Genome Informatics 2022

Lesson 5 - Single cell RNA & Spatial transcriptomics analysis

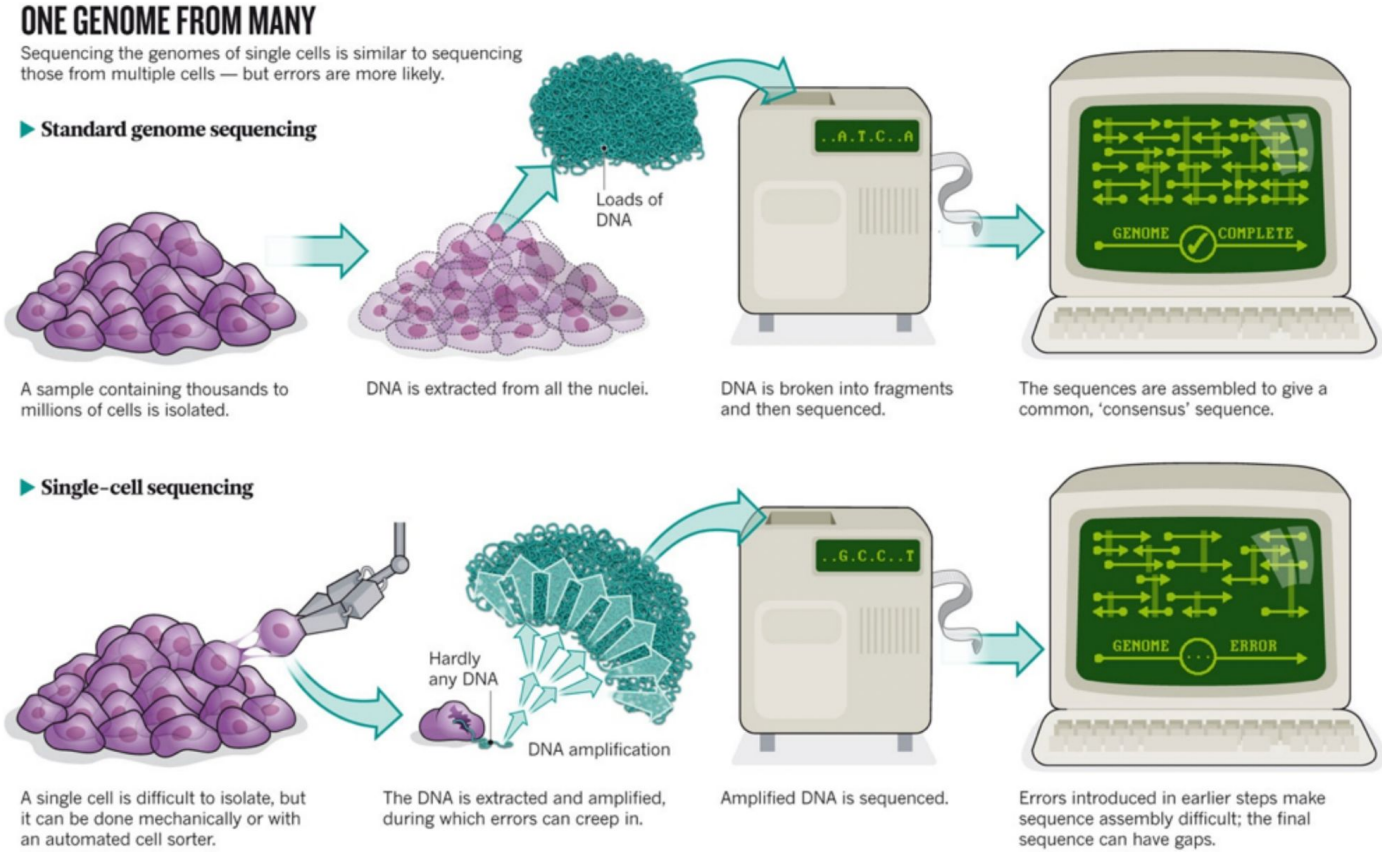
Why Single Cell study?

- Hidden variation in gene expression
- Regulatory process of biotechnological or medical relevance
- Relationship between cellular processes and external stimuli

Why Single Cell study?

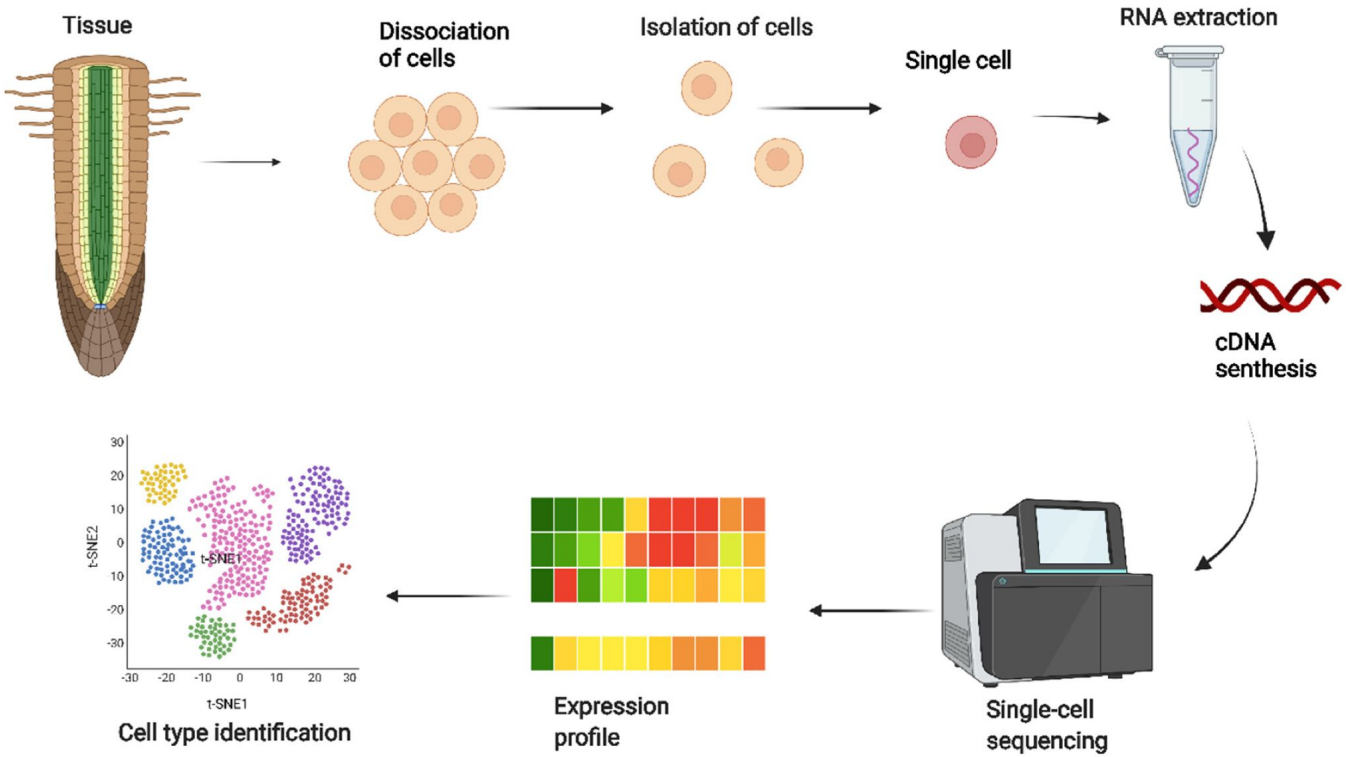
- Developmental biology
  - Discover more complicated mechanisms in cellular development
  - Confirm the distinct gene expression signatures across different cell types
  - Identify functional differences among the same cell cell type
- Cancer biology
  - Find evidence for models of cancer
  - Infer timing of mutations and the drivers
  - Evaluate effectiveness of targeted therapy
- Microbiology
  - Discover low-abundance species that are are difficult to culture in vitro
  - Monitor transcriptional gene activation mechanisms for functional annotation

