SINGLE-CELL analysis with 10x genomics

Introduction to scRNA-seq analysis

25.10.2023

Joana P. Bernardes

Institute for Clinical Molecular Biology-Kiel

IKMB

Institute of **Clinical**
Molecular Biology Kiel

C COMPETENCE
C CENTRE FOR
G GENOMIC
A ANALYSIS KIEL

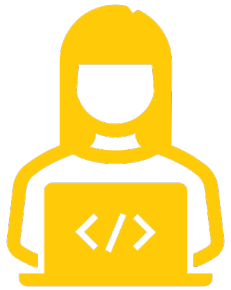
What do you know so far?

- Overview single-cell technologies
- Single-cell techniques in understanding disease
- Precision medicine: Present and future

Objectives for the next 3 days

Theory: *Introduction to scRNA-seq analysis*

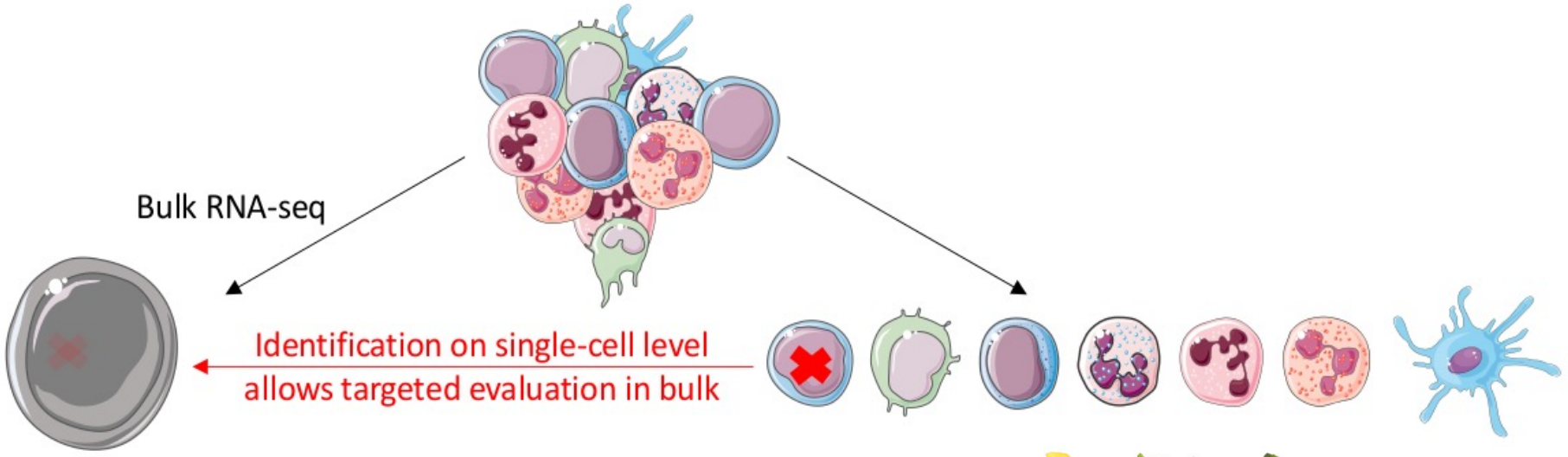
- Concepts of single cell RNA-seq technology
- Workflow of single cell analysis
- Limitations of the technology



Workshop: *Practical Single cell transcriptome analysis 101*

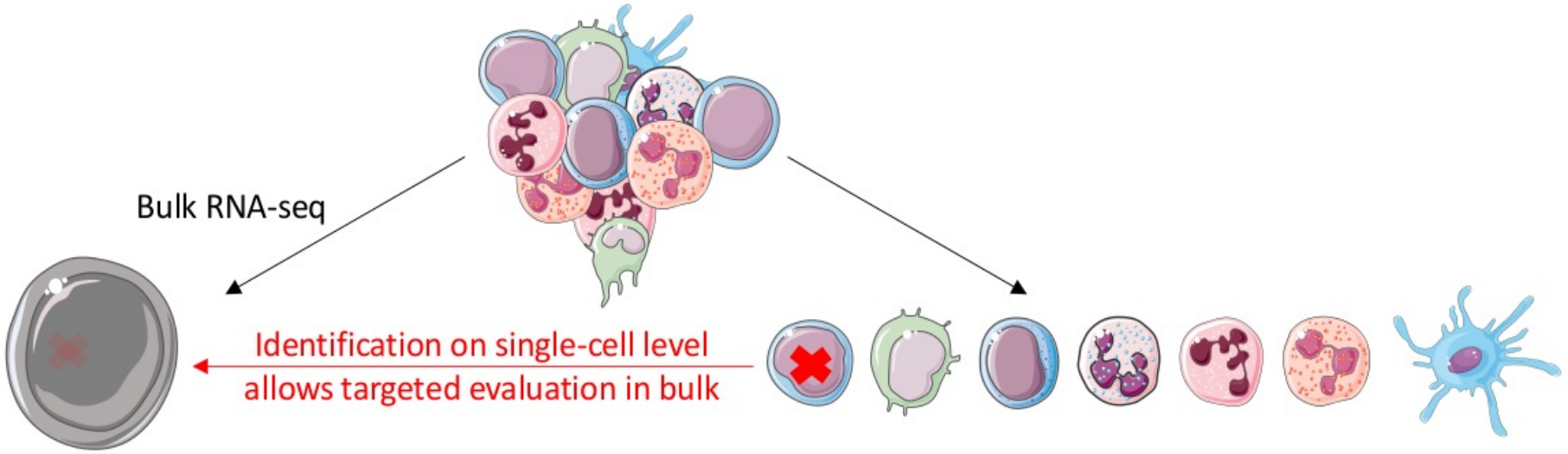
- Advances in medical life sciences using single cell technology
- COVID-19 pandemic as example

In simple ‘words’:



Modified from Servier Medical Art by Jonas Schulte-Schrepping
http://365-smoothie-rezepte.de/obst-beeren-und-gemuese-sorgfaeltig-auswaehlen/

In simple ‘words’:



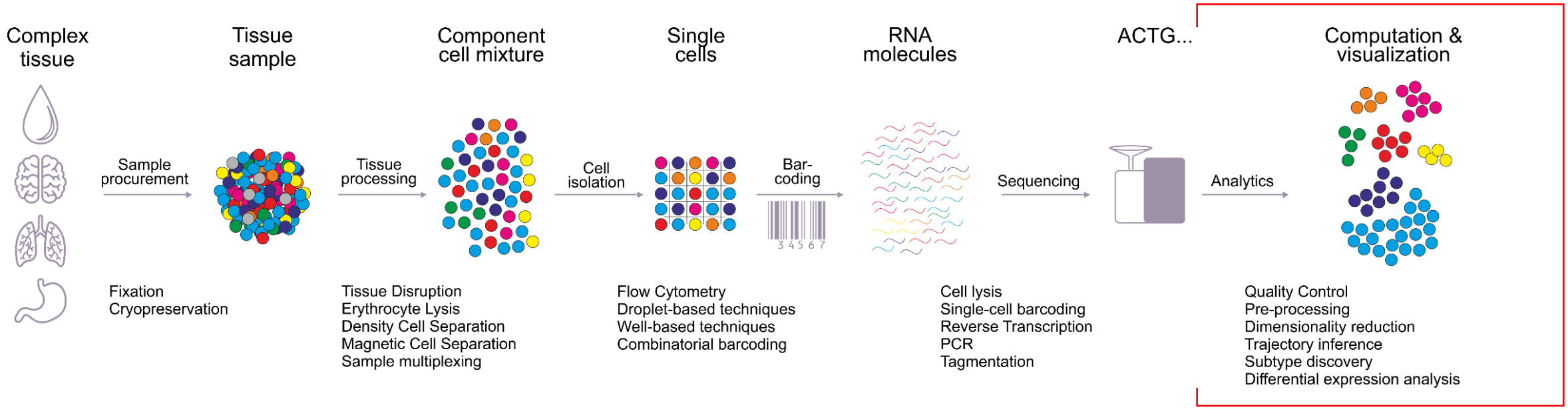
e.g. bulk RNA-seq

- Average expression
- ‘*Homogeneous*’ cell population

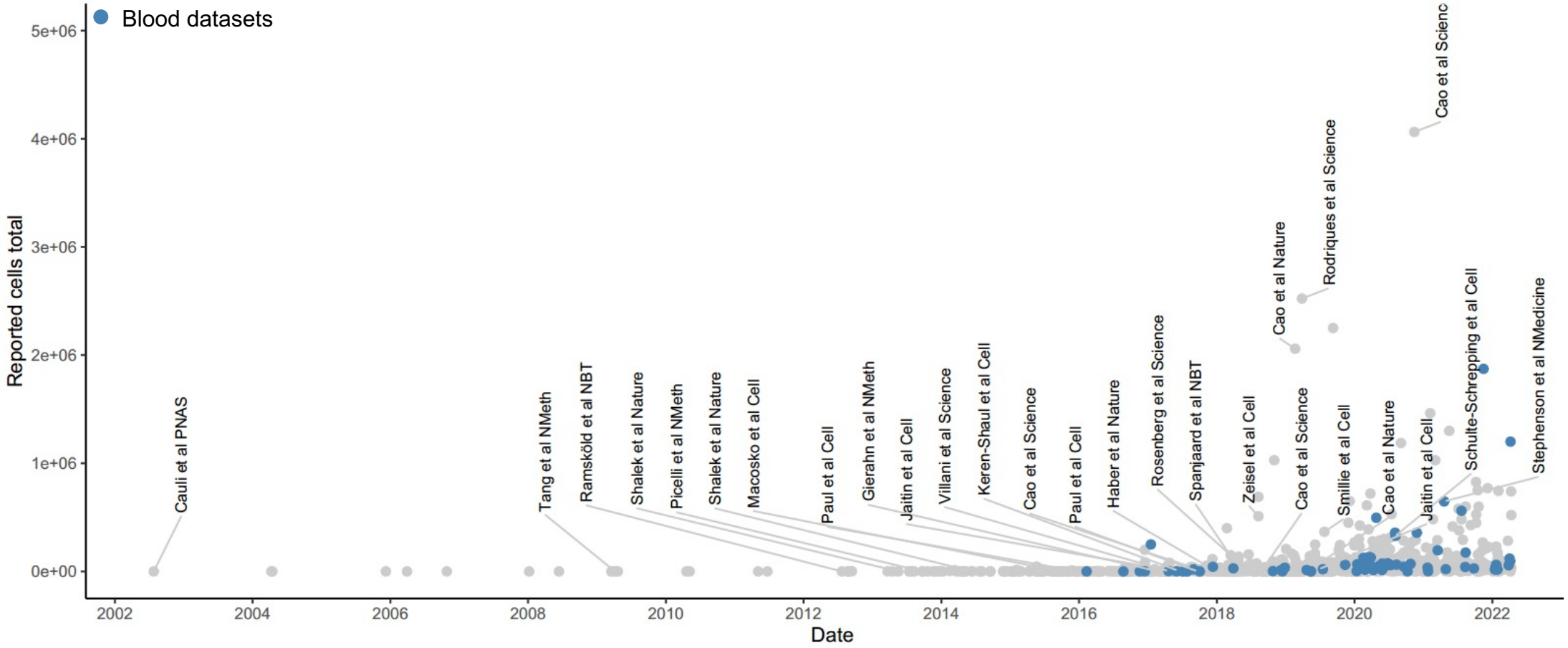
e.g. single cell RNA-seq

- Cell specific expression
- Heterogenous cell population

Single-cell RNA-seq experiments are highly complex



The evolution of single-cell RNA-seq



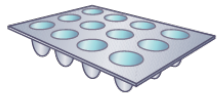
https://www.nxn.se/single-cell-studies

Single-cell: what can we do?

