Genome Informatics 2024

Lesson 6 - Single cell RNA sequencing

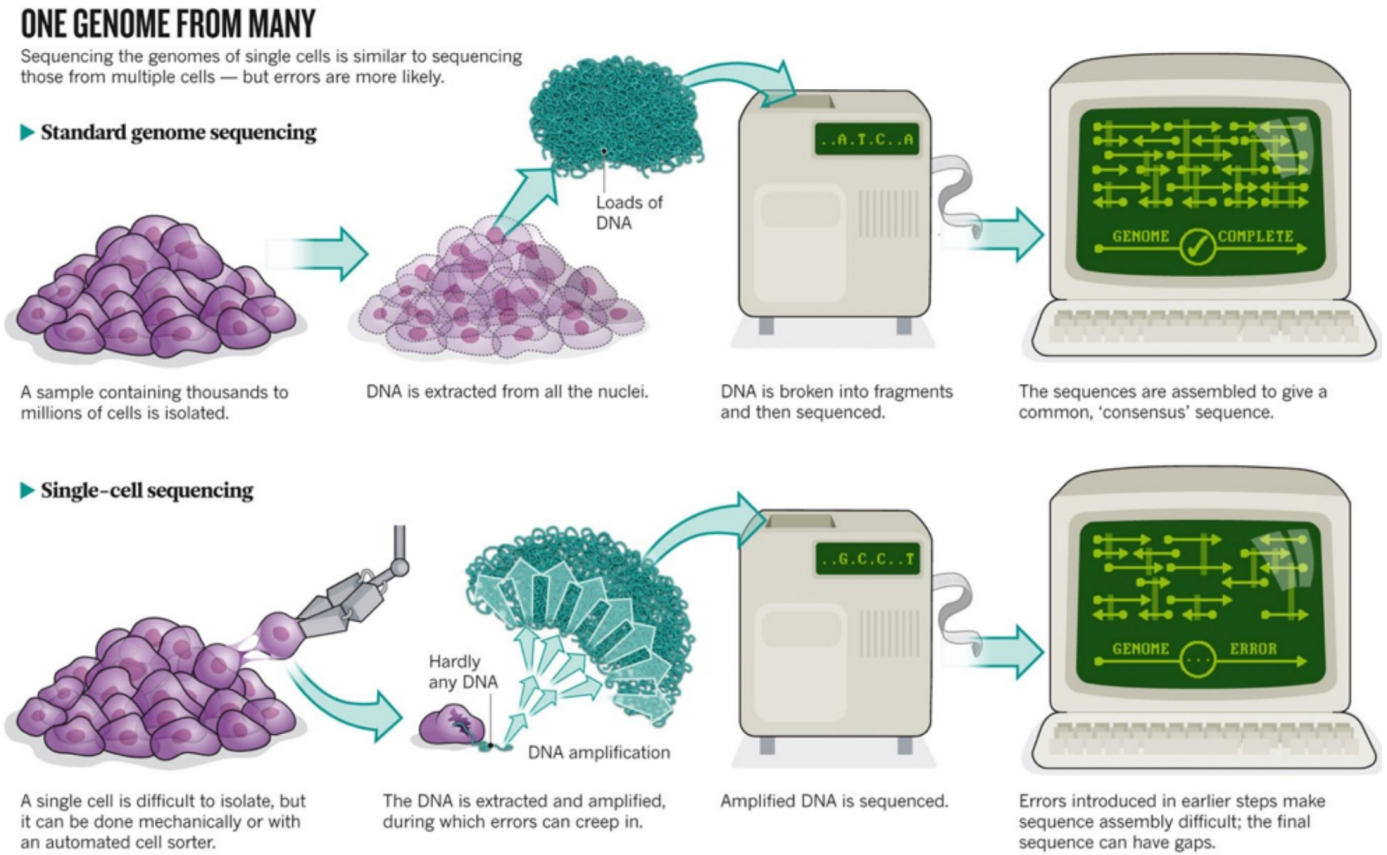
Why Single Cell study?

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

Why Single Cell study?

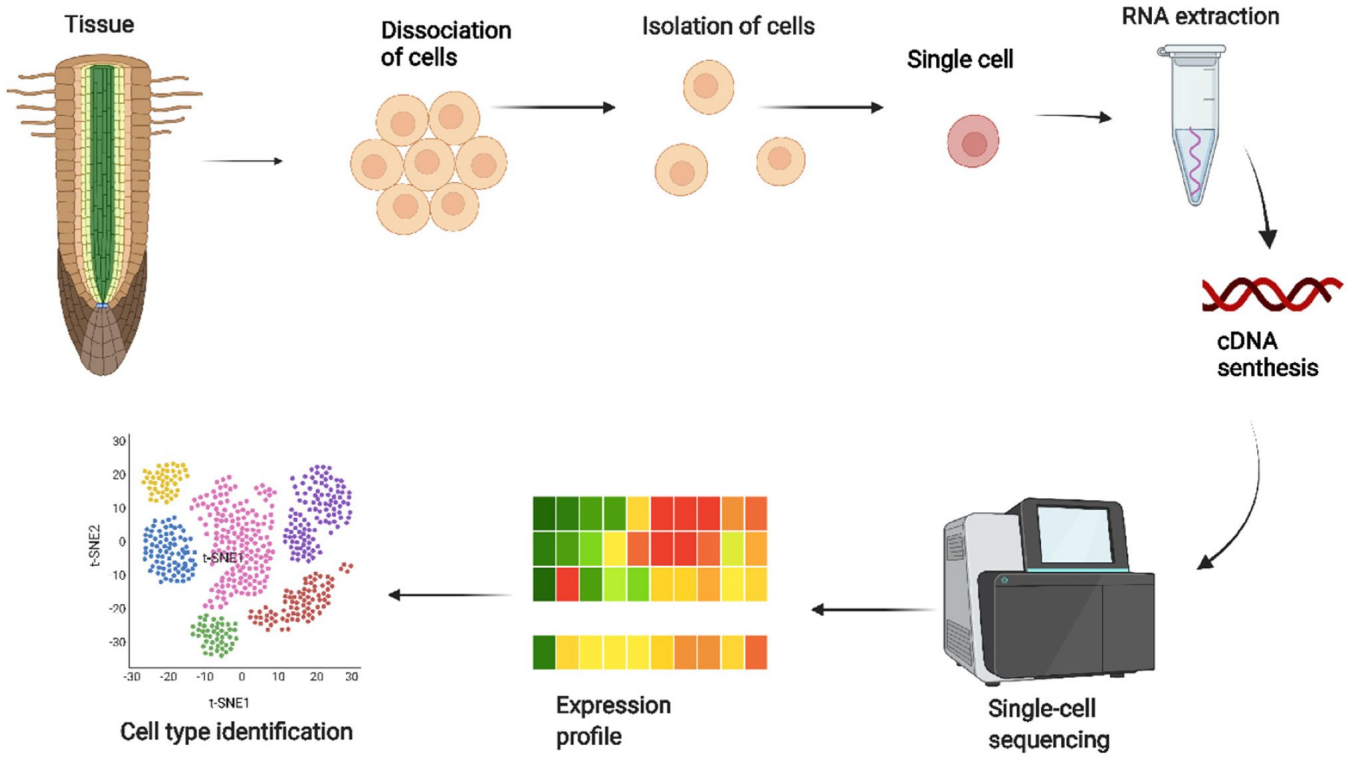
- Developmental biology
  - Discover more complicated mechanisms in cellular development
  - Confirm the distinct gene expression signatures across different cell types
  - Identify functional differences among the same cell cell type
- Cancer biology
  - Find evidence for models of cancer
  - Infer 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