Bulk RNA sequencing vs Single cell



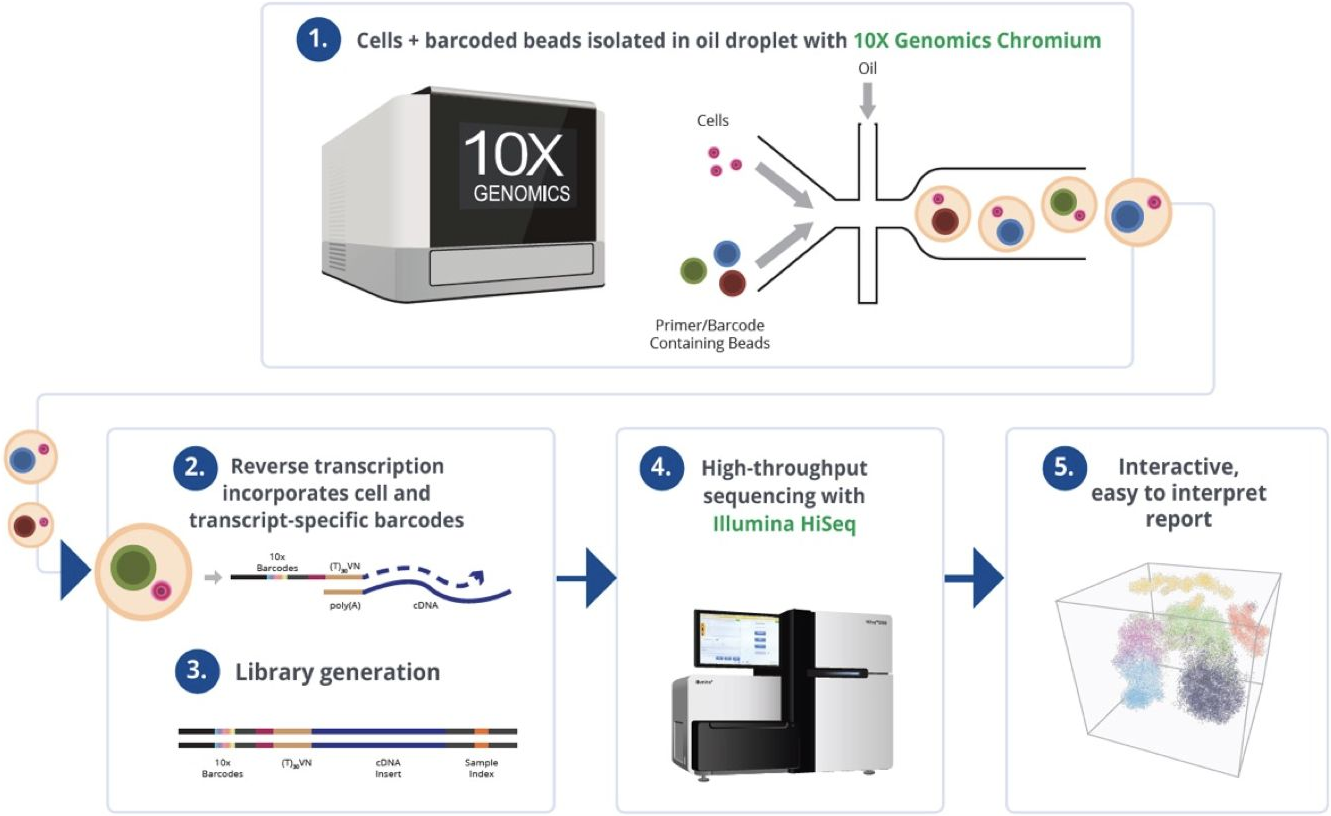
https://scitechdaily.com/images/one-genome-from-many.jpg

Single cell sequencing

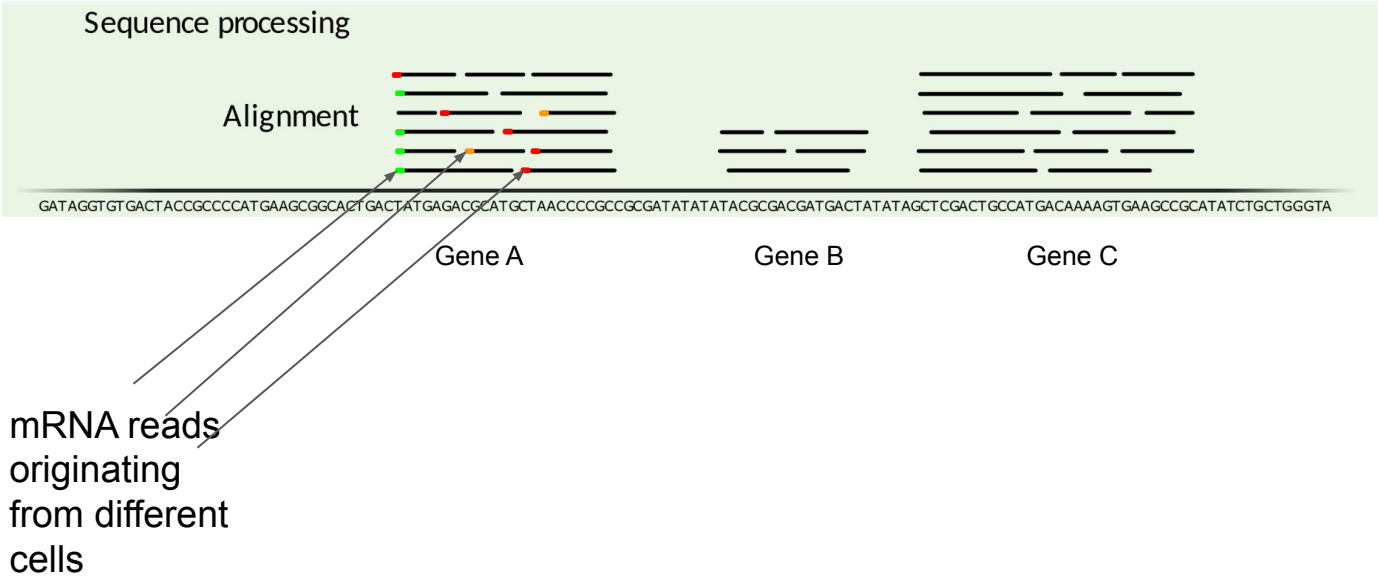


Single-Cell RNA Sequencing for Plant Research: Insights and Possible Benefits by George Bawa

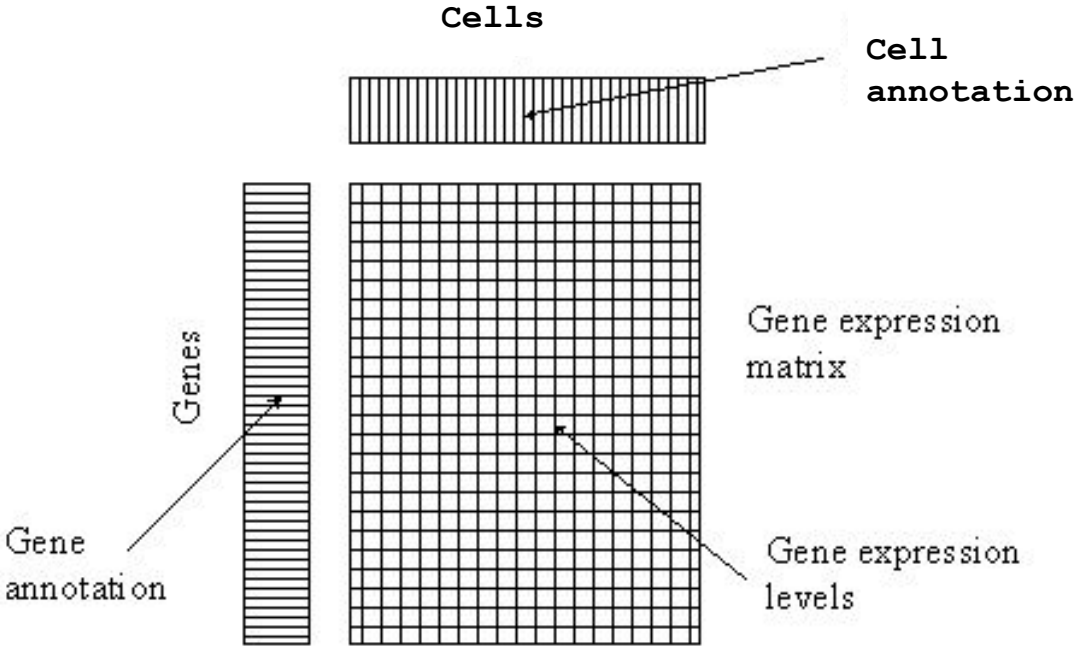
Single Cell RNA-seq: Easy as 1,2, 3, … 5