Manual

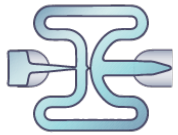
Tang et al. 2009

Multiplexing



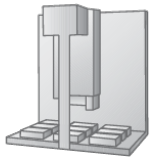
Islam et al. 2011

Integrated fluidic circuits



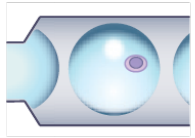
Brennecke et al. 2013

Liquid-handling robotics



Jaitin et al. 2014

Nanodroplets

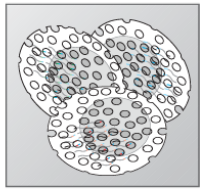


Klein et al. 2015
Macosko et al. 2015

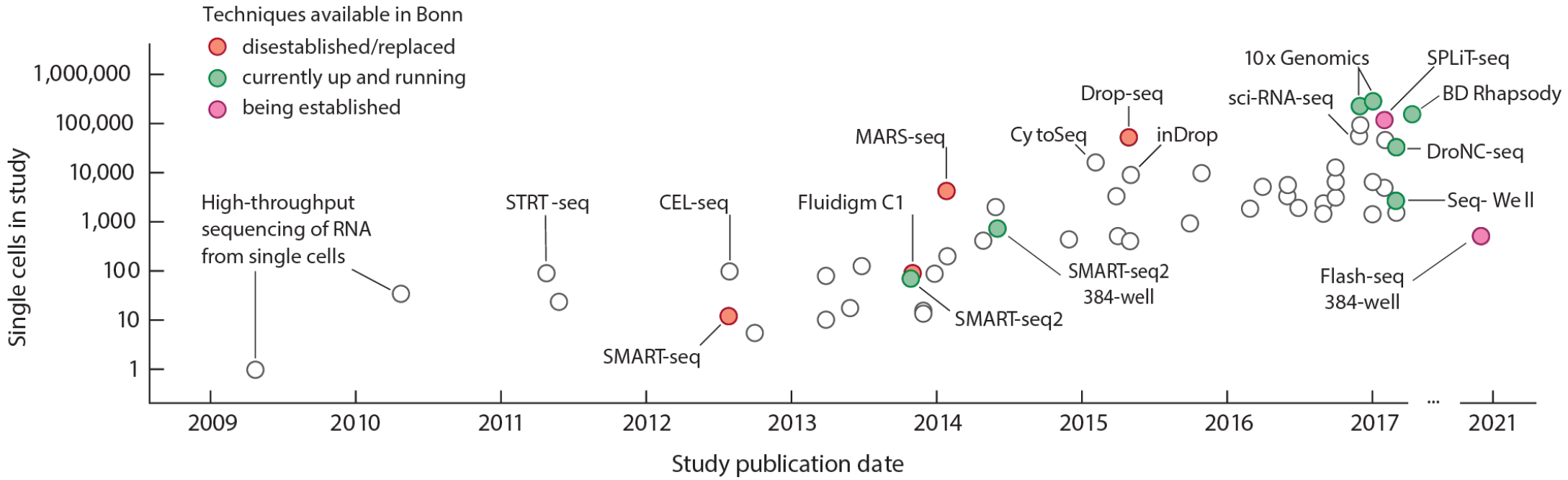
Picowells

Bose et al. 2015

In situ barcoding



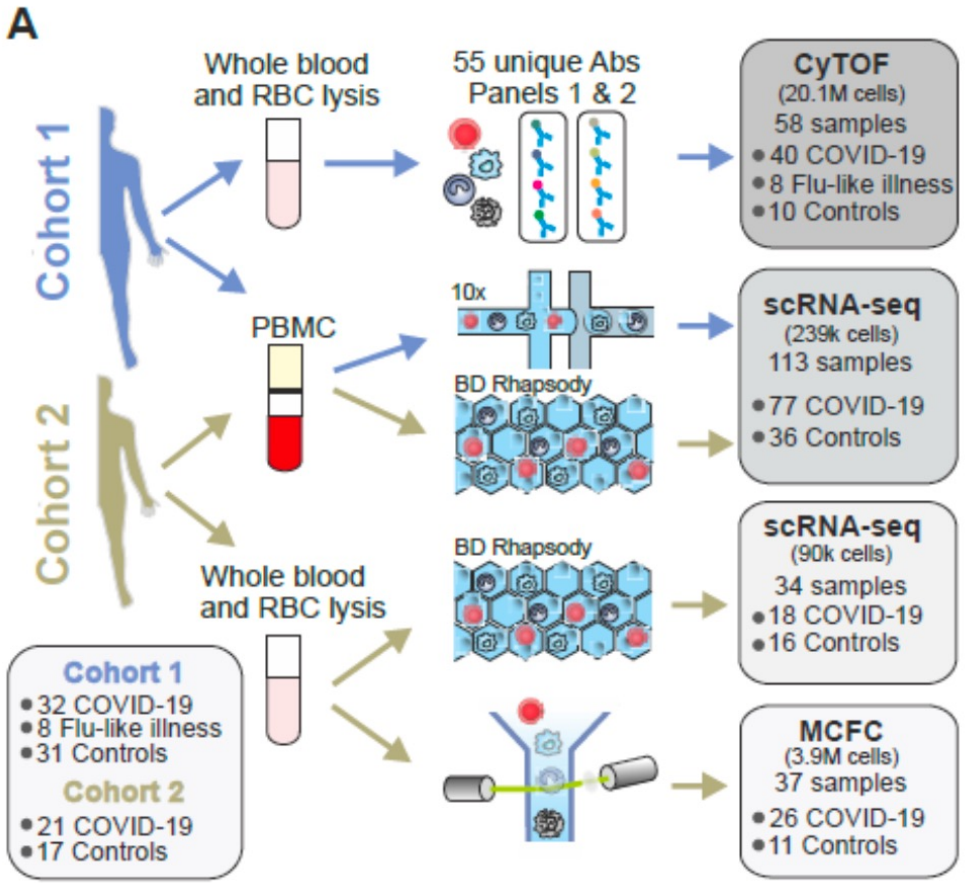
Cao et al. 2017
Rosenberg et al. 2017



Modified from Svensson *et al.*, Nature Protocols, 2018

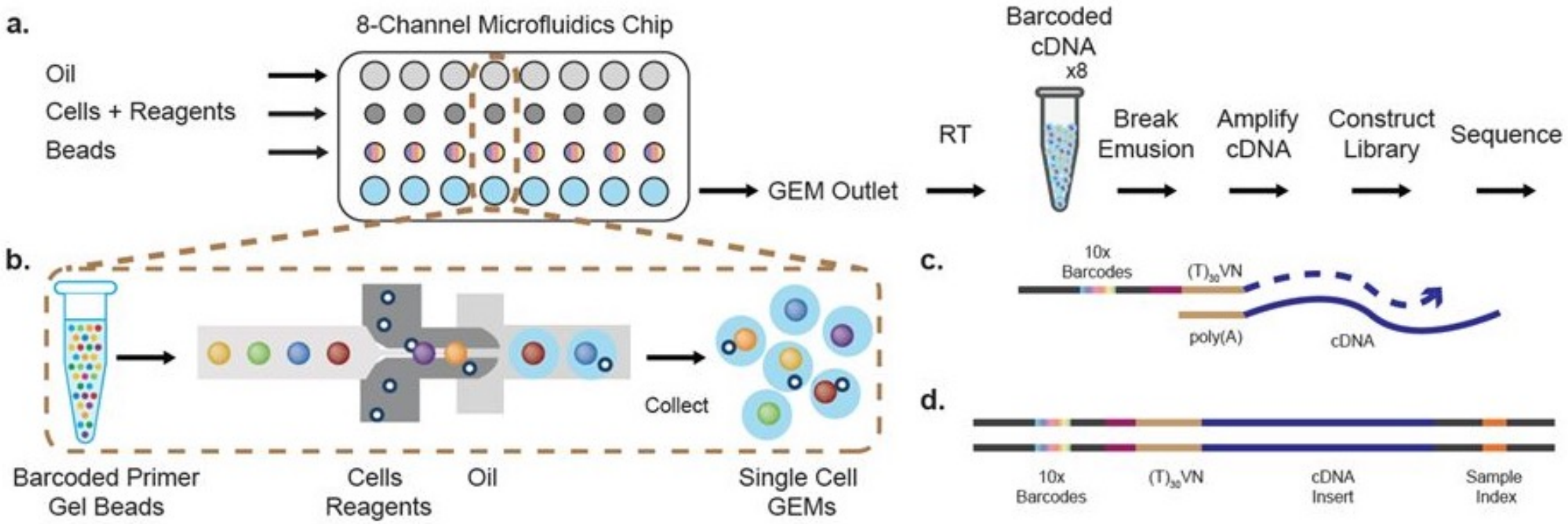
Single cell technology available

- BD Rhapsody
- **10x Genomics**
  - **Gene expression: 3'** and 5'
  - VDJ' (TCR and BCR)
  - CITE-seq
  - ATAC-seq
  - Multiome
  - **Spatial transcriptomics**



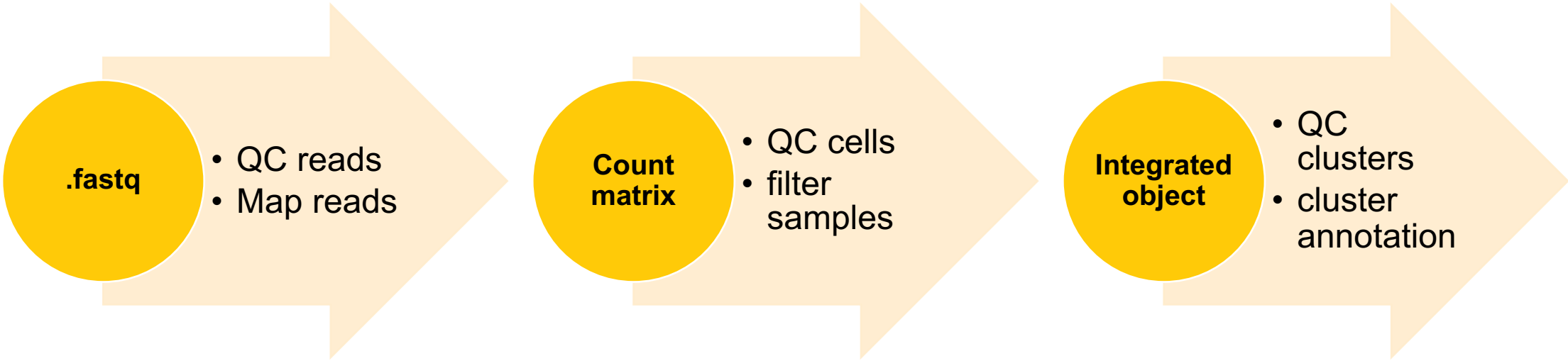
Jonas Schulte-Schrepping et al.,Cell 2020