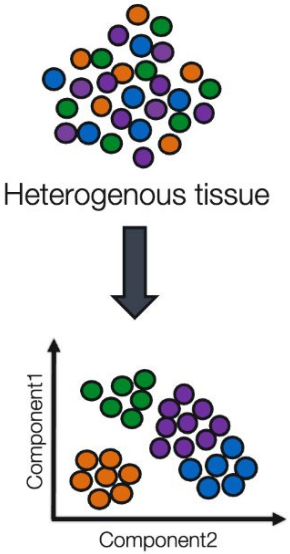
Single cell sequencing



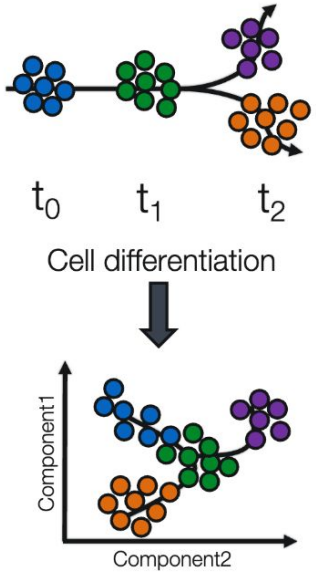
Single-Cell RNA Sequencing for Plant Research: Insights and Possible Benefits by George Bawa

Single cell sequencing applications

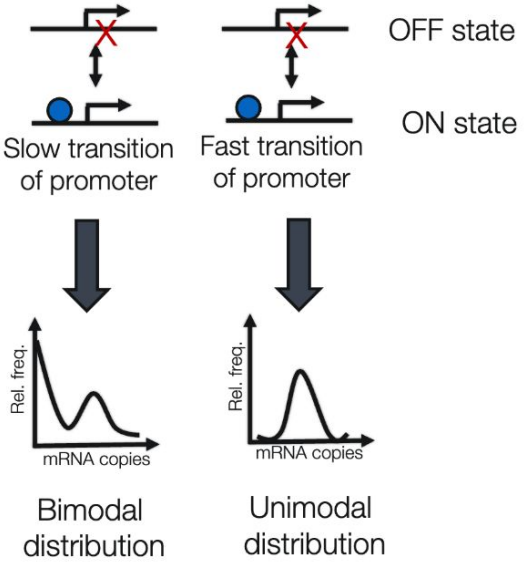
Studying heterogeneity



Lineage tracing study



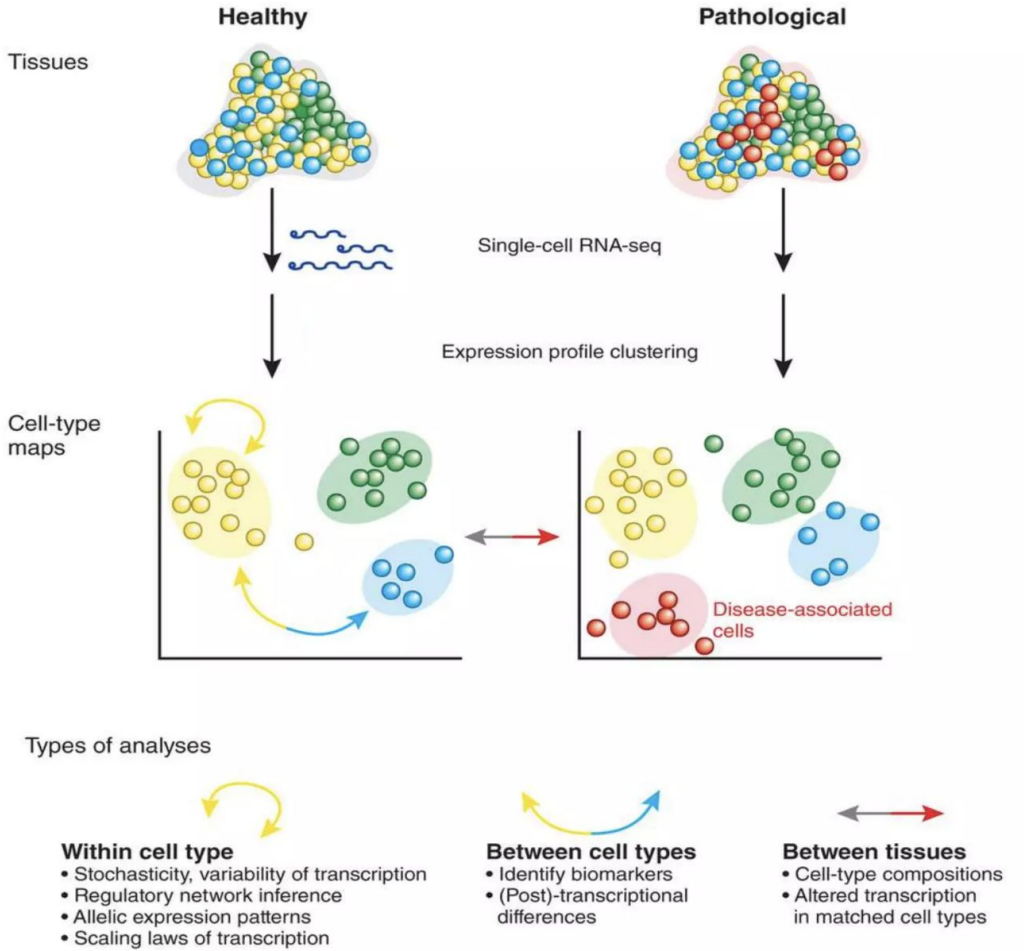
Stochastic gene expression



Liu S and Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges, F1000 Research 2016 (doi: 10.12688/f1000research.7223.1)
Junker and van Oudenaarden; Every Cell Is Special: Genome-wide Studies Add a New Dimension to Single-Cell Biology, Cell 2014 (doi: 10.1016/j.cell.2014.02.010)

Single cell sequencing applications

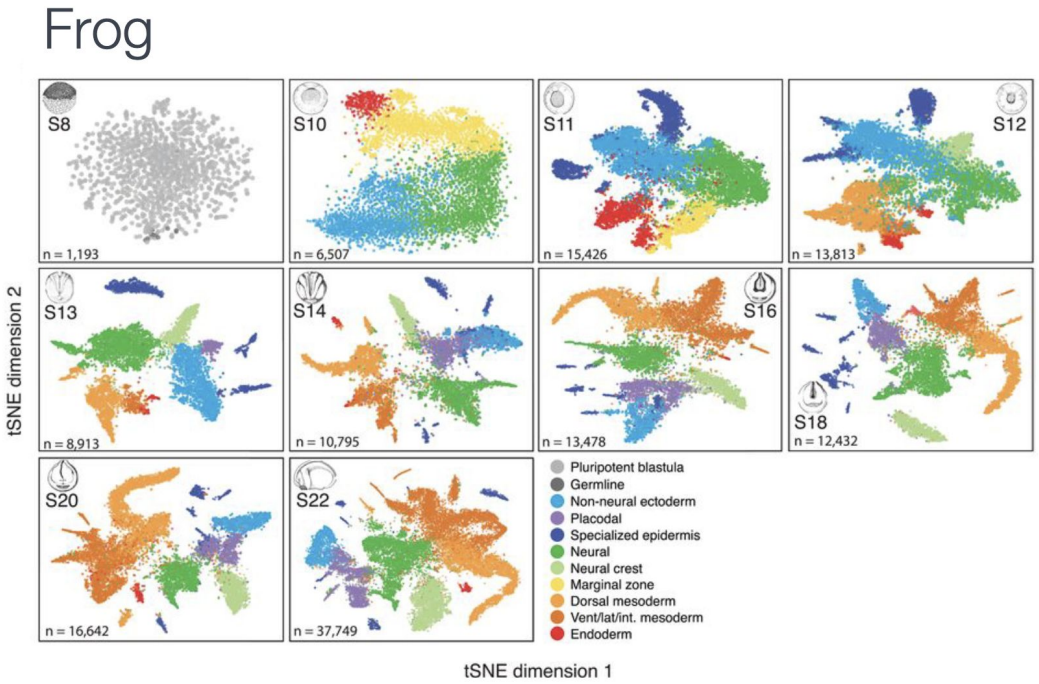
- Non-invasive way to monitor the progress of the disease
- Monitor biological process
- Knockout gene studies