Single cell sequencing alignment and gene count

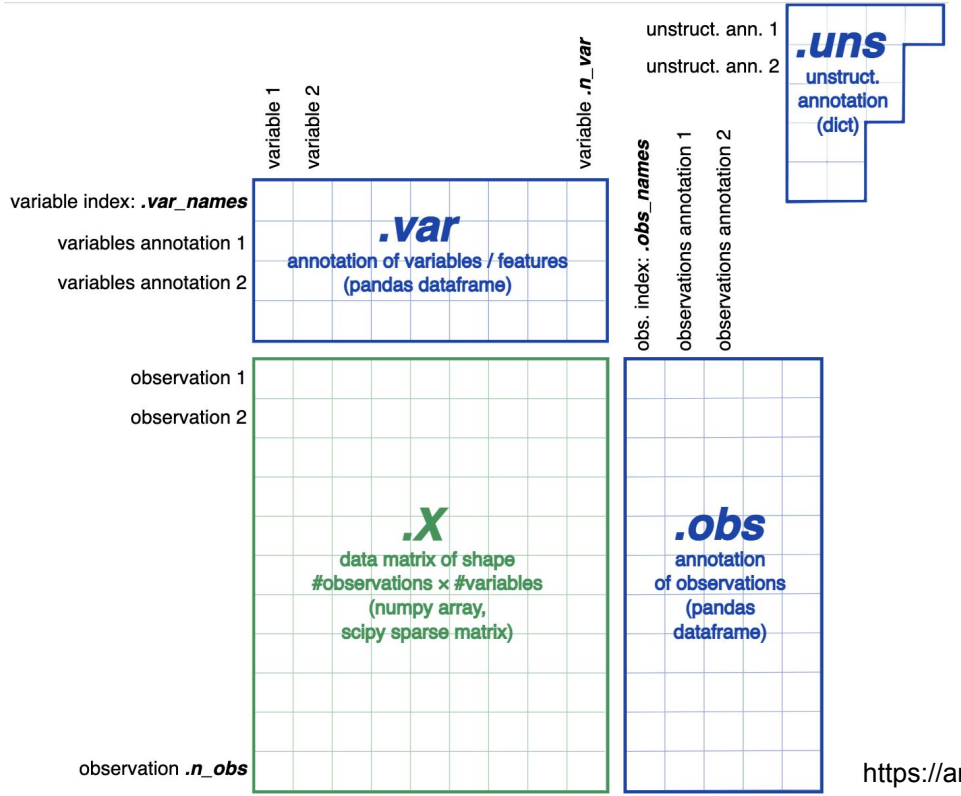


Cell-gene matrix



http://www.people.vcu.edu/~mreimers/OGMDA/gene_expression_matrix.gif

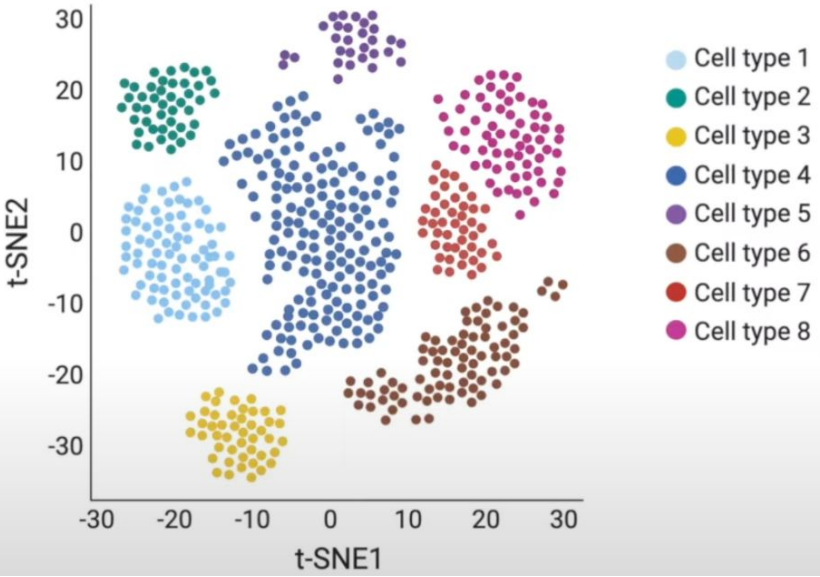
Annotated data object



https://anndata-tutorials.readthedocs.io/en/latest/getting-started.html

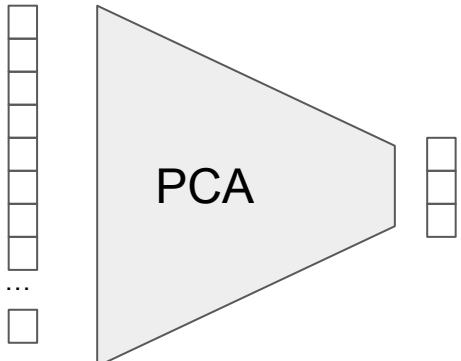
Cell-gene matrix

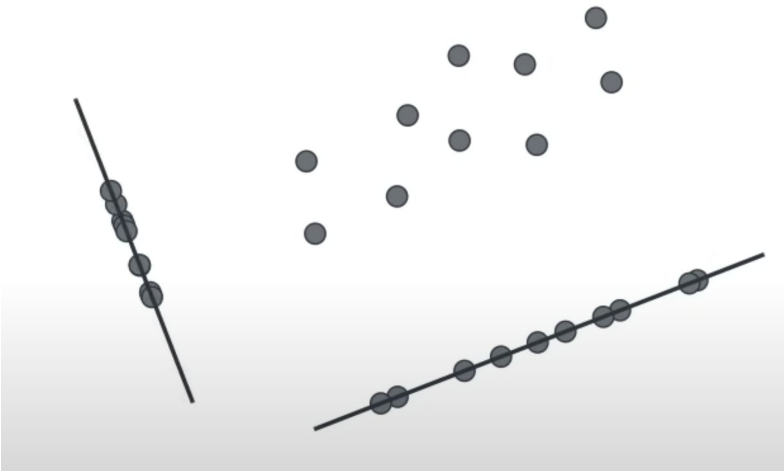
	Cell 1	cell 2	cell 3	...	cell X
Gene a	3	5	6	...	3
gene b	3	5	3	...	2
gene c	5	6	5	...	4
gene d	5	6	7	...	8
...	...	...	...	...	
gene z	7	8	4	...	3