10x Genomics- single cell gene expression

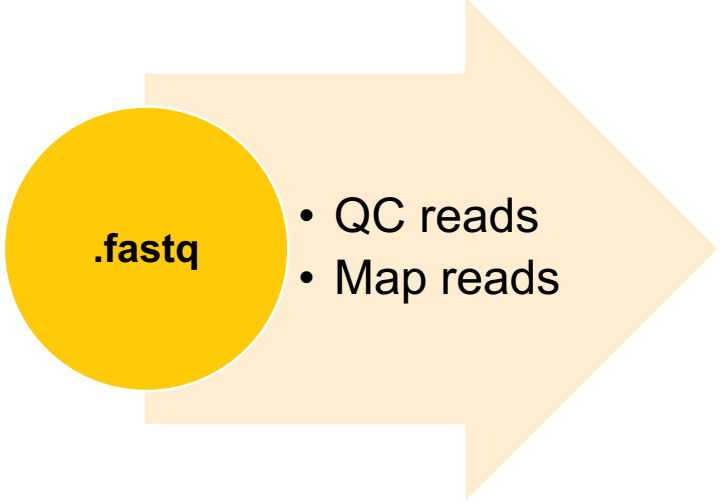


https://www.10xgenomics.com/products/single-cell-gene-expression

Processing workflow

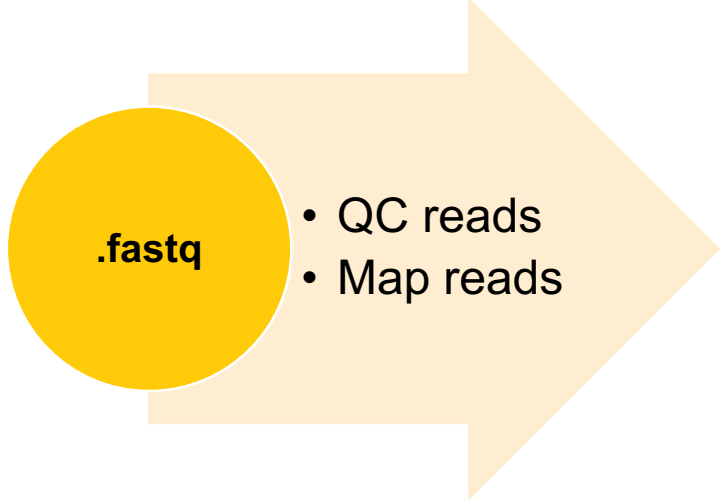


From reads to count matrix

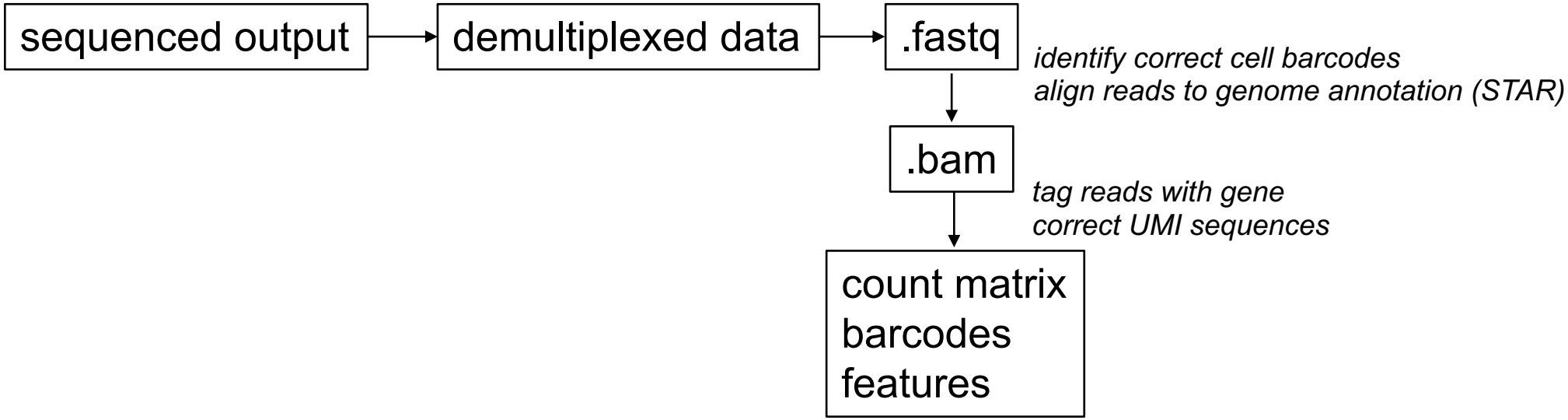


- cell ranger

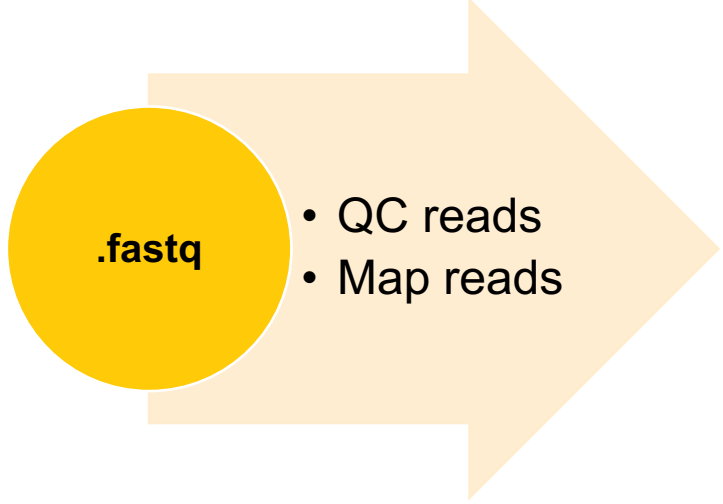
From reads to count matrix



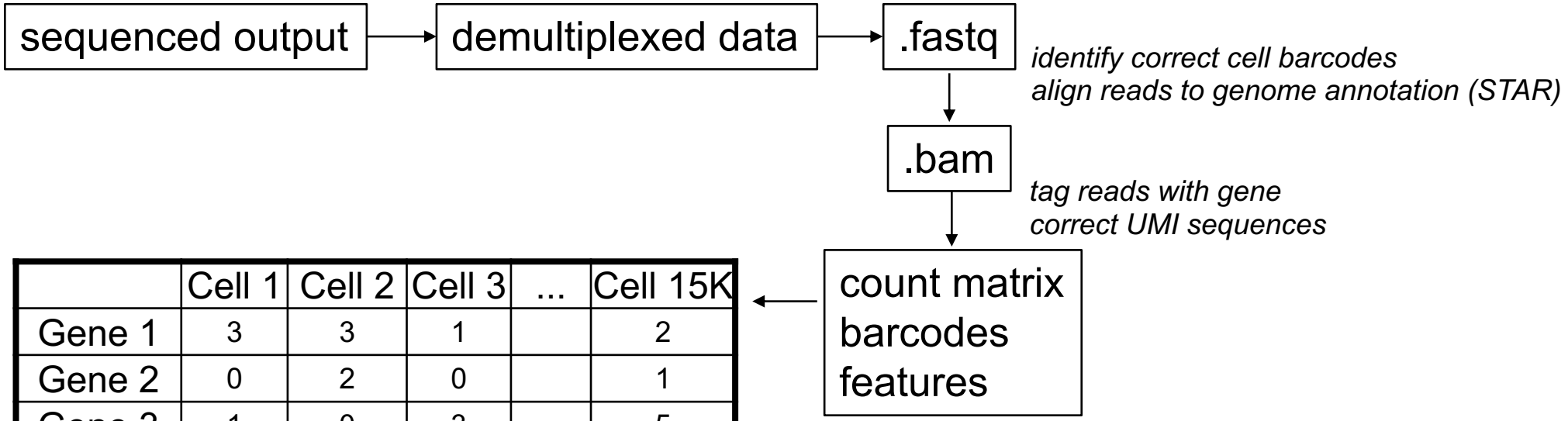
- cell ranger



From reads to count matrix



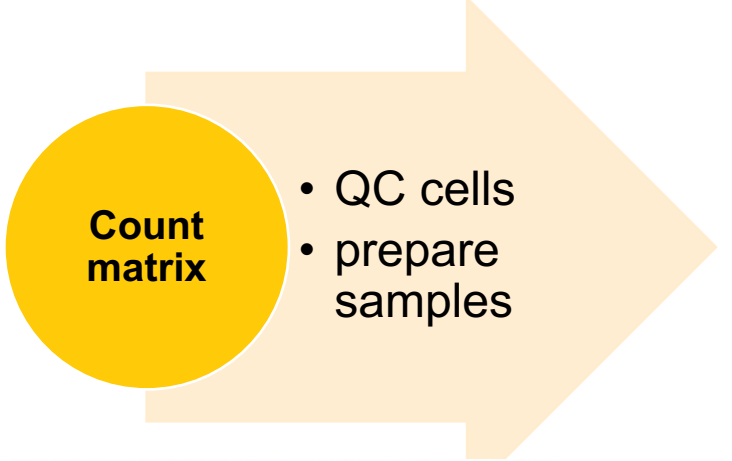
- cell ranger



	Cell 1	Cell 2	Cell 3	...	Cell 15K
Gene 1	3	3	1		2
Gene 2	0	2	0		1
Gene 3	1	0	3		5
...					
Gene K	14	7	1		0
...					
Gene 25K	0	13	1		0

- Lots of zeros due to dropout or no expression
- Variable total counts across cells – cell specific bias
- Variable counts per gene – part biological, part technical

From matrix to sample objects



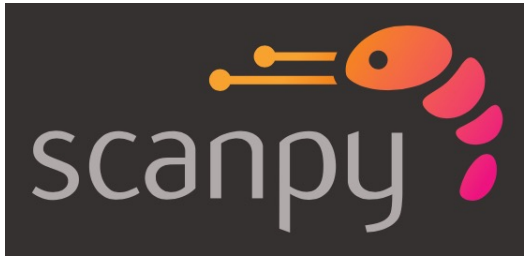
- Create object

R packages

- **Seurat V4 > S4 seurat object**
- scater V1.18.3 > SingleCellExperiment
- Monocle3 > derived SingleCellExperiment

Python packages

- scanpy V1.7.0 > Anndata
- loom 3.0.0 > h5ad



The Seurat Object

- Each type of data has it‘s own slot
- Facilitates streamlined analysis, easy execution of functions

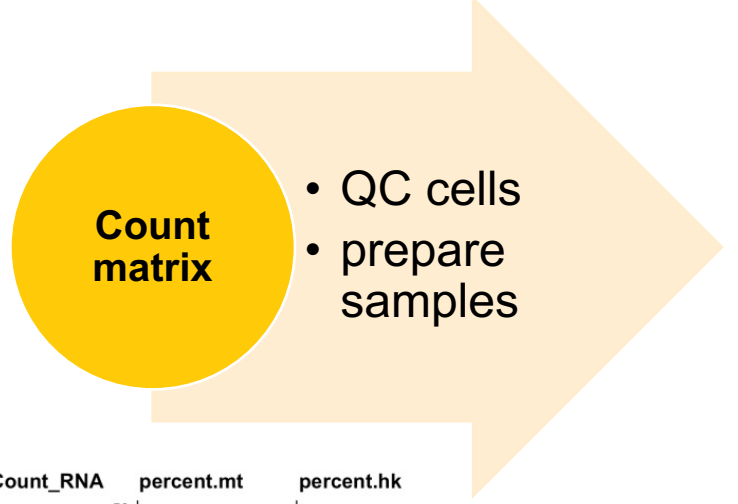
e.g. raw counts:

pbmc@assay$counts

GetAssayData(object = pbmc,
slot = 'counts')

```
> str(pbmc)
Formal class 'Seurat' [package "Seurat"] with 12 slots
  ..@ assays      :List of 1
  .. ..$ RNA:Formal class 'Assay' [package "Seurat"] with
  .. .. .. ..@ counts       :Formal class 'dgCMatrix' [p
  .. .. .. .. .. ..@ i       : int [1:2491554] 21 26 28 
  .. .. .. .. .. ..@ p       : int [1:1177] 0 2618 4423 
  .. .. .. .. .. ..@ Dim     : int [1:2] 15246 1176
  .. .. .. .. .. ..@ Dimnames:List of 2
  .. .. .. .. .. .. ..$ : chr [1:15246] "AL627309.1" "AL
  .. .. .. .. .. .. ..$ : chr [1:1176] "AAACCCAAGGAGAGTA
  .. .. .. .. .. ..@ x       : num [1:2491554] 1 1 1 1 2
  .. .. .. .. .. ..@ factors : list()
  .. .. .. ..@ data         :Formal class 'dgCMatrix' [p
  .. .. .. .. .. ..@ i       : int [1:2491554] 21 26 28 
  .. .. .. .. .. ..@ p       : int [1:1177] 0 2618 4423 
  .. .. .. .. .. ..@ Dim     : int [1:2] 15246 1176
  .. .. .. .. .. ..@ Dimnames:List of 2
  .. .. .. .. .. .. ..$ : chr [1:15246] "AL627309.1" "AL
  .. .. .. .. .. .. ..$ : chr [1:1176] "AAACCCAAGGAGAGTA
  .. .. .. .. .. ..@ x       : num [1:2491554] 1 1 1 1 2
  .. .. .. .. .. ..@ factors : list()
  .. .. .. ..@ scale.data   : num[0 , 0 ]
  .. .. .. ..@ key          : chr "rna_"
  .. .. .. ..@ var.features : logi(0)
  .. .. .. ..@ meta.features:'data.frame':      15246 obs
  .. .. .. ..@ misc         : NULL
  ..@ meta.data   :'data.frame':        1176 obs. of  3 
  .. ..$ orig.ident  : Factor w/ 1 level "10X_PBMC": 1 1
  .. ..$ nCount_RNA  : num [1:1176] 8286 5509 4280 2754 
  .. ..$ nFeature_RNA: int [1:1176] 2618 1805 1559 1225 
  ..@ active.assay: chr "RNA"
  ..@ active.ident: Factor w/ 1 level "10X_PBMC": 1 1 1 
  .. ..- attr(*, "names")= chr [1:1176] "AAACCCAAGGAGAGTA
  ..@ graphs      : list()
  ..@ neighbors   : list()
  ..@ reductions  : list()
  ..@ project.name: chr "10X_PBMC"
  ..@ misc        : list()
  ..@ version     :Classes 'package_version', 'numeric_v
  .. ..$ : int [1:3] 3 0 2
  ..@ commands    : list()
  ..@ tools       : list()
```

From matrix to sample objects



- Quality control/filtering of cells:

  Number of genes [200; 5.000]

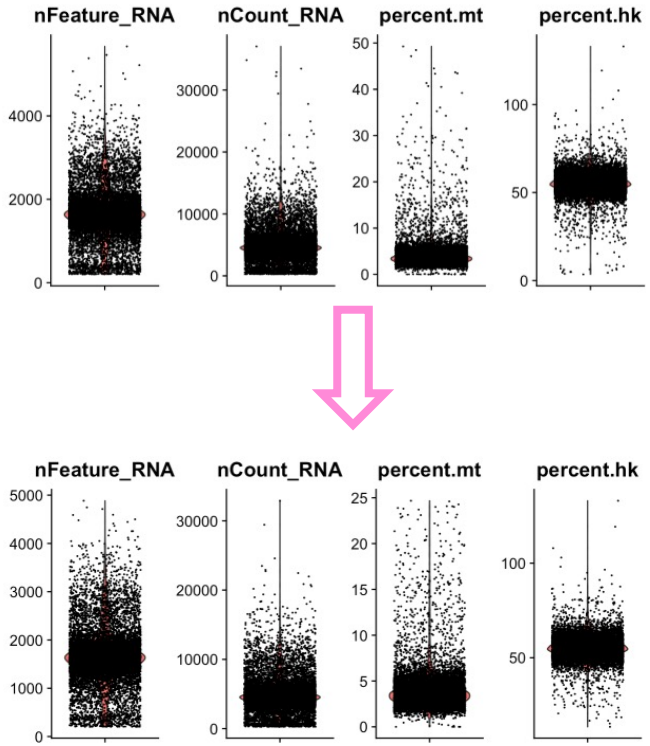
  - remove ambient RNA
  - remove doublets

  % Mitochondria (<25%)

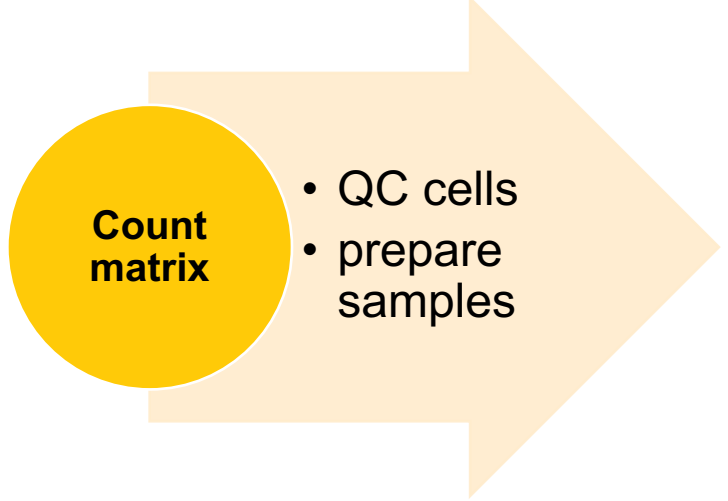
  - remove disrupted cells

  % housekeeping genes (>30%)

  - global quality measure



From matrix to sample objects



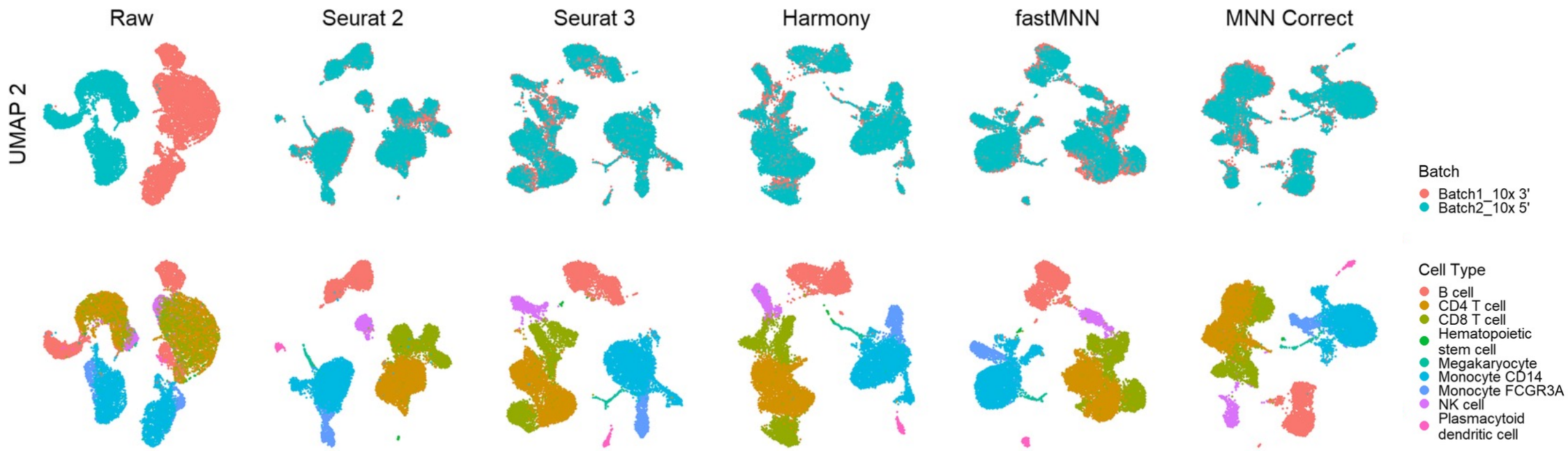
- Add the metadata- *each cell with information*
  - % mitochondria
  - % housekeeping genes
  - Sample code
  - Gender
  - Treatment
  - Disease
  - Timepoint
  - ....
  - Cell annotation

