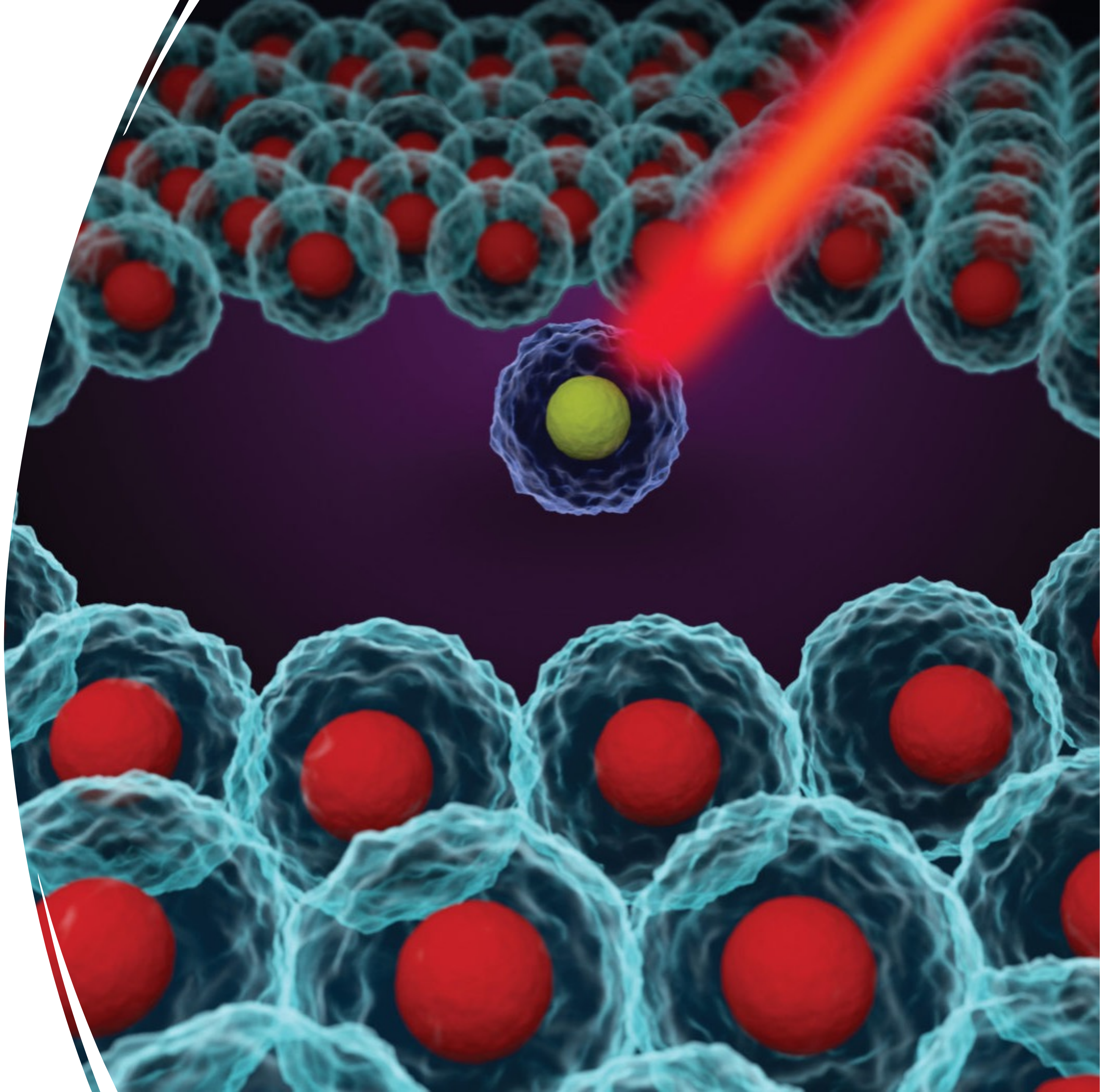
Exploring the world of Single Cell Technology