Single cell RNA sequencing; Methods and applications, Bushra Arif, University of Panjab

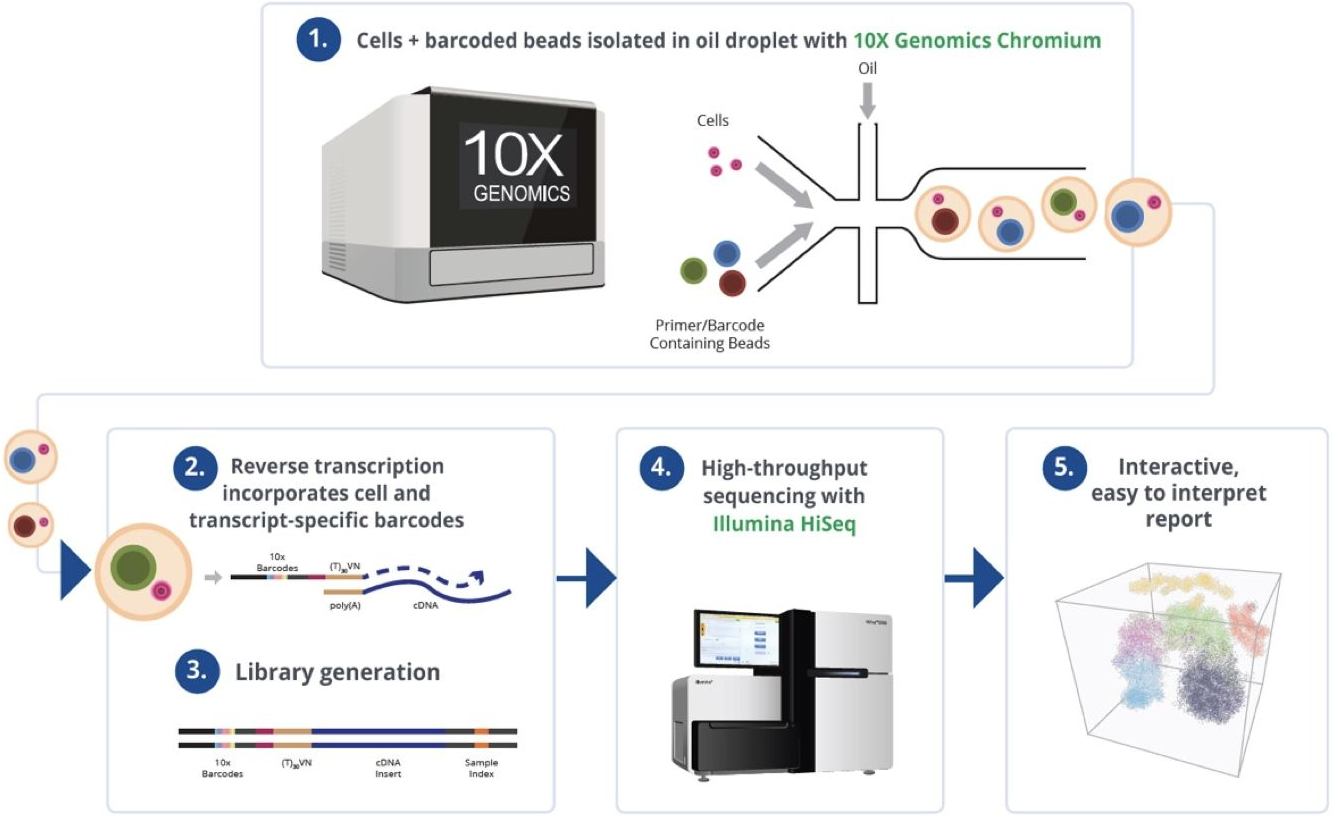
Single cell sequencing applications

- Development Lineage Tracing

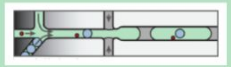


JA. Briggs et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution,

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

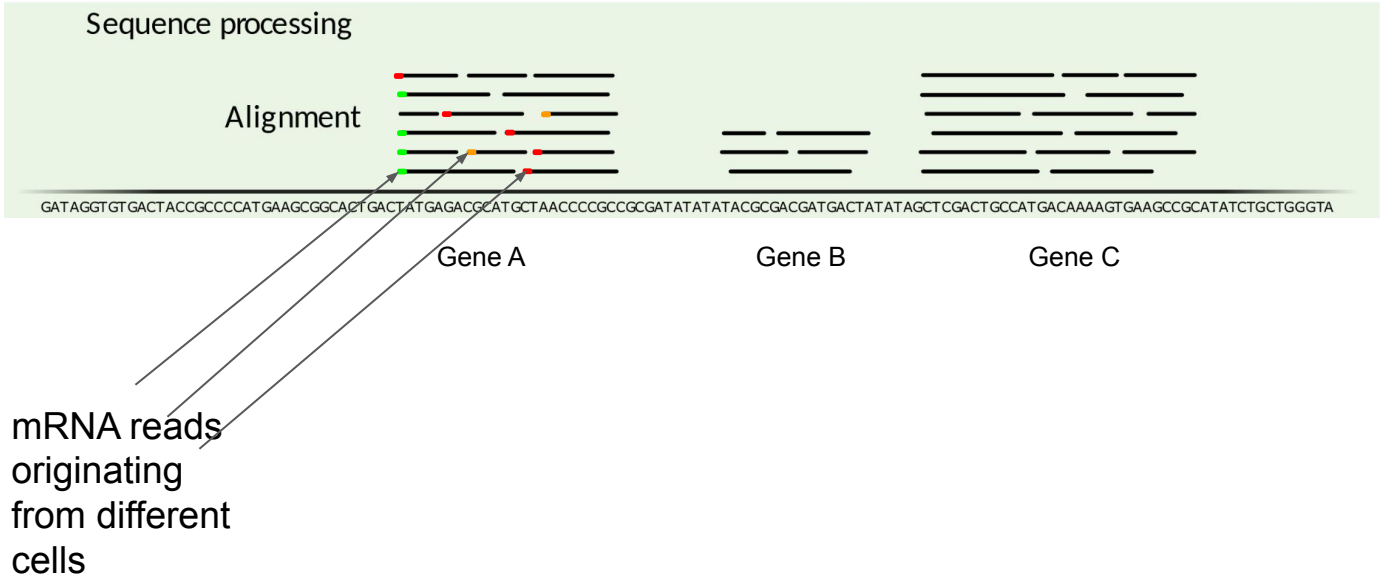


Comparison of single cell sequencing methods

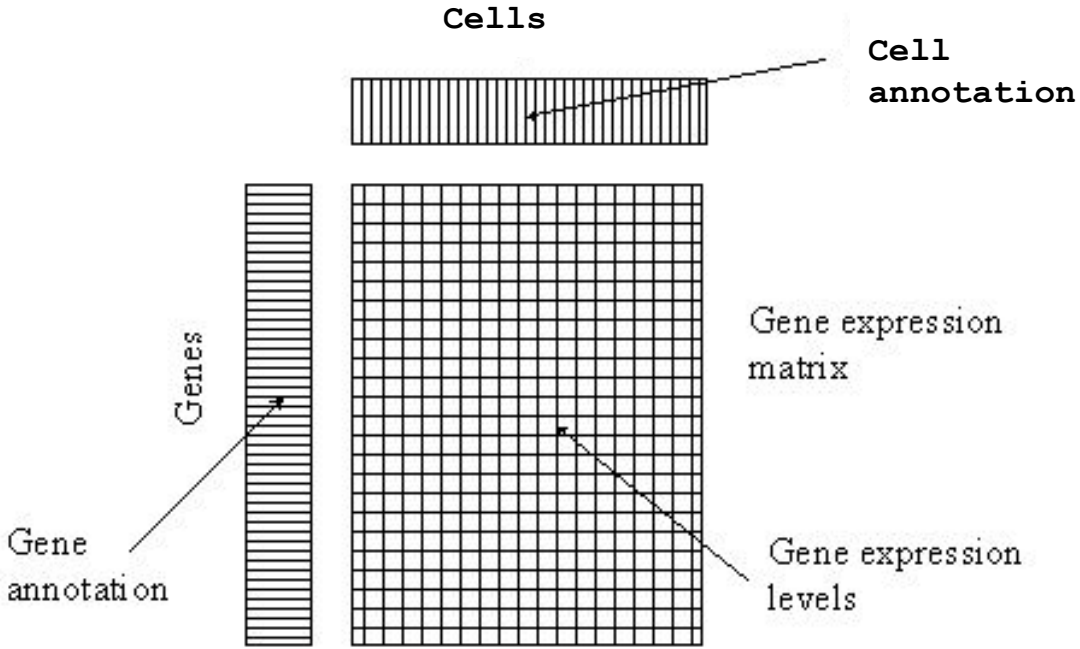
	inDrops	10x	Drop-seq	Seq-well	SMART-seq
Cell capture efficiency	~70-80%	~50-65%	~10%	~80%	~80%
Time to capture 10k cells	~30min	10min	1-2 hours	5-10min	--
Encapsulation type	Droplet 	Droplet 	Droplet 	Nanolitre well 	Plate-based 
Library prep	CEL-seq Linear amplification by IVT	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification
Commercial	Yes	Yes	--	--	Yes
Cost (~$ per cell)	~0.06	~0.2	~0.06	--	1
Strengths	• Good cell capture • Cost-effective • Real-time monitoring • Customizable	• Good cell capture • Fast and easy to run • Parallel sample collection • High gene / cell counts	• Cost-effective • Customizable	• Good cell capture • Cost-effective • Real-time monitoring • Customizable	• Good cell capture • Good mRNA capture • Full-length transcript • No UMI
Weaknesses	Difficult to run	Expensive	Difficult to run & low cell capture efficiency	Still new!	Expensive

C. Ziegenhain et al., Comparative Analysis of Single-Cell RNA Sequencing Methods, Molecular Cell 2017 (doi: 10.1016/j.molcel.2017.01.023)

