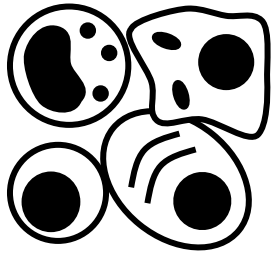
Summer School 2020 in
Computational RNA Biology

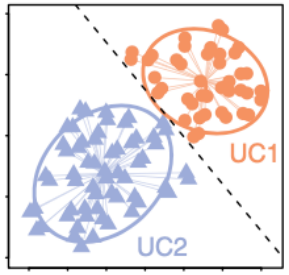
Mini-Project:

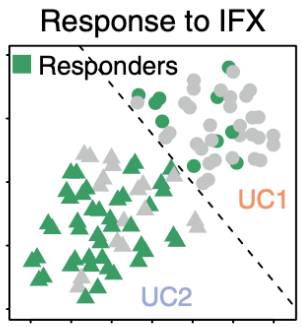
Single cell RNA-seq analysis

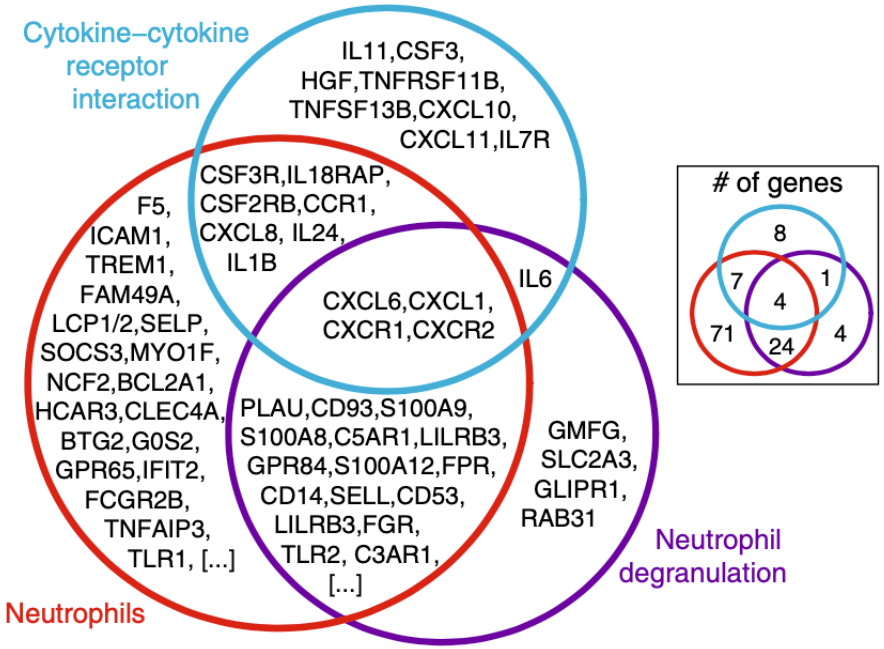


Paulo Czarnewski
Erik Fasterius

Project1 - background





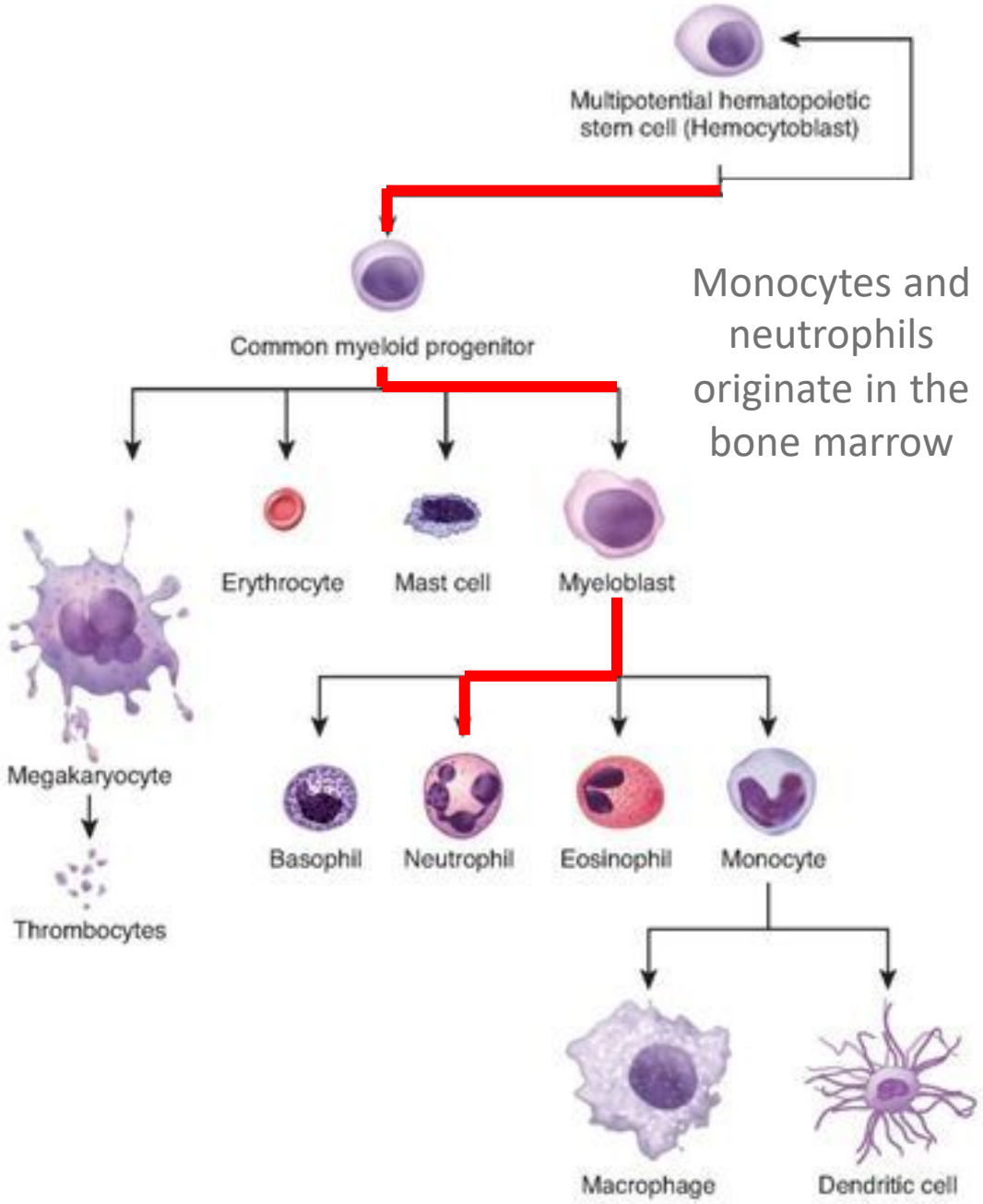


Czarnewski et al (2019) *Nat Communications*
Skatteborg et al (2020) *J Crohn and Colitis*

Ulcerative colitis patients can be subdivided in UC1 and UC2 subgroups

Monocytes and neutrophils drive inflammation in UC1 patients

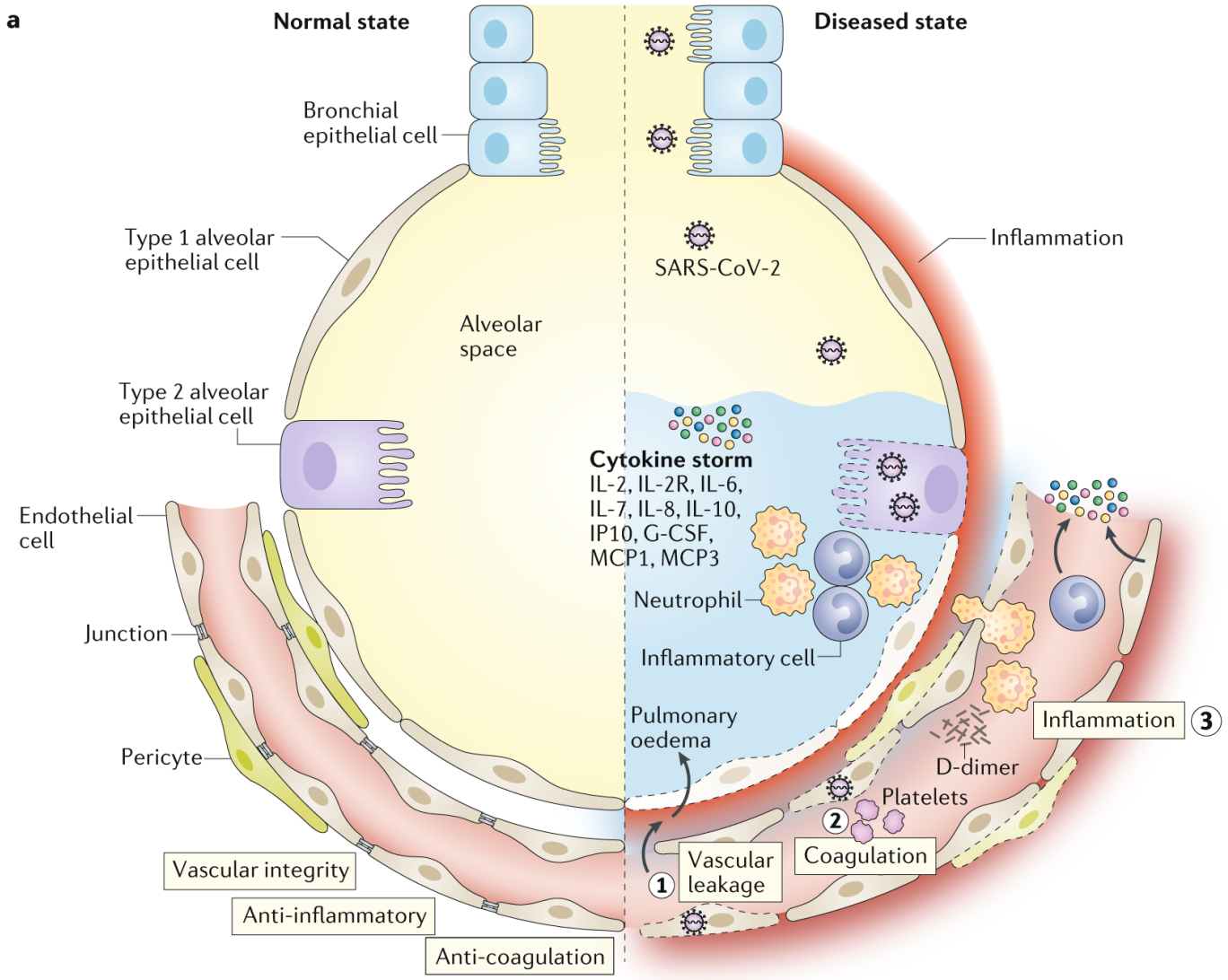
UC1 patients are refractory to current therapy



Monocytes and neutrophils originate in the bone marrow

GOAL: Which genes specifically drive the differentiation of Neutrophils?

Project2 - background



Teuwen et al (2020) *Nat reviews Immunology*

Elderly patients usually develop severe lung inflammation and lung disfunction.

Many cell types orchestrate the immune response to the virus.

Their relative contribution at the single-cell resolution is still unclear

GOAL: Which cell types and genes are altered when comparing blood immune cells from healthy versus COVID-19 patients?

Project-based learning (PBL)

Report.Rmd

Load and merge datasets

- Consult the Glossary or additional sources for help
- Which file format do we have the data in?
- Describe in form of text the rational for this step in your markdown report.

Glossary

Reading files

There are many formats available in which one can store single cell information, many of which cannot all be listed here. The most common formats are:

[...]

How to run it:

```
# From .csv .tsv .txt format
raw_matrix <- read.delim(
          file = "data/folder_sample1.csv",
          row.names = 1 )

# From .mtx format
sparse_matrix <- Seurat::Read10X(
          data.dir = "data/folder_sample1")

# From .h5 format
sparse_matrix <- Seurat::Read10X_h5(
          filename = "data/matrix_file.h5",
          use.names = T)
```

[...]

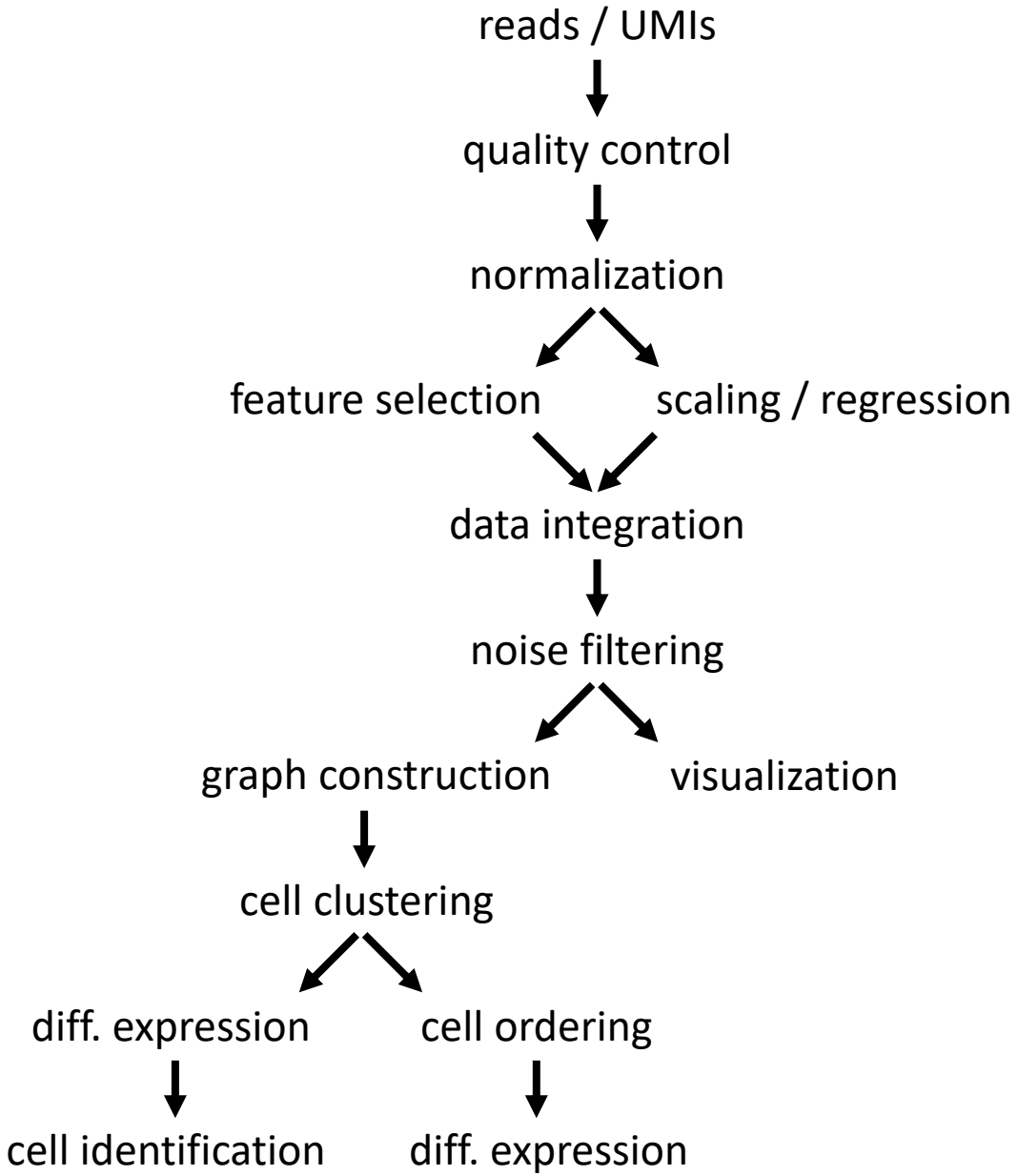
Report.Rmd

```
# Loading data

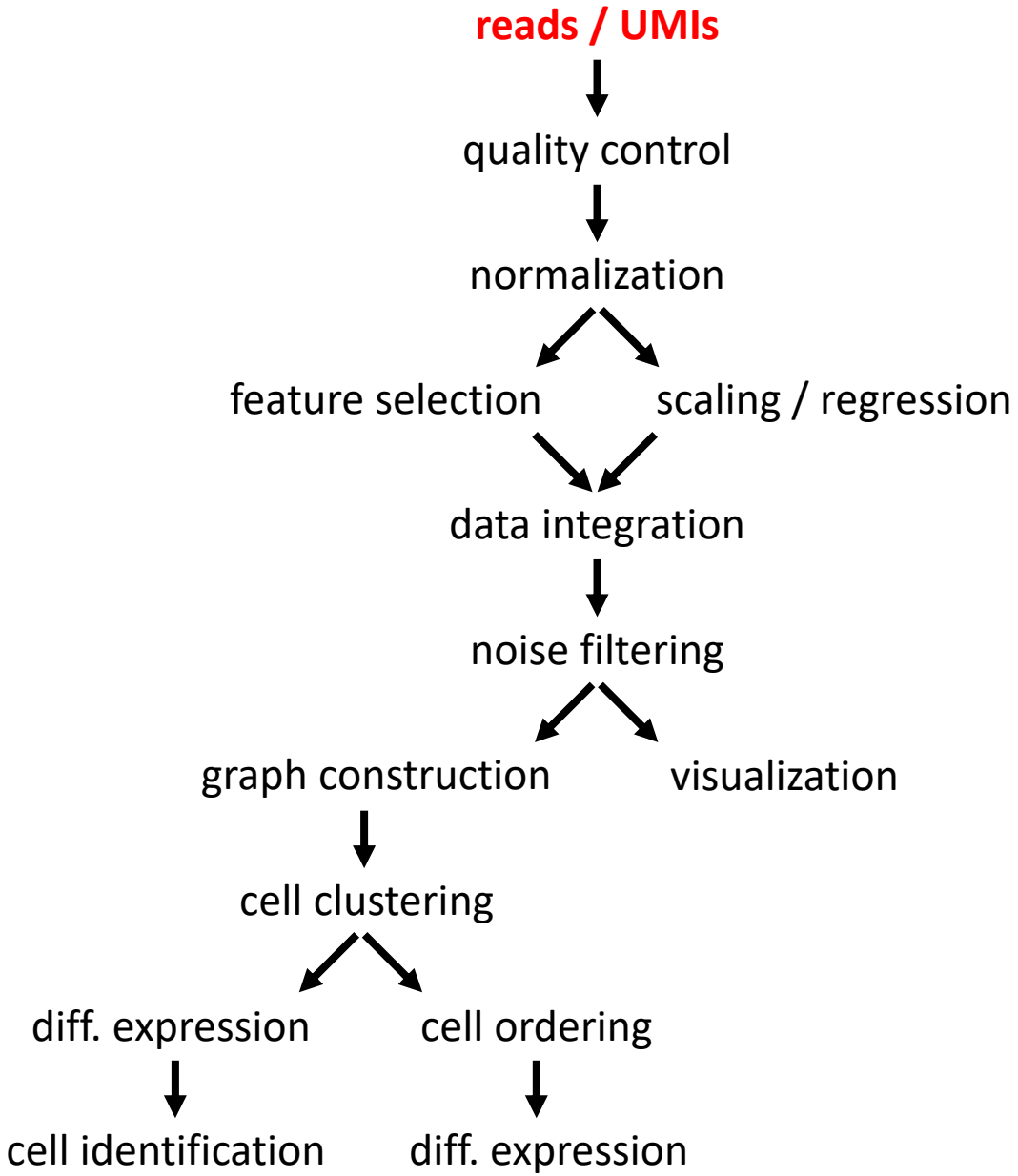
We first load the single cell RNA-seq dataset
supplied from the `.h5` format in order to create
a Seurat object.