Single Cell Analysis Boot Camp

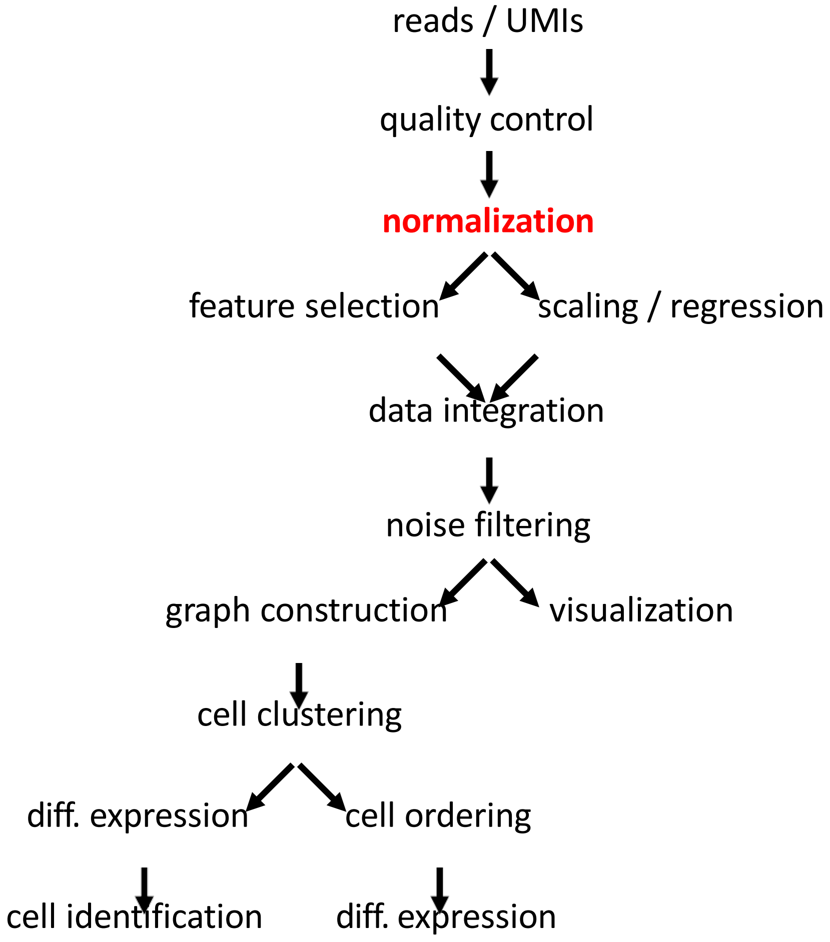
Day 2

September 2023



Normalization

scRNA-seq analysis workflow



reads / UMIs

quality control

normalization

feature selection

scaling / regression

data integration

noise filtering

graph construction

visualization

cell clustering

diff. expression

cell ordering

cell identification

diff. expression

scRNA-seq normalization

Count normalization (UMI and read counts)
for uneven sequencing depth

- CPM - log[CP10K+1]

Gene length normalization (read counts)
for differences in gene detection due to gene length

- TPM (closer to UMI counts)
- FPKM

Drop-out rate normalization (UMI and read counts)
for differences in RNA content / drop-out rates

- Deconvolution/Scran(Pooling-Across-Cells)
- SCnorm(Expression-DepthRelation)
- SCTransform
- Census
- Linnorm
- ZINB-WaVE
- …

bulk

$$\boldsymbol{CPM} = \log(\frac{counts}{library_{si\square}} \cdot 10^{\mathbf{6}} + 1)$$

$$\downarrow$$

single-cell

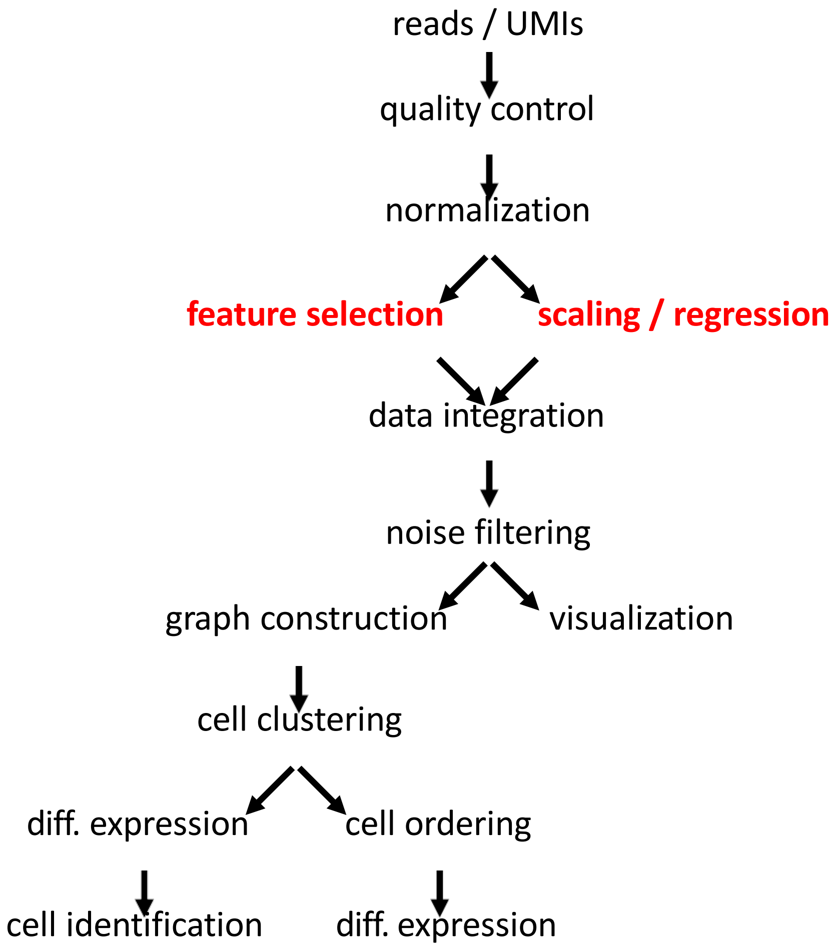
$$log[TP10K + 1] = \log(\frac{counts}{library_{si\square}} \cdot 10^{\%} + 1)$$