Latent (low-dimensional) representation of data

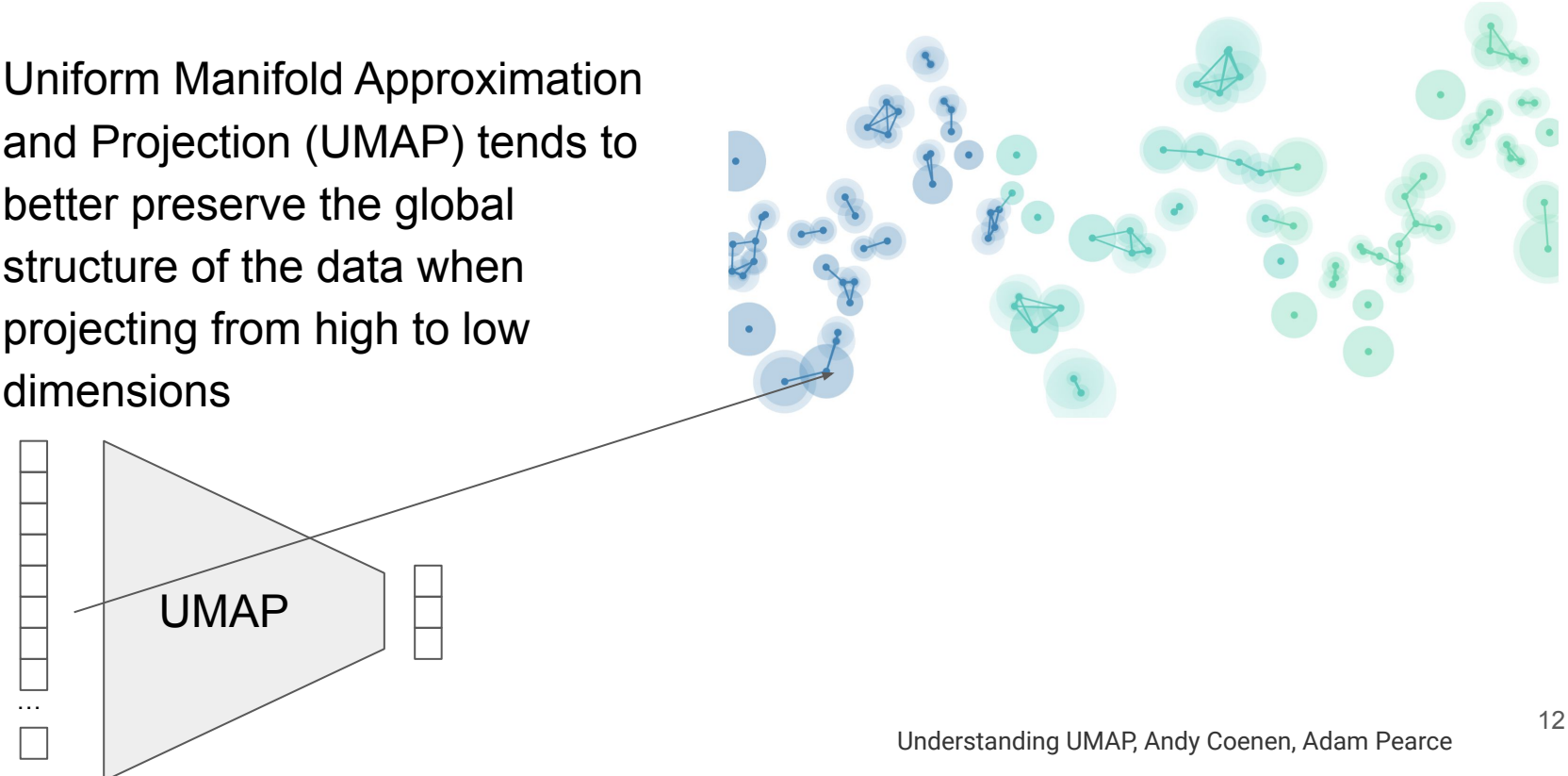
- Principal Component analysis - linearly transforming the data into a new coordinate system where (most of) the variation in the data can be described with fewer dimensions than the initial data





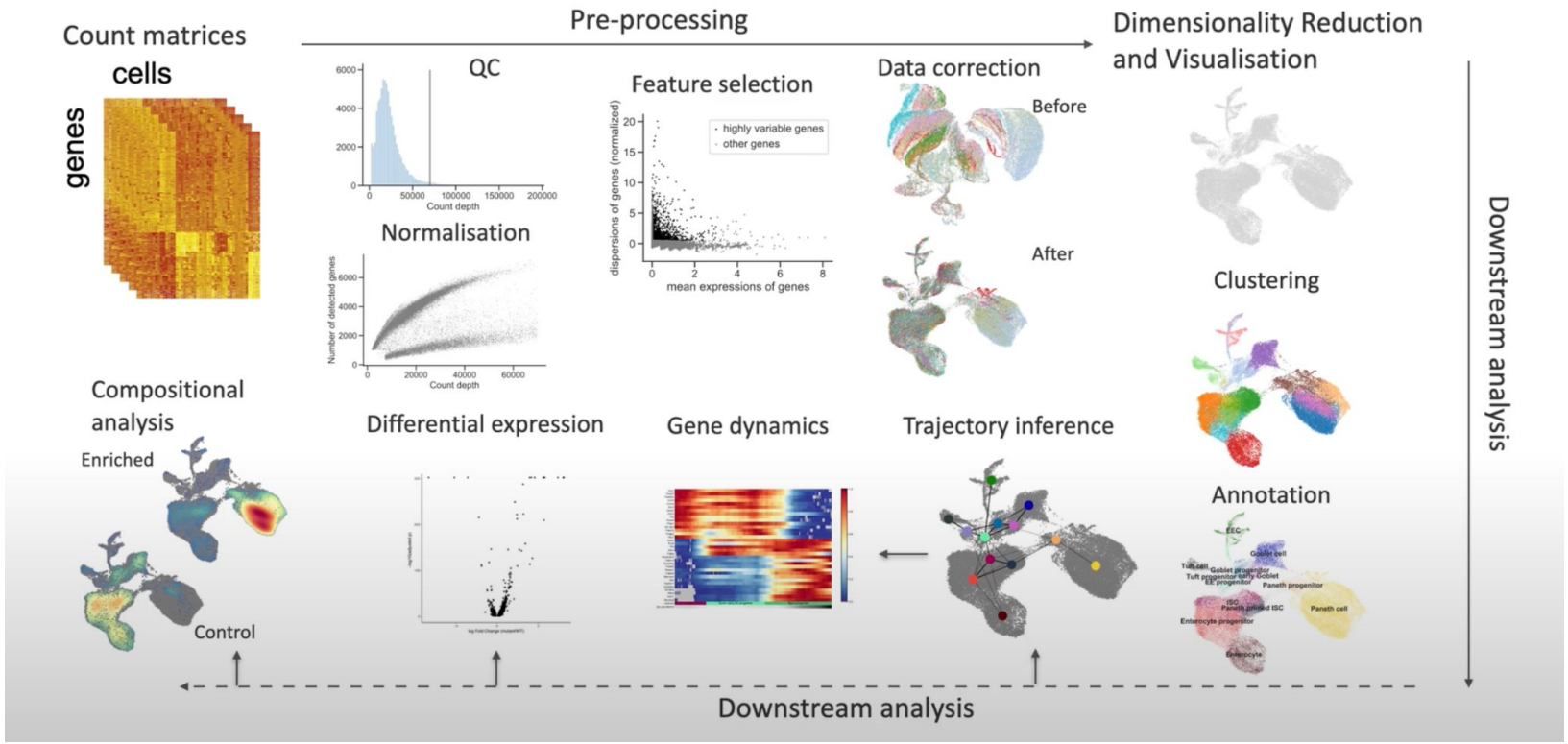
Latent (low-dimensional) representation of data

- Uniform Manifold Approximation and Projection (UMAP) tends to better preserve the global structure of the data when projecting from high to low dimensions



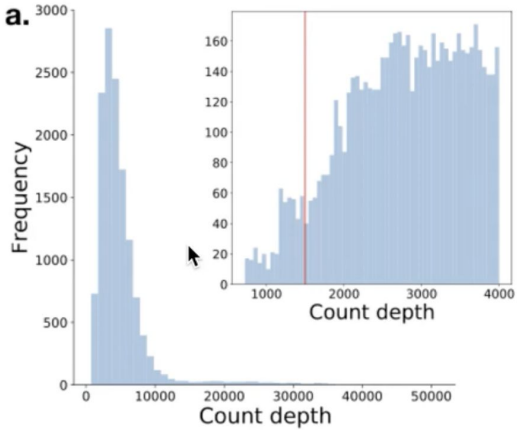
Understanding UMAP, Andy Coenen, Adam Pearce

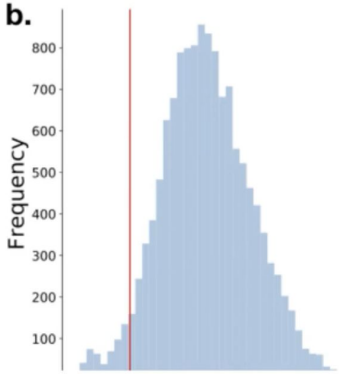
Single-cell RNA downstream analysis workflow