```{r}
data <- Seurat::Read10X_h5( filename =
"data/colon_dataset.h5", use.names = T)
```
```

scRNA-seq analysis workflow

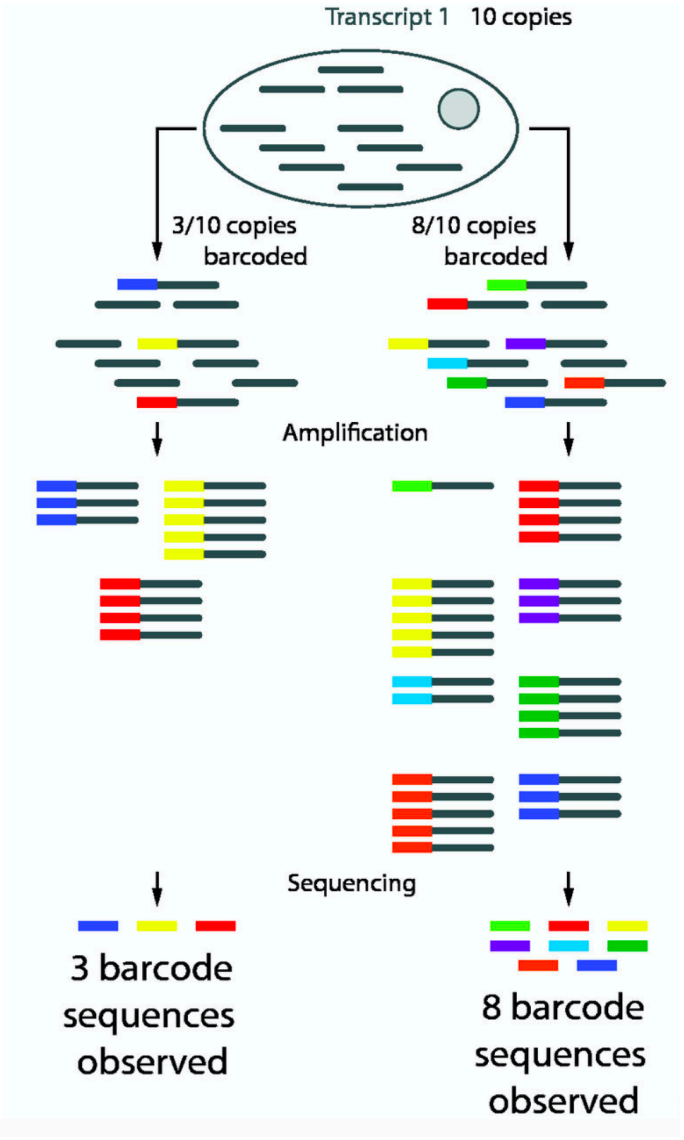


scRNA-seq analysis workflow

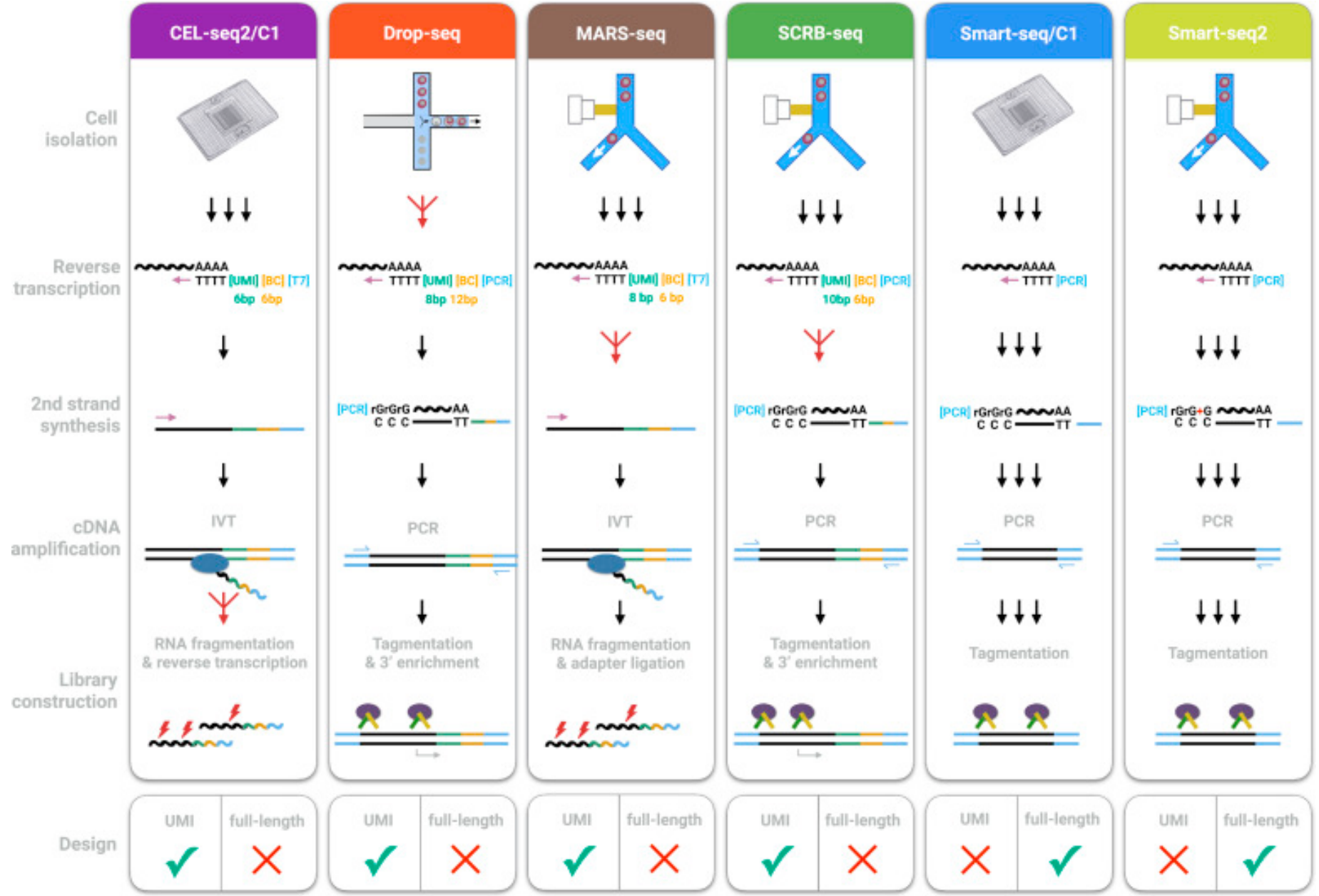


scRNA-seq - technologies

UMI (unique molecular identifier)

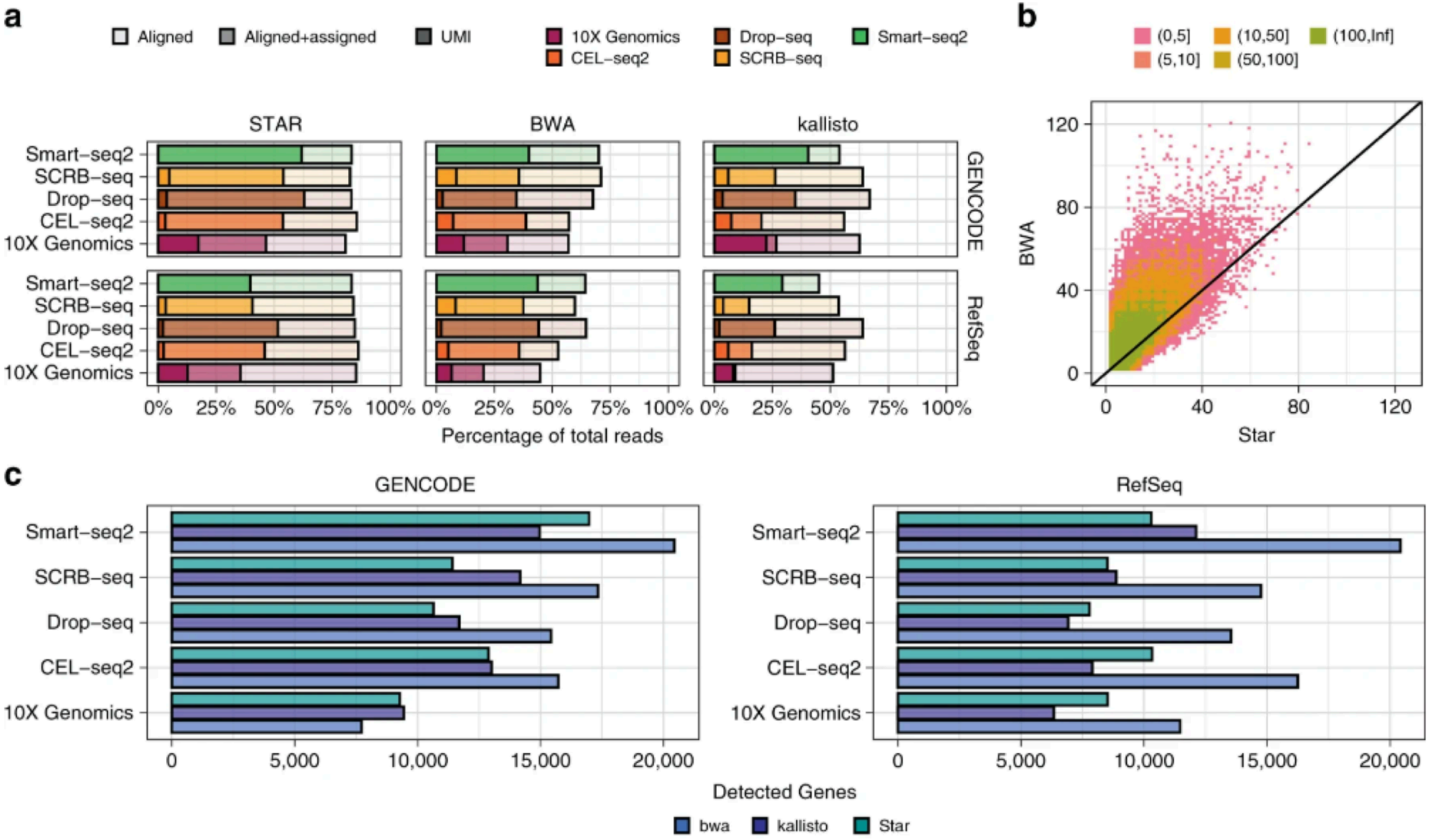


Zajac et al (2013) *Plos One*



Ziegenhain et al (2015) *Molecular Cell*

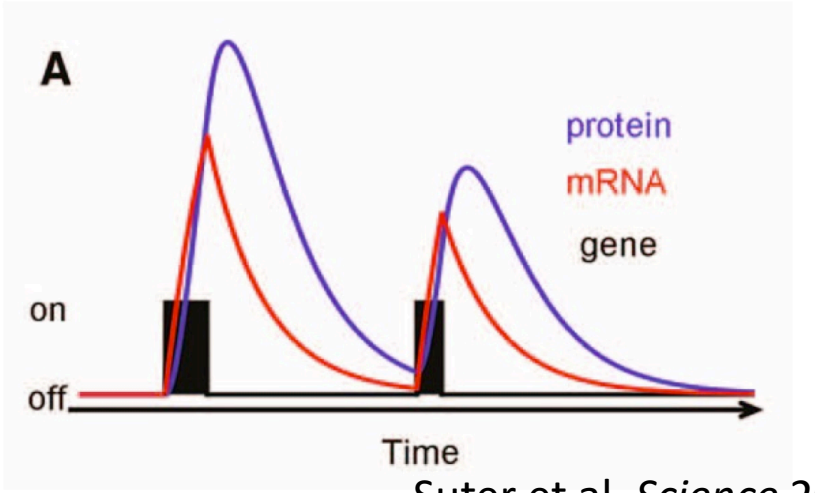
scRNA-seq – alignment methods



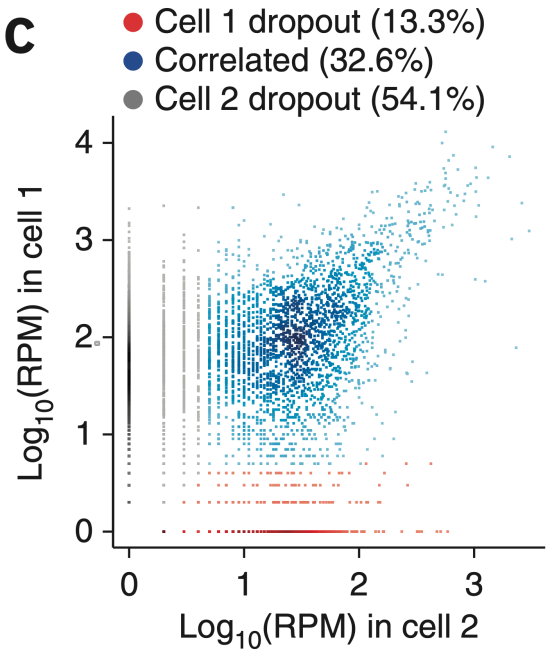
Vieth et al (2019) *Nature Comm*

scRNA-seq - problems compared to bulk RNA-seq

- Amplification bias
- Drop-out rates
- Transcriptional bursting
- Background noise
- Bias due to cell-cycle, cell size and other factors
- Often clear batch effects
- Dissociation protocols may introduce transcriptional artifacts
- Ambient RNA