Most common for UMI data / fast

$$FPKM = \log(\frac{counts}{library_{size} \; ; \; transcript_{len)t+}} \cdot 10^{\%} + 1)$$

$$TPM = \log(\frac{counts}{transcript_{len)\,\square}} \; ; \; \frac{10^{\%}}{\sum \frac{counts}{transcript_{len)\,\square}}} + 1)$$

scRNA-seq analysis workflow



reads / UMIs

quality control

normalization

feature selection **scaling / regression**

data integration

noise filtering

graph construction visualization

cell clustering

diff. expression cell ordering

cell identification diff. expression

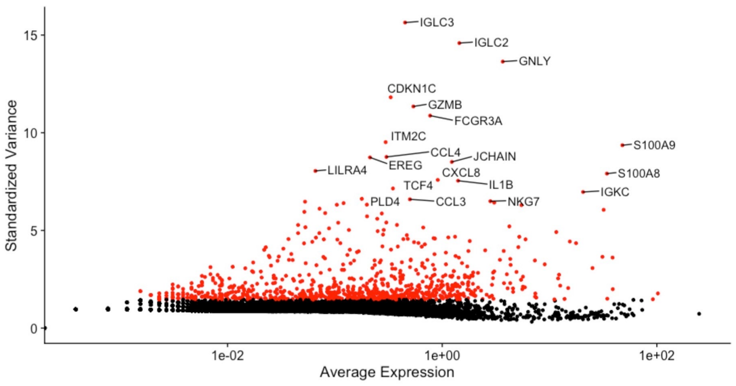
scRNA-seq feature selection

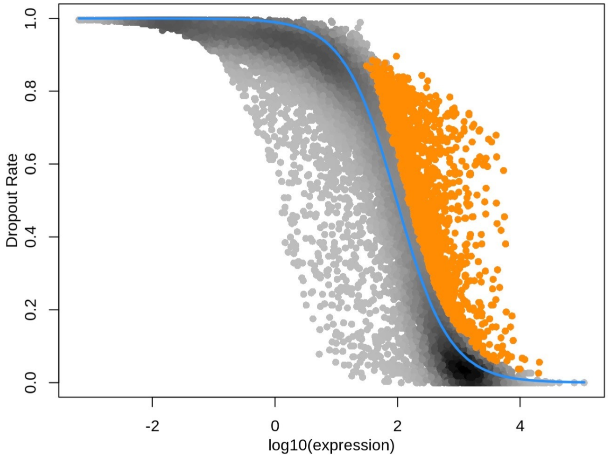
Not all genes are important to define you cell types

Hyper-variable genes are typically characterized by large differences in expression levels between cells, indicating distinct functional roles or cellular states. They can reflect diverse biological processes such as cell cycle stages, cell type-specific markers, or genes associated with cellular responses and regulatory networks.