From samples to one object

Integrated object

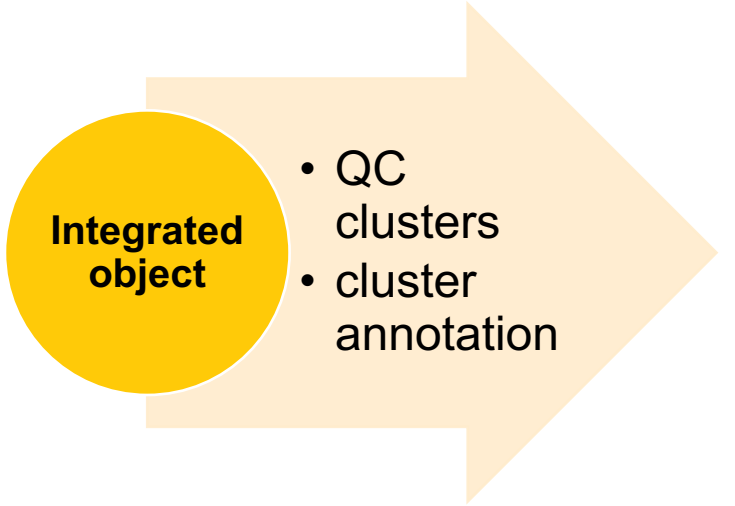
- QC clusters
- cluster annotation

- Sample integration
  1. Simple merge

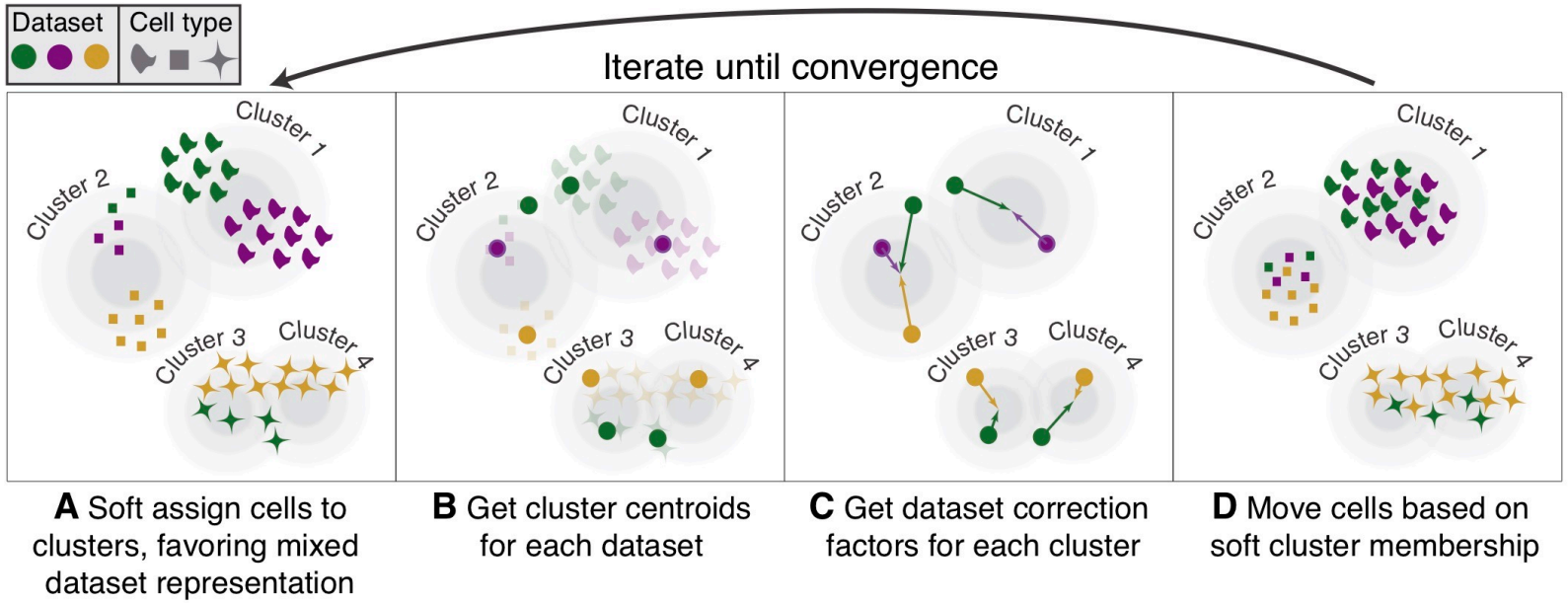


Adapted: Tran, H.T.N., Ang, K.S., Chevrier, M. *et al. Genome Biol* (2020).

From samples to one object



- Sample integration
  1. Integration with correction
     a. Harmony integration



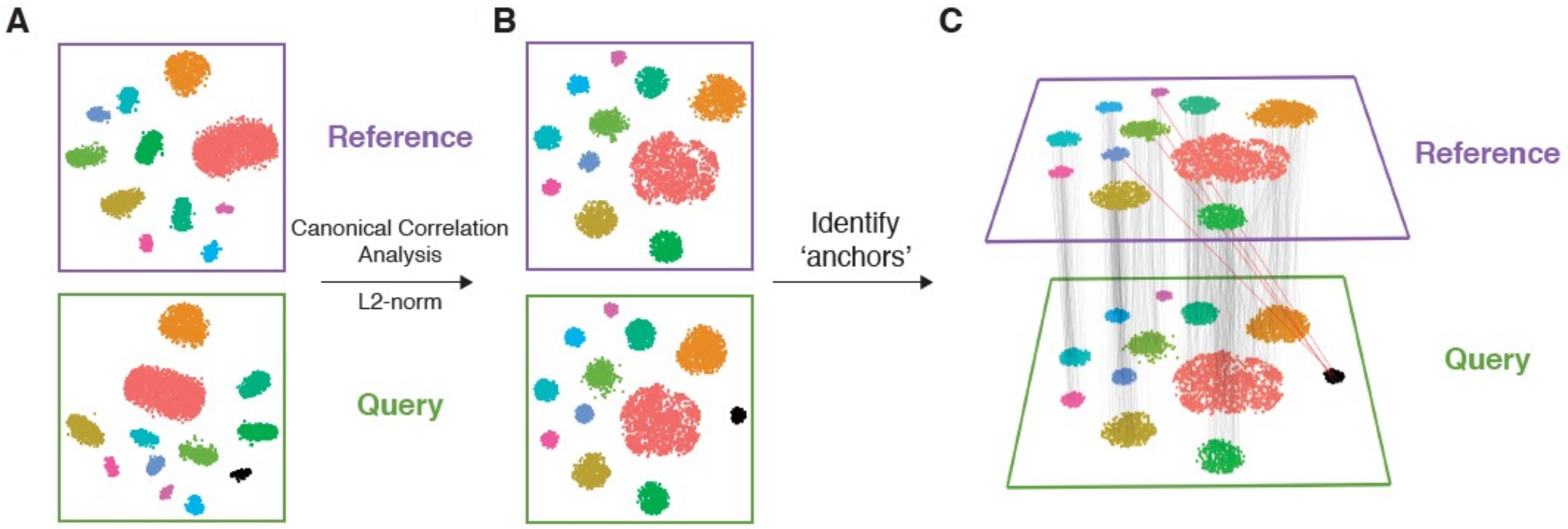
https://portals.broadinstitute.org/harmony/articles/quickstart.html

From samples to one object

Integrated object

- QC clusters
- cluster annotation

b. Seurat sample integration



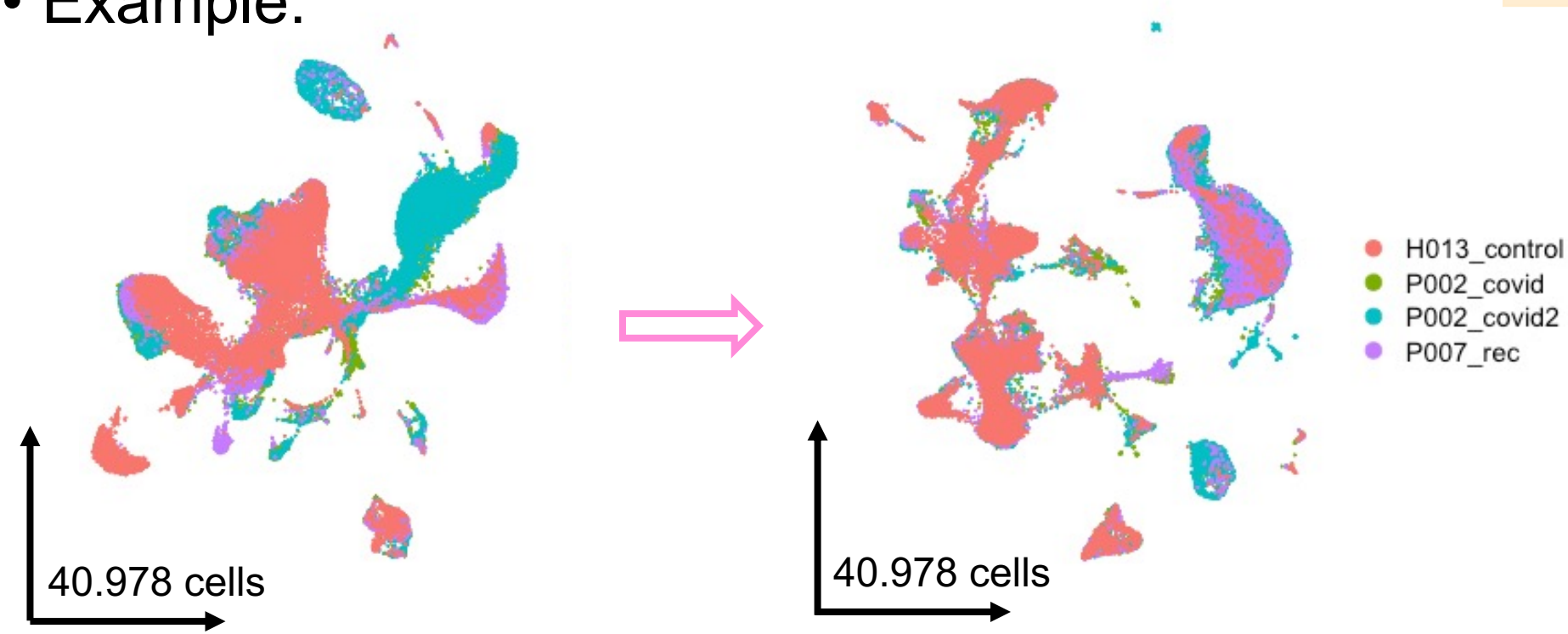
https://satijalab.org/seurat/articles/integration_introduction.html

From samples to one object

Integrated object

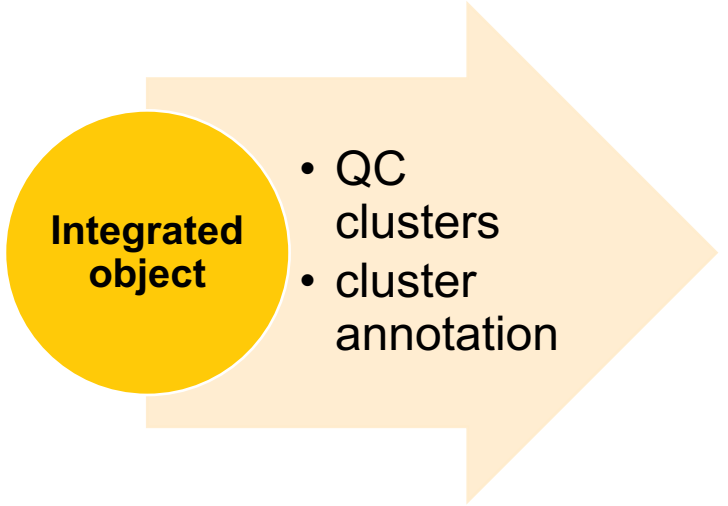
- QC clusters
- cluster annotation

- Example:



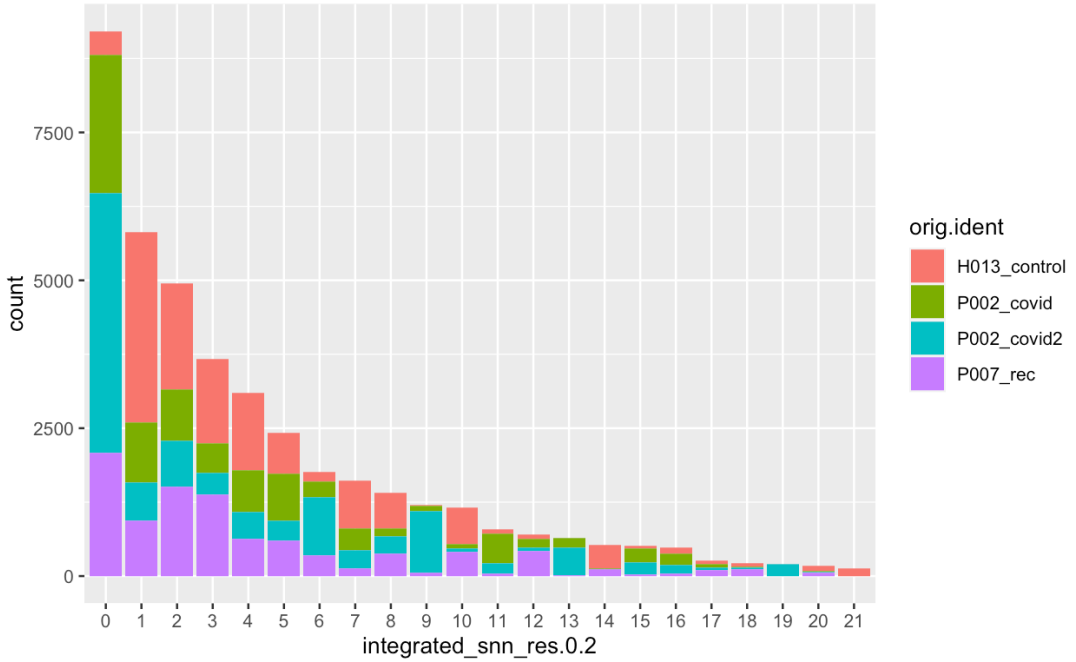
! Goal is not to lose biological relevant features.

