|---|---|---|
| Samples | 1063 | 305 |
| Tissues | 184 | 74 |
| Cells | 4,459,768 | 1,126,580 |
| Clusters | 8,651 | 1,748 |




Chen et al. eNeuro 8 January 2020

# Summary: Analysis Workflow




Modified plot from-

Andrews et al. Molecular Aspects of Medicine
Volume 59, February 2018, Pages 114-122

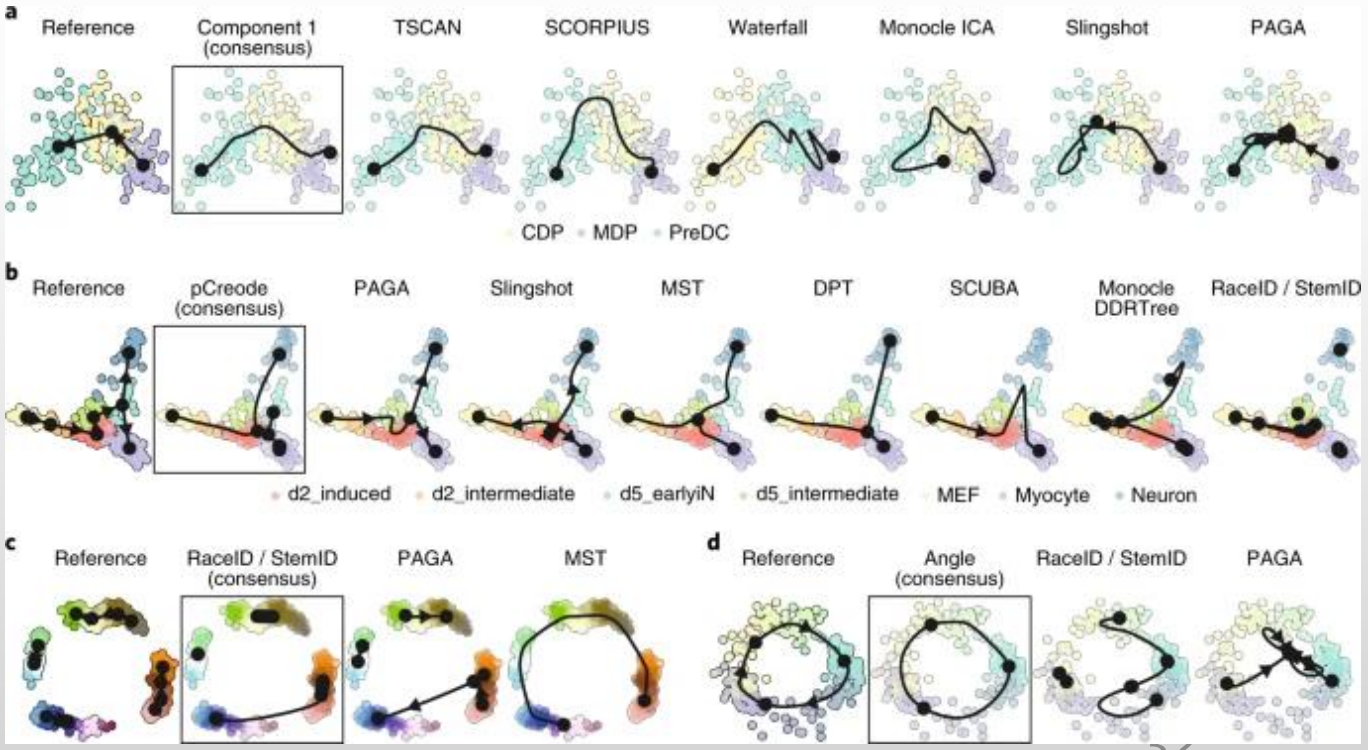
# Trajectory Inference Analysis

- Clustering is a discrete classification approach and therefore lacks in the ability to capture:
  - Transitions between cell identities
  - Branching differentiation processes
- Trajectory inference methods interpret single-cell data as a snapshot of a continuous process
- Interpretation of a trajectory requires additional data sources