$$HVG = \frac{variance}{\log(meanExpression)}$$

$$HVG = \frac{\log(meanExpression)}{dropout_{rate}}$$





Dimensionality Reduction

Dimensional reduction compared

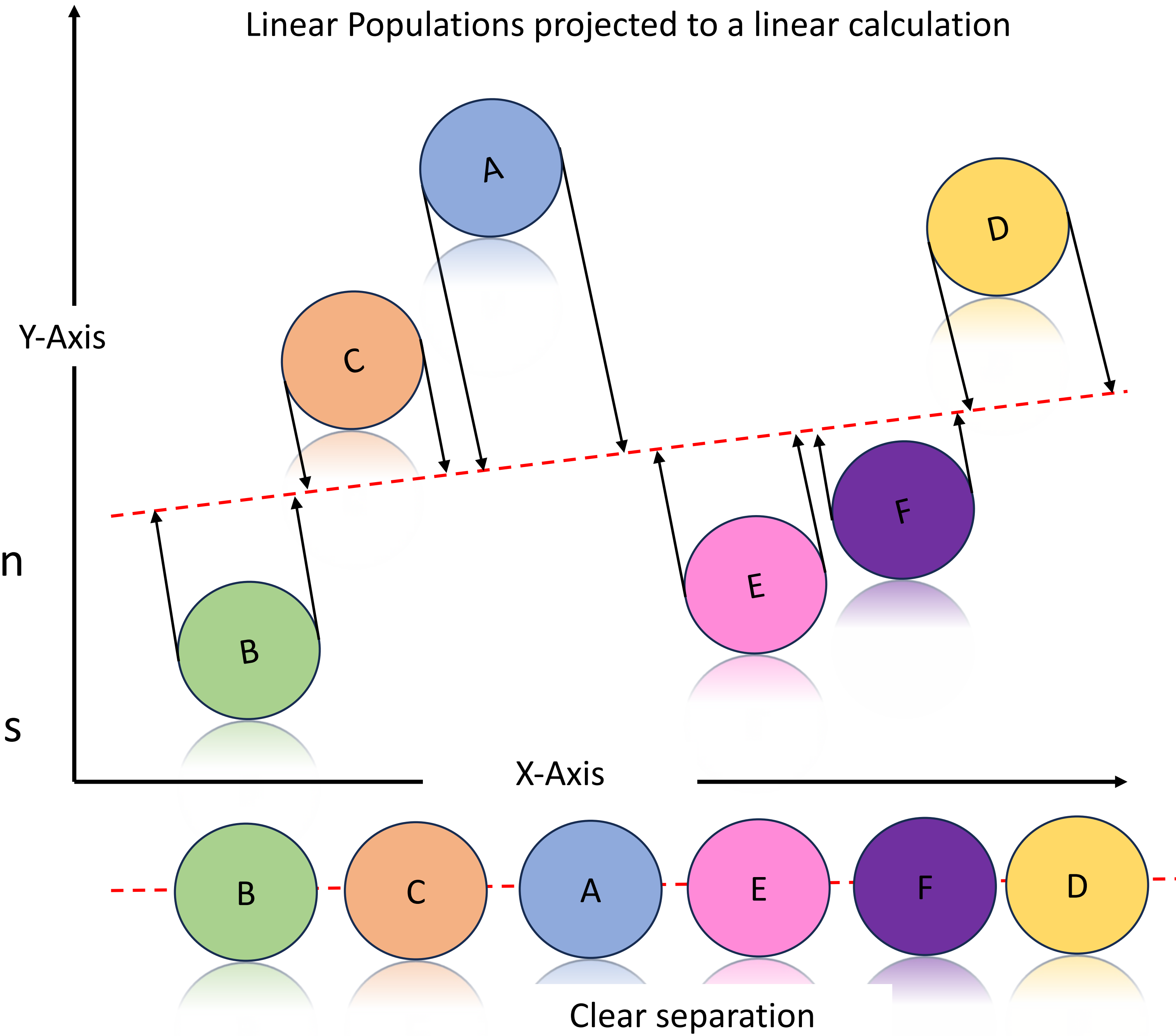
	t-SNE (2018)	UMAP (2018)	PCA (1901)
Type	Non-Linear	Non-Linear	Linear
Suitability for Cytometry	Good	Good	Poor
Can make prediction on new data	No	Yes	Yes
Calculation Speed	Slow	Medium	Extremely Fast
Interpret the axex?	No Preserve local, rather than global structures	Sometimes Preserve local and global structures	Yes Impact of new variable on the new axes can be quantified