From samples to one object



- Doublet removal:
  - Non-informative cells removal

    - actual doublets
    - cells sequenced too shallow
    - clusters with sample enrichment



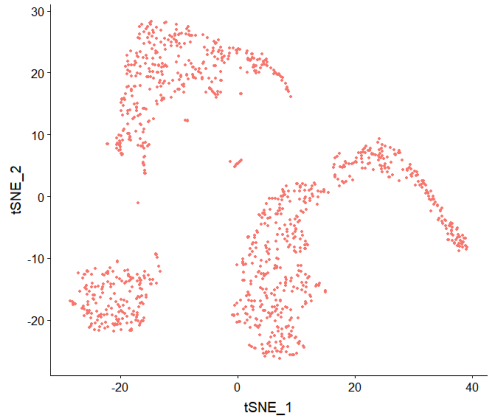
Dimensionality reduction

~15000 genes

2000 variable genes

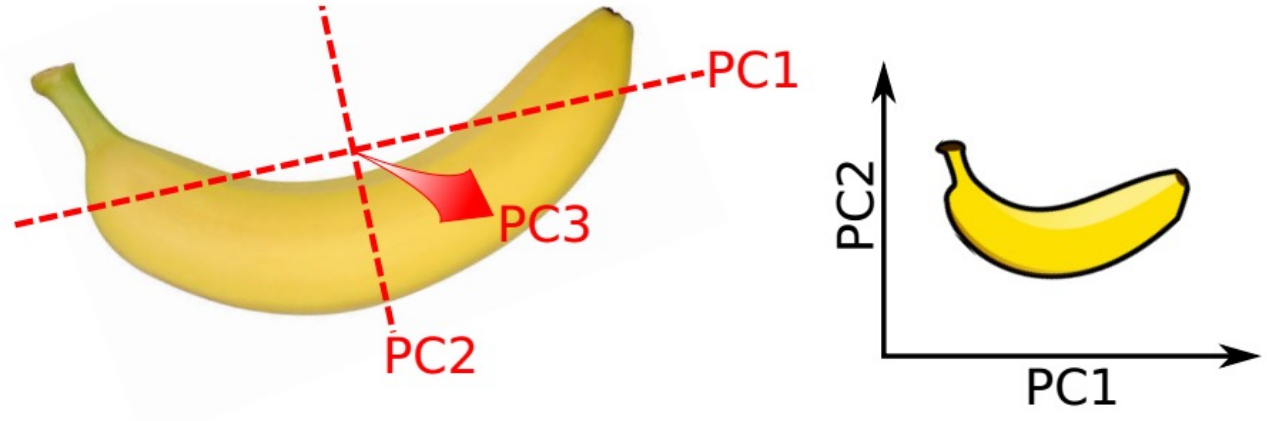
~50 principal components

2 tSNE/UMAP dimensions



Dimensionality reduction with PCA

- Identifies axes of maximal variance in high-dimensional data
- Each subsequent principal component explains less variance



Use the first few PCs as a “summary” of the relevant variance in the data

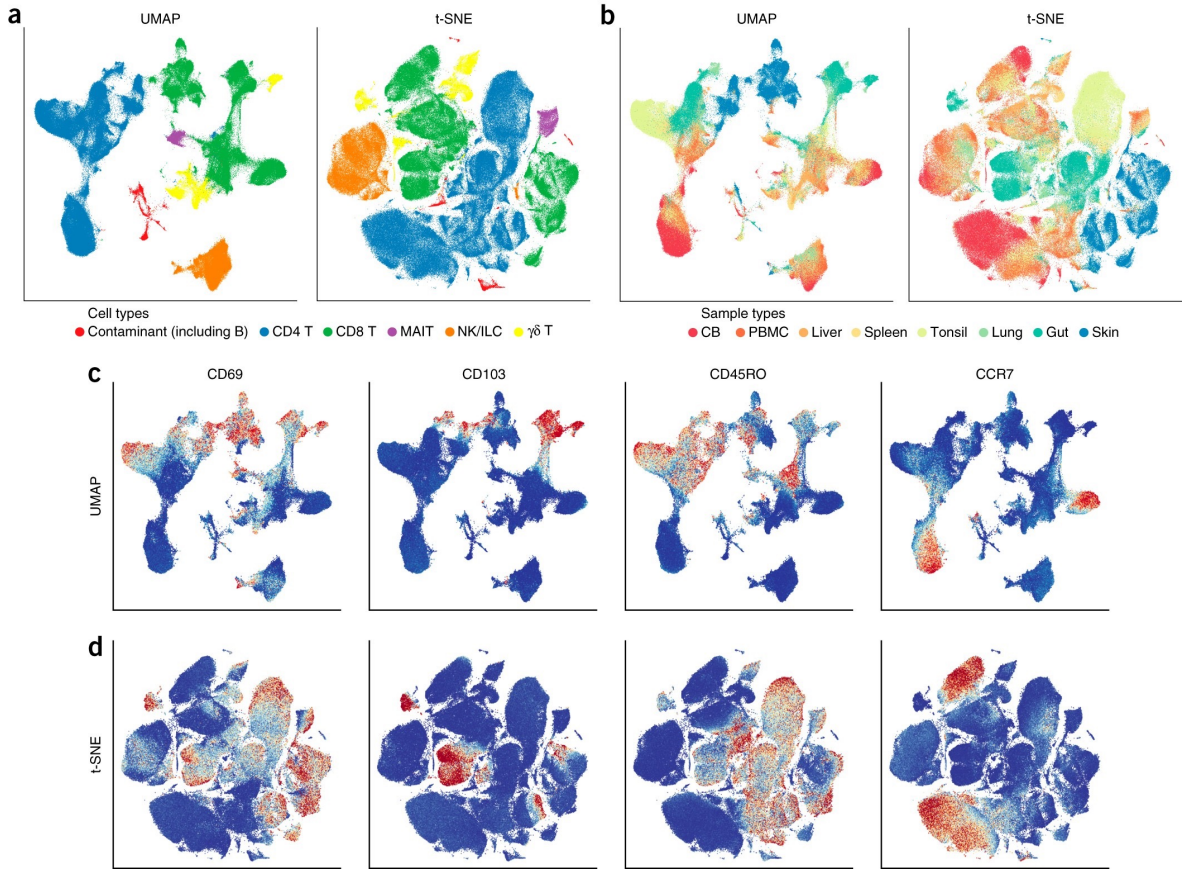
- Speed up downstream analyses by reducing dimensionality
- Focus on biology, remove random noise in later PCs

https://www.youtube.com/watch?v=FgakZw6K1QQ

Slide adapted from Single-cell Analysis Workshop by Aaron Lun, CRUK Cambridge Institute

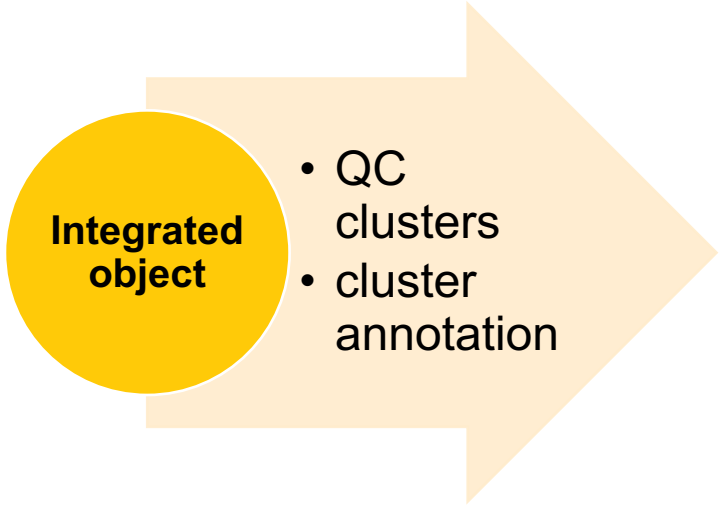
Uniform Manifold Approximation and Projection for Dimension Reduction

UMAP by McInnes et al. is an incredibly powerful tool in the data scientist's arsenal, and offers a number of advantages over t-SNE. While both UMAP and t-SNE produce somewhat similar output, the increased speed, better preservation of global structure, and more understandable parameters make UMAP a more effective tool for visualizing high dimensional data. Finally, it's important to remember that no dimensionality reduction technique is perfect - by necessity, we're distorting the data to fit it into lower dimensions - and UMAP is no exception.

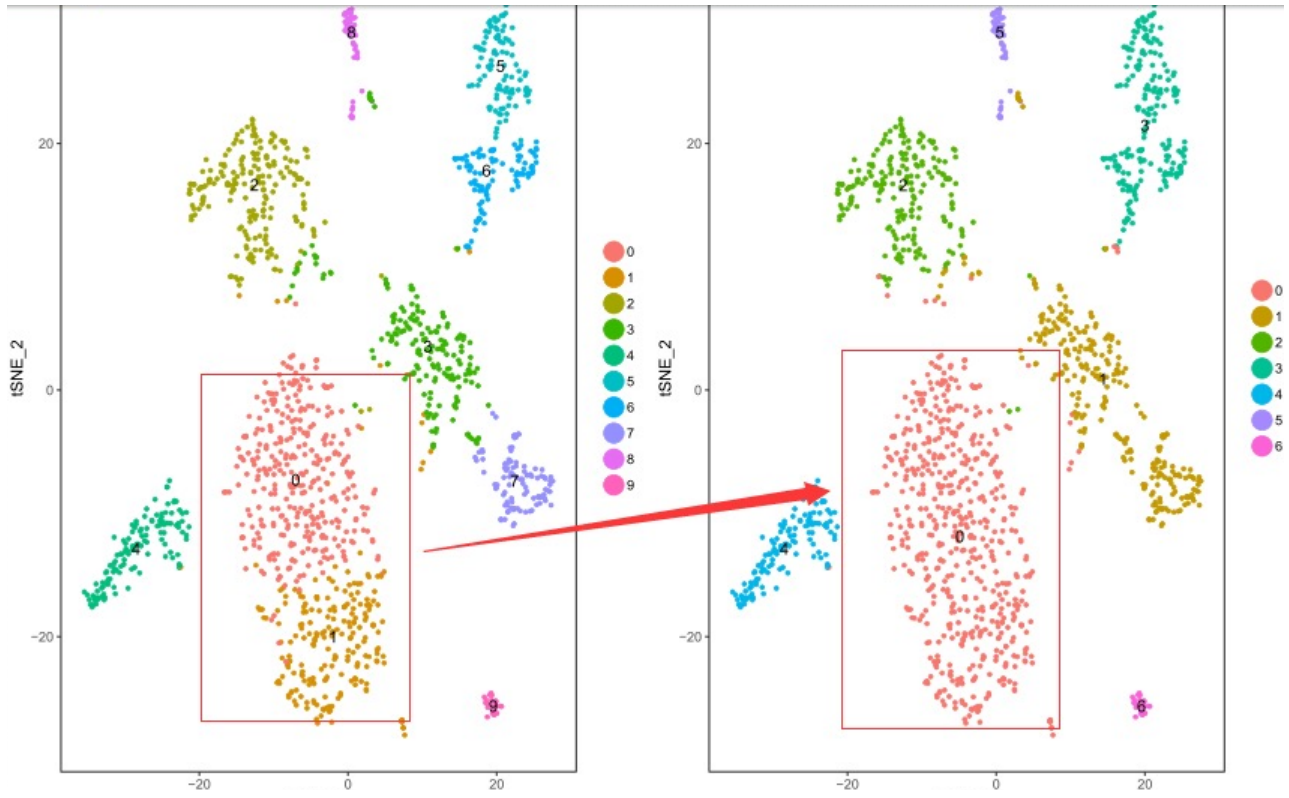


https://www.nature.com/articles/nbt.4314
https://pair-code.github.io/understanding-umap/

How many clusters?