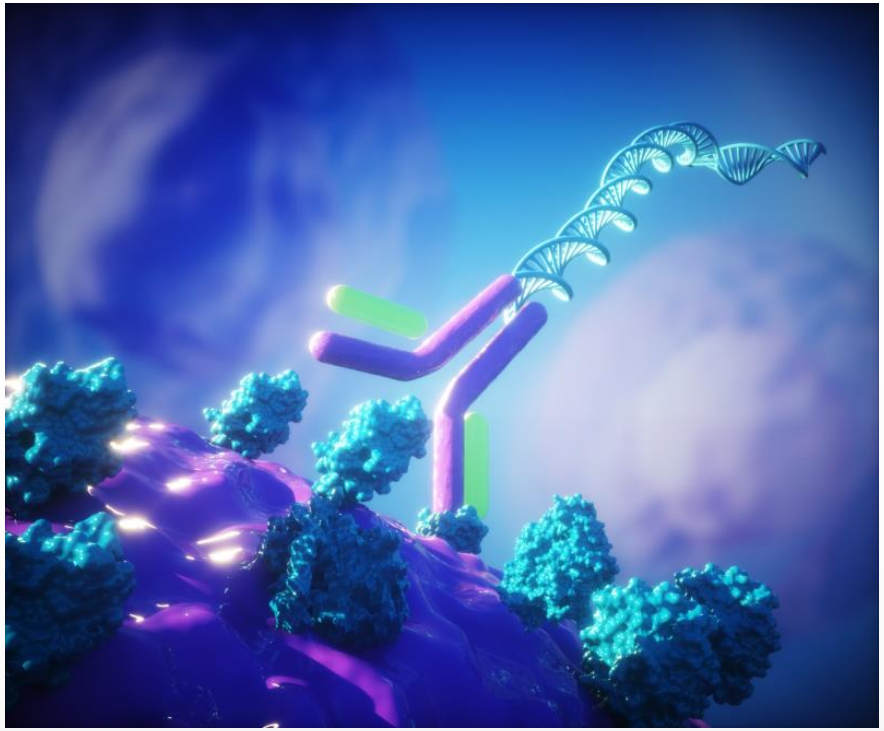
**A comparison of single-cell trajectory inference methods**