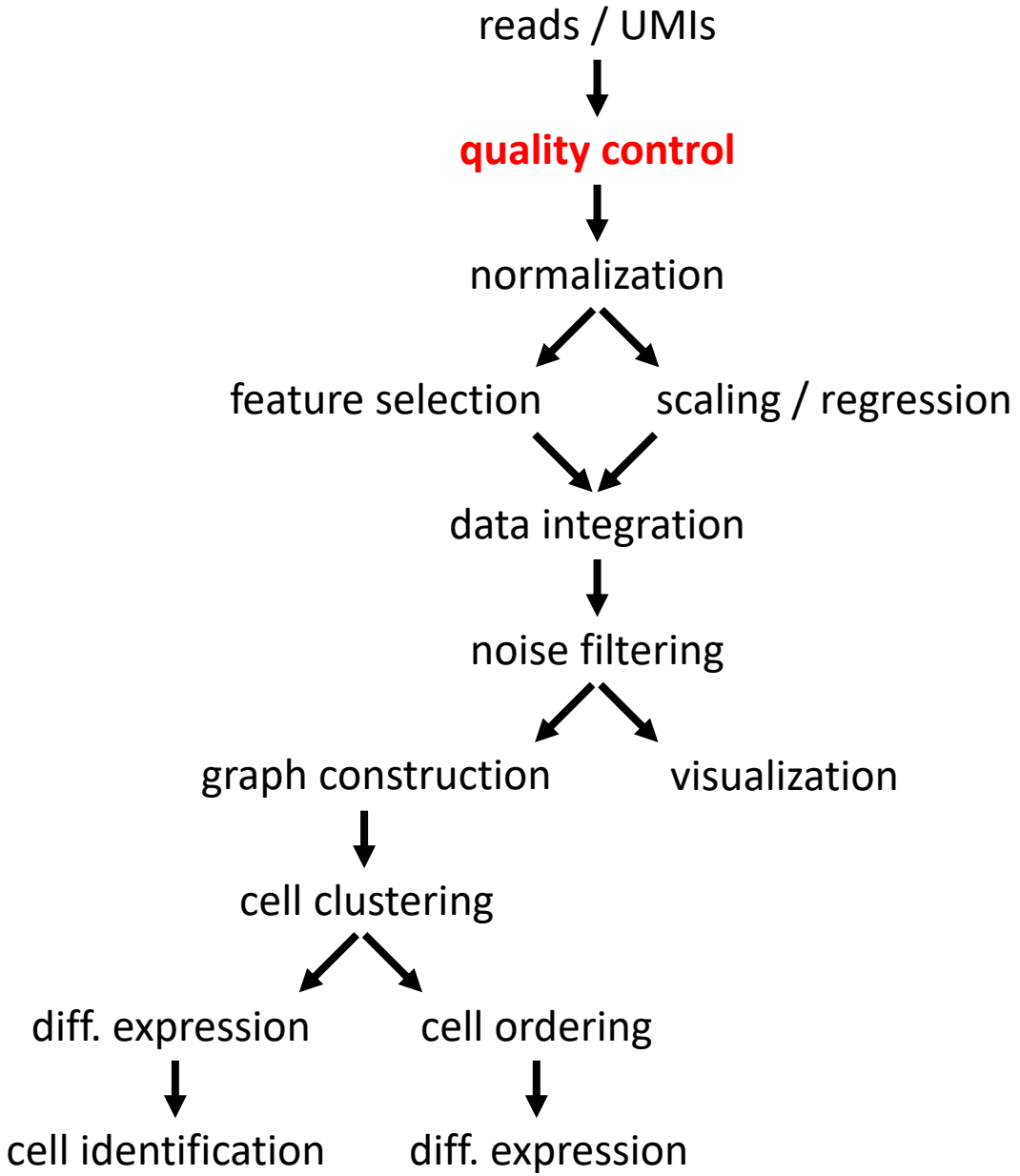


Suter et al. *Science* 2011



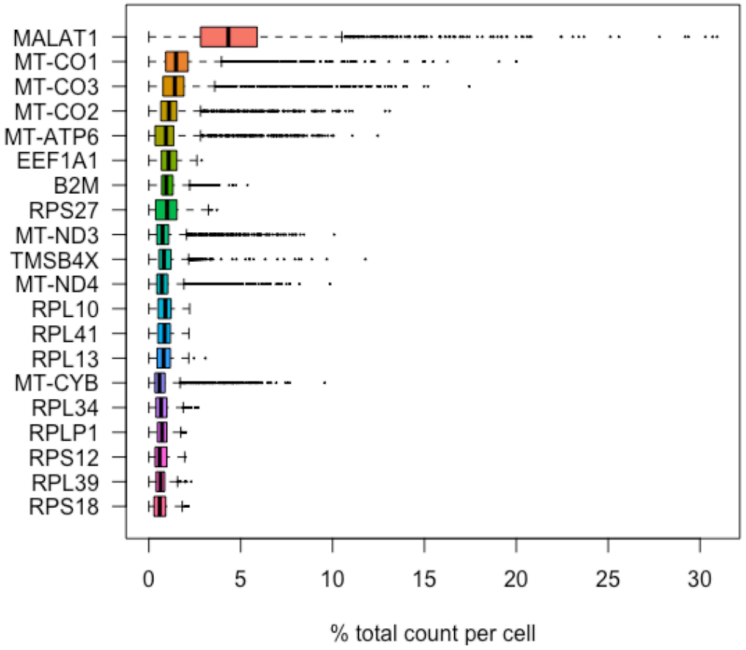
Karchenko et al. *Nature Methods* 2014

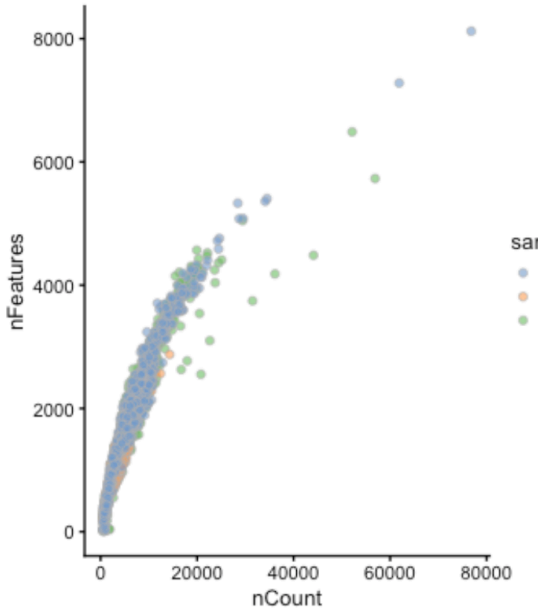
scRNA-seq analysis workflow

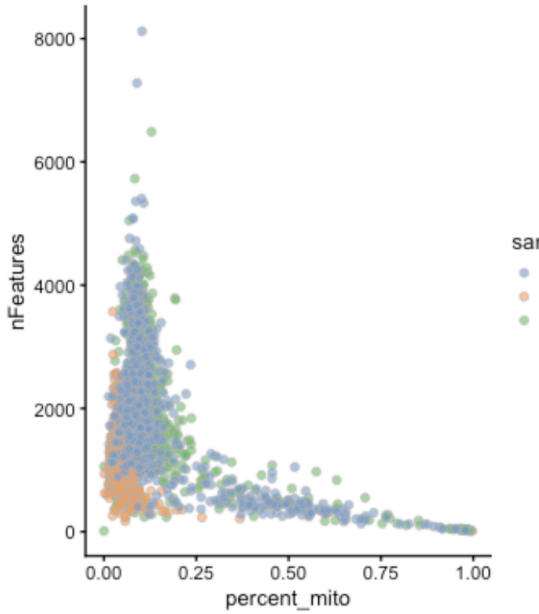


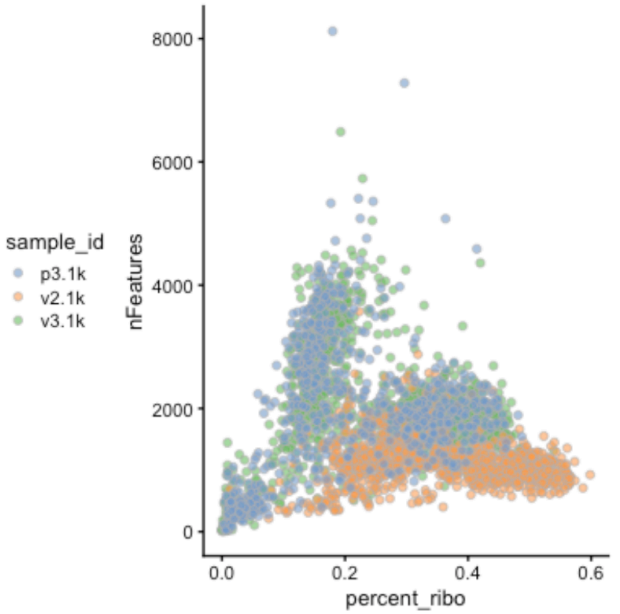
scRNA-seq – quality control

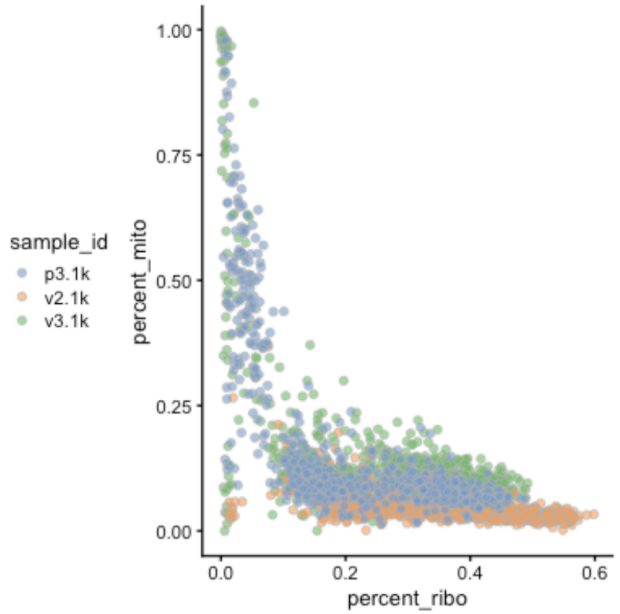
- Mapping statistics (% uniquely mapping)
- Cell cycle biases
- 3’ bias – for full length methods like SS2
- mRNA-mapping read percentage
- Number of UMIs/read counts
- Number of detected genes
- Spike-in detection
- Mitochondrial percentage
- ribosomal percentage
- Protein-coding percentage



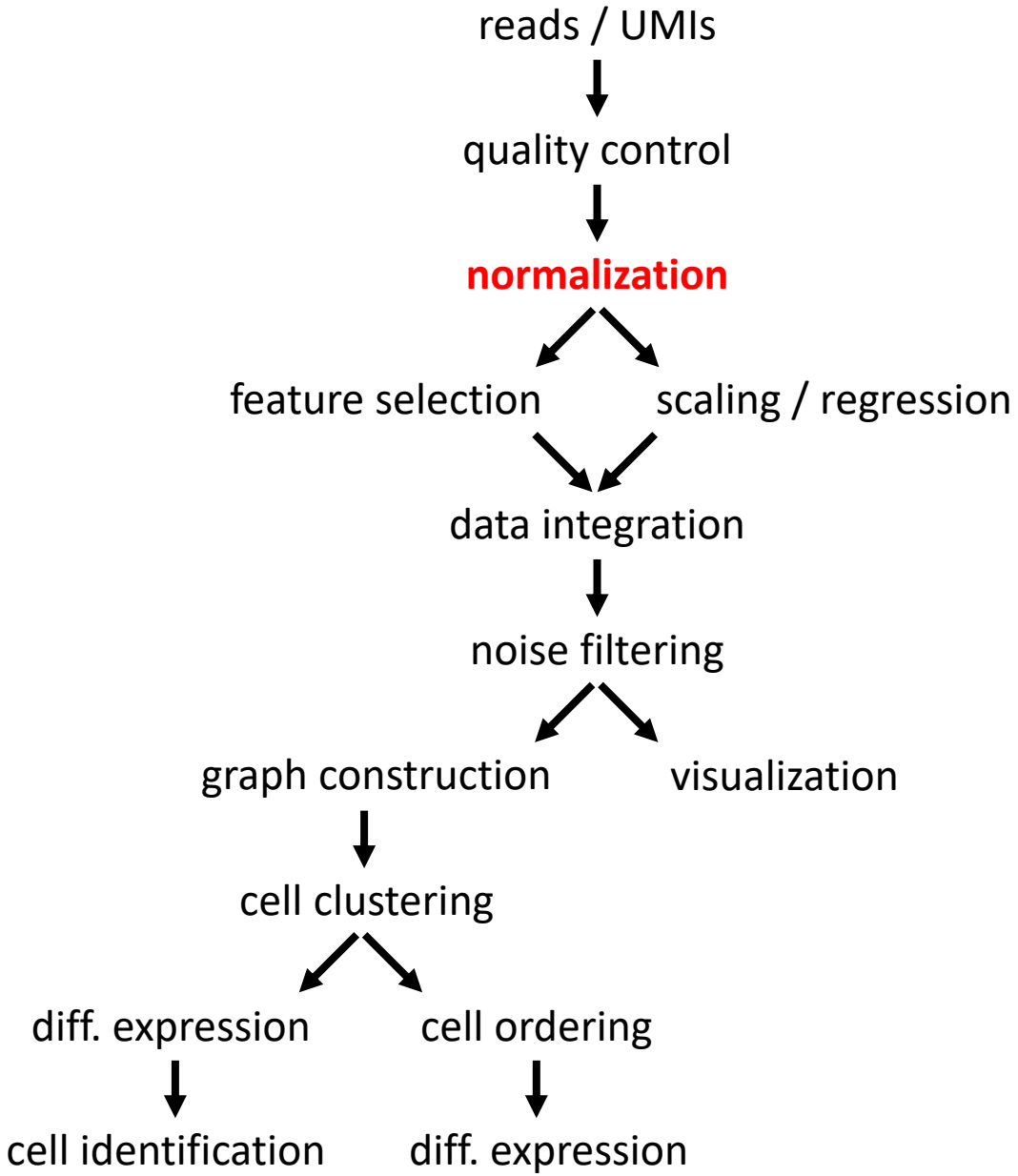








scRNA-seq analysis workflow



scRNA-seq - normalization

Count normalization (UMI and read counts)
for uneven sequencing depth
- CPM - log[CP10K+1]

bulk

$$\boldsymbol{CPM} = \log\left(\frac{counts}{library_{size}} \cdot 10^{6} + 1\right)$$

single-cell

$$log[TP10K + 1] = \log\left(\frac{counts}{library_{size}} \cdot 10^{4} + 1\right)$$

Most common for UMI data / fast

Gene length normalization (read counts)
for differences in gene detection due to gene length
- TPM (closer to UMI counts)
- FPKM

$$FPKM = \log\left(\frac{counts}{library_{size} \cdot transcript_{length}} \cdot 10^{4} + 1\right)$$

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
- …

$$TPM = \log\left(\frac{counts}{transcript_{length}} \cdot \frac{10^{4}}{\sum \frac{counts}{transcript_{length}}} + 1\right)$$