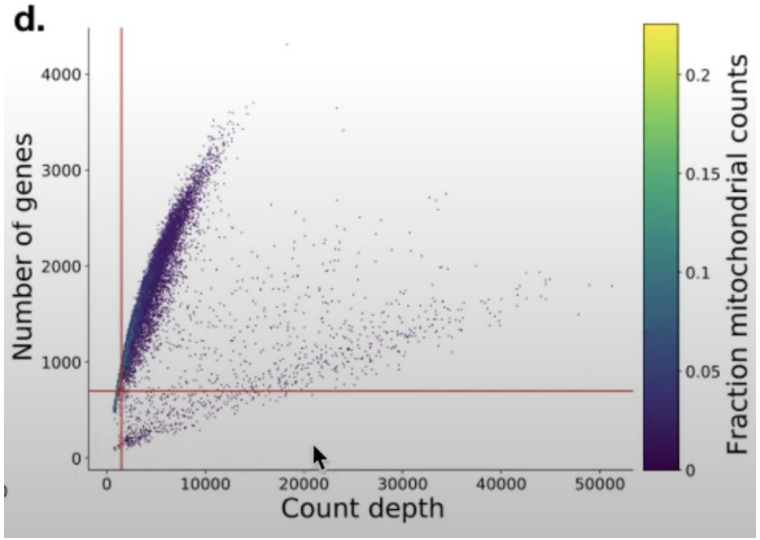


https://www.youtube.com/watch?v=HUGuHf6M0yg

Preprocessing



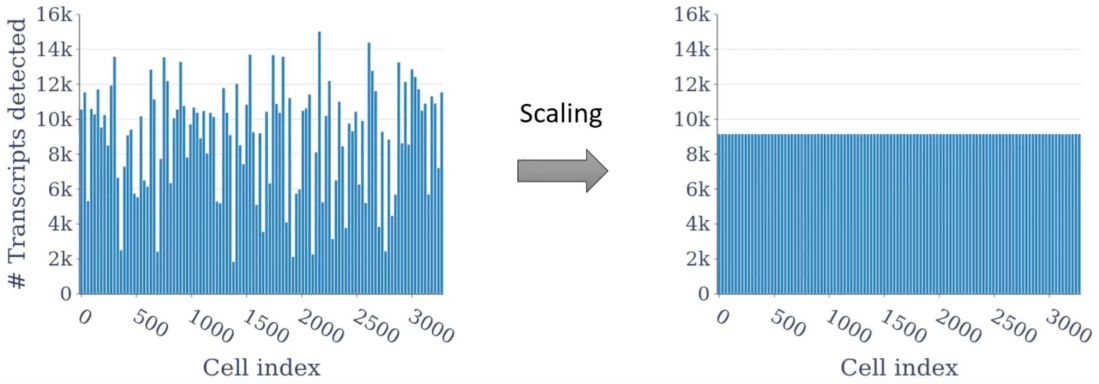




https://www.youtube.com/watch?v=HUGuHf6M0yg

Normalisation

- Gene length might affect the number of captured reads
- Scaling
- Transformation
  - Log
  - Square root
  - Pearson residual (scTeansform)



Raw data

	Cell Type A	Cell Type B	Δ
Gene 1	1	2	1
Gene 2	100	200	100

Log_2 transform

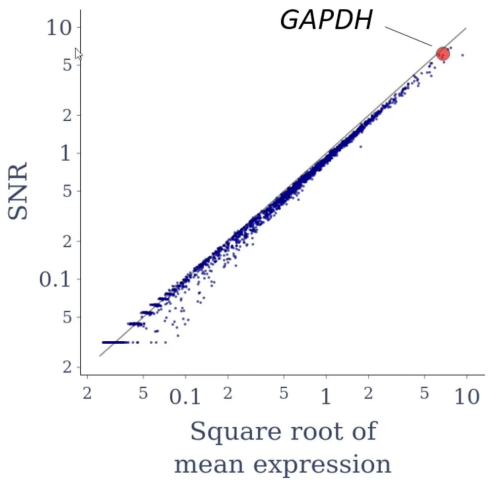
Cell Type A	Cell Type B	Δ
0	1	1
6.64	7.64	1

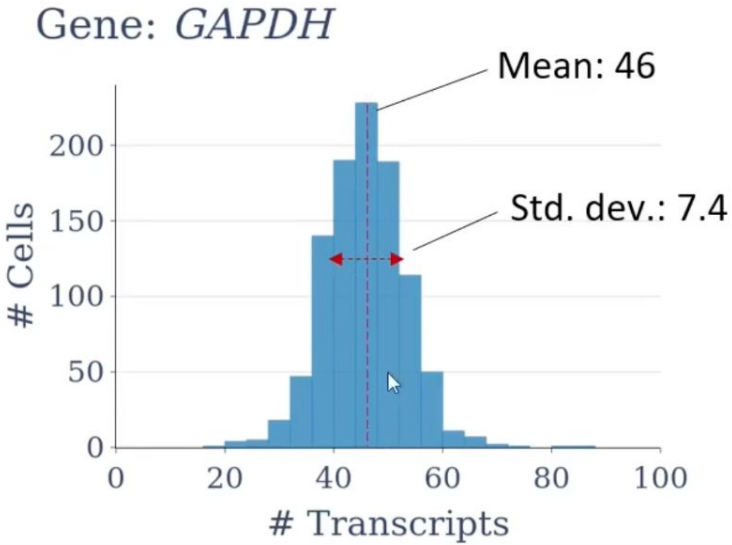
Square root transform

Cell Type A	Cell Type B	Δ
1	1.41	0.41
10	14.1	4.1

Florian Wagner - Normalization methods for single-cell RNA-Seq data

Normalisation





⇒ Let's quantify the measurement accuracy for *GAPDH* using the **signal-to-noise ratio** (SNR):

$$SNR = \frac{\mu}{\sigma} = \frac{46}{7.4} = 6.2$$

mean

standard deviation

Normalisation - Pearson residual

Pearson residuals

	Cell Type A (50%)	Cell Type B (50%)	Δ
Gene 1	0.816	1.63	0.814
Gene 2	8.16	16.3	8.14

1. Simple transformations → $y_{ij} = f(x_{ij})$
   - Log transform
   - Square root transform
2. **Pearson residuals** → $y_{ij} = w_j * x_{ij}$ $\quad w_i = \frac{1}{\sqrt{\mu_i}}$

⇒ Instead of transforming each measurement individually, Pearson residuals apply a weight to **all** measurements of a gene.

⇒ This makes it so that each gene contributes to the analysis according to **how much evidence** there is that it is non-uniformly expressed.

⇒ This favors genes that are expressed in **only a small fraction of cells**.