Interpret distance between clusters	No	Yes	Yes
Hyperparameters	Perplexity	Number of neighbours	Scale
	Distance Metrics	Distance metric	Center
	Maximum iterations	Maximum distance	
	Theta (for Barnes Hut)	Minimum distance	

Dimension Reduction

Dimension reduction discards redundant information

The Population progression through X and Y axes is the **redundant** information

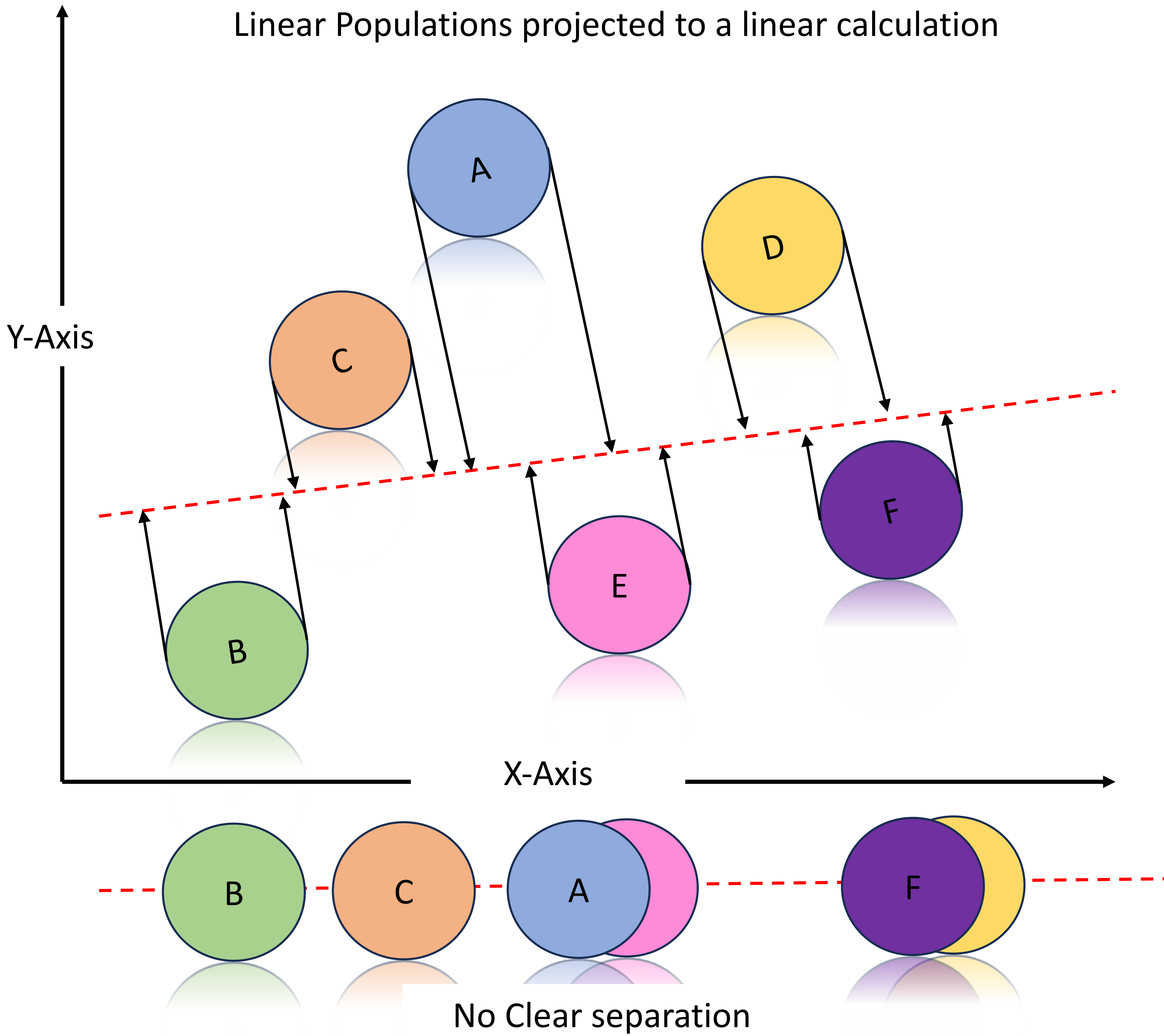
Degree of Separation from each other is the relevant information