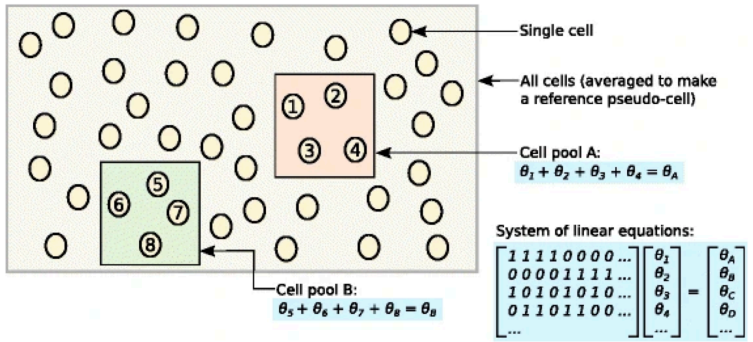
scRNA-seq - normalization

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

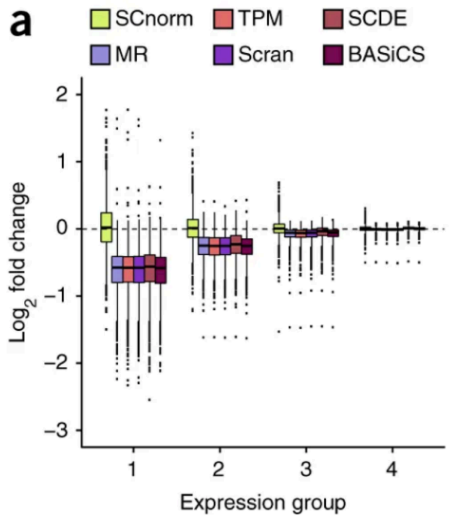
Drop-out rate normalization (UMI and read counts)
for differences in RNA content / drop-out rates
- Deconvolution/Scran(Pooling-Across-Cells)
- SCnorm(Expression-DepthRelation)
- SCTransform
- Census
- Linnorm
- ZINB-WaVE
- …

Deconvolution



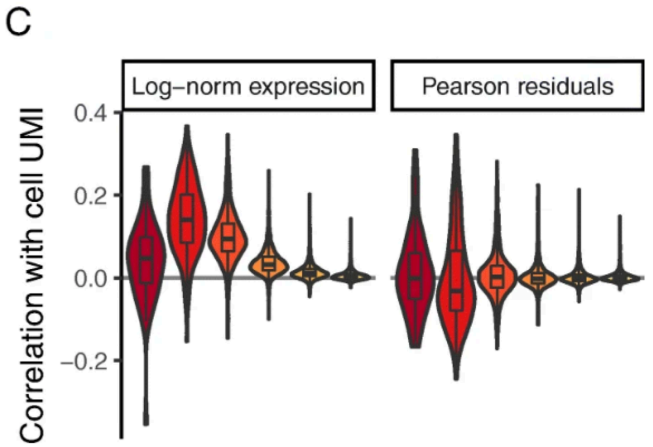
Lun et al. Genome Biol. 2016

SCnorm



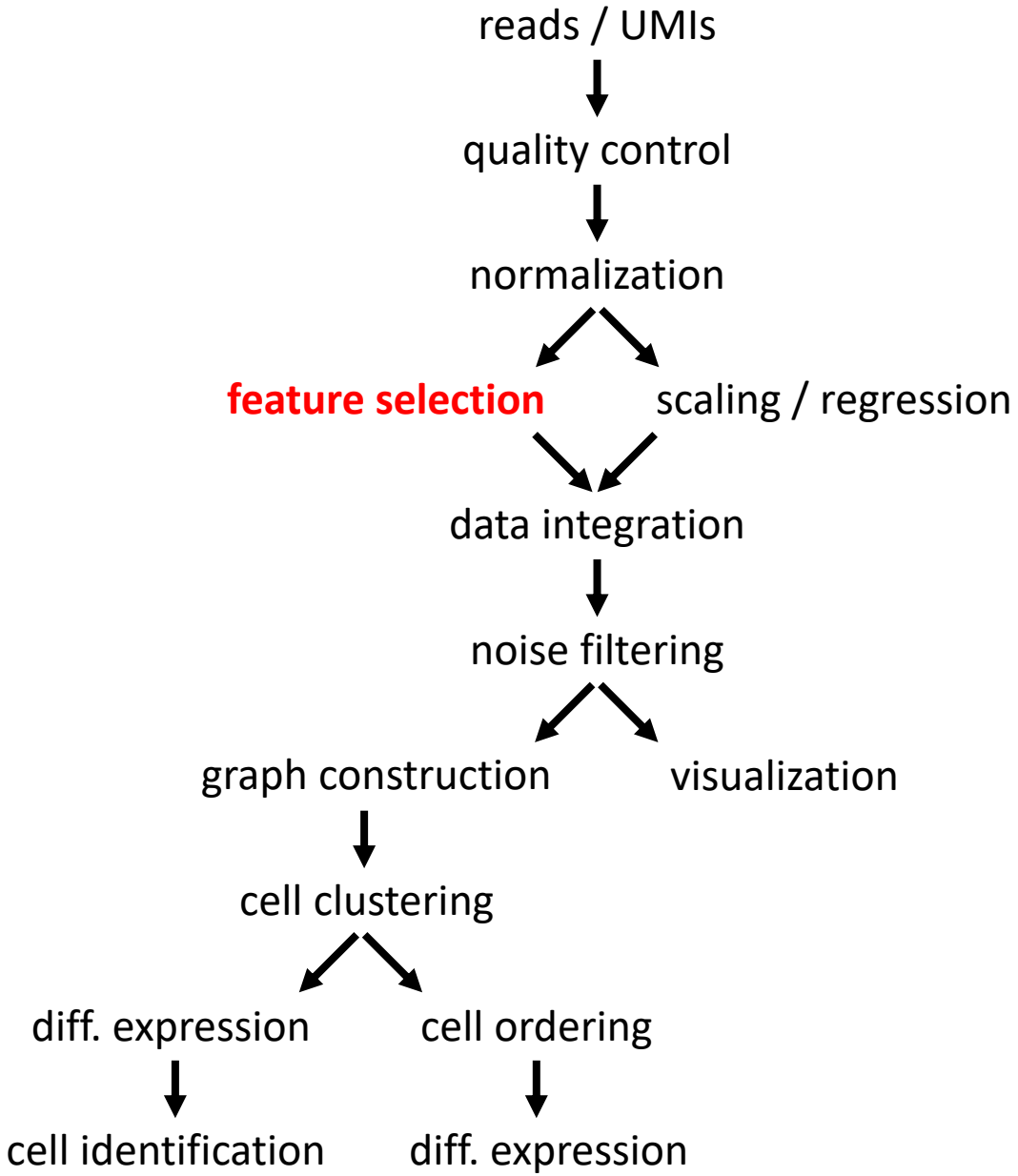
Bacher et al. Nature Methods 2017

SCTransform



Hafmeister & Satija Genome Biology 2019

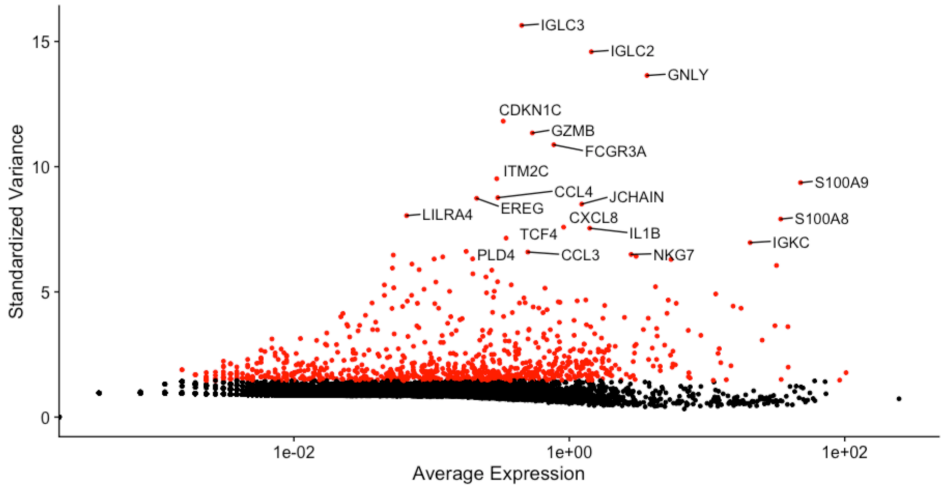
scRNA-seq analysis workflow



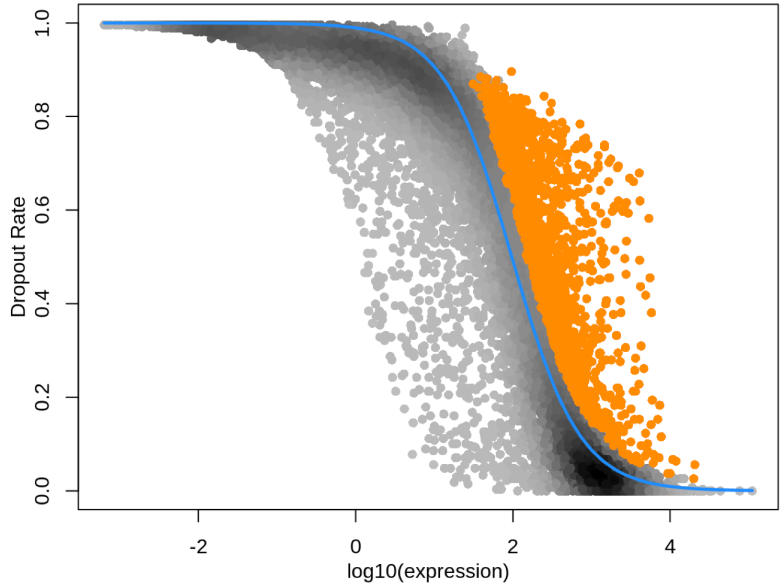
scRNA-seq - dimensionality reduction

Not all genes are important to define you cell types

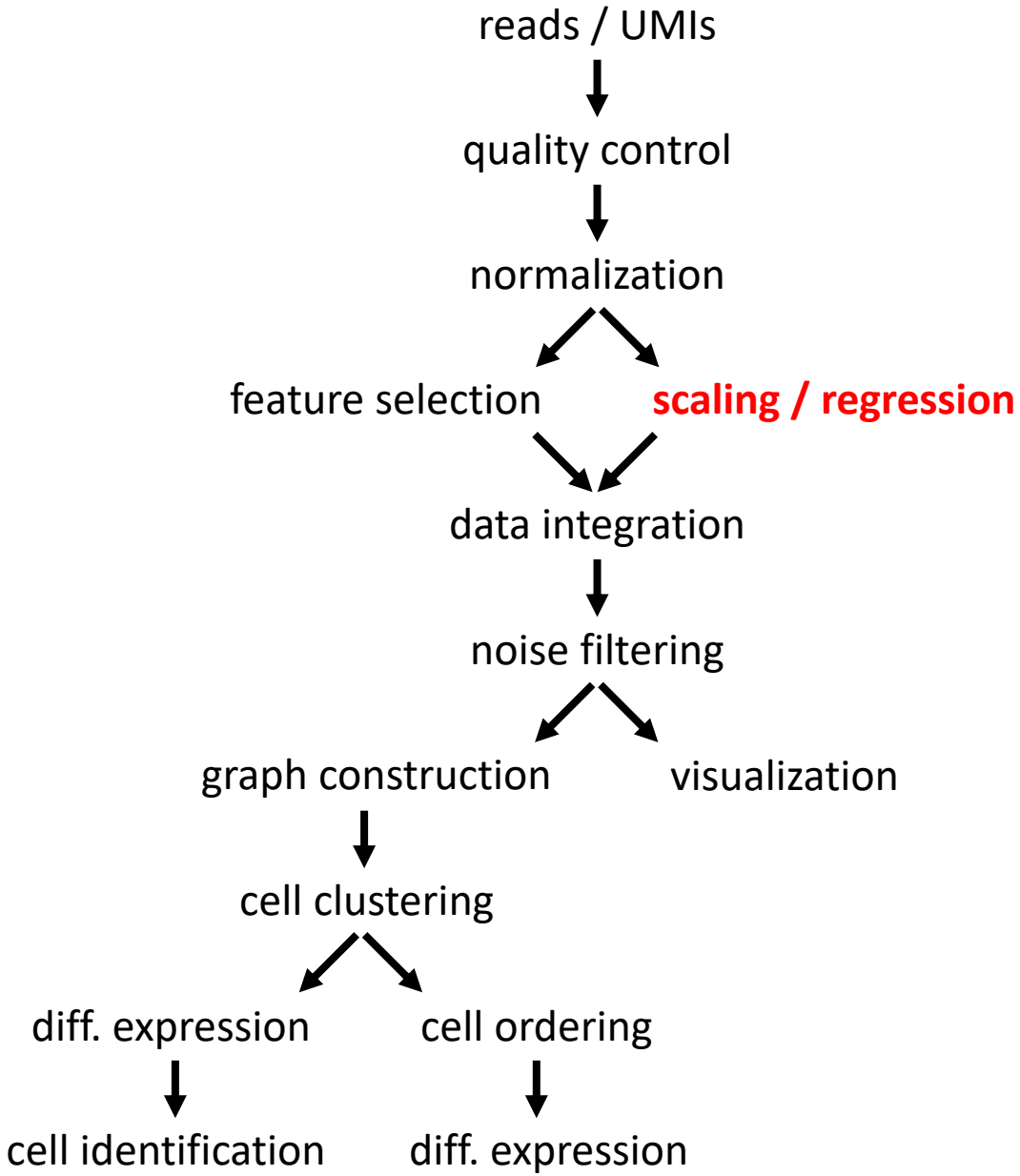
$$HVG = \frac{variance}{\log(meanExpression)}$$



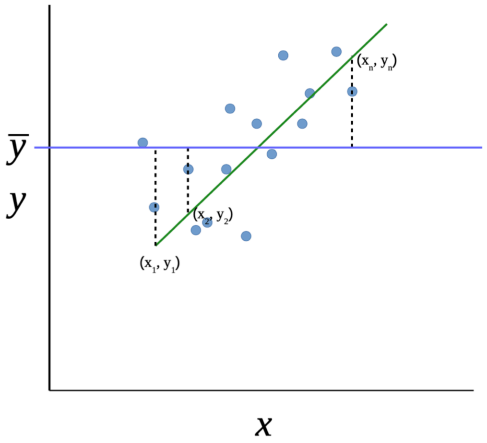
$$HVG = \frac{\log(meanExpression)}{dropout_{rate}}$$



scRNA-seq analysis workflow



scRNA-seq – scaling and regression



Any source of variation that you do not expect to give separation of the cell types should be regressed out.