Wouter Saelens, Robrecht Cannoodt, Helena Todorov & Yvan Saeys

*Nature Biotechnology* **37**, 547–554(2019) | Cite this article



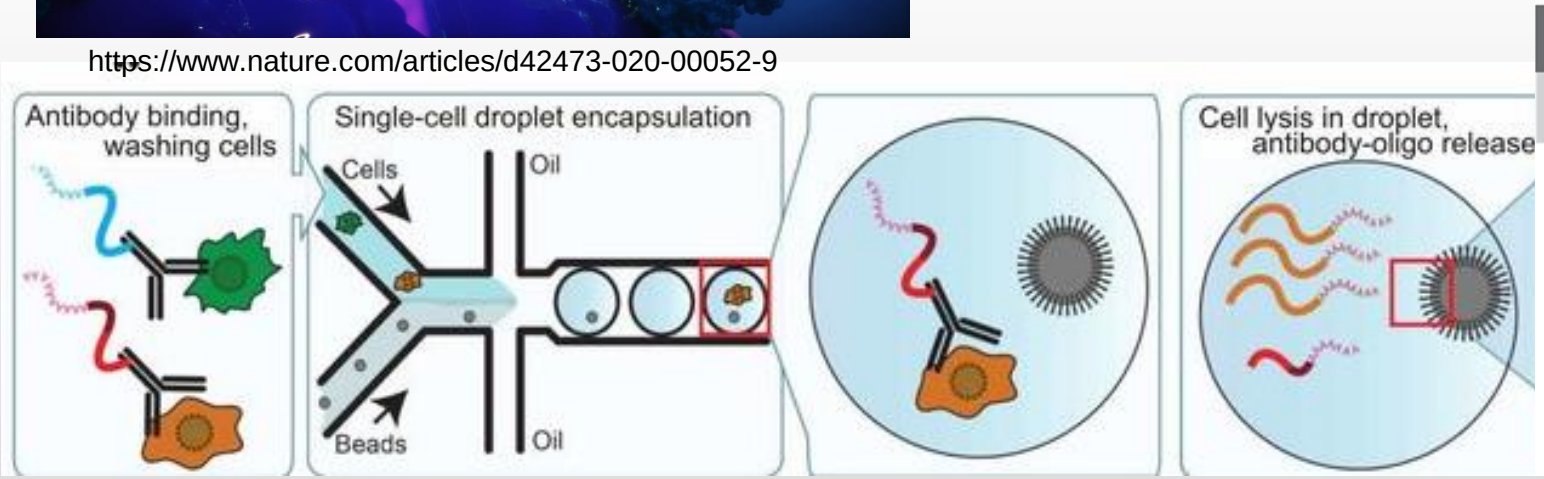
# Multi-omics: CITE-seq

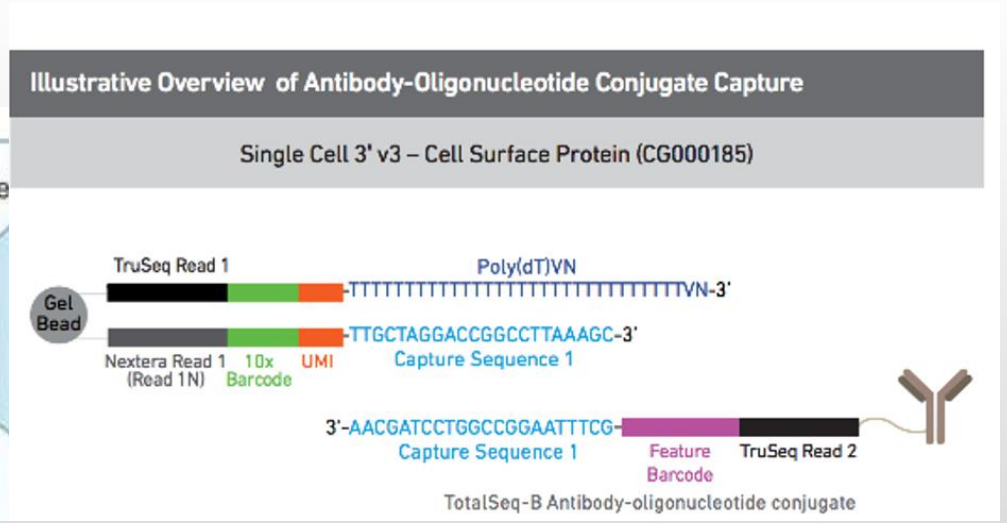


https://www.nature.com/articles/d42473-020-00052-9

Cellular Indexing of Transcriptomes and Epitopes by sequencing

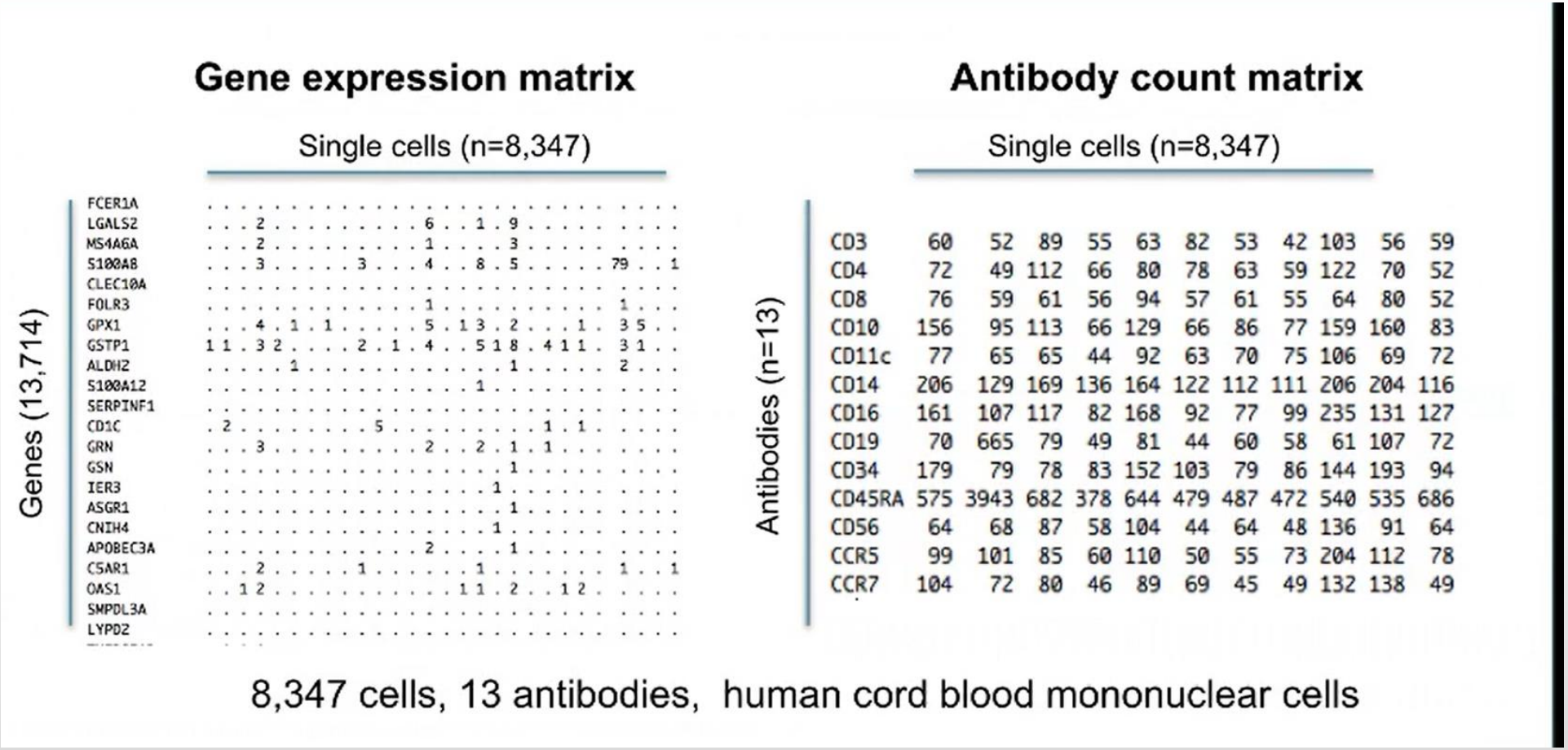
CITE-seq enables simultaneous detection of single-cell transcriptomes and protein markers