Raw data

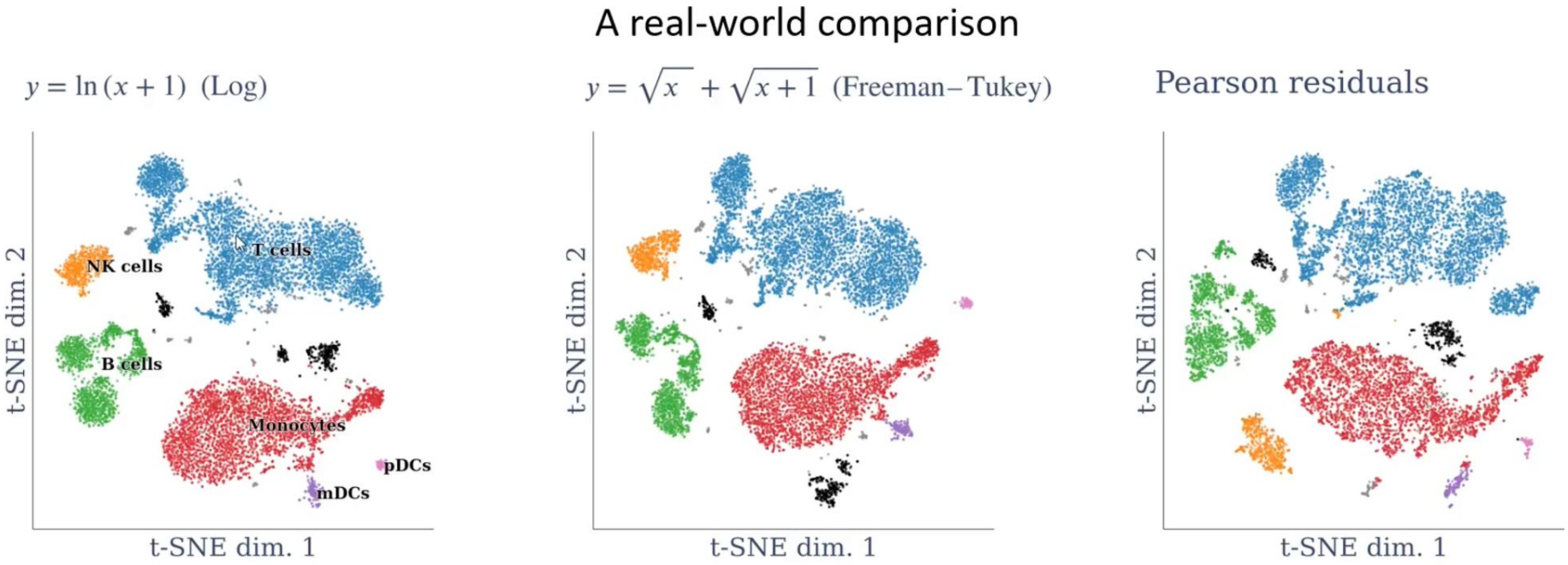
	Cell Type A (50%)	Cell Type B (50%)	
		Subtype 1 (48%)	Subtype 2 (2%)
Gene 1	0	8	8
Gene 2	0	0	4.5

Pearson residuals

	Cell Type A (50%)	Cell Type B (50%)	
		Subtype 1 (48%)	Subtype 2 (2%)
Gene 1	0	4	4
Gene 2	0	0	15

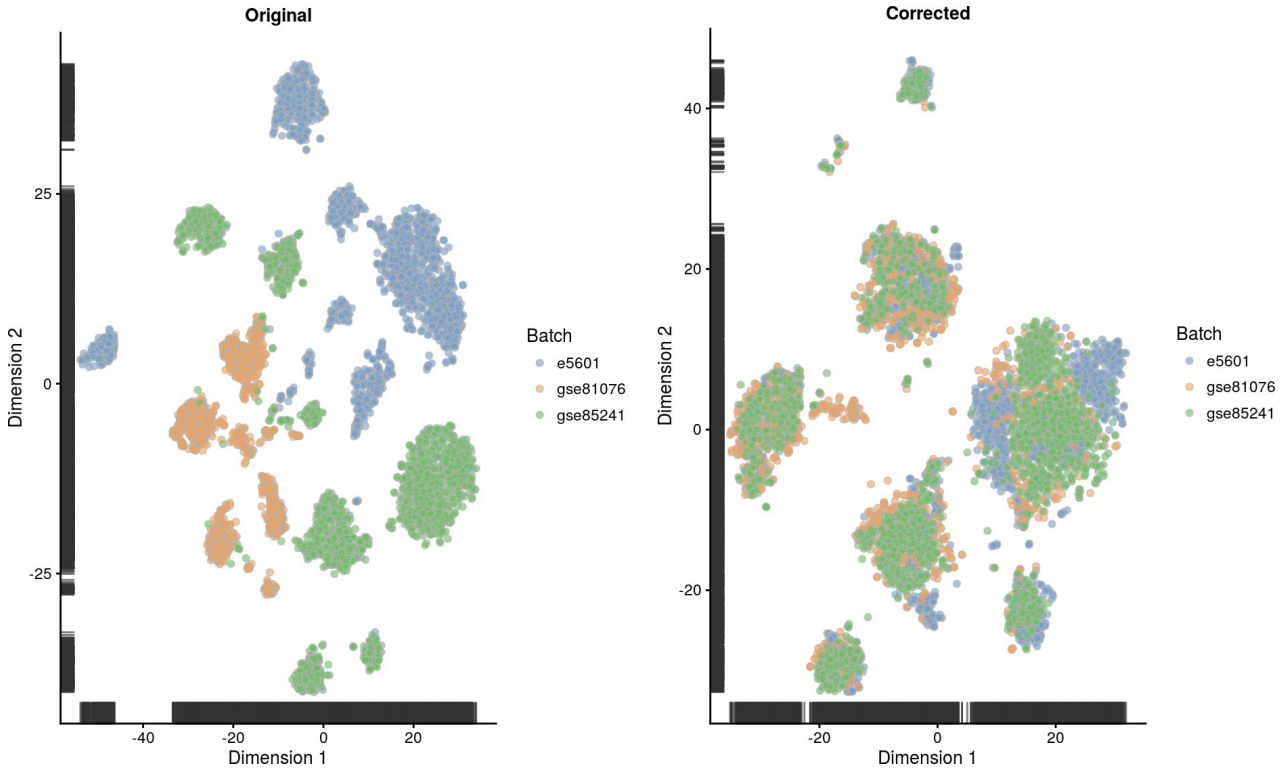
Florian Wagner - Normalization methods for single-cell RNA-Seq data

Normalisation - Log, square root and Person residual



Florian Wagner - Normalization methods for single-cell RNA-Seq data

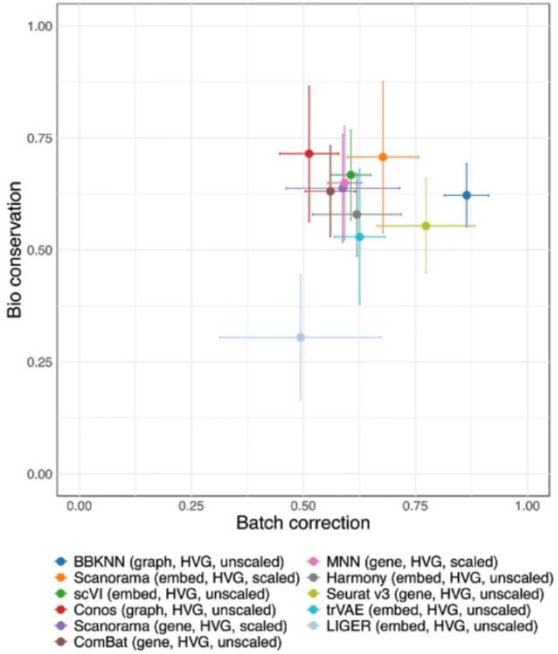
Batch effects correction

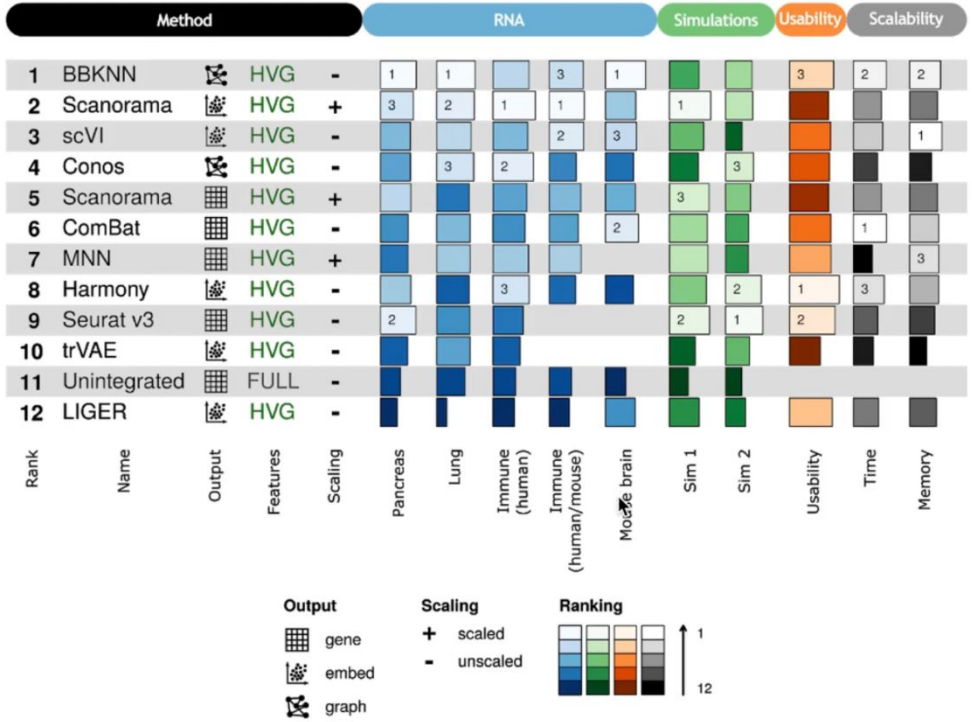


t-SNE plots of the pancreas datasets, before and after MNN correction
Each point represents a cell and is coloured by the batch of origin.