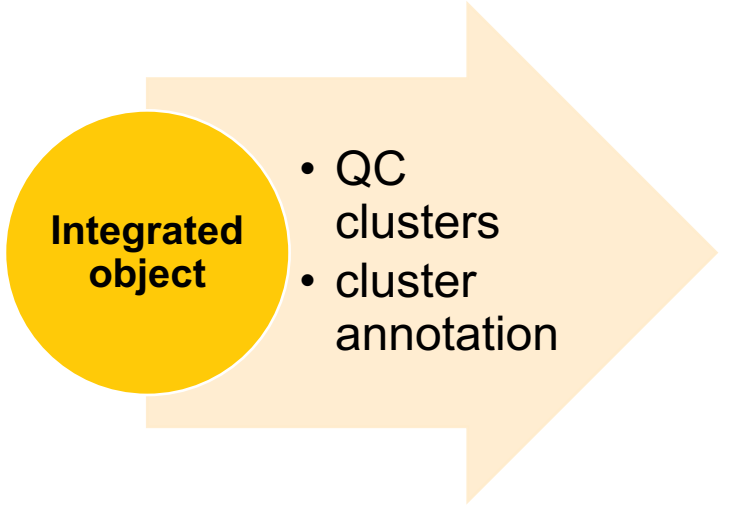


Cluster resolution is a parameter that controls the granularity of clustering in Seurat.
Higher resolution > more clusters
Lower resolution > fewer clusters



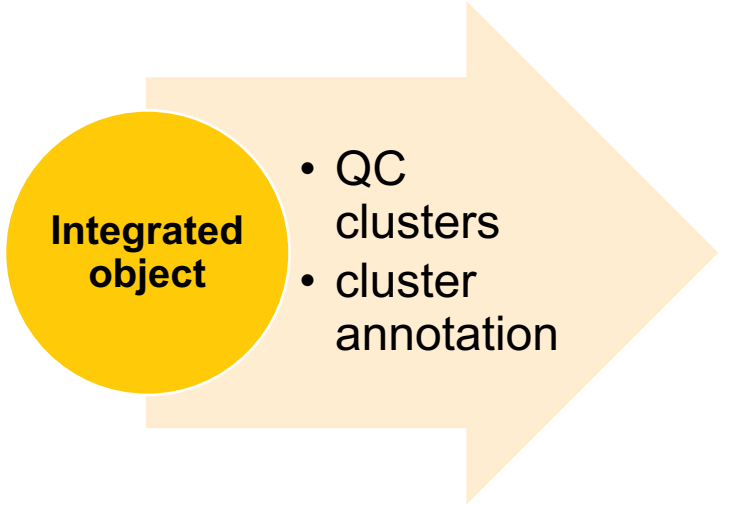
Keep in mind that the optimal cluster resolution may vary between datasets and experiments, and it's often a matter of experimentation and fine-tuning to strike the right balance between granularity and interpretability of the results.

From samples to one object



- Cluster annotation:
  1. Reference based method
  2. Package based method
  3. Literature or *in-house* biomarkers

From samples to one object



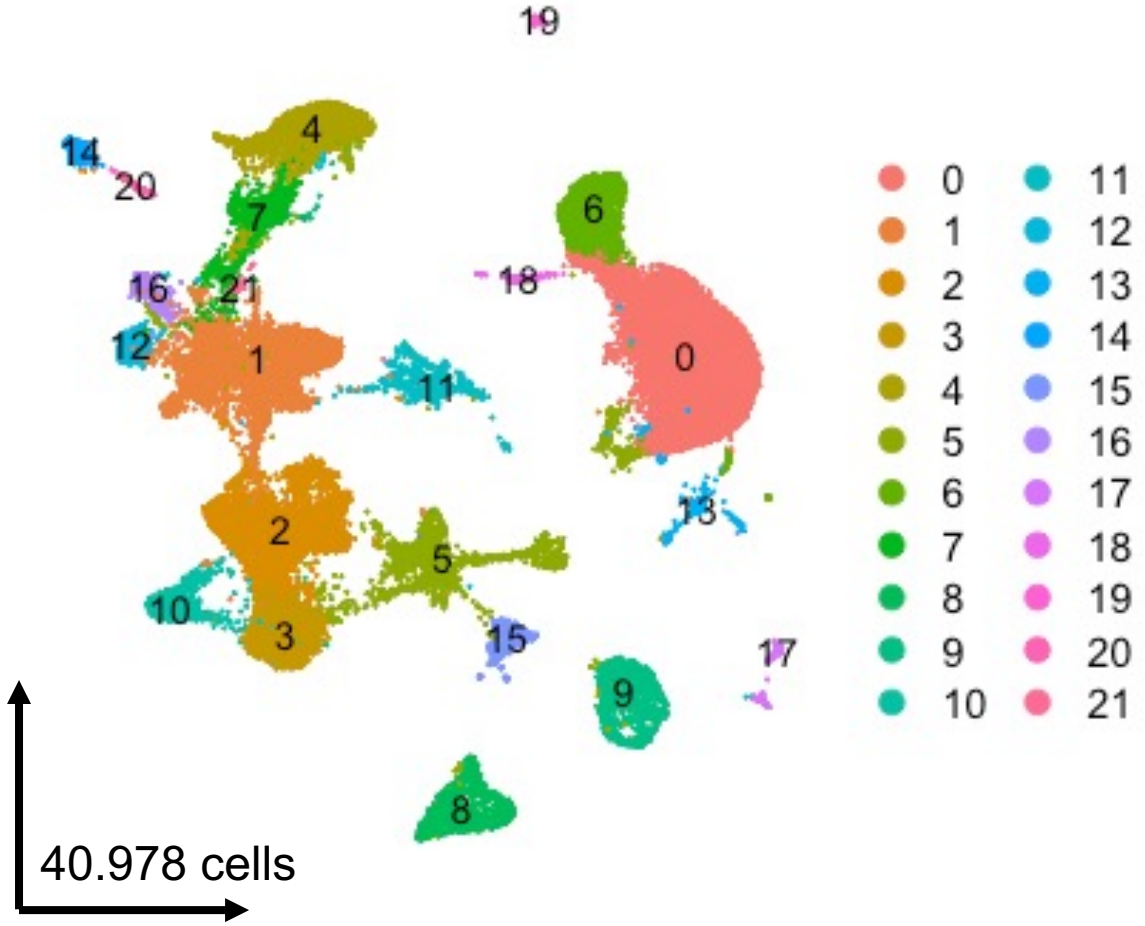
- Reference based method

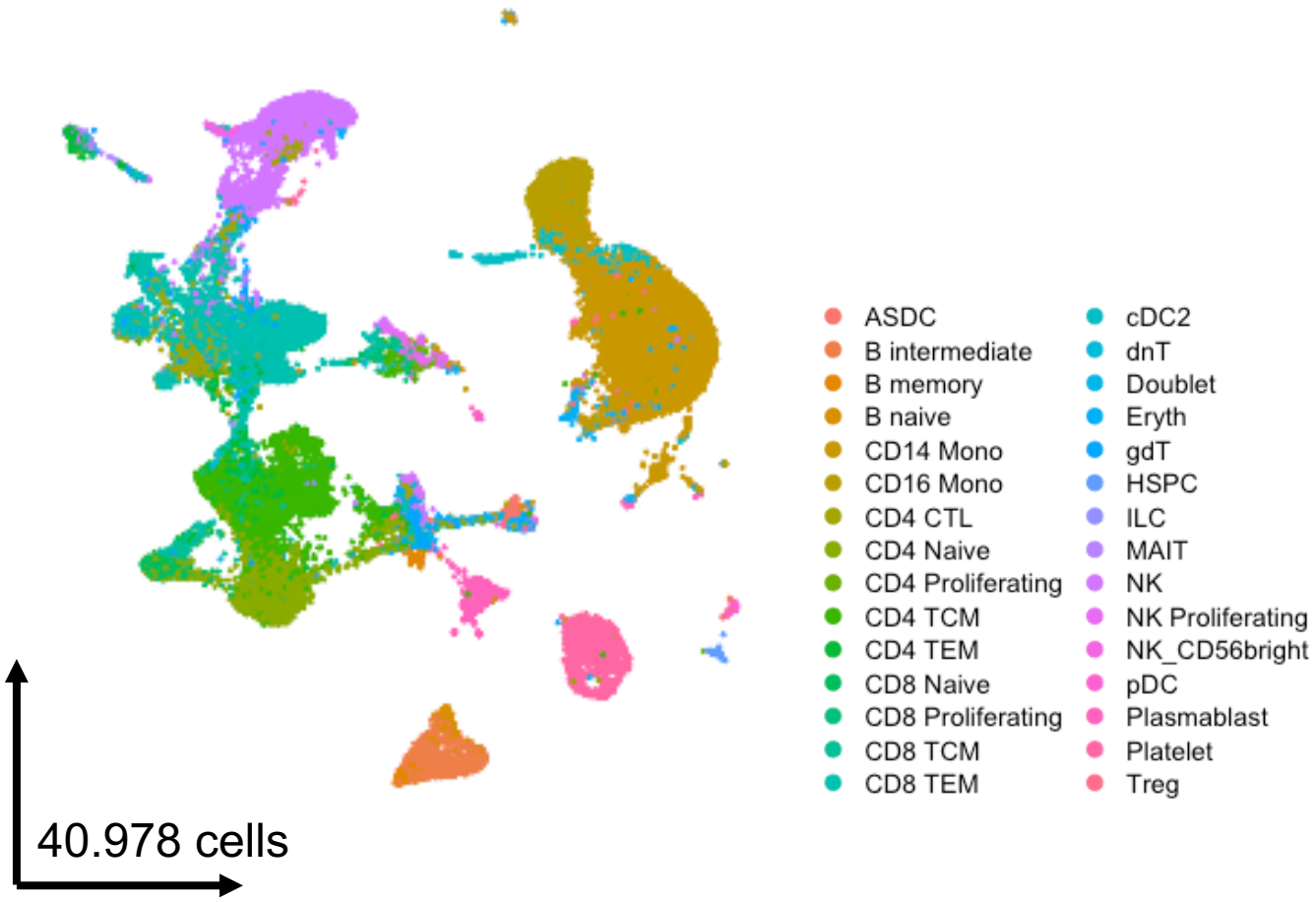
  - Choose a reference available in literature (e.g. spleen or PBMCs)
  - Use the manuscript annotated cells as anchors/references for your own object/query.

  Advantages:

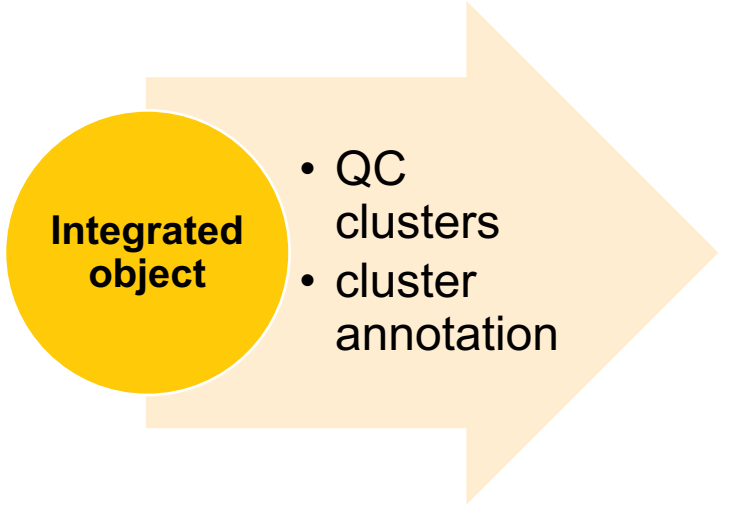
  Disadvantages:

Reference based annotation (Seurat V4)





From samples to one object



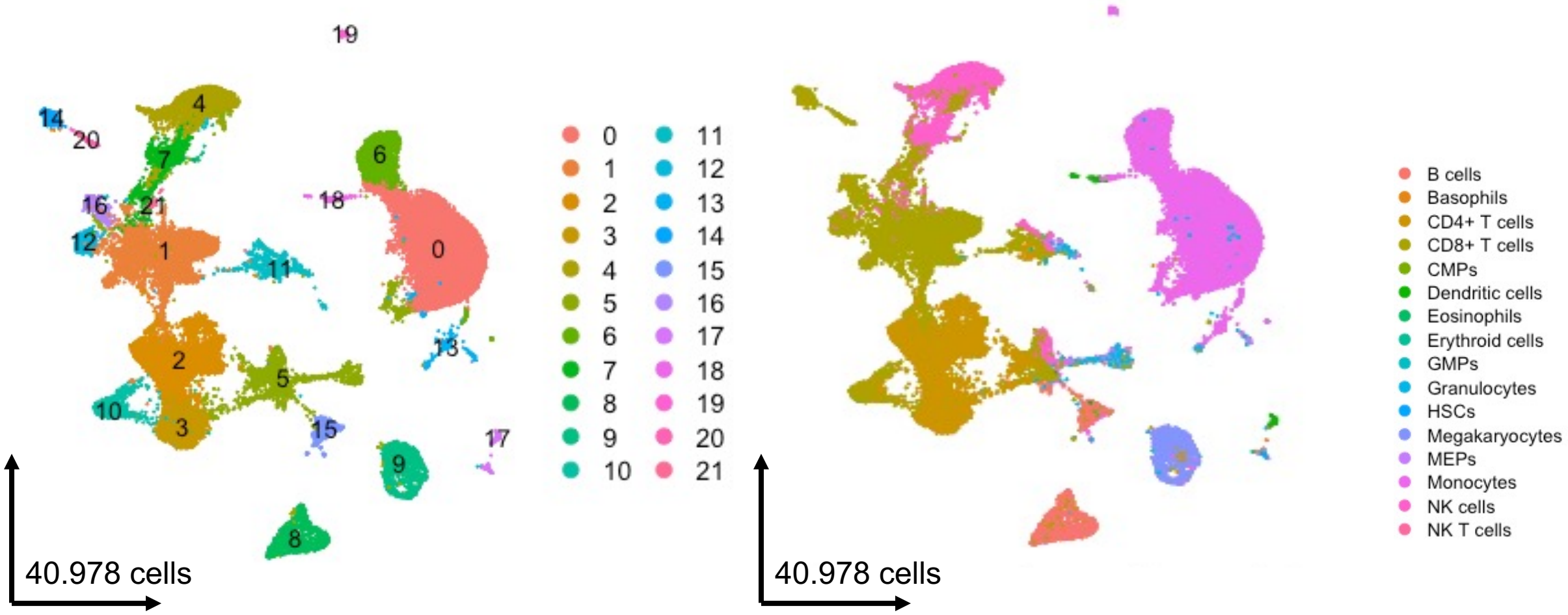
- Package based method

  - Use packages available (SingleR)
  - Use your cells expression and map to databases (single cell and flow cytometry)