- Fit a line to the gene expression *vs* variable of interest
- Calculate residuals
- Remove variance explained by the variable of interest by taking the residuals.
- Linear / Negative Binomial / Poisson distributions

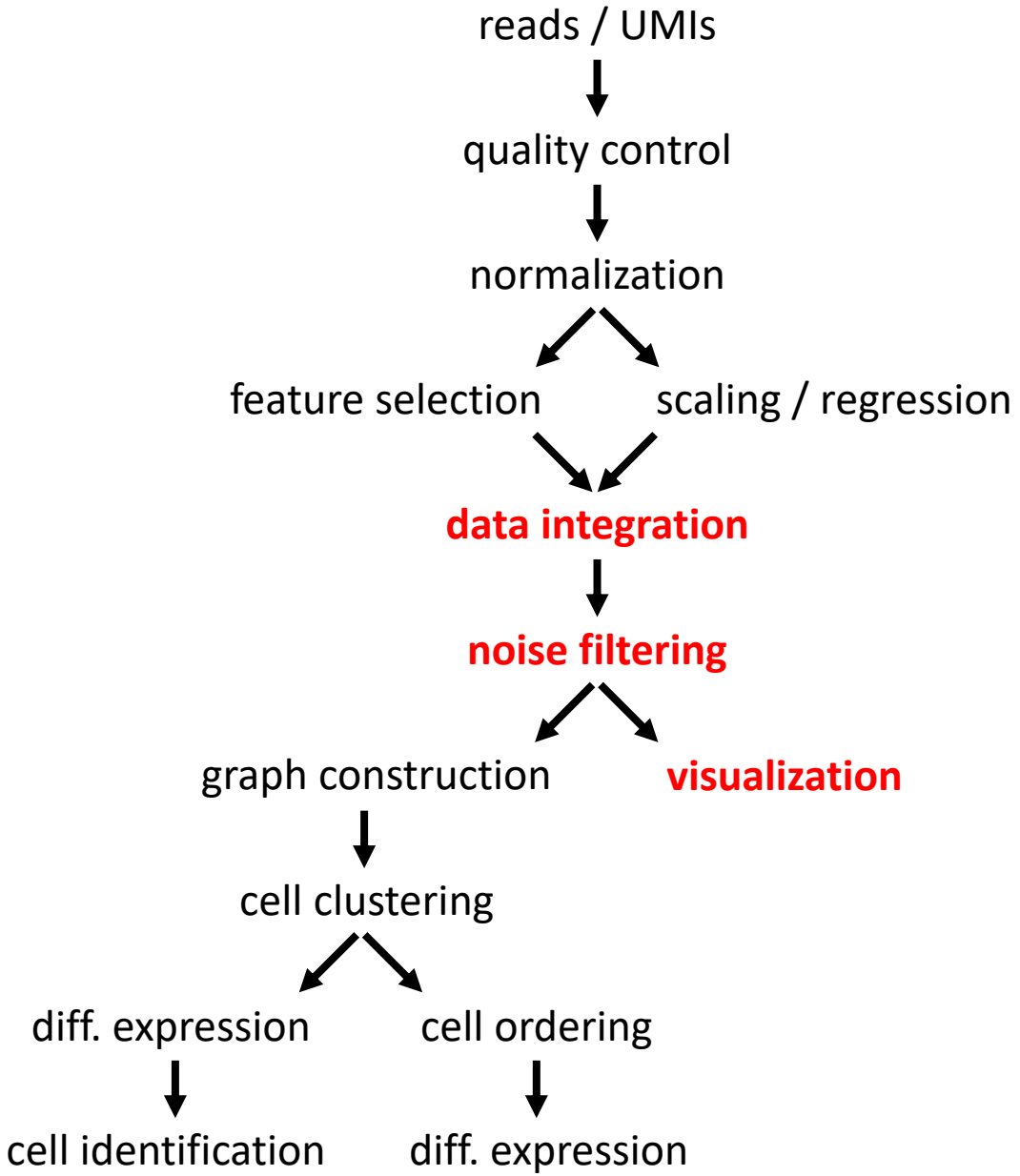
fast

logNormalized counts follows a log-linear distribution

slower (but more accurate)

Regressing counts directly is better with count-based distributions

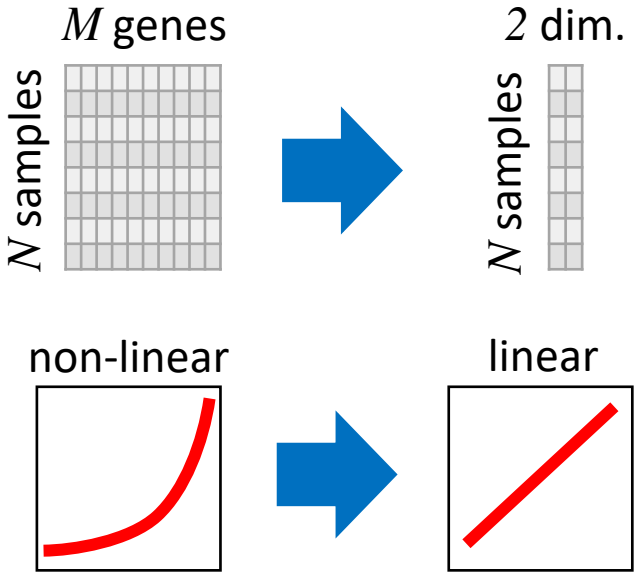
scRNA-seq analysis workflow

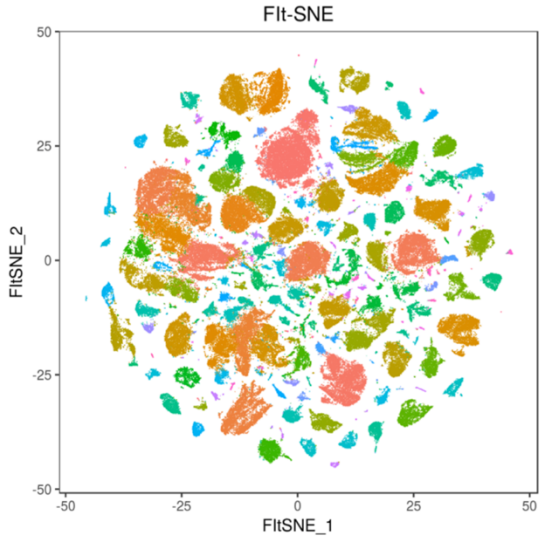


scRNA-seq - dimensionality reduction

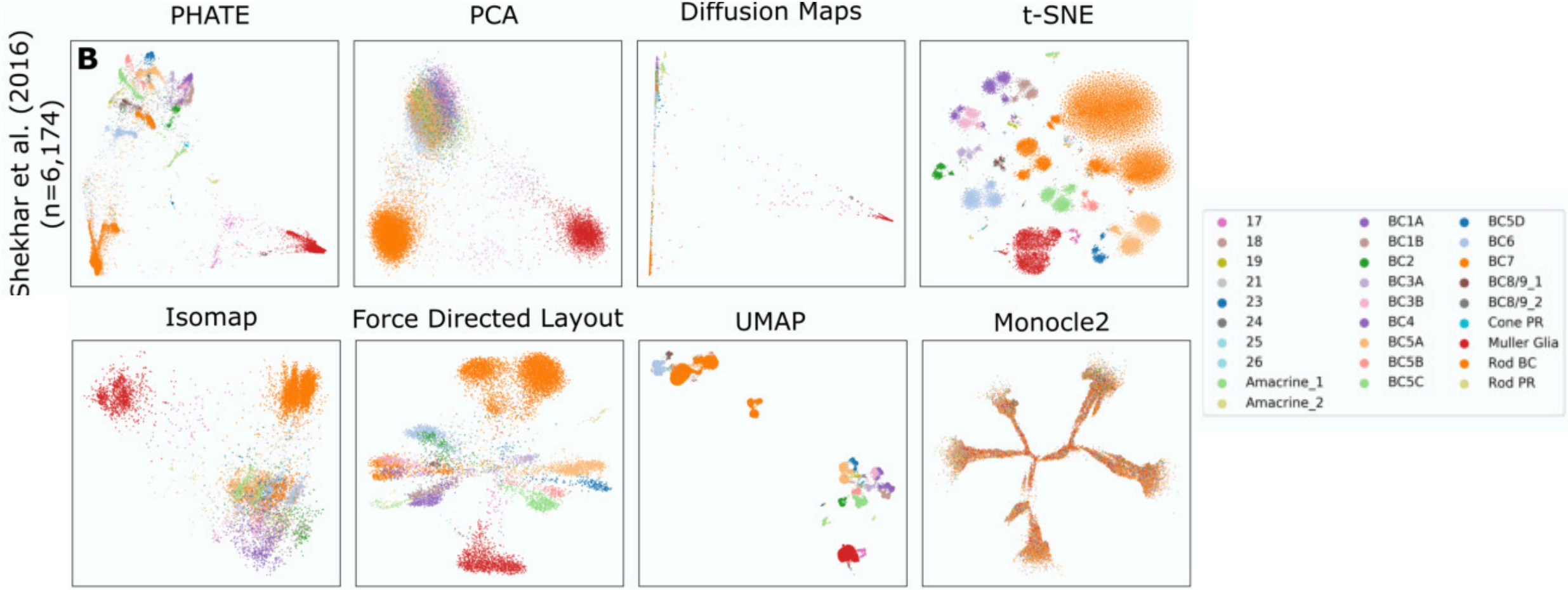
- Simplify complexity, so it becomes easier to work with.
  Reduce number of features (genes)
  In some: Transform non-linear relationships to linear
- "Remove" redundancies in the data
- Identify the most relevant information (find and filter noise)
- Reduce computational time for downstream procedures
- Facilitate clustering, since some algorithms struggle with too many dimensions
- Data visualization

… and more …



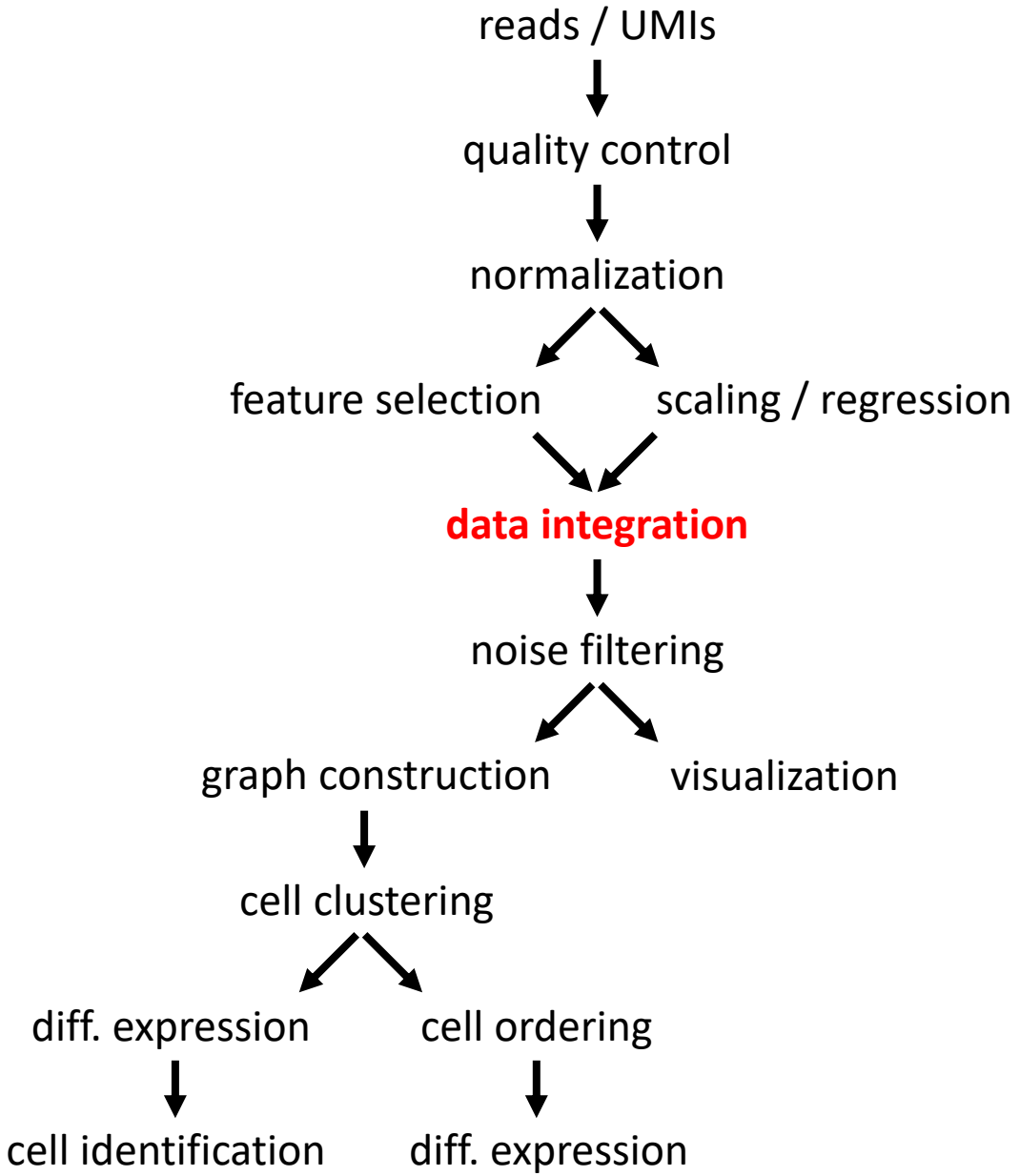


scRNA-seq dimensionality reduction



Moon et al (2019) *BioRxiv*

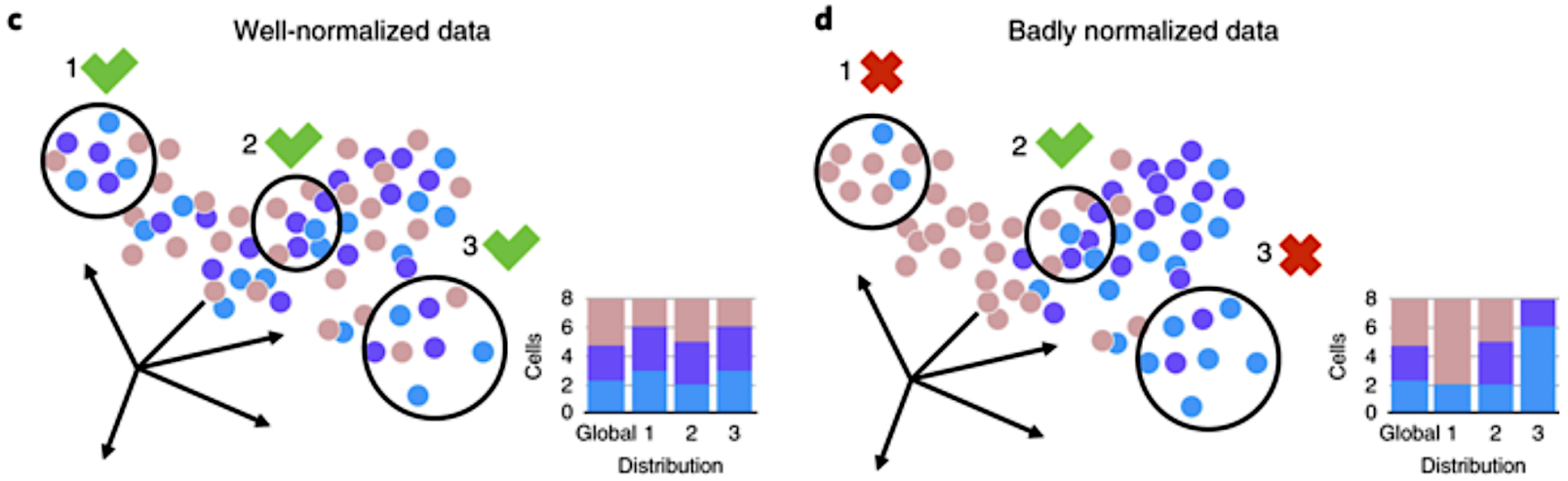
scRNA-seq analysis workflow



scRNA-seq – data integration

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

Goal:	**What it practically means:**
1.The batch-originating variance is erased	Similar cell types are intermixed across batches
2.Meaningful heterogeneity is preserved	We are not mixing distinct cell types (across or within batches)
3.No artefactual variance is introduced	We do not separate similar cells within batches