Single cell sequencing alignment and gene count

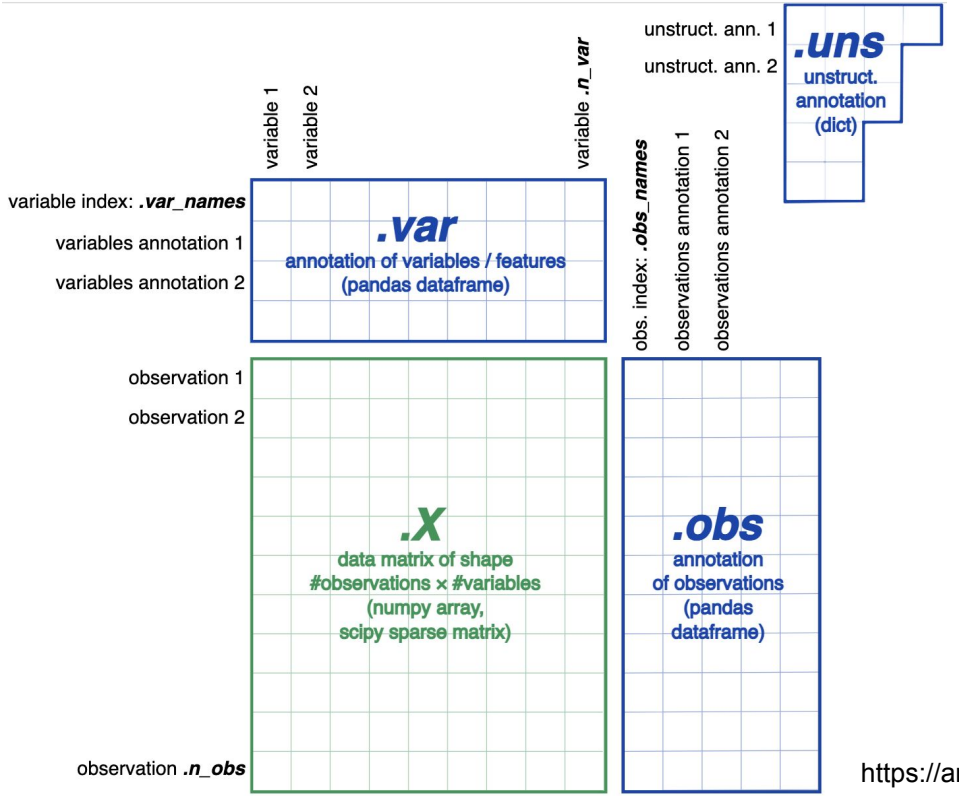


Cell-gene matrix



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

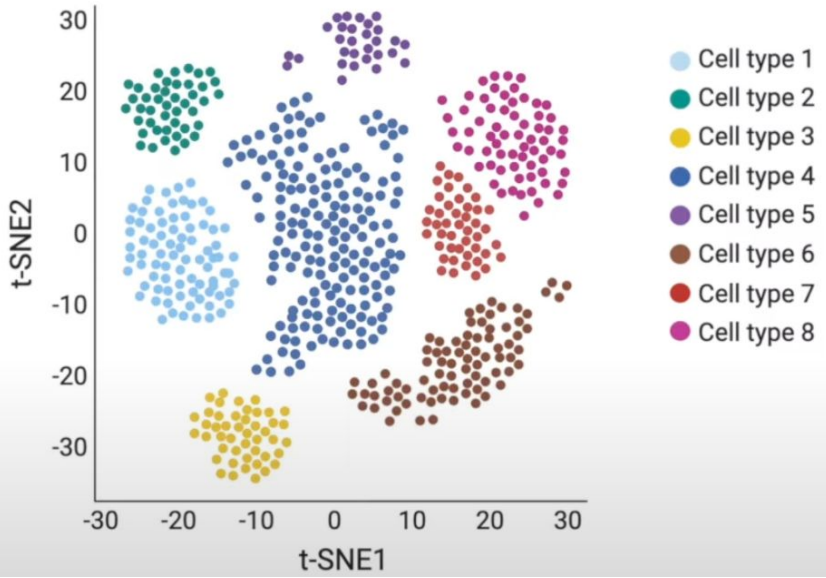
Annotated data object



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

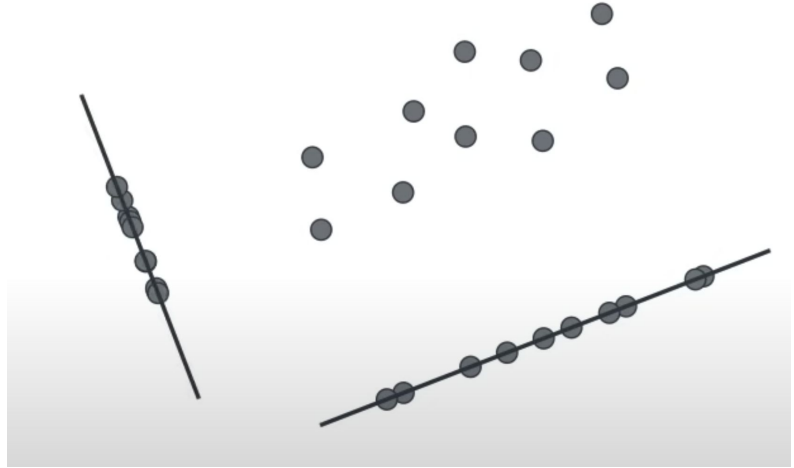
Cell-gene matrix

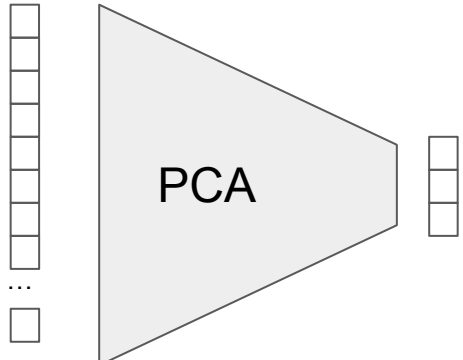