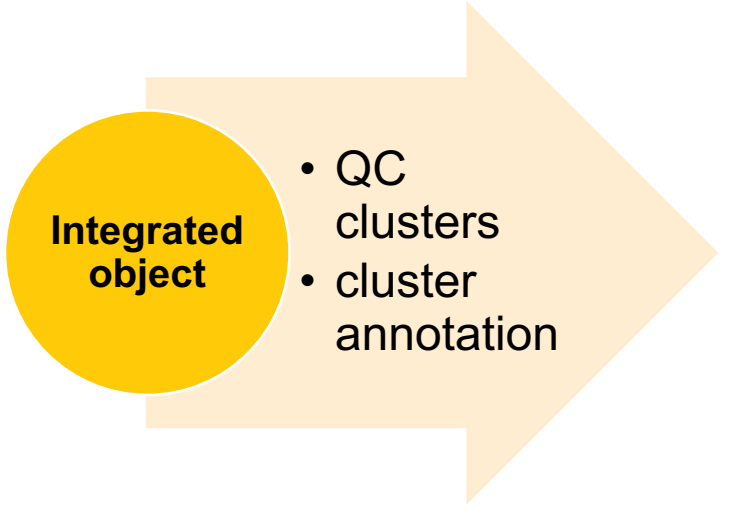
  Advantages:

  Disadvantages:

Package based annotation (SingleR- dmap)



From samples to one object



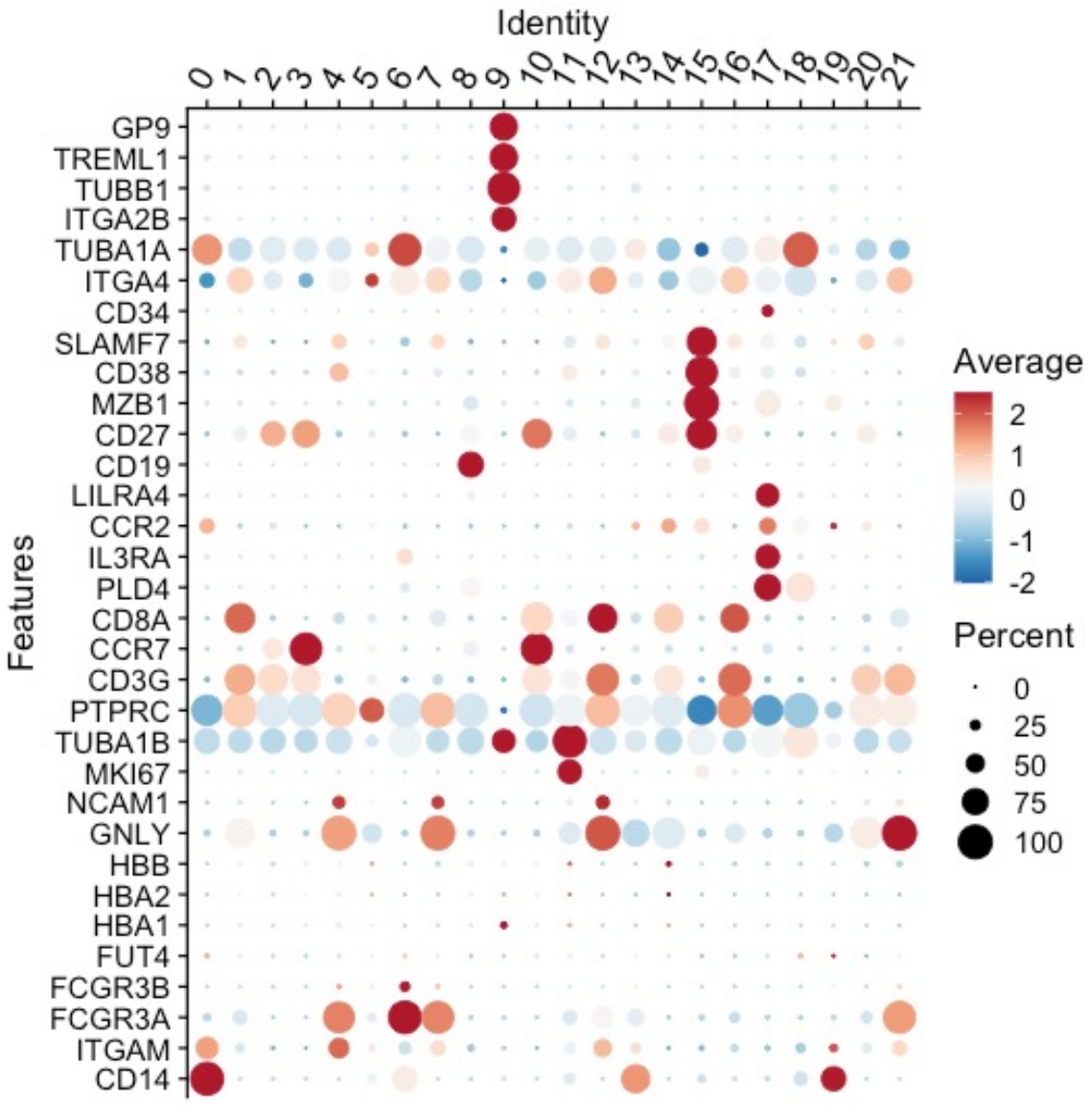
- Literature or *in-house* biomarkers

  - Scientists have been using cell expression to identify cells for decades.
  - Use (e.g. flow cytometry) markers to identify your cells based on cluster expression.

  Advantages:

  Disadvantages:

Cluster based annotation (in-house/literature knowledge)



0: CD14 Monocytes

1: CD8 T cells

...

8: B cells

9: Megakaryocytes

Your object

Integrated object

- QC clusters
- cluster annotation

40.978 cells

- CD4+ T cells
- CD8+ Tcells
- other T cells
- NK cells
- B cells
- Plasmablasts
- Prolif Lymphocytes
- Megakaryocytes
- HPSC
- DC
- CD14 Monocytes
- CD16 Monocytes

Recommendations

The object is big- depending on the cell number

- Work smart: parallelize operations
- Let your IT department/responsible person know

The focus should be on individual cell types/ compartments

- We normally do not want to ‘discover new cell types’
- Base the cell annotations on the experts in the field
- Histology books can be very useful

Transcriptome analysis

- Finally we can start dissecting the biological signals:

cell proportion analysis	Differential gene expression	Coexpression analysis	pseudotime analysis
• scCODA • ANOVA • Mixed linear models	• Wilcoxon • MAST	• hdWGCNA	• monocle3 • RNA velocity

e.g. treated vs. non-treated or throughout **timepoints** (covid-19 paper)

Cell proportion analysis

cell proportion analysis

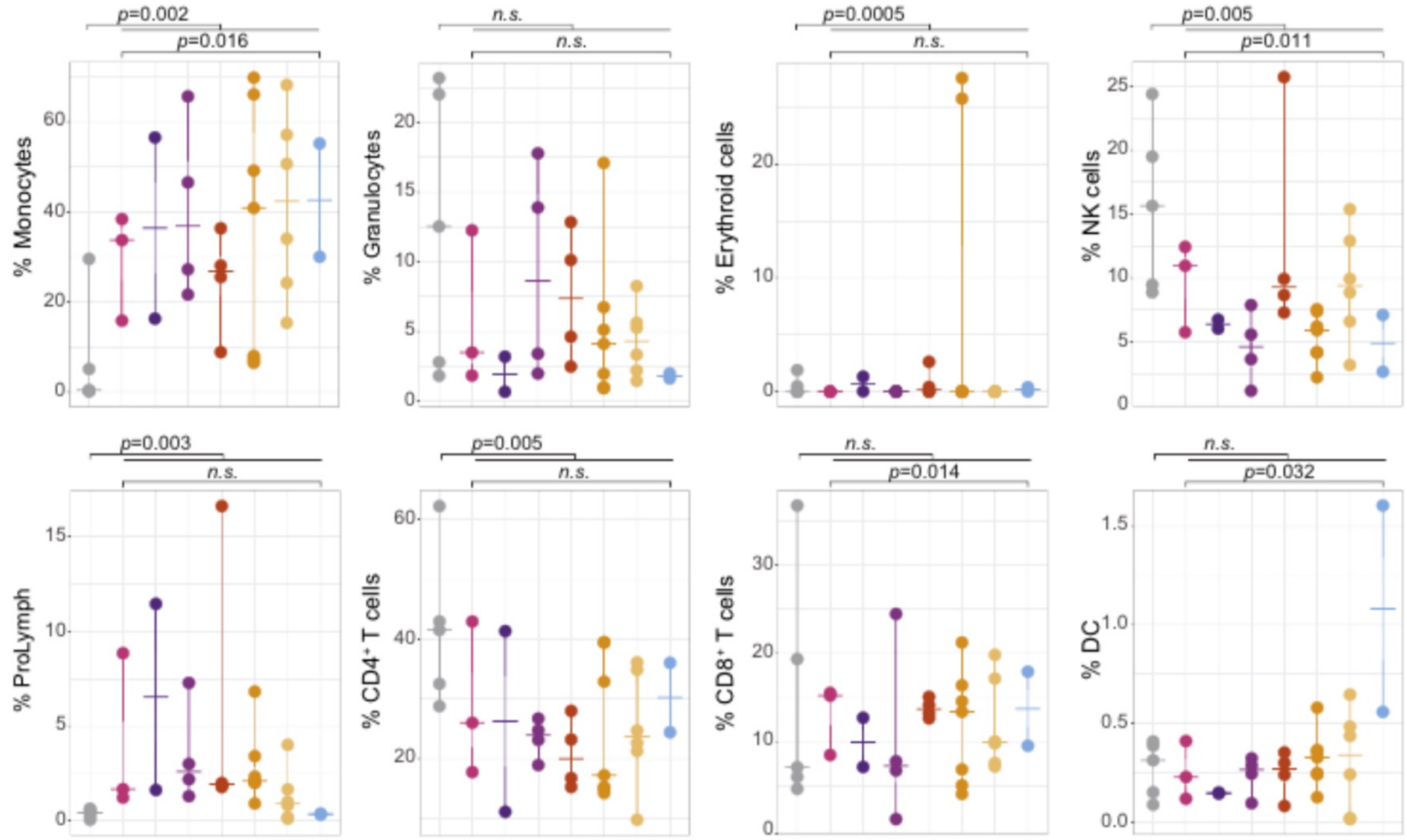
- scCODA
- ANOVA
- Mixed linear models

• Identifies significant changes in cell proportions based on treatment or other parameters.

Disadvantages:

Cell proportion change through time: Mixed linear model

Cell proportion change between healthy and covid19: ANOVA



Bernardes JP, Mishra N Tran F et al. Immunity 2020

Differential expression analysis

Differential gene expression

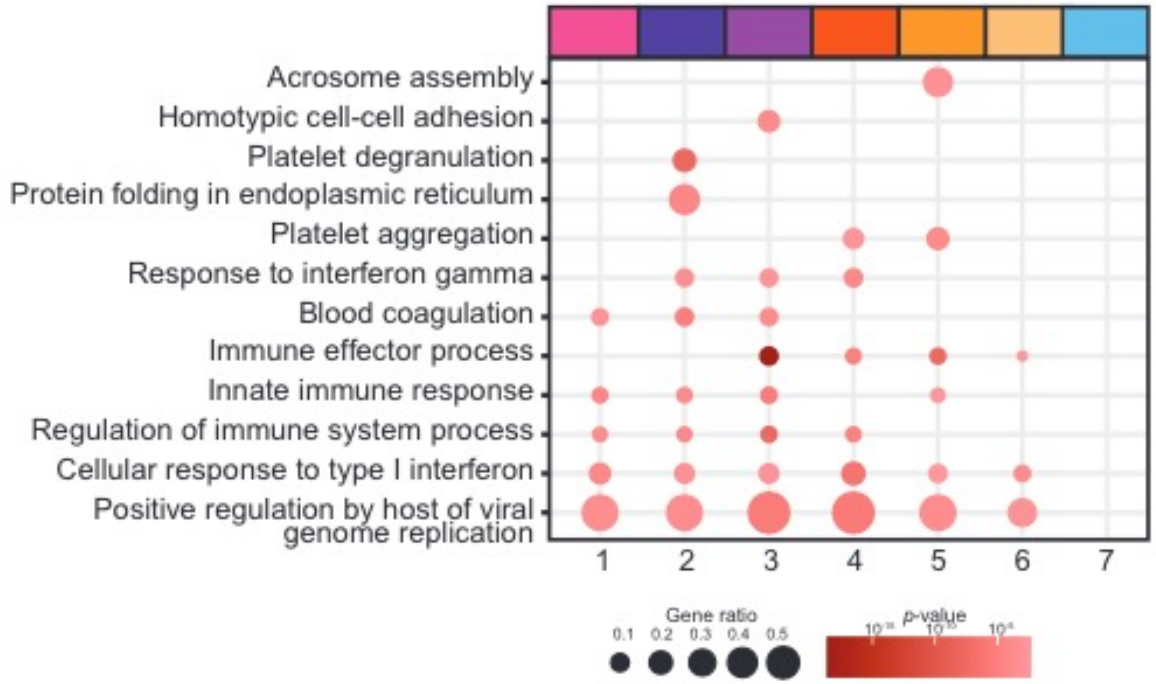
- GO enrichment
- KEEG enrichment

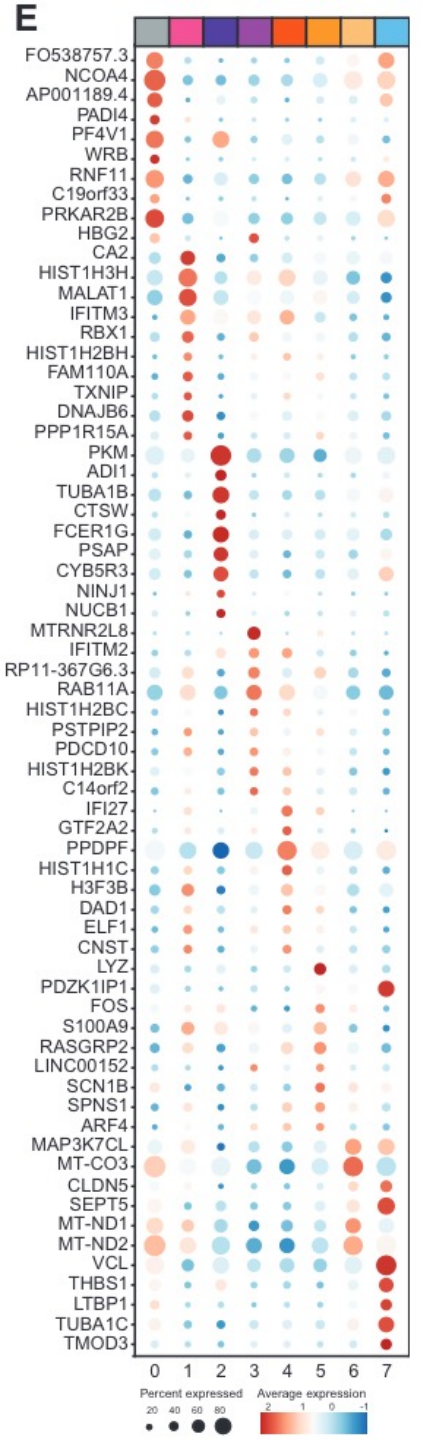
• Identifies expression changes between treatments, for individual genes within a cell type. Differential expression can be based on:

- Signature genes: compare one against all other groups/clusters.
- Pairwise comparisons: compare one group (e.g. treatment) to another group.