Linear vs Non Linear

Flow cytometry population tend to clump in spiral formation

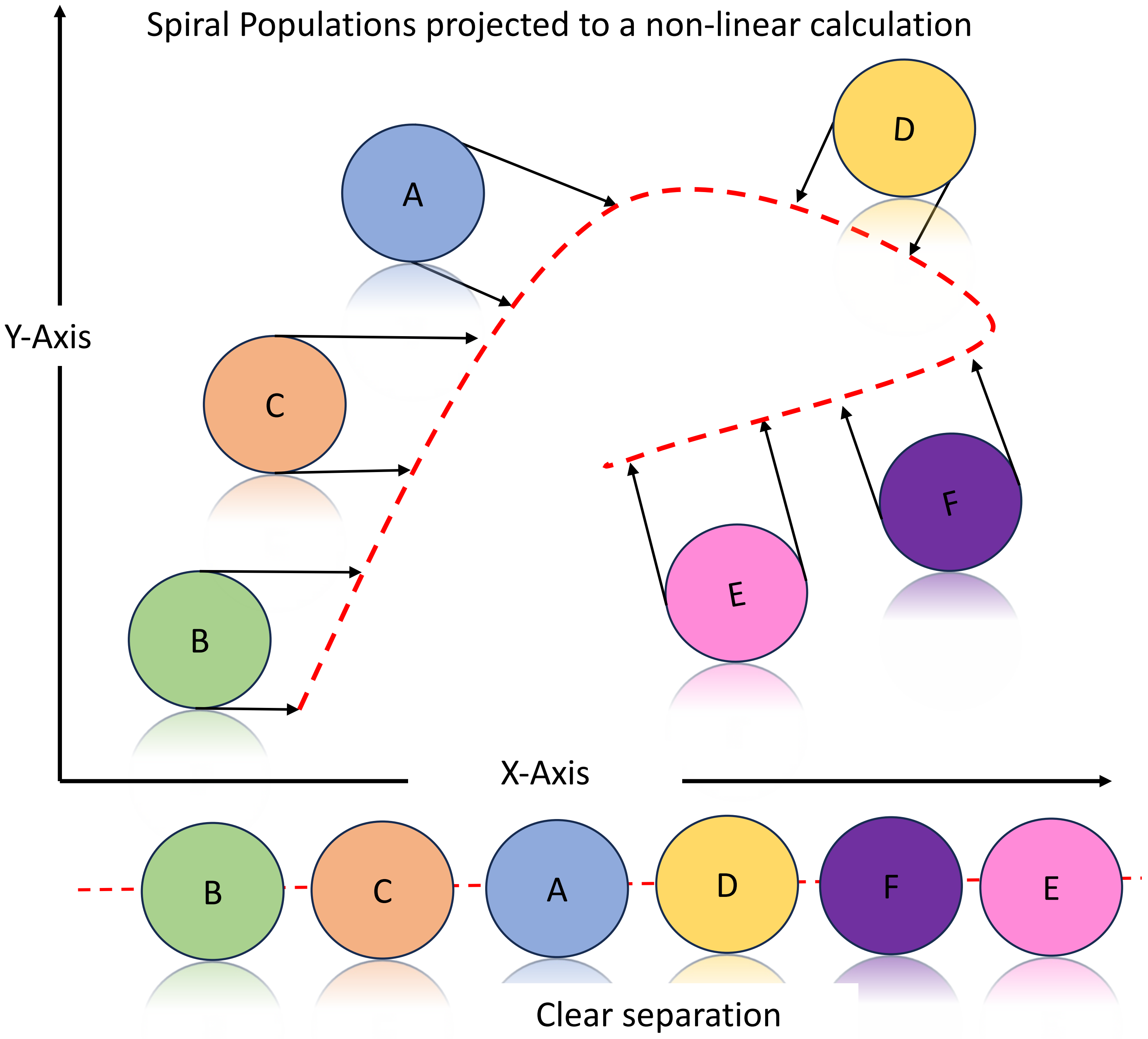
Linear Dimension reduction techniques (e.g PCA) can struggle o separate these.