	Cell 1	cell 2	cell 3	...	cell X
Gene a	3	5	6	...	3
gene b	3	5	3	...	2
gene c	5	6	5	...	4
gene d	5	6	7	...	8
...	...	...	...	...	
gene z	7	8	4	...	3



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

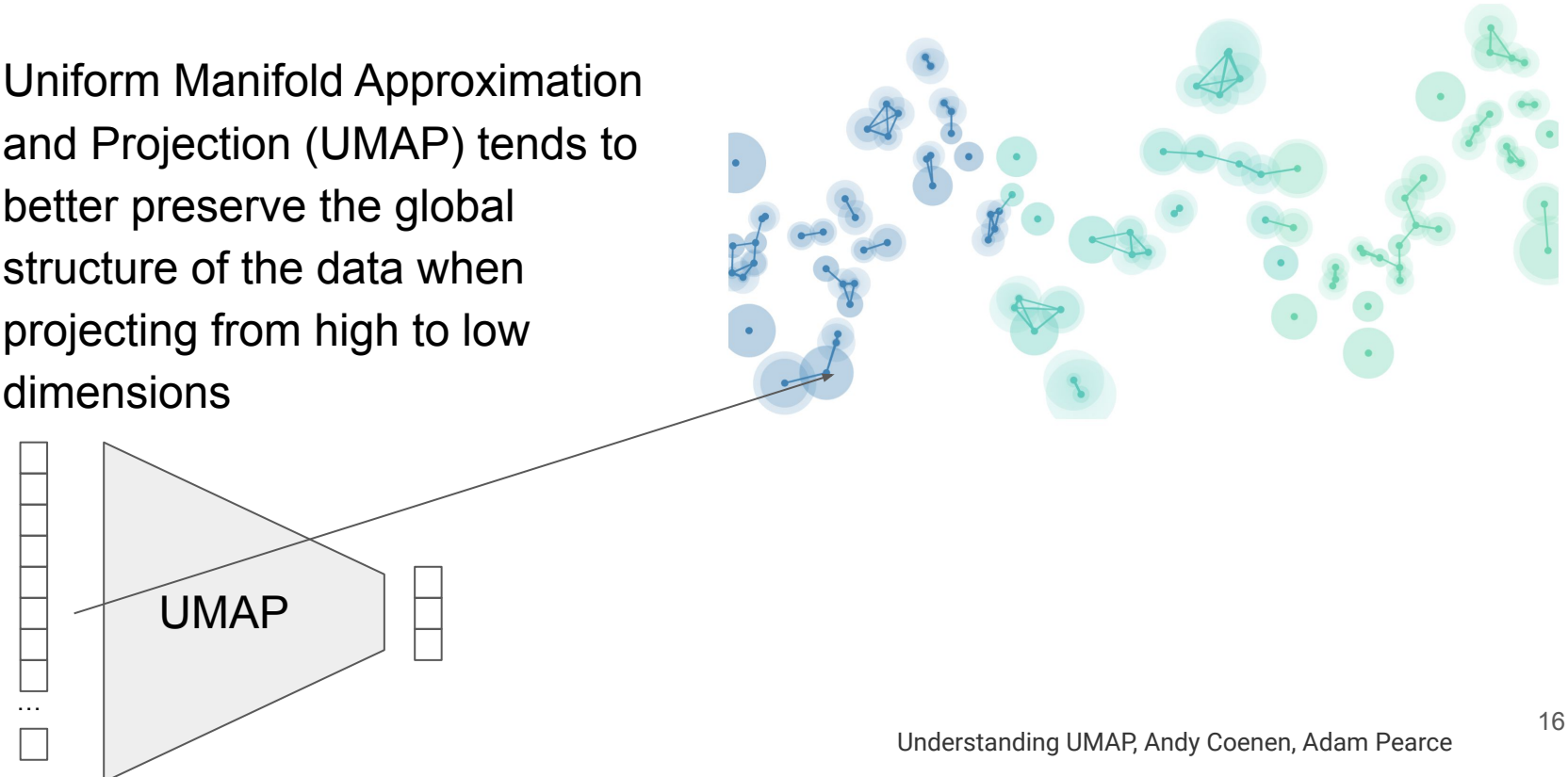




Luis Serrano - Principal component analysis

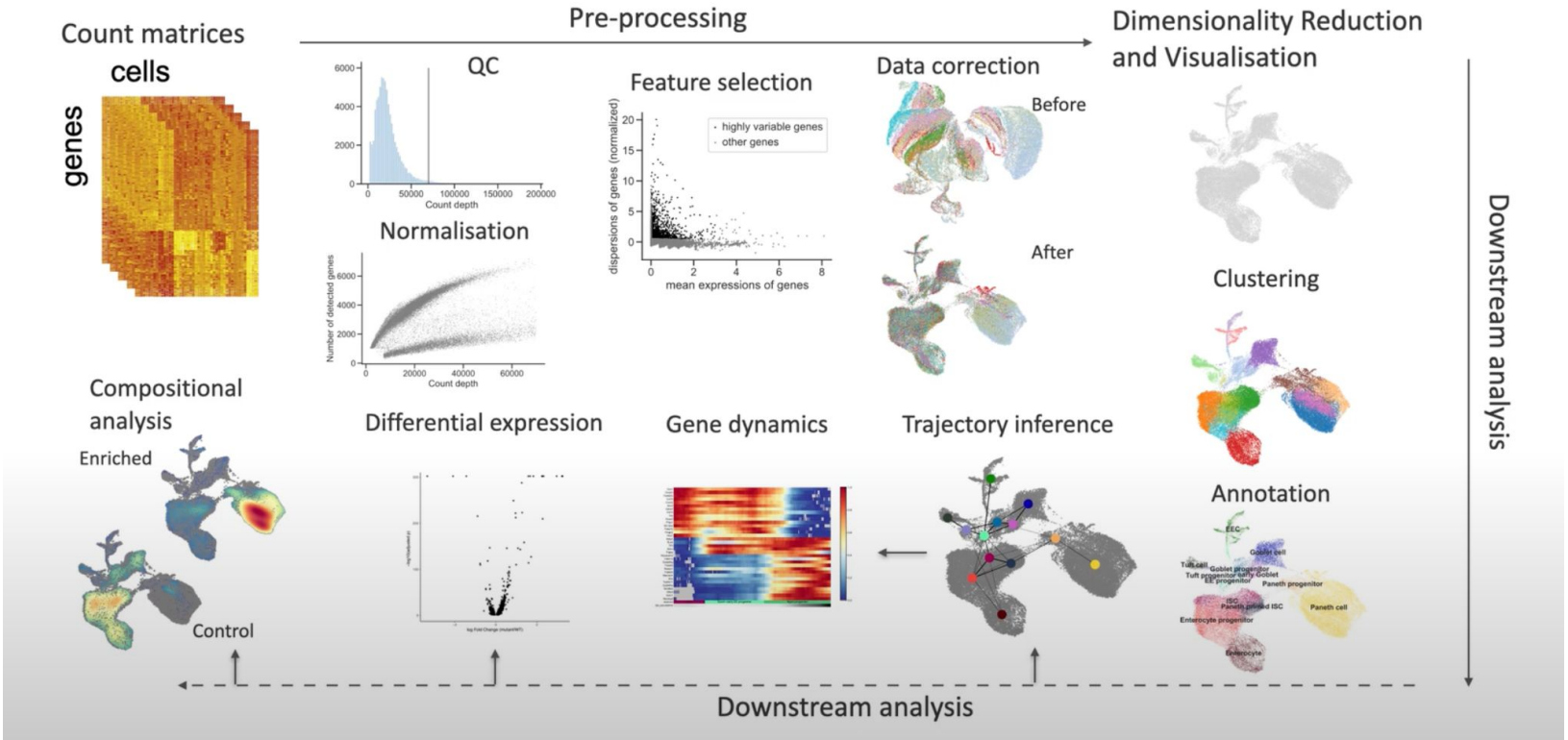
Latent (low-dimensional) representation of data