Correcting batch effects in single-cell RNA-seq data, Aaron T. L. Lun1 and Michael D. Morgan

Batch effects correction

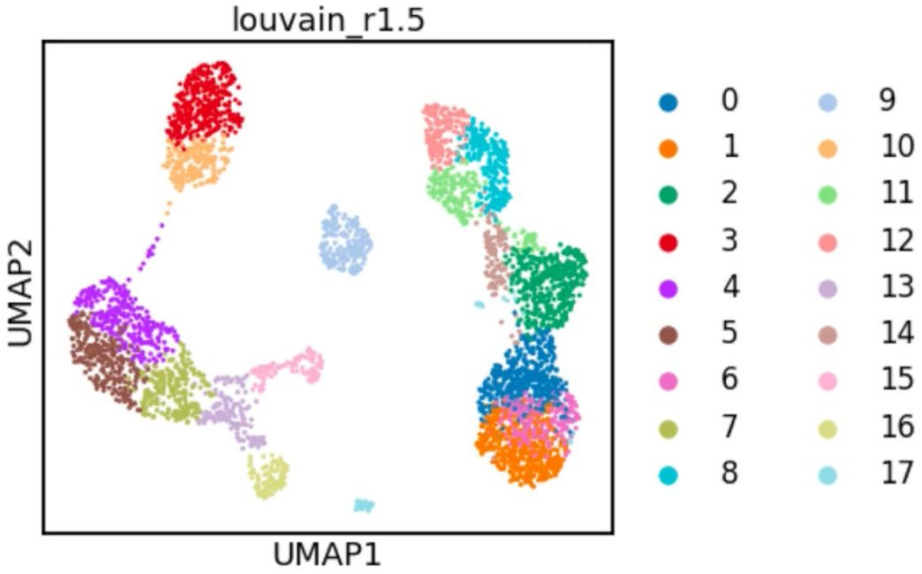


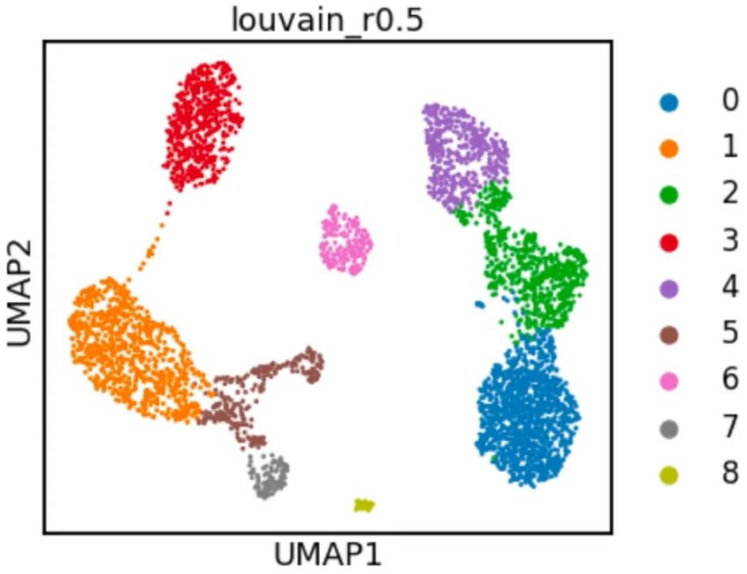


Luecken et al., biorxiv (2020)

Clustering

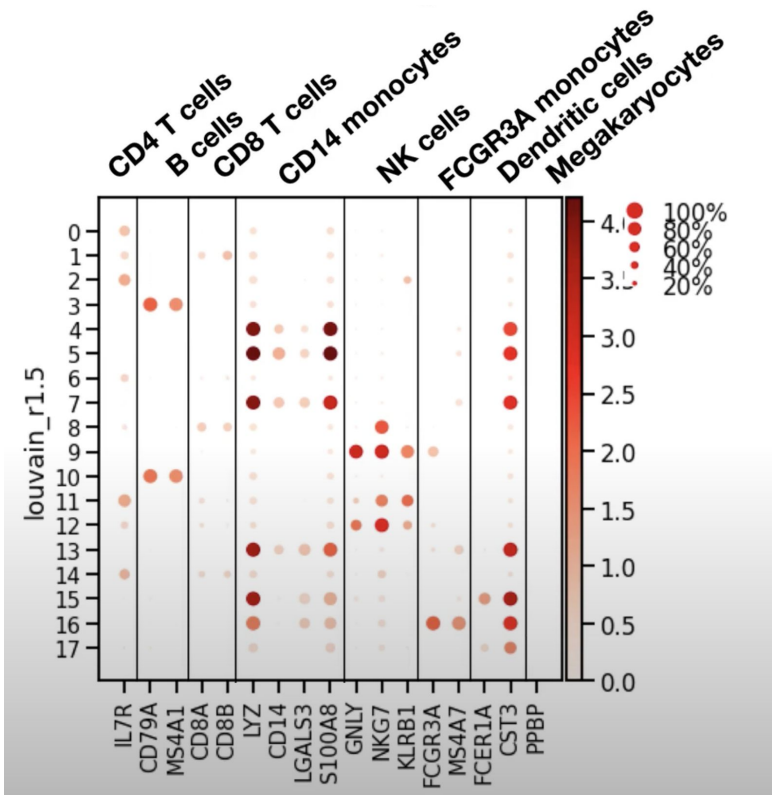
Louvain clustering





Clustering and cell type annotation

- Marker genes - genes with statistically significant variation in the specific cluster comparing to the rest of the cells



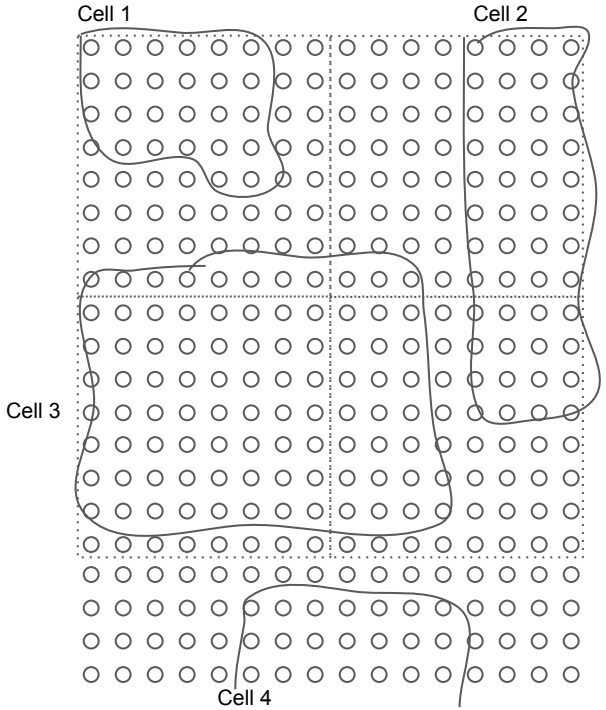
Scanpy library

- Scalable toolkit for analyzing single-cell gene expression data
- Annotated data object
- Tutorial: Preprocessing and clustering 3k PBMCs