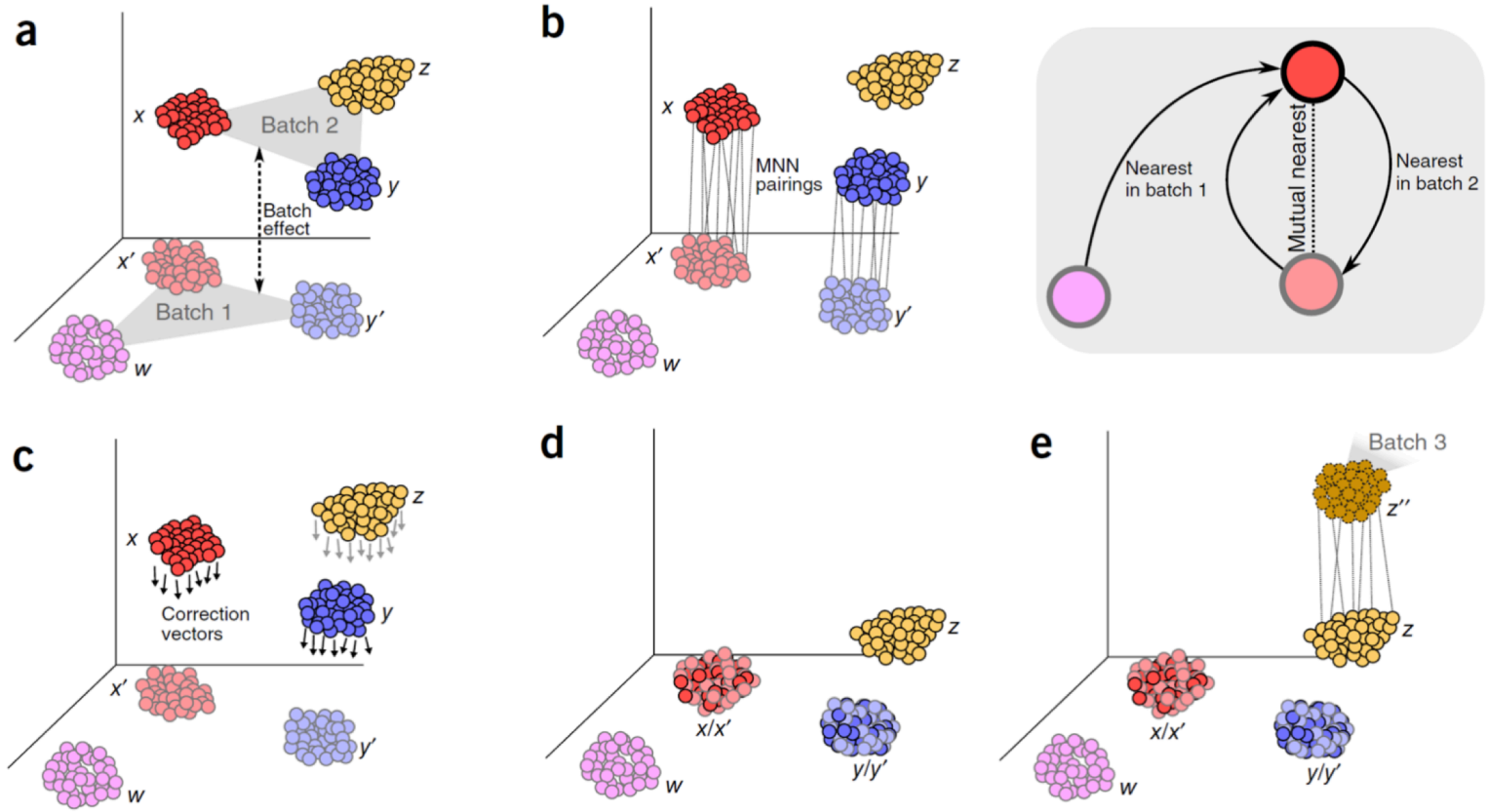
Büttner, et al (2019) *Nat Methods*

scRNA-seq – data integration

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

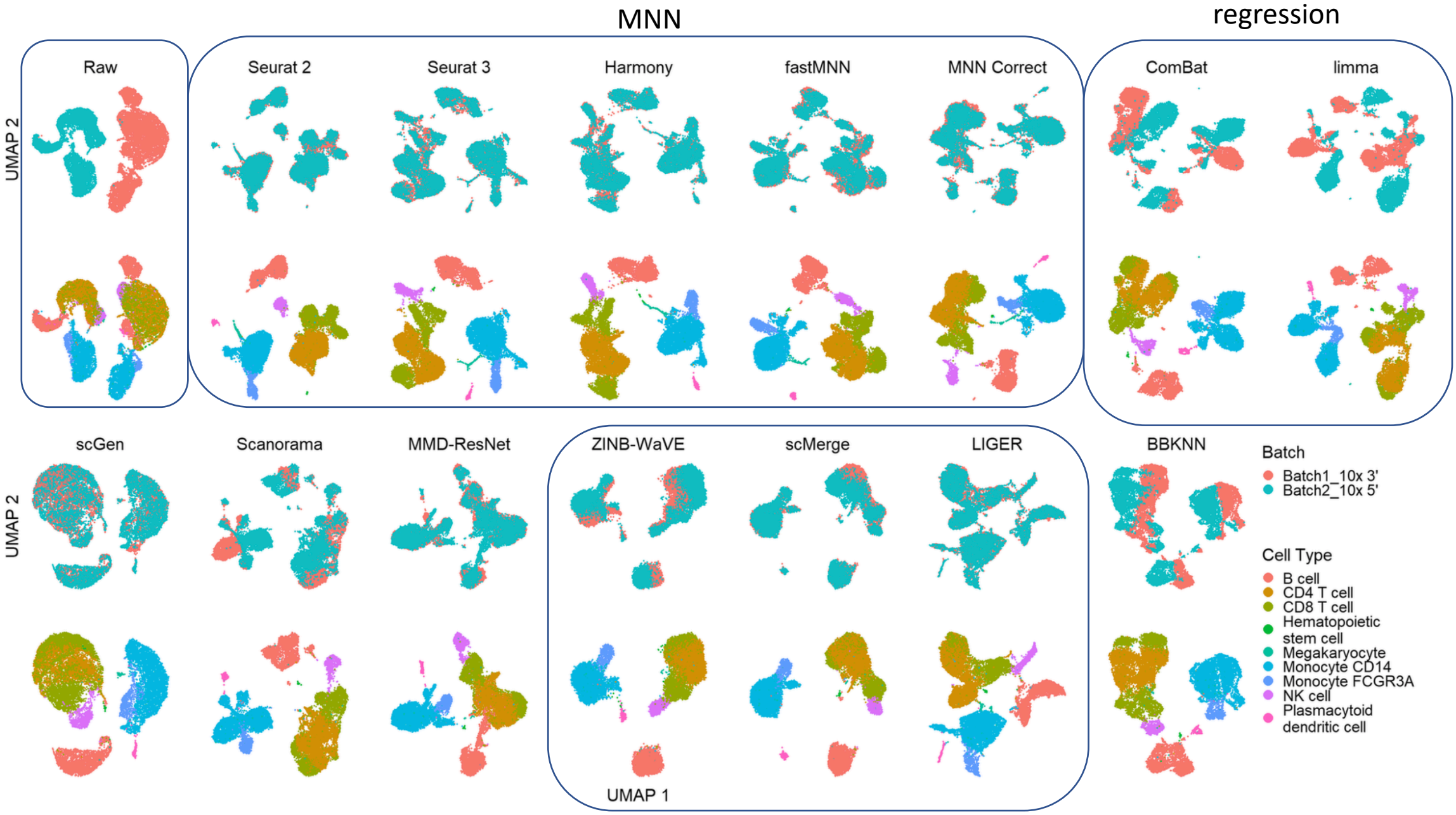
Modern data integration methods are based on the same principle:

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



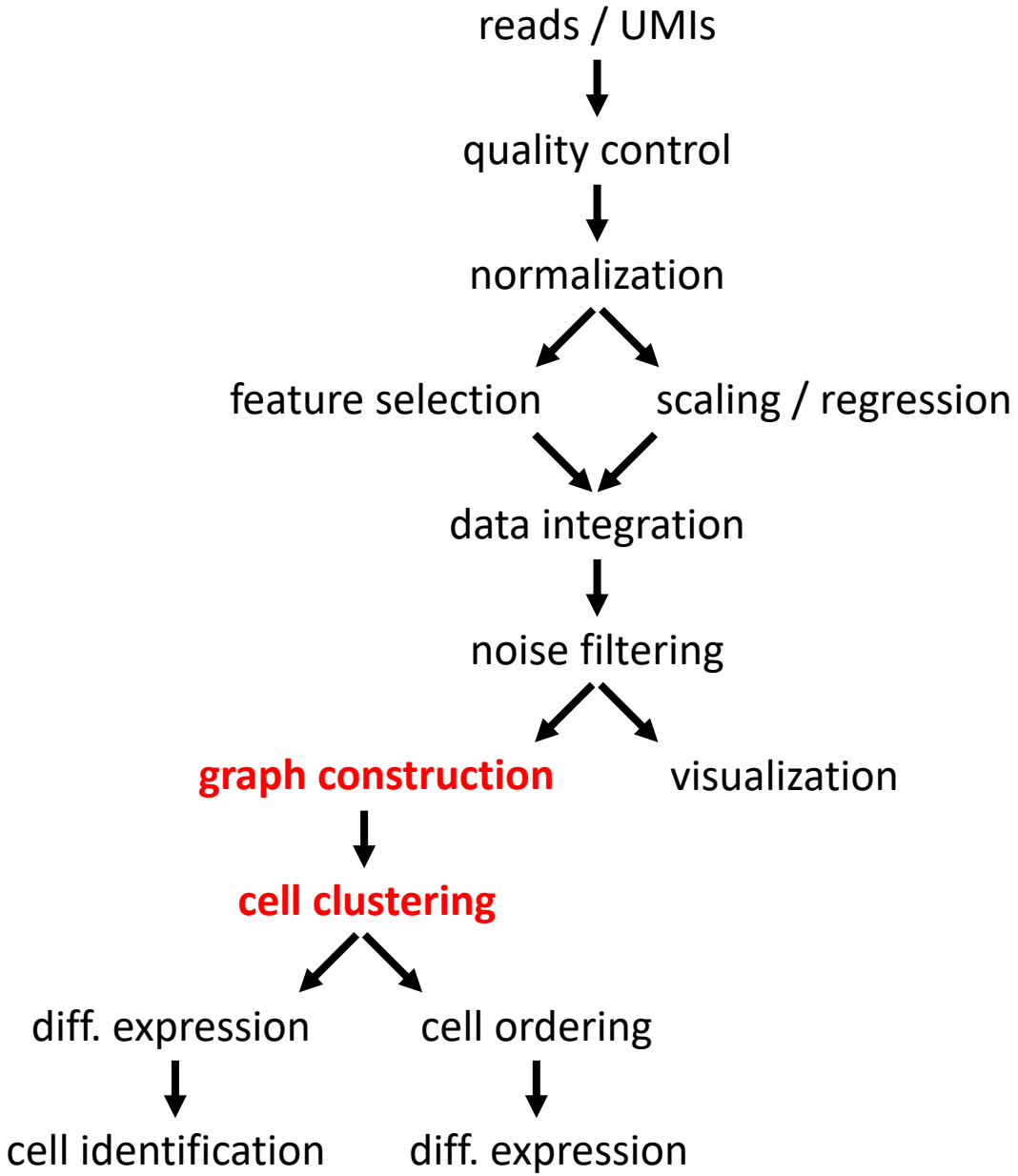
Haghverdi et al (2017) Nat Biotechnology

scRNA-seq - data integration

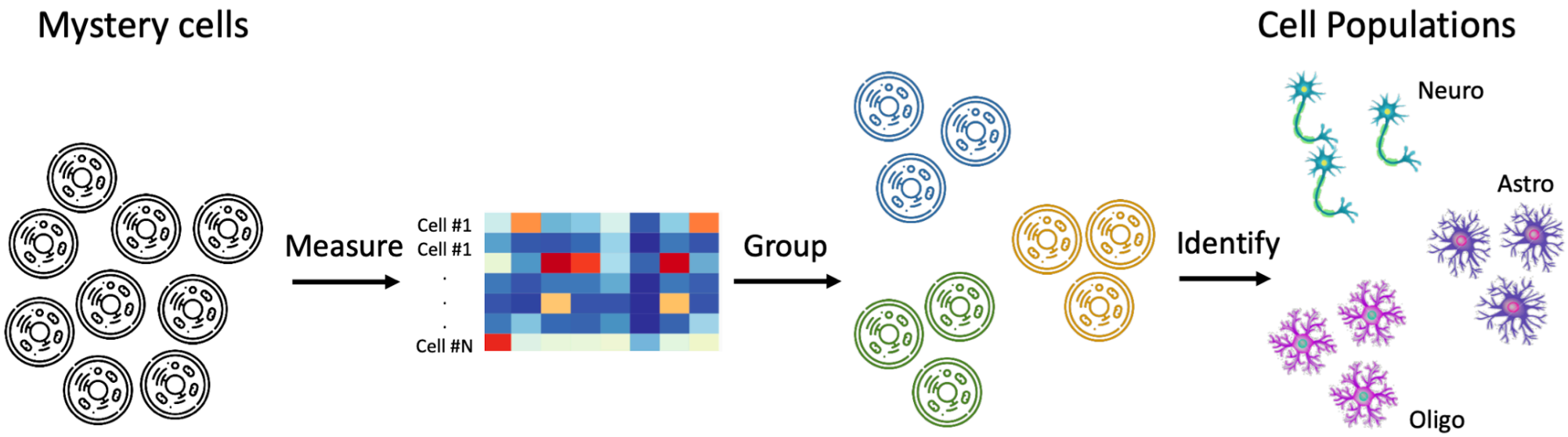


Tran et al (2020) *Genome Biology*

scRNA-seq analysis workflow

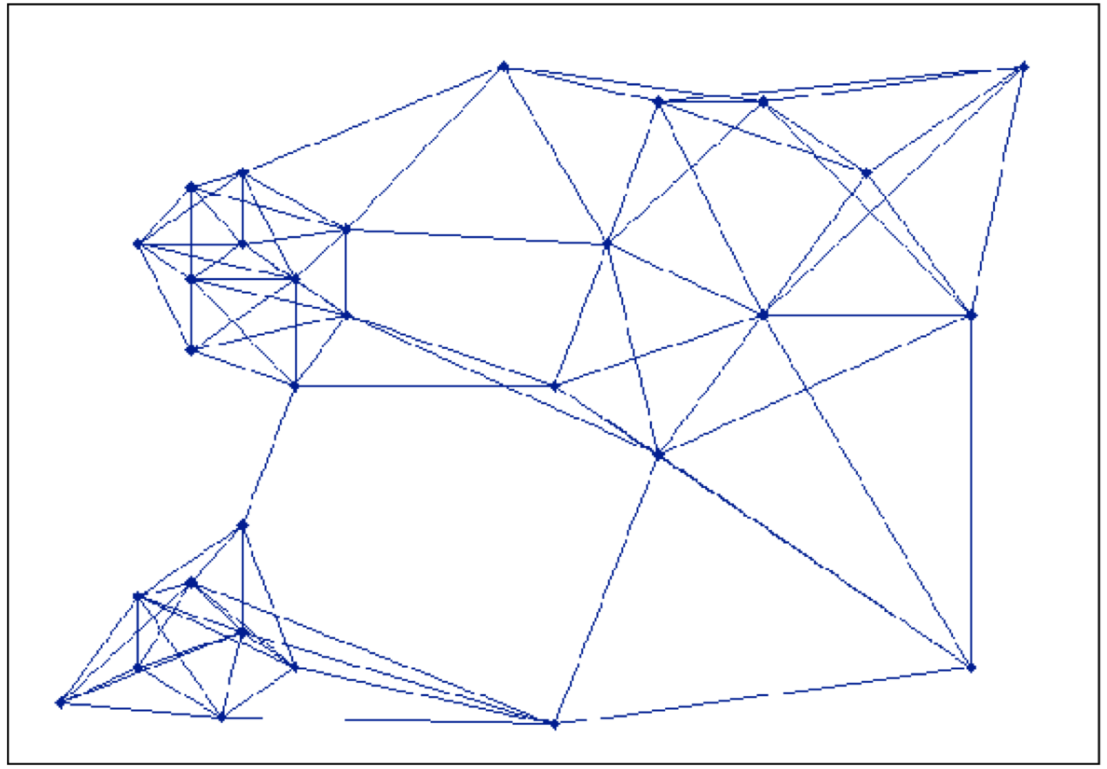


scRNA-seq – graph construction and clustering

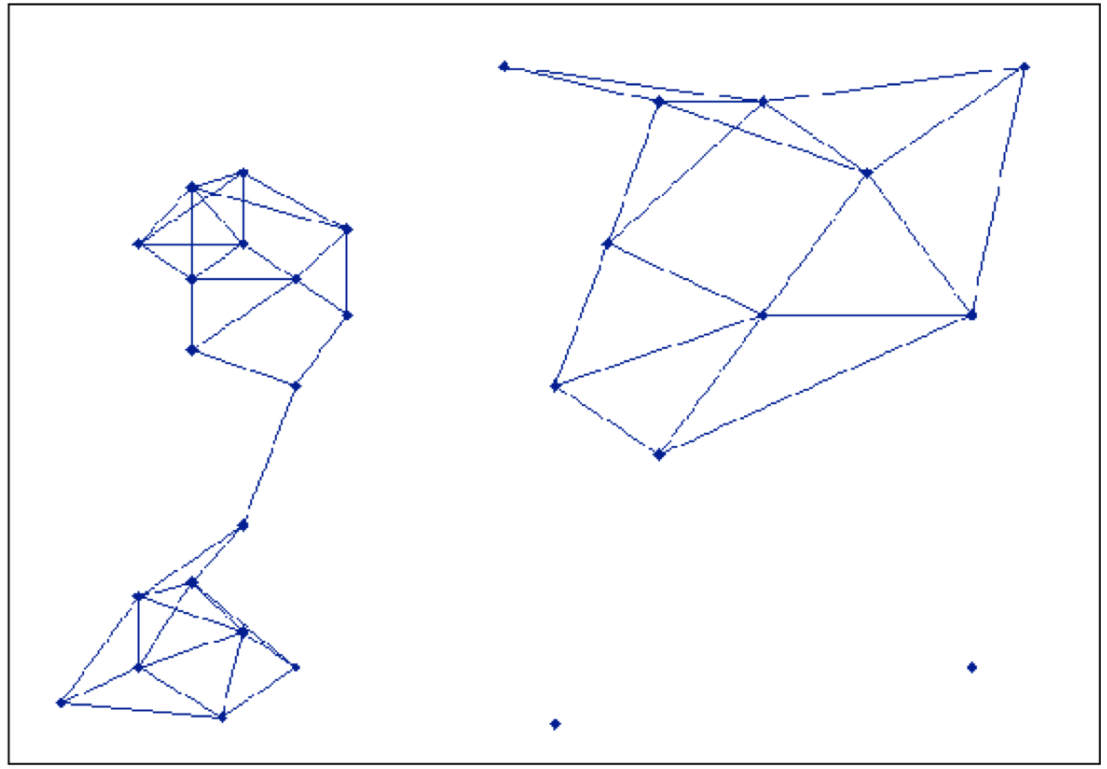


scRNA-seq – graph construction and clustering

The ***k*-Nearest Neighbor** (***k*NN**) graph is a graph in which two vertices *p* and *q* are connected by an edge, if the distance between *p* and *q* is among the *k*-th smallest distances from *p* to other objects from *P*.