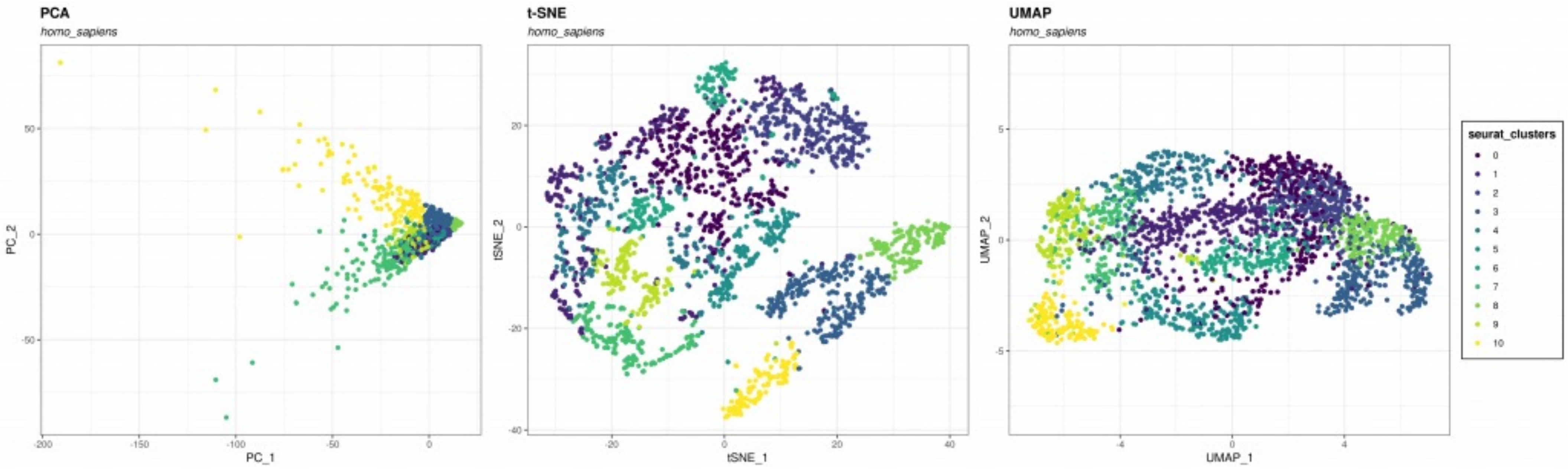
Linear vs Non Linear

tSNE and UMAP make non-linear calculations to project onto

They are most suited to perform dimension reduction on scRNA-Seq Data



Dimension Reduction

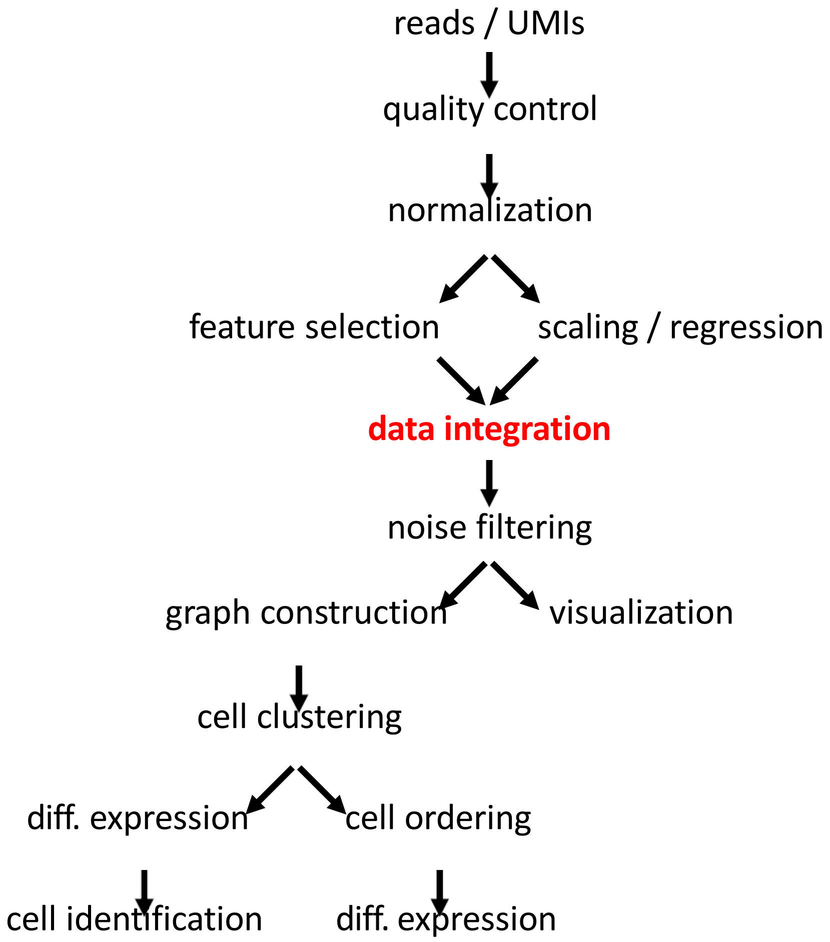


PCA, t-SNE and UMAP create representations of data

Data is de-formed to make t easier to visualise

Data Integration

scRNA-seq analysis workflow



reads / UMIs

quality control

normalization

feature selection

scaling / regression

data integration

noise filtering

graph construction

visualization

cell clustering

diff. expression

cell ordering

cell identification

diff. expression

scRNA-seq data integration

We wish to obtain corrected data where the following goals are met:

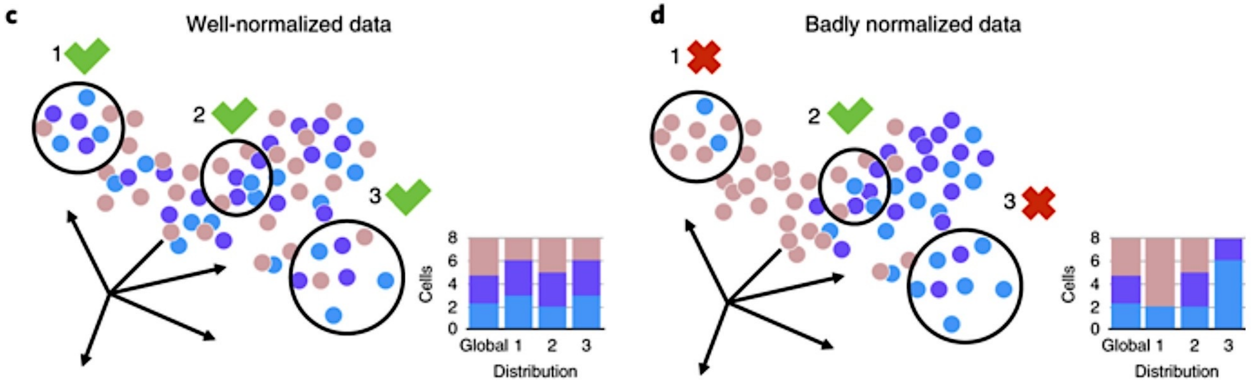
Goal:

1.The batch-originating variance is erased 2.Meaningful heterogeneity is preserved

3.No artefactual variance is introduced

What it practically means:

Similar cell types are intermixed across batches We are not mixing distinct cell types (across or within batches) We do not separate similar cells within batches