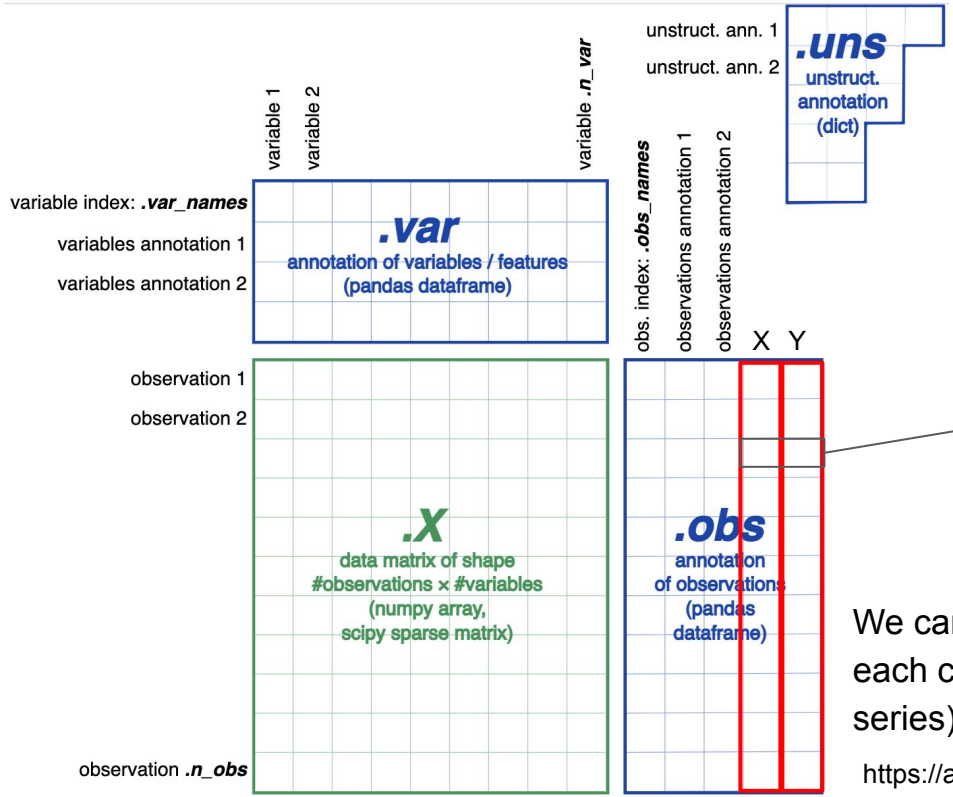
SCANPY: large-scale single-cell gene expression data analysis

Spatial transcriptomics

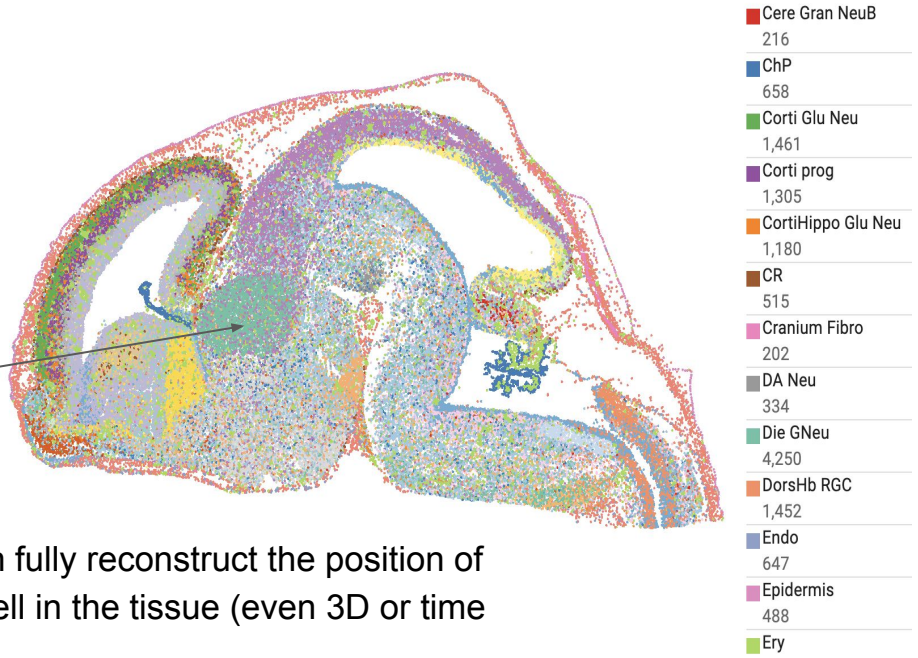
- What if we obtain for every cell beside expression its spatial coordinates
- Each spot is 250 nm in diameter and the center-to-center distance between neighboring spots is 500 nm



Annotated data object



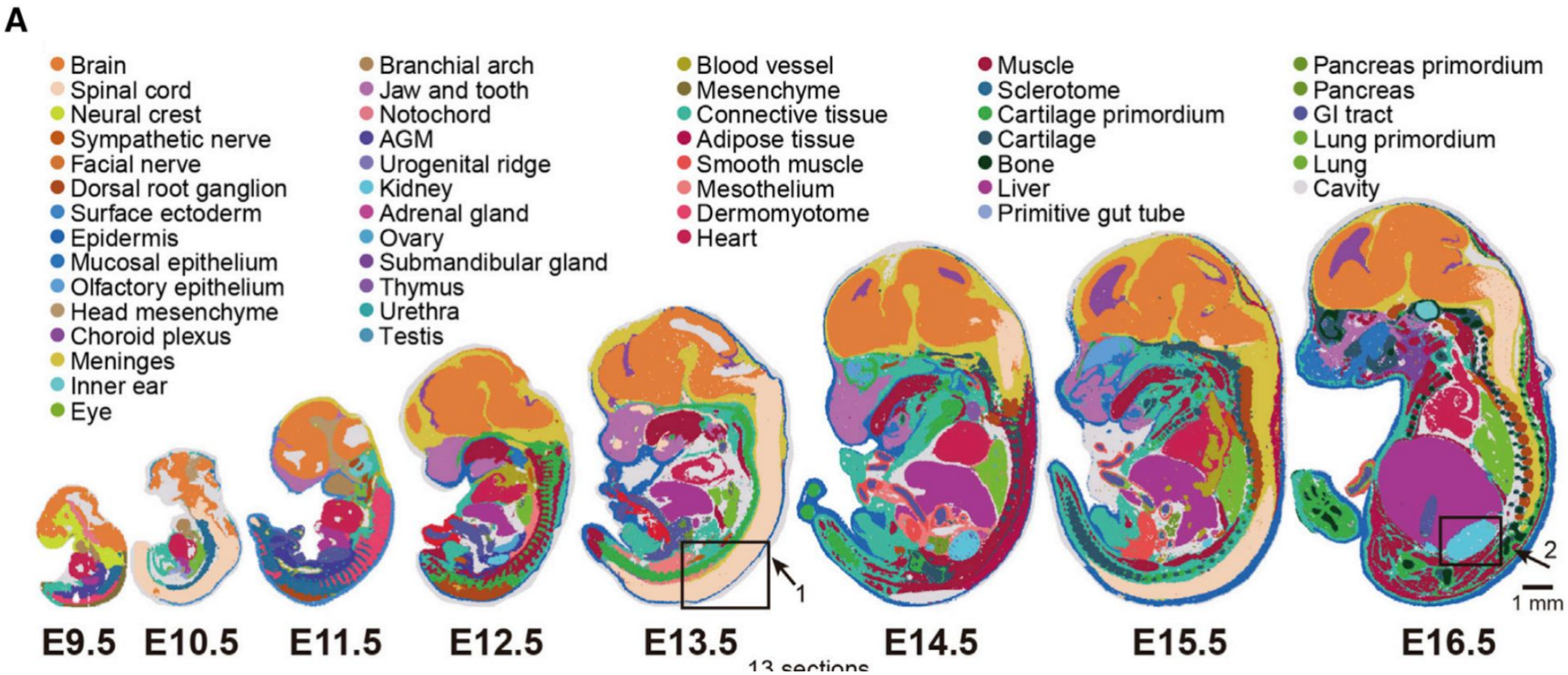
Mouse embryo brain, day 16



We can fully reconstruct the position of each cell in the tissue (even 3D or time series)

https://anndata-tutorials.readthedocs.io/en/latest/getting-started.html

Stereo-Seq: Spatiotemporal transcriptomic atlas of mouse organogenesis



Source paper. Mosta database

Spatial transcriptomics data analysis

- Stereopy library
- Quick start tutorial