Disadvantages:

Differentially expressed genes can also be used to inform GO terms enriched throughout timepoints





Bernardes JP, Mishra N Tran F et al. Immunity 2020

Coexpression analysis

Coexpression analysis

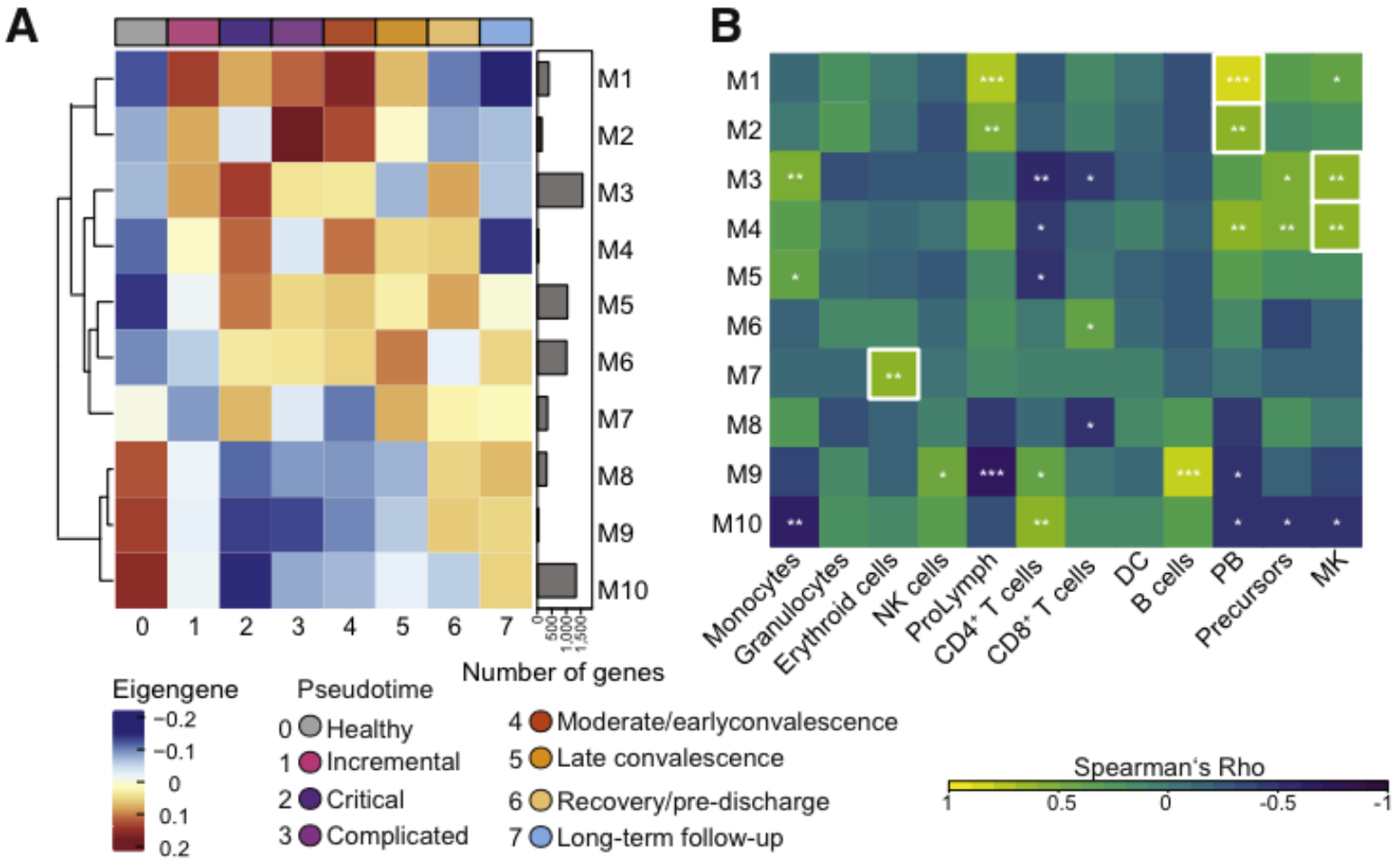
- GO enrichment
- KEEG enrichment

• Specially useful for longitudinal studies. Groups genes into modules that change similarly throughout time for a particular cell-type.

- Based on differentially expressed genes
- Relate modules with clinical parameters

Disadvantages:

Coexpression modules can be calculated in bulk RNA-seq and correlated with cell-type proportions



Bernardes JP, Mishra N Tran F et al. Immunity 2020

Pseudotime or trajectory analysis

pseudotime analysis

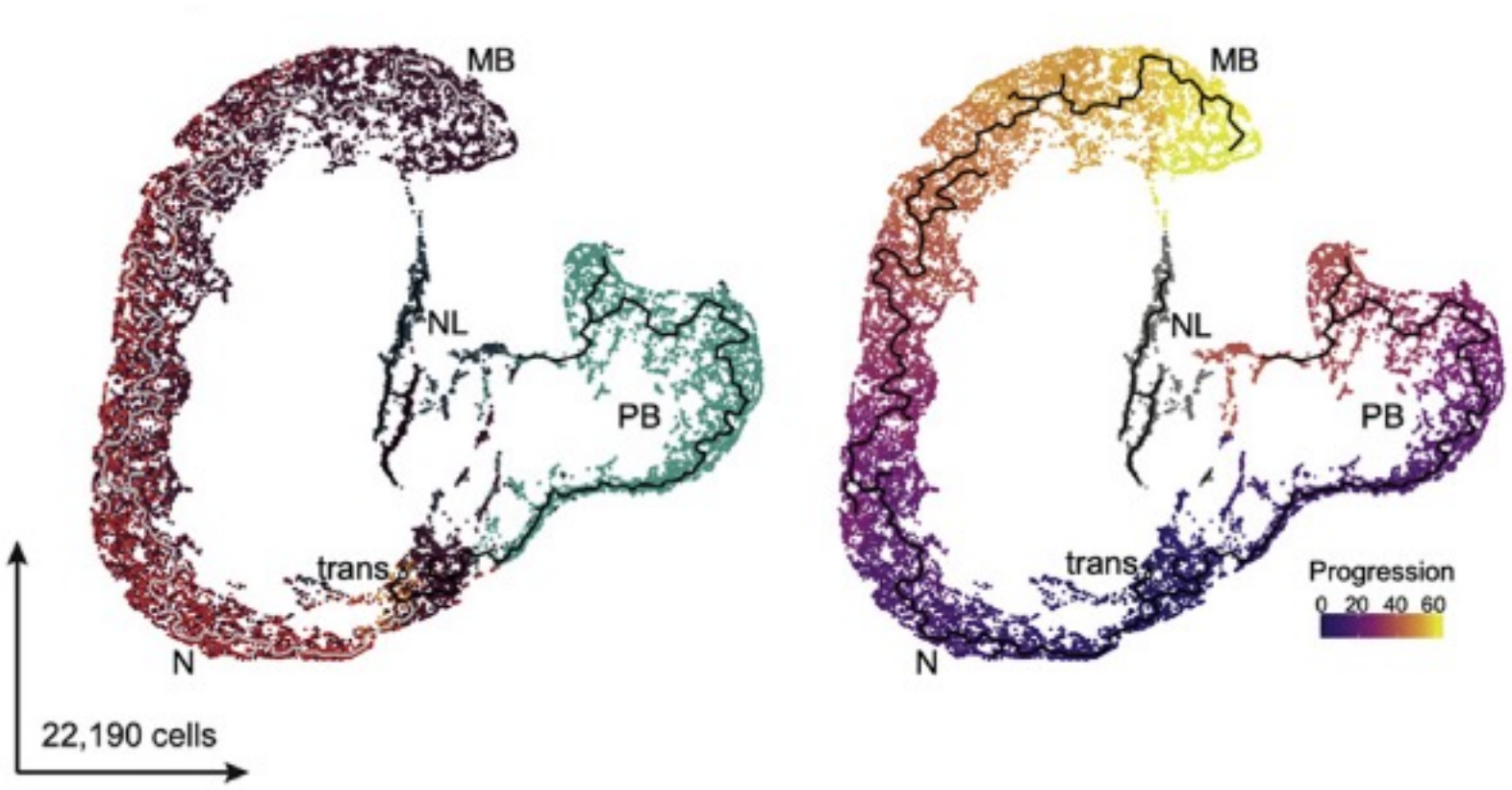
- monocle3
- RNA velocity

• Allocate cells to lineages and then orders them based on pseudotimes within these lineages.

- Rooted (Monocle3)
- Unrooted: scVelo

Disadvantages:

Cell-type lineage can resolve differentiation



Bernardes JP, Mishra N Tran F et al. Immunity 2020

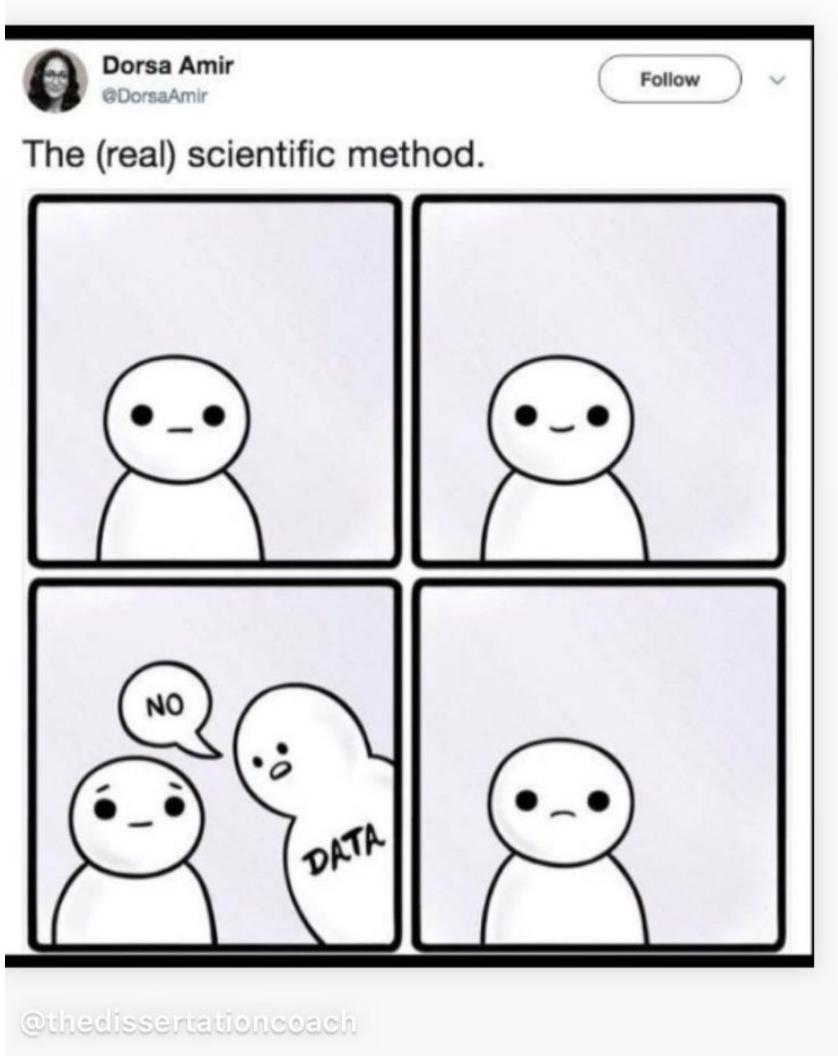
What we learned

Preparation of samples into the single cell object.

How to annotate the single cell object.

How to analyze the cell types annotated.

Questions?


Dorsa Amir
@DorsaAmir
Follow
The (real) scientific method.
NO
DATA
@thedissertationcoach

Where can I get public data?

- www.fastgenomics.org
- https://www.ncbi.nlm.nih.gov/geo/
- https://support.10xgenomics.com/single-cell-gene-expression/datasets/
- https://portals.broadinstitute.org/single_cell
- https://preview.data.humancellatlas.org/
- http://jinglebells.bgu.ac.il/
- ….