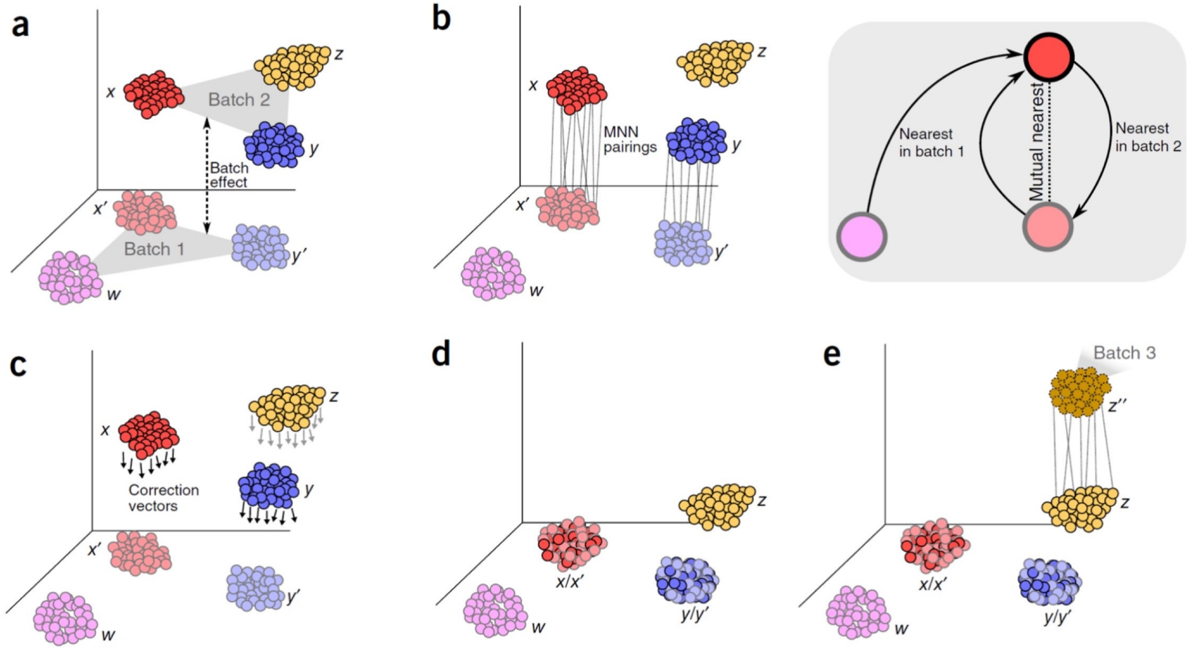
Büttner, et al (2019) *Nat Methods*

scRNA-seq analysis workflow

Regression based bulk-RNAseq batch correction methods are slow and assume the batch is constant across cells

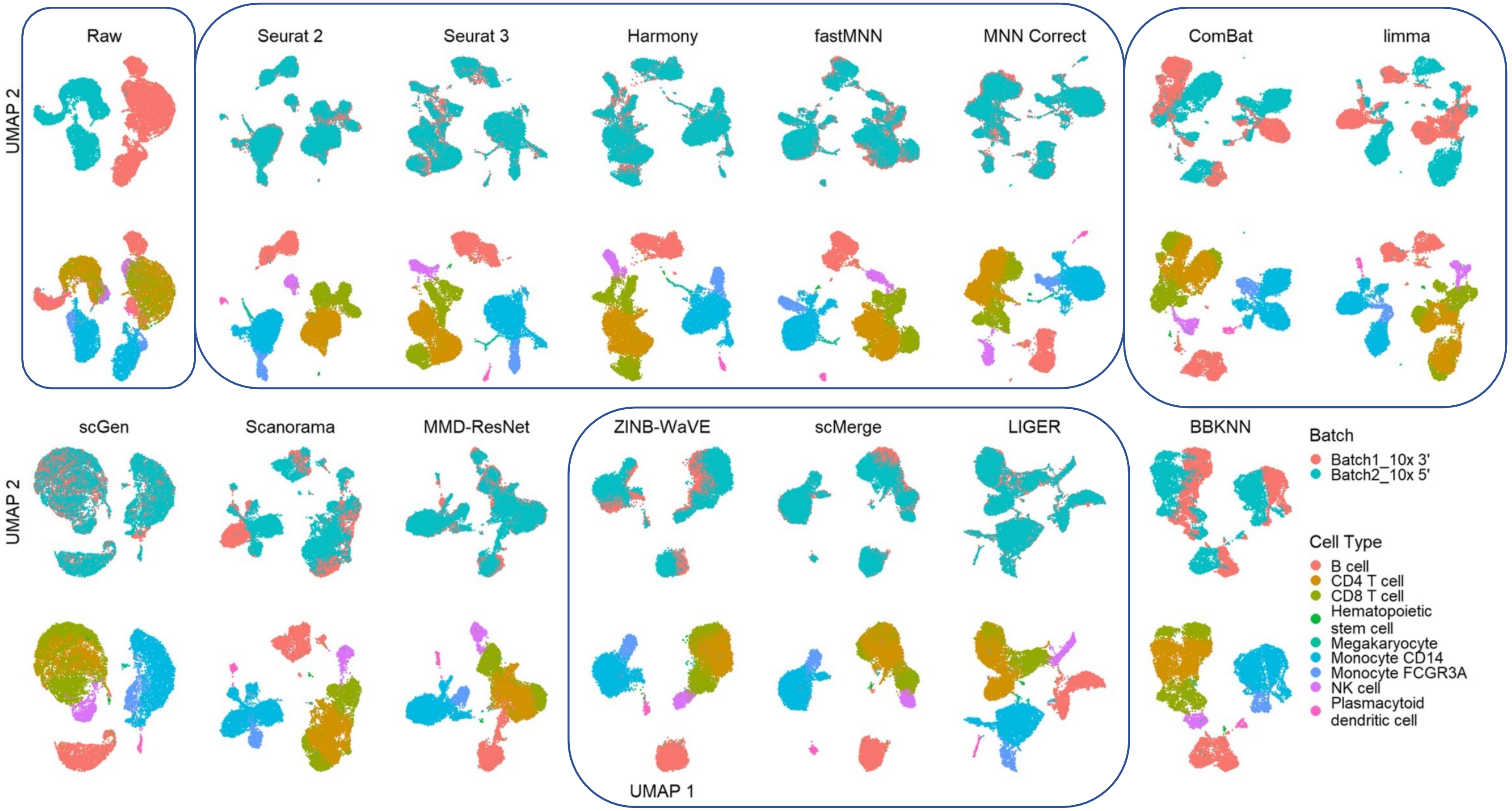
Modern data integration methods are based on the same principle:

- find MNN (mutual nearest neighbours) across datasets and correct each cell individually
- Done on a graph: much faster



Haghverdi et al (2017) Nat Biotechnology

scRNA-seq analysis workflow



MNN

regression

matrix factorization

Tran et al (2020) *Genome Biology*