Modified DOI: 10.1126/sciadv.aax8978

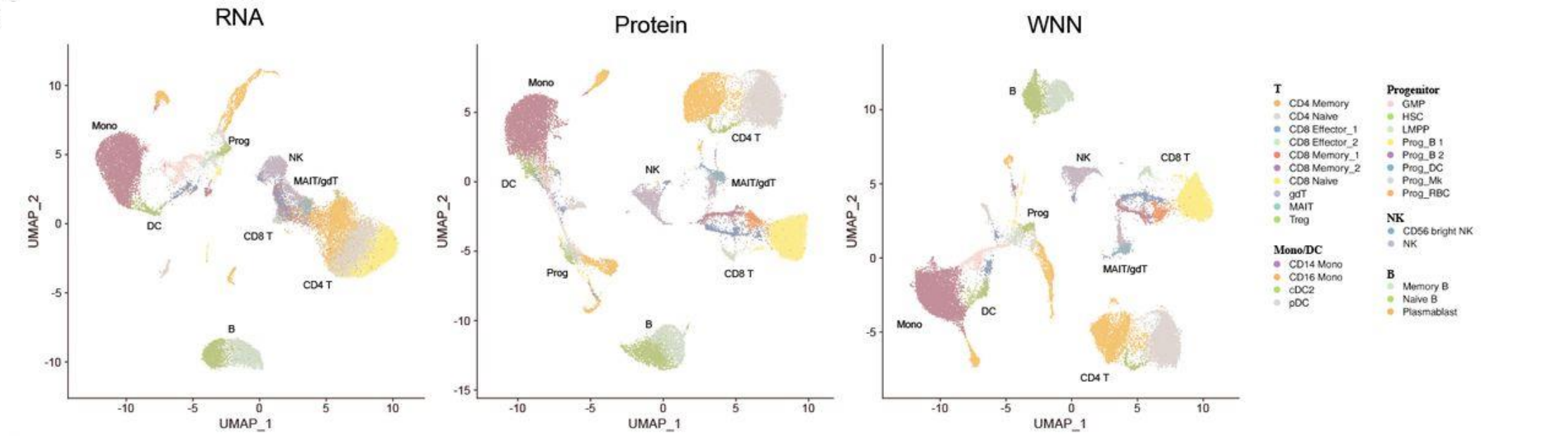
# Count matrix of CITE-Seq




Taken from a presentation of Rahul Satija

# Integrated Analysis of Multimodal Single-Cell Data

- Hai et al. Satija doi: https://doi.org/10.1101/2020.10.12.335331 BioRxiv
- Introduce ‘weighted-nearest neighbor’ (WNN) analysis
- They demonstrate that WNN analysis substantially improves the ability to define cellular states



# References

- Current best practices in single-cell RNA-seq analysis: a tutorial, Luecken et al. Mol Syst Biol. (2019) 15: e8746