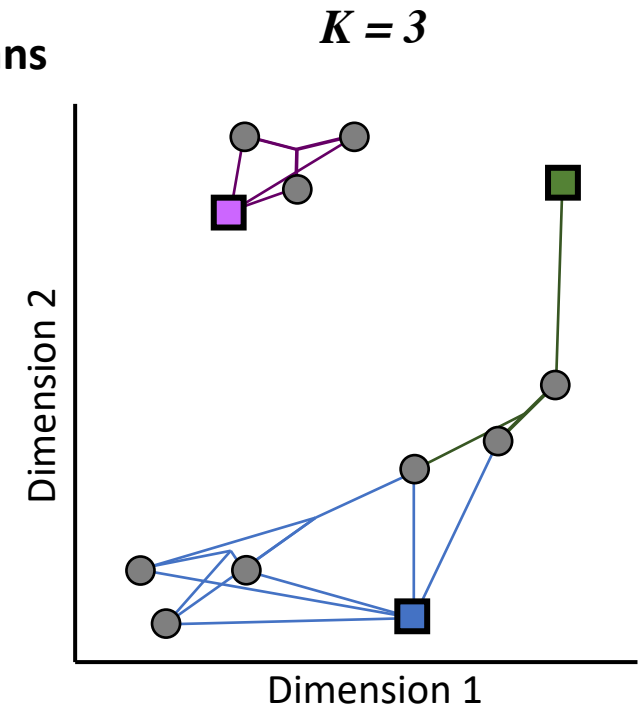


The **Shared Nearest Neighbor** (**SNN**) graph has weights that defines proximity, or similarity between two edges in terms of the number of neighbors (i.e., directly connected vertices) they have in common.

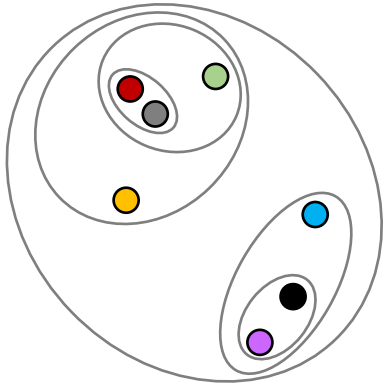


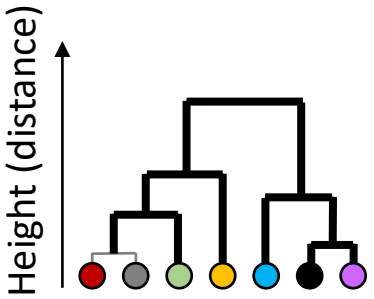
scRNA-seq – graph construction and clustering

K-means



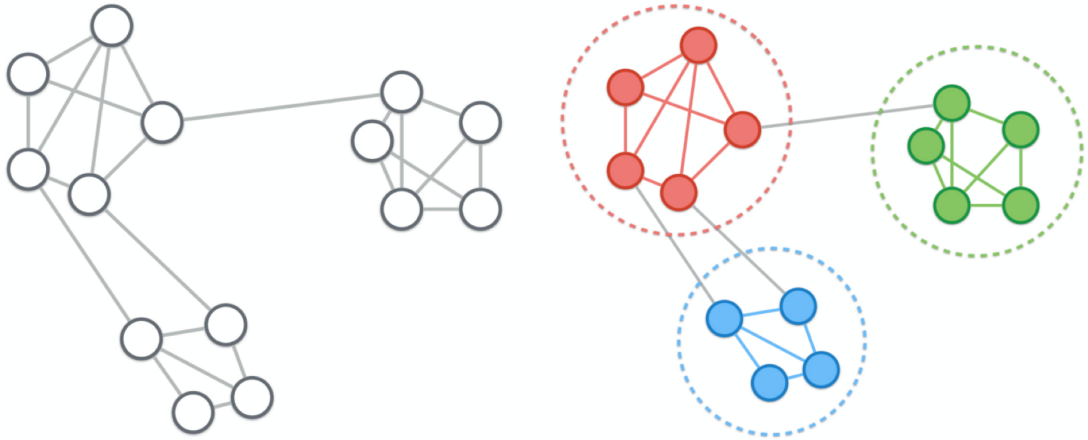
Hierarchical Clustering



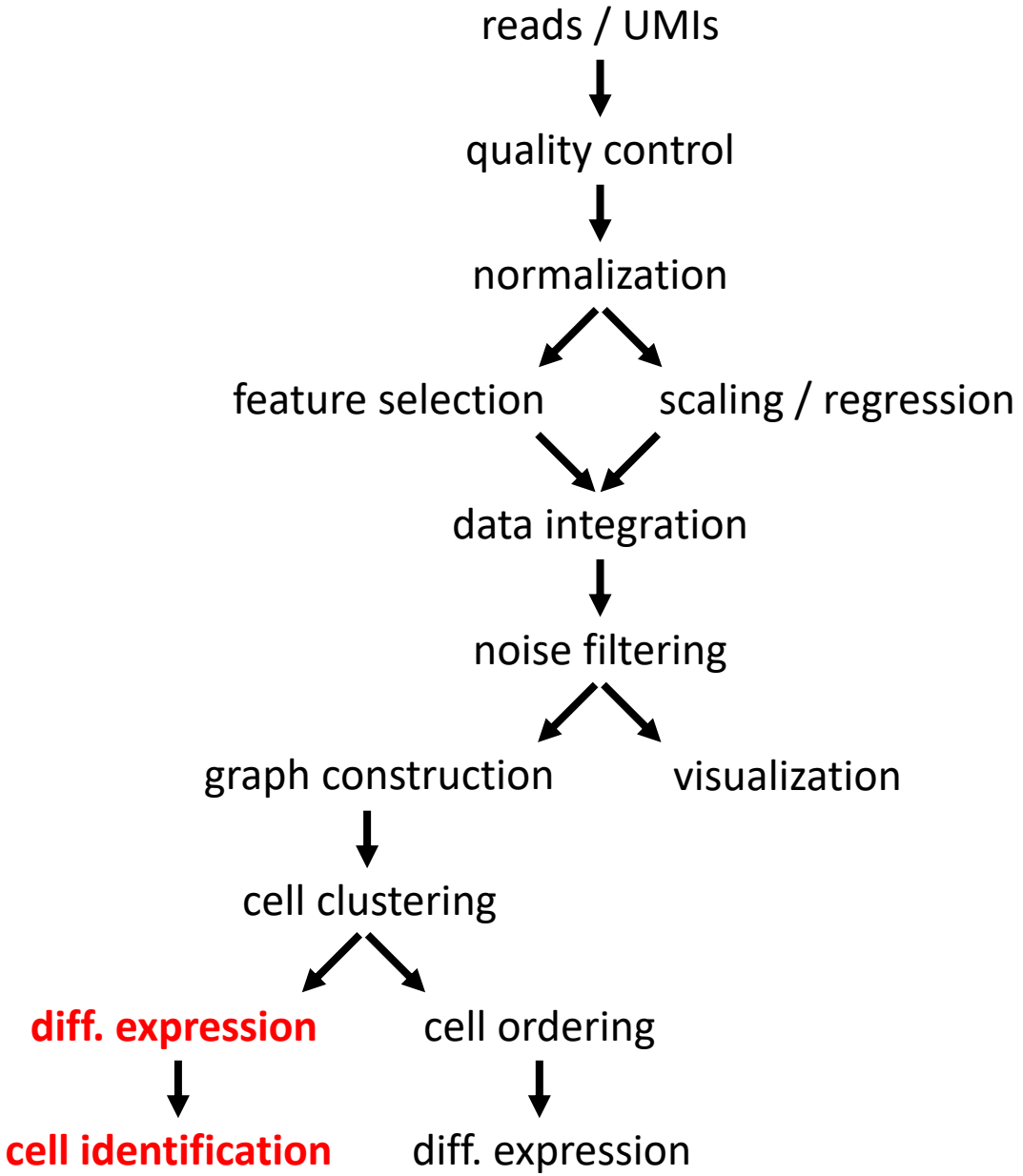


GRAPH
Louvain / Leiden community detection

Communities, or clusters, are usually groups of vertices having higher probability of being connected to each other than to members of other groups.



scRNA-seq analysis workflow



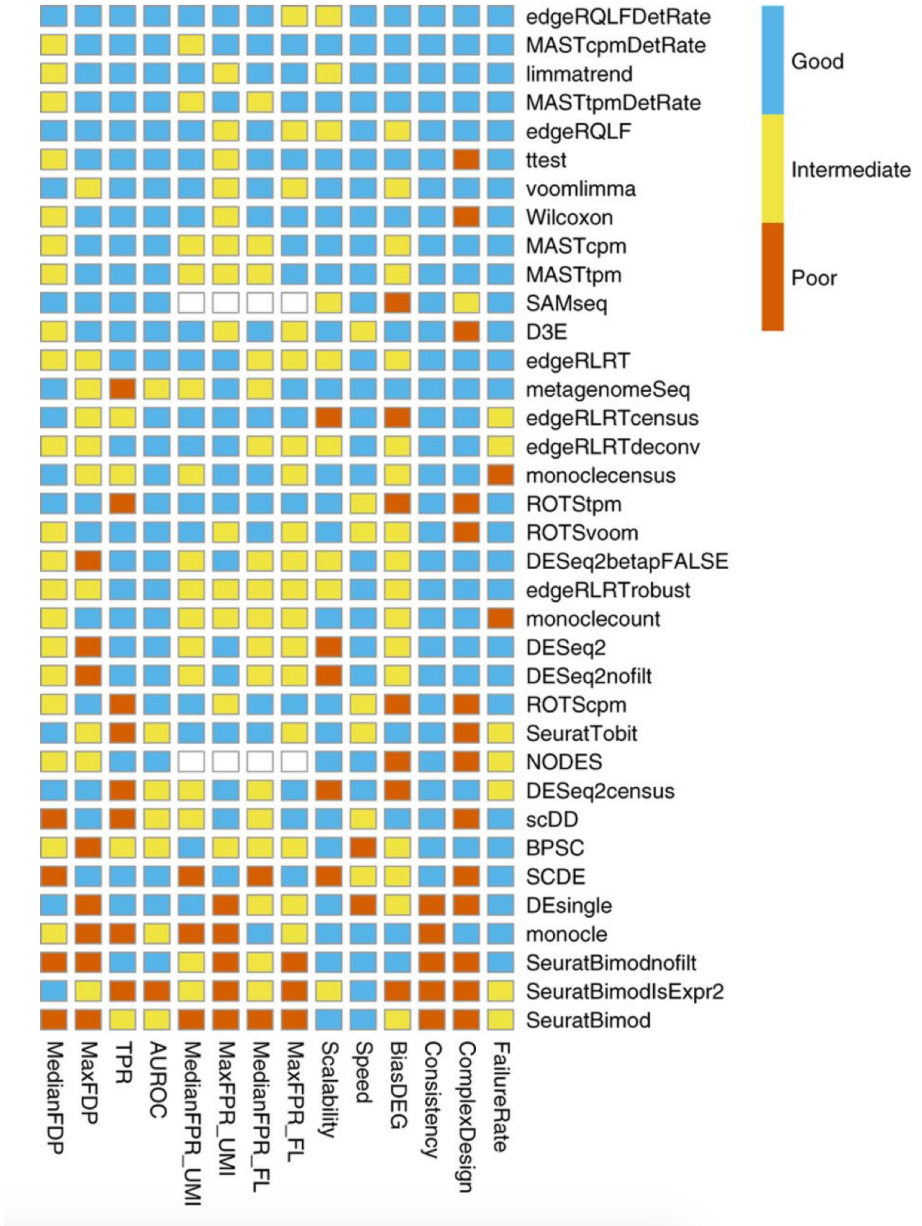
scRNA-seq - differential gene expression

Typically we have more than two clusters in a data set

For a given cluster, are we interested in “marker genes” that are:

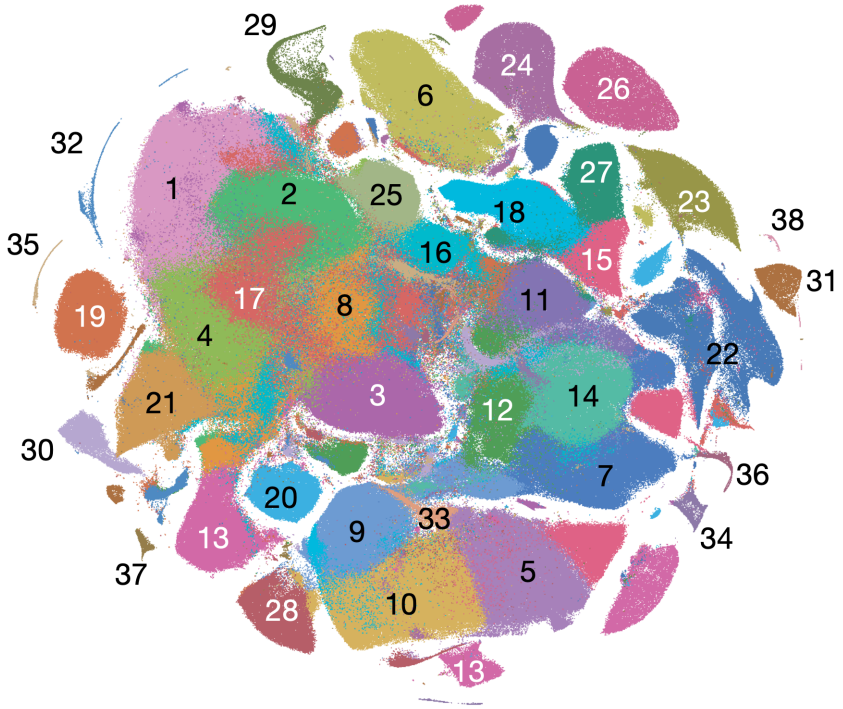
- **DE compared to all cells outside of the cluster (most common)**
- DE compared to at least one other cluster
- DE compared to *each* of the other clusters
- DE compared to “most" of the other clusters