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



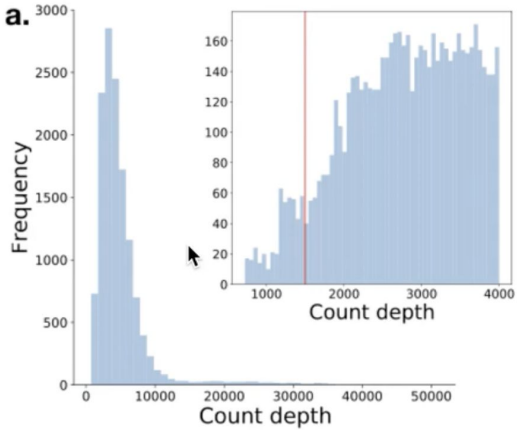
Understanding UMAP, Andy Coenen, Adam Pearce

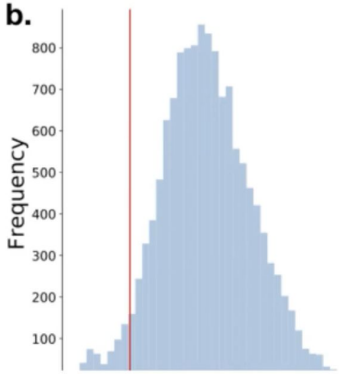
Single-cell RNA downstream analysis workflow

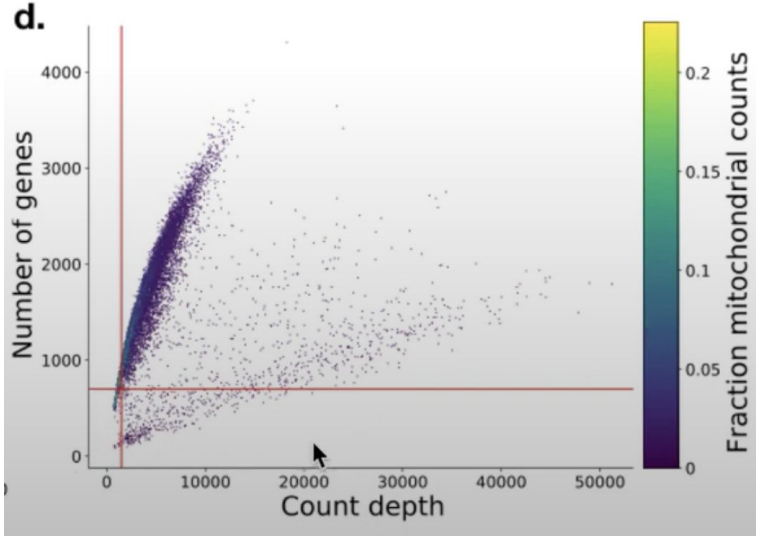


https://www.youtube.com/watch?v=HUGuHf6M0yg

Preprocessing



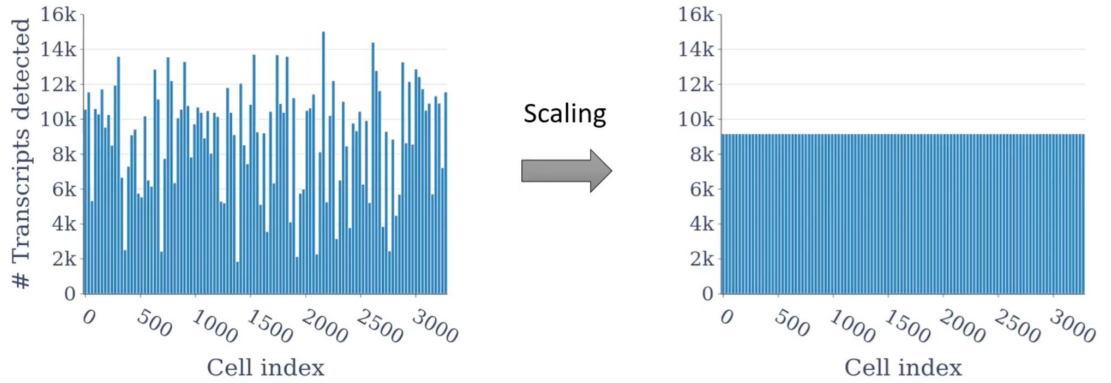




https://www.youtube.com/watch?v=HUGuHf6M0yg

Normalisation

- Gene length might affect the number of captured reads
- Normalize each cell by total counts over all genes, so that every cell has the same total count after normalization.
- Scaling transformation
  - Log
  - Square root
  - Pearson residual (scTransform)



Raw data

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

Log_2 transform

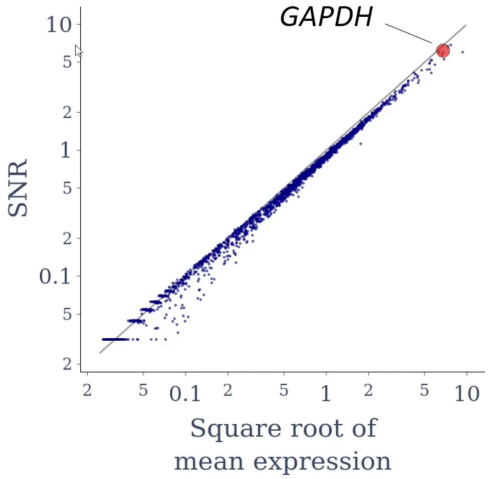
Cell Type A	Cell Type B	Δ
0	1	1
6.64	7.64	1

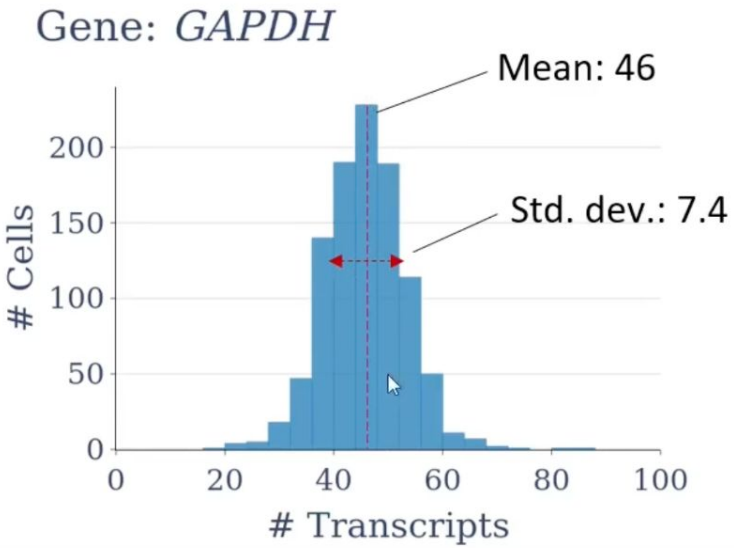
Square root transform

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):

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

(μ: mean; σ: 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**.

