

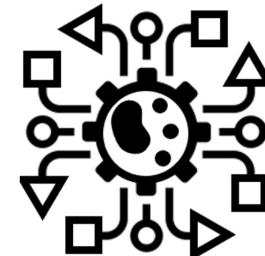
Advanced Topics in Single Cell Omics



Spatial
transcriptomics



RNA
velocity

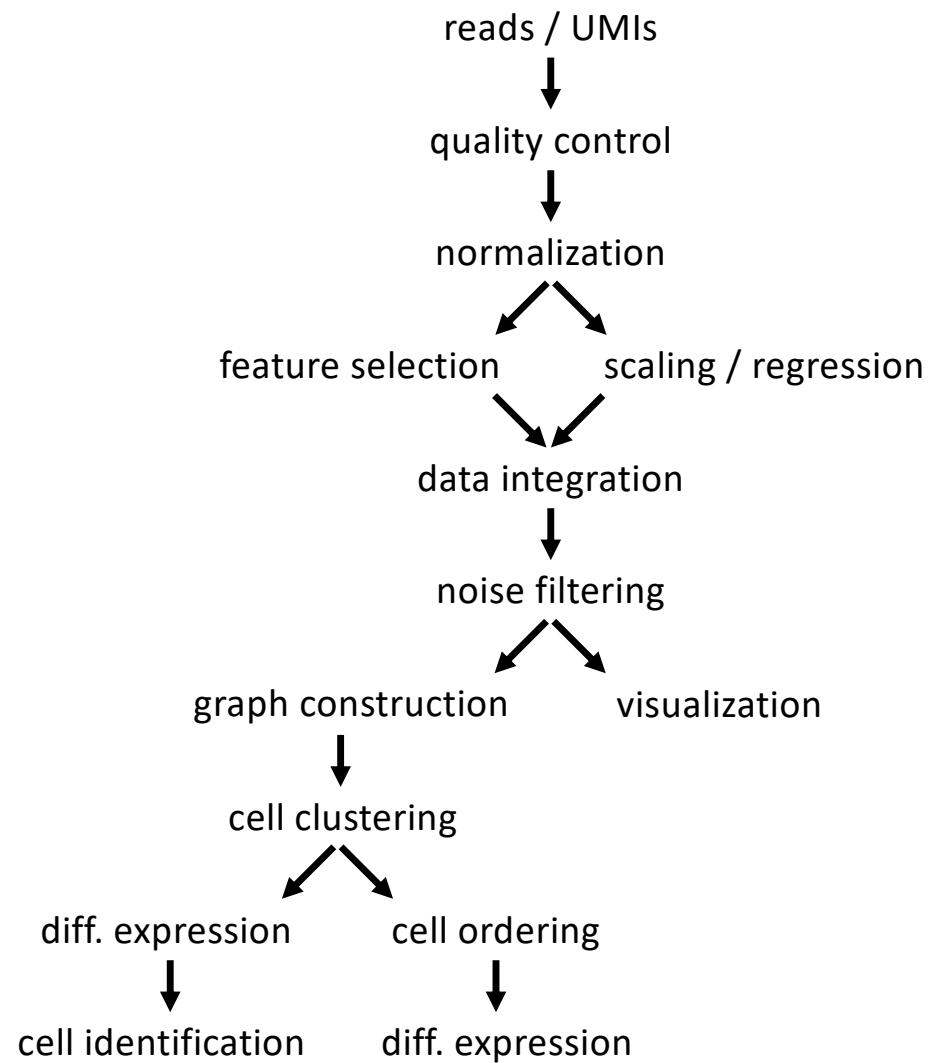


Omics
integration