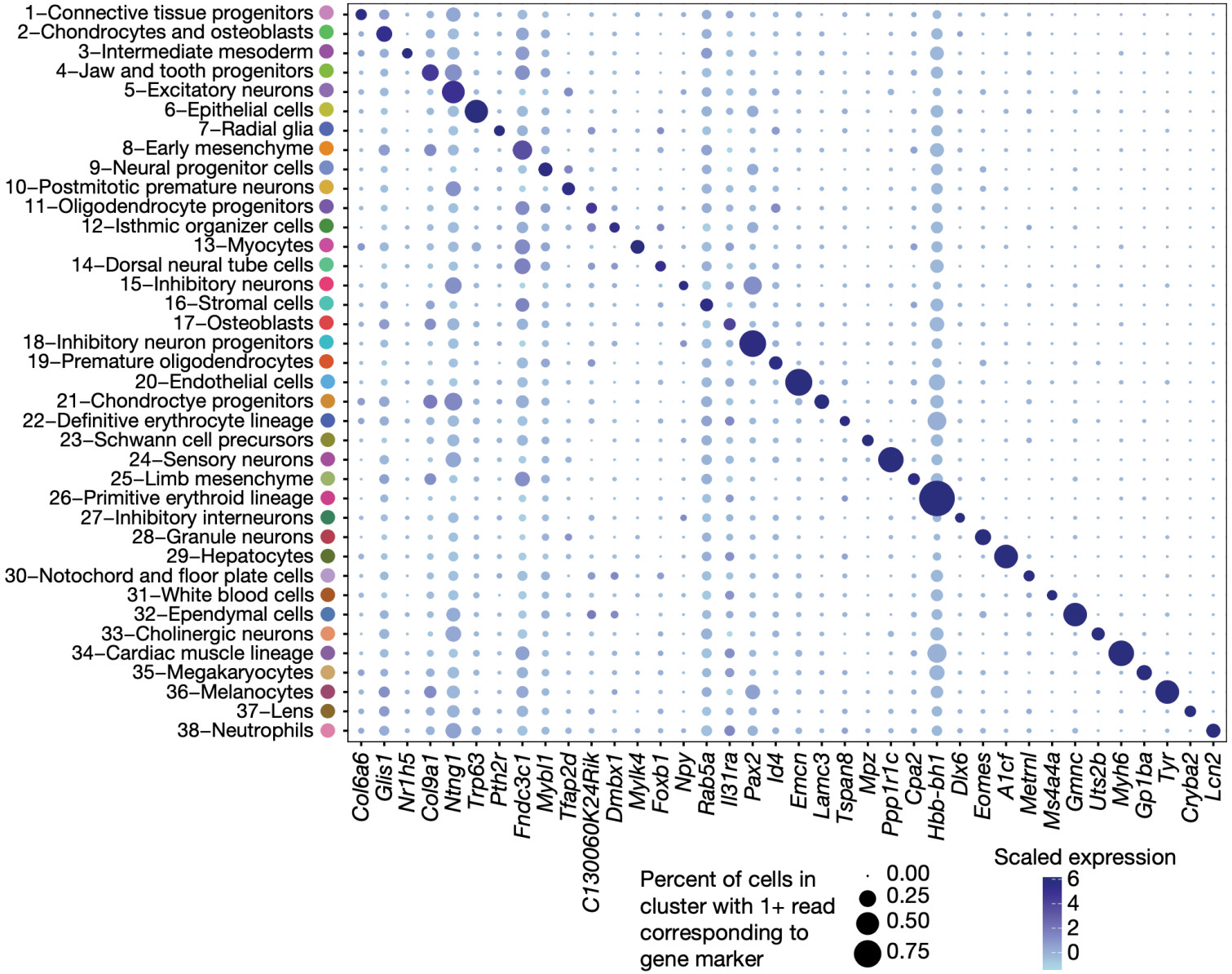
Even the t-test and the Wilcoxon test work well (assuming that you have at least a few dozen cells to compare)



Soneson et al 2018 *Nat Methods*

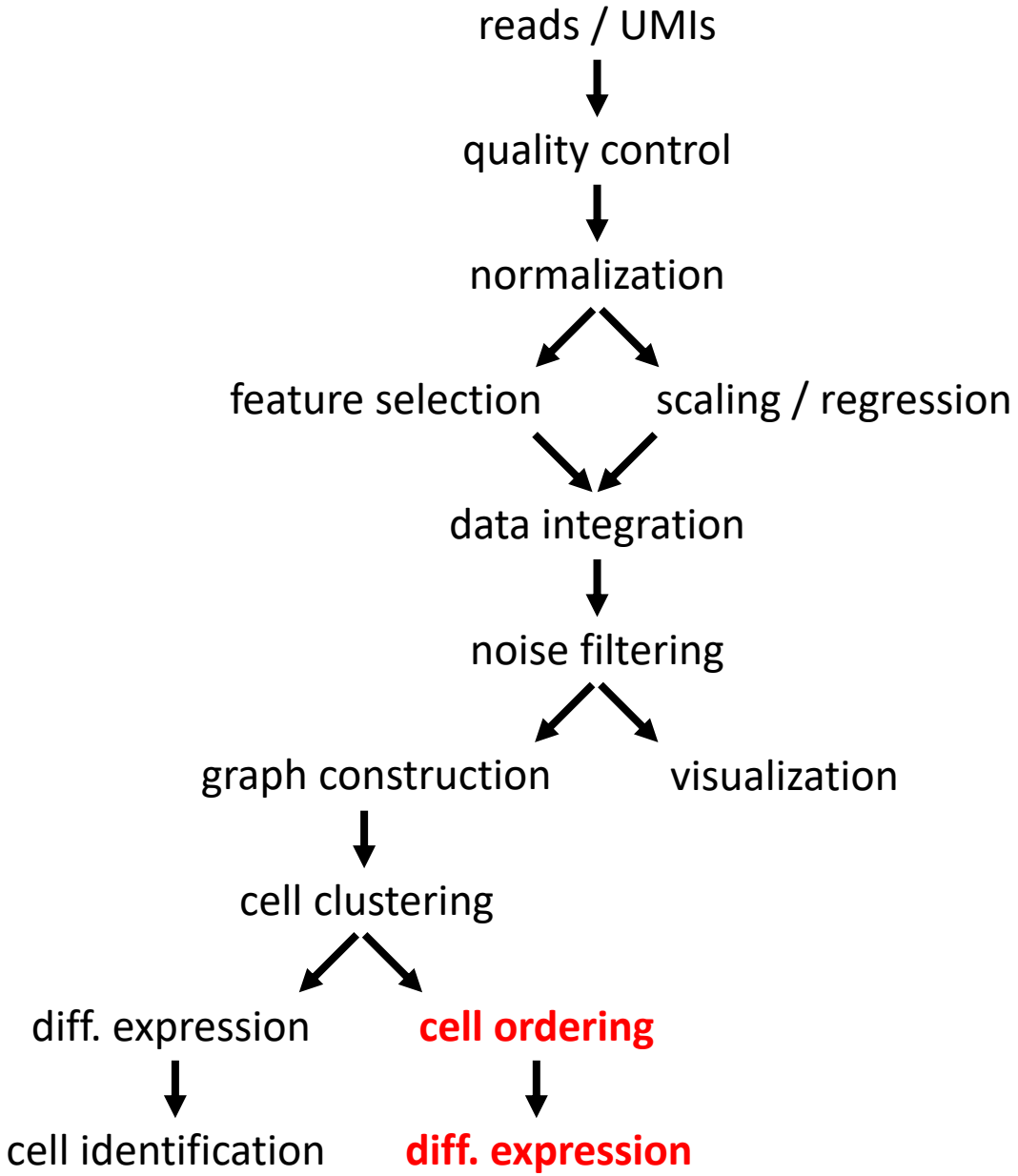
scRNA-seq - differential gene expression



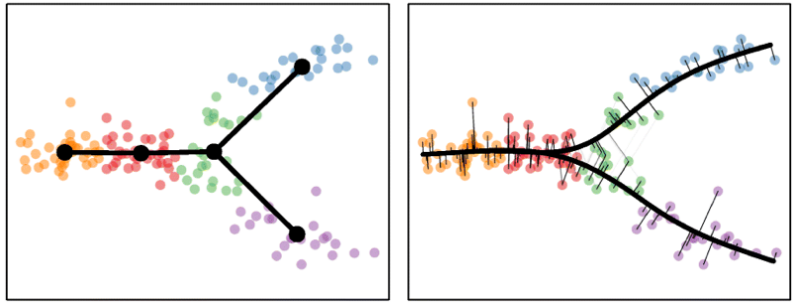


Cao et al 2019 *Nature*

scRNA-seq analysis workflow



scRNA-seq – trajectory inference



Street et al (2018) *BMC Genomics*

Are you sure that you have a developmental trajectory?

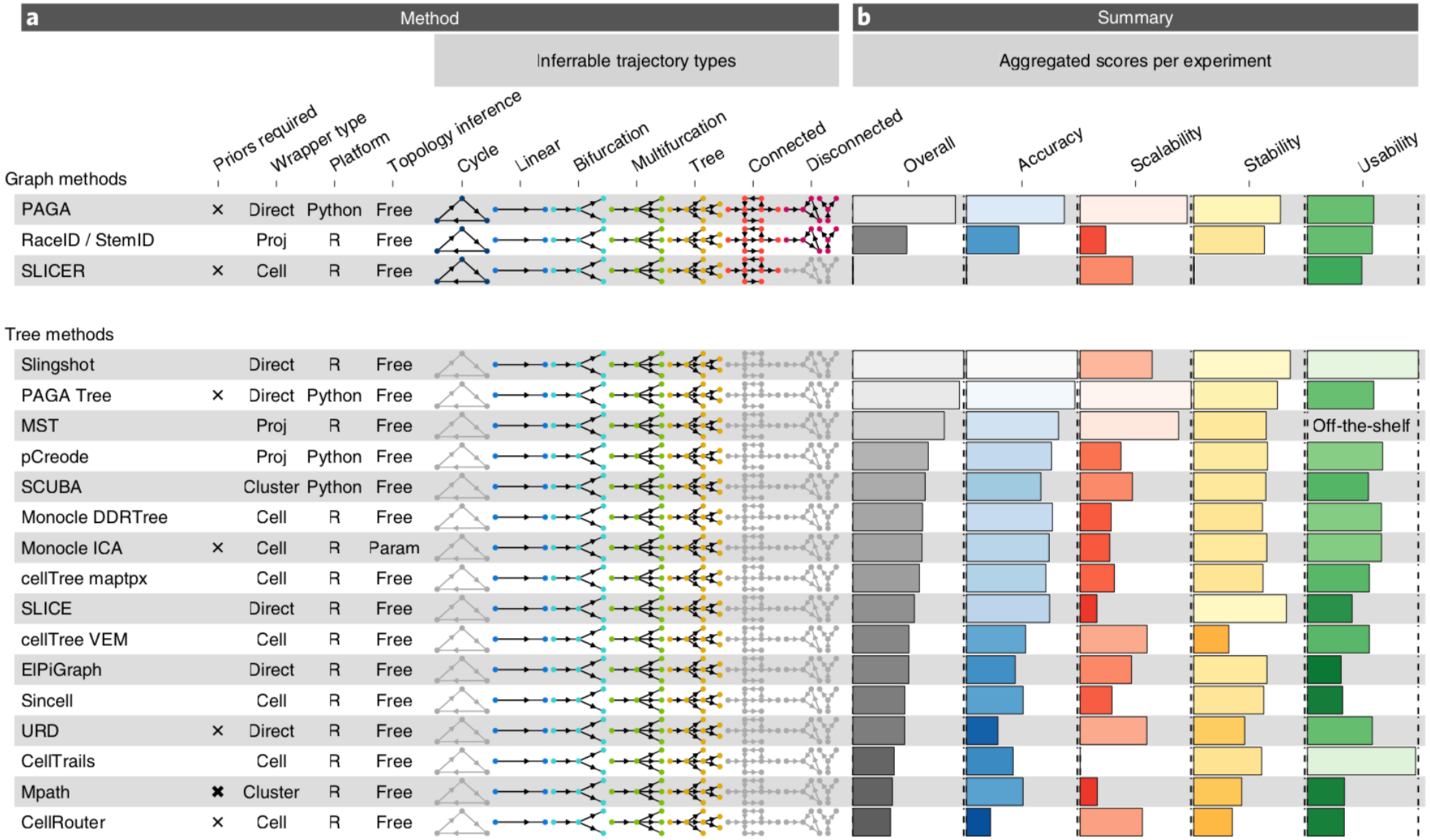
Do you have intermediate states?

Do you believe that you have branching in your trajectory?

❗ Be aware, any dataset can be forced into a trajectory without any biological meaning!

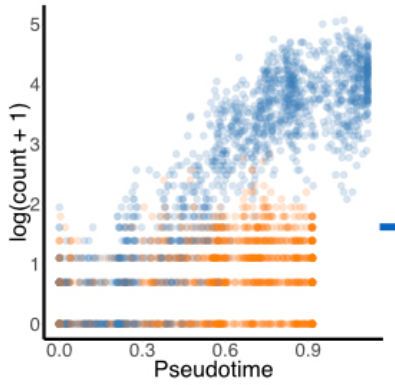
❗ First make sure that gene set and dimensionality reduction captures what you expect.

scRNA-seq – trajectory inference methods

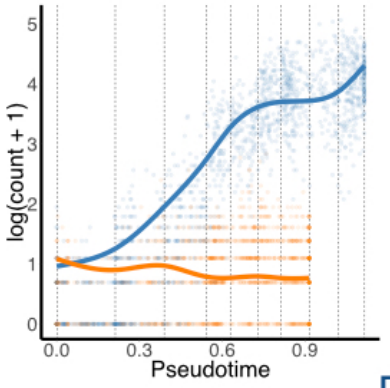


Saelens et al 2019 *Nat Biotechnology*

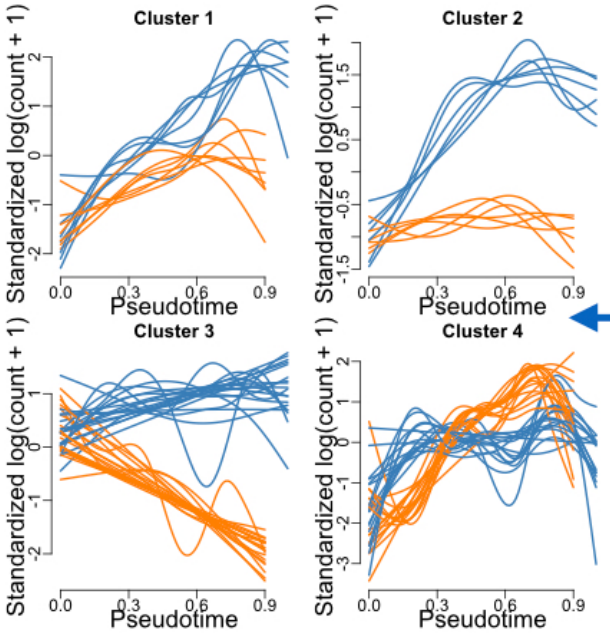
scRNA-seq - differential gene expression in trajectories



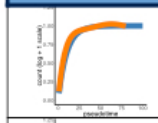
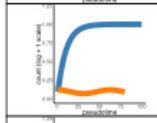
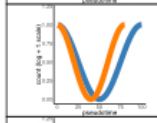
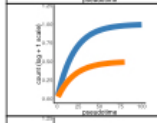
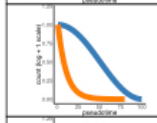
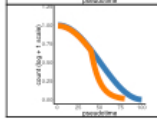
fitGAM



Clustering



Differential Expression Tests

	Within the orange lineage		Between the orange and blue lineages		
Lineages	**association Test**	**startVsEnd Test**	**diffEnd Test**	**pattern Test**	**earlyDE Test**
	DE	DE	Not DE	Not DE	Not DE
	Not DE	Not DE	DE	DE	DE
	DE	Not DE	Not DE	Not DE	Not DE
	DE	DE	DE	DE	Not DE
	DE	DE	Not DE	DE	DE
	DE	DE	Not DE	DE	Not DE

Berge et al 2020 *Nat Communication*

scRNA-seq analysis workflow