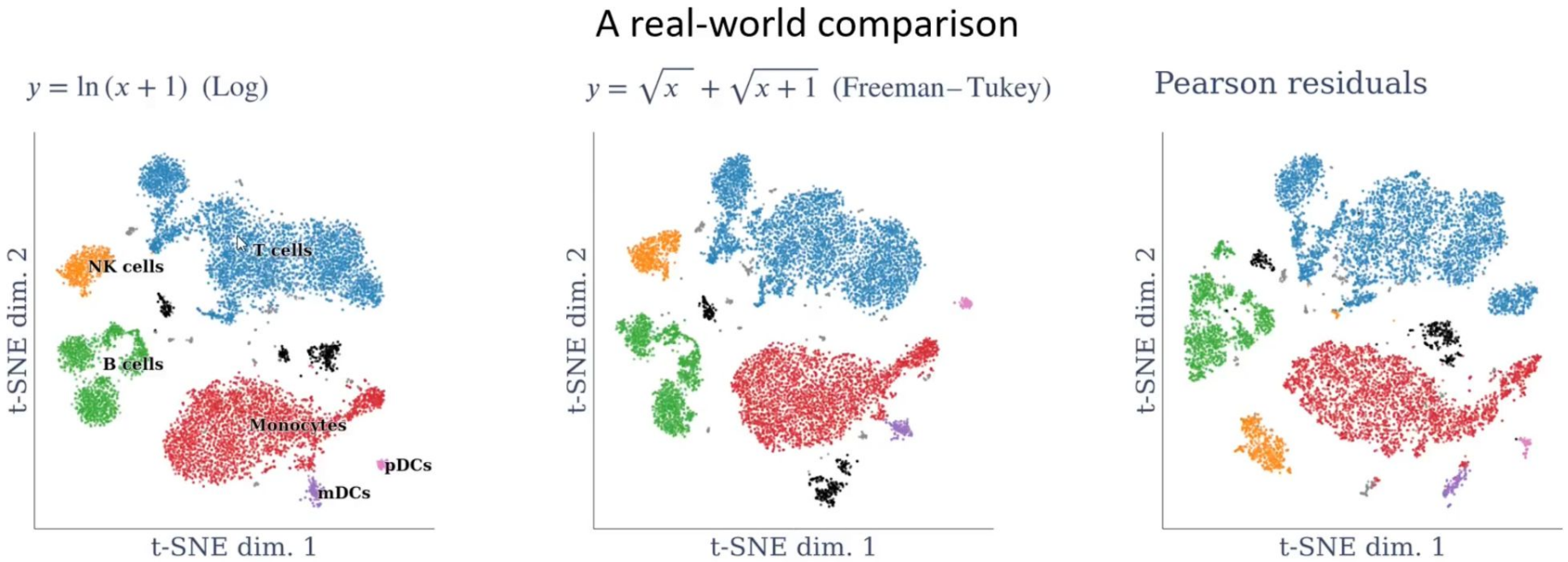
Raw data

	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

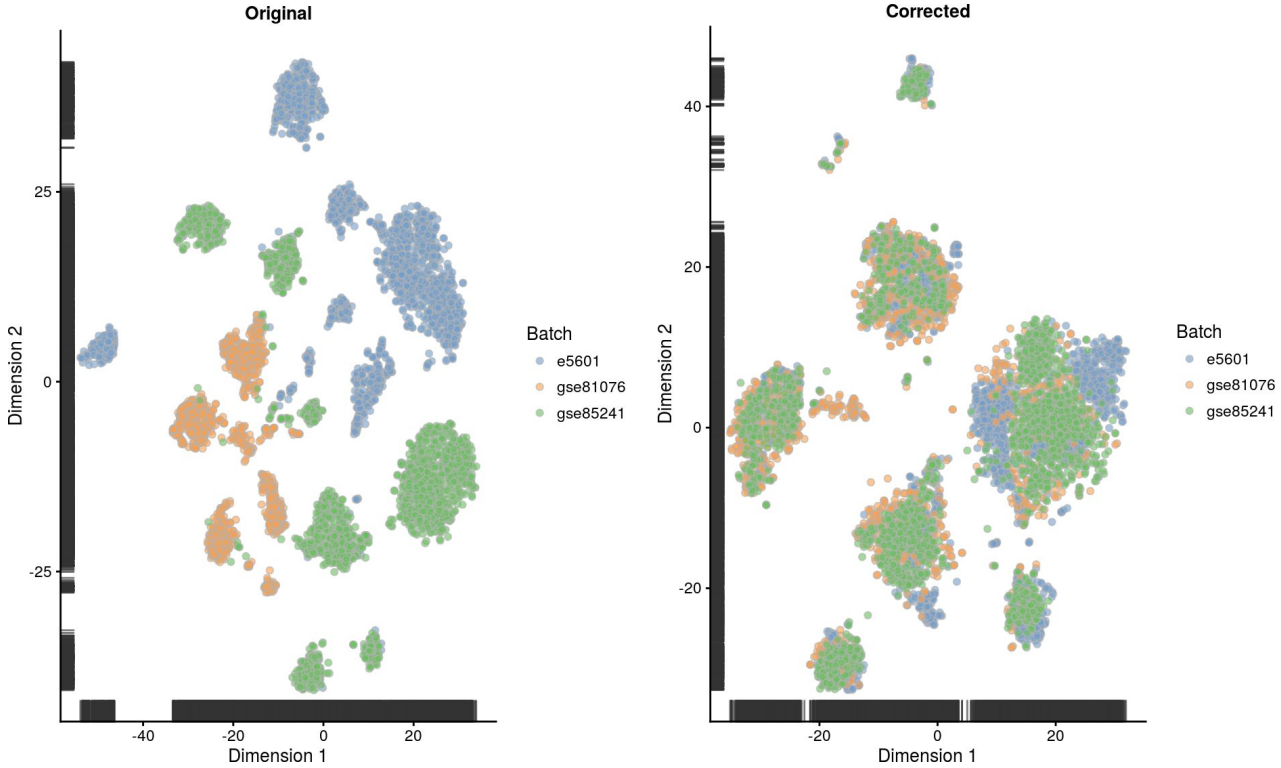
	Cell Type A (50%)	Cell Type B (50%)	
		Subtype 1 (48%)	Subtype 2 (2%)
	0	4	4
	0	0	15