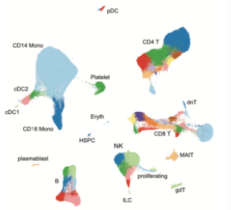
SATIJA LAB

HOME NEWS PEOPLE RESEARCH PUBLICATIONS SEURAT JOIN/CONTACT SINGLE CELL GENOMICS DAY

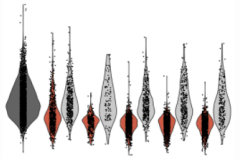
**2020**

Stuart T, Srivastava A, Lareau C, Satija R
**Multimodal single-cell chromatin analysis with Signac**
**bioRxiv.** 2020

Hao Y*, Hao S*, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zagar M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LB, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P*, Satija R*
**Integrated analysis of multimodal single-cell data**
**bioRxiv.** 2020



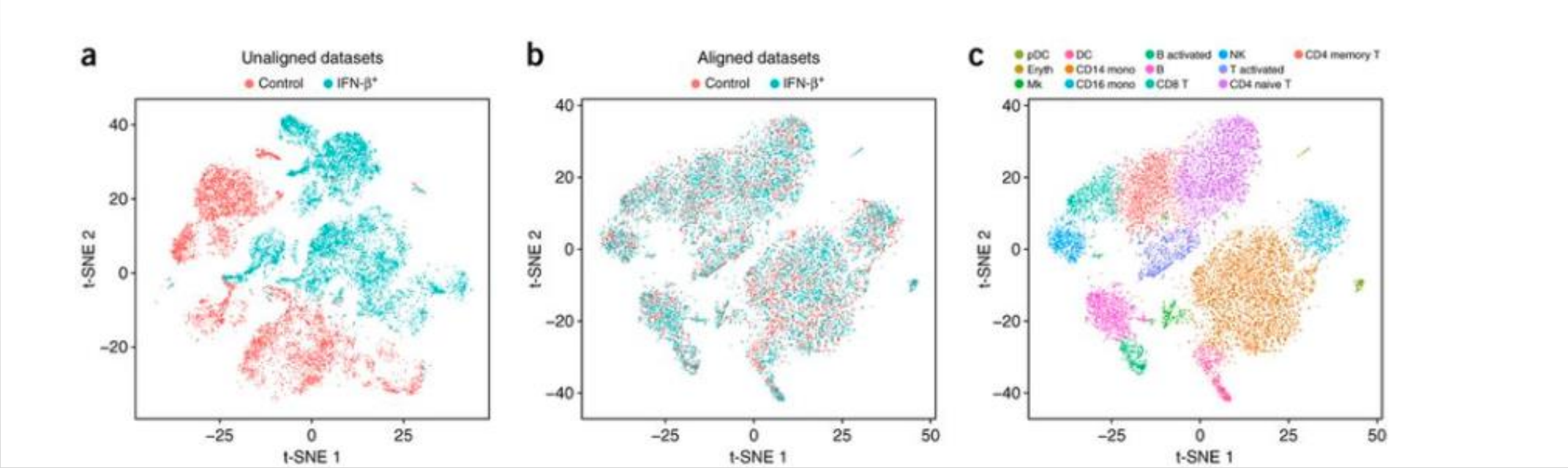
Papalexi E, Mimitou E, Butler A, Foster S, Bracken B, Mauck WM, Wessels H, Yeung BZ, Smibert P, Satija R
**Characterizing the molecular regulation of inhibitory immune checkpoints with multi-modal single-cell screens**
**Nature Genetics (In press).** 2020



# The End
# Thanks for listening
# Questions?

# CCA - canonical correlation analysis (CCA)
# A method to align datasets

An analytical strategy for integrating scRNA-seq data sets based on common sources of variation, enabling the identification of shared populations across data sets and downstream comparative analysis.



# New Method for Normalization

**Seurat new normalization method package -ctransform performs** 'regularized negative binomial regression' for the normalization and variance stabilization of single-cell data,

The method has a few attractive properties when compared to previous method of log-normalization:

- We learn a statistical model of technical noise directly from the data, and remove this without dampening biological heterogeneity
- We do not assume a constant size or global 'scaling factor' for single cells
- We do not apply heuristic steps, such as log-transformation, pseudocount addition, or z-scoring

Hafemeister et al. bioRxiv preprint doi: https://doi.org/10.1101/576827