Normalisation - Log, square root and Person residual



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

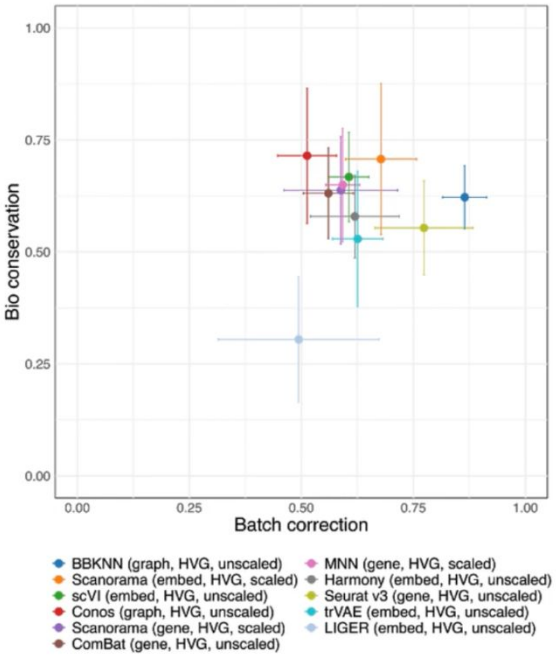
Batch effects correction

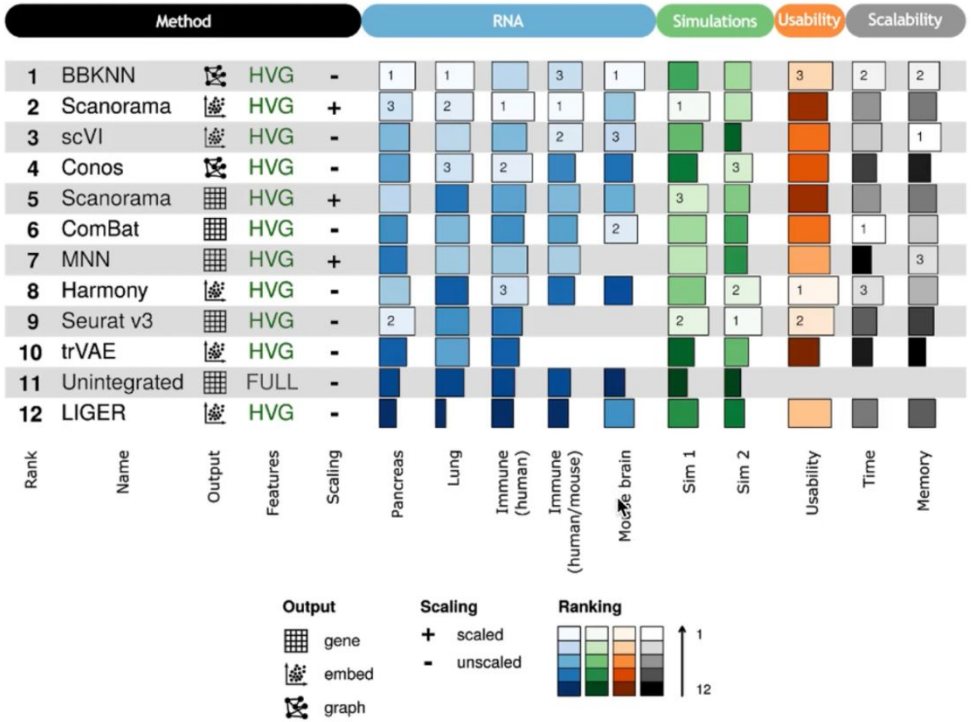


t-SNE plots of the pancreas datasets, before and after MNN correction
Each point represents a cell and is coloured by the batch of origin.

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

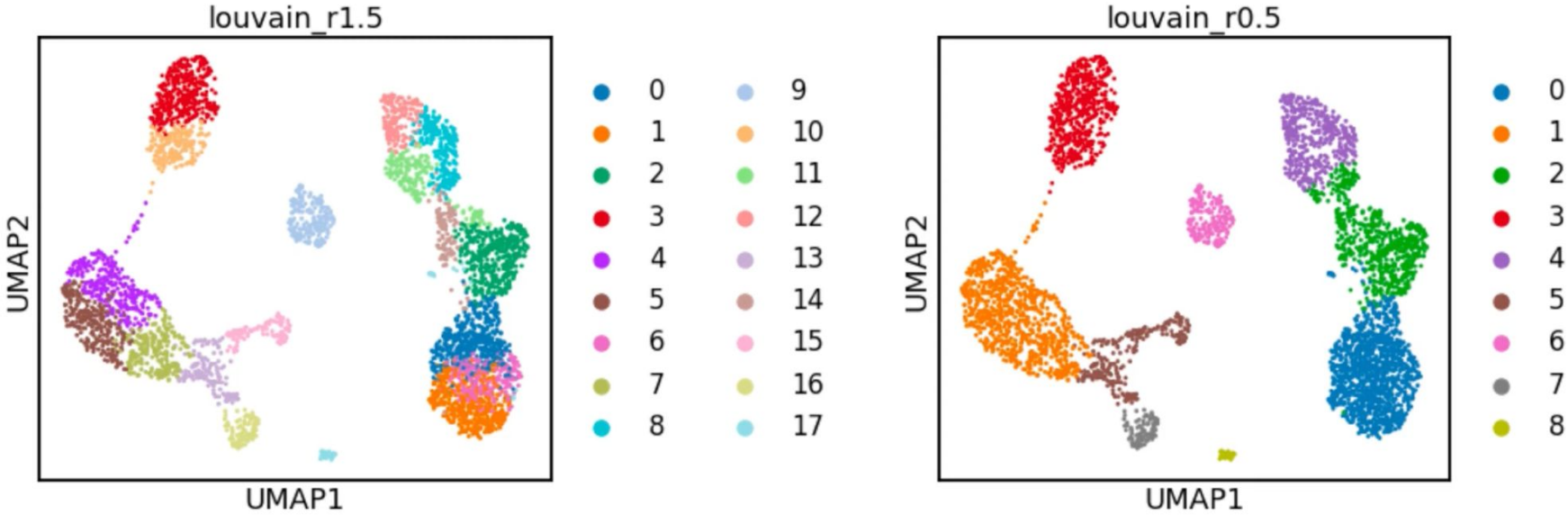
Batch effects correction





Luecken et al., biorxiv (2020)

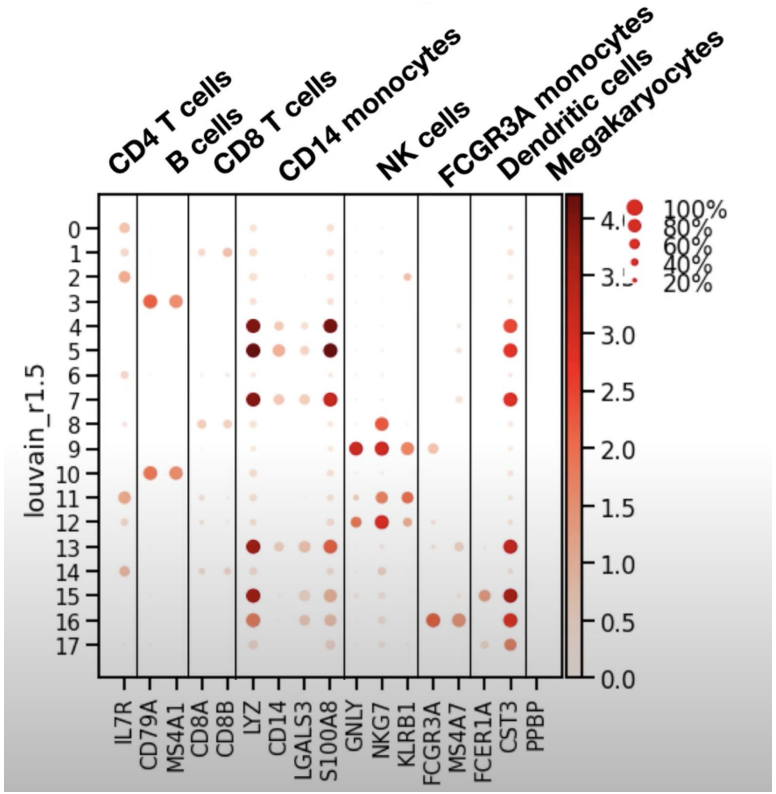
Clustering - Leiden and Louvain



- The algorithm separates nodes into disjoint communities so as to maximize a modularity score for each community
- “From Louvain to Leiden: guaranteeing well-connected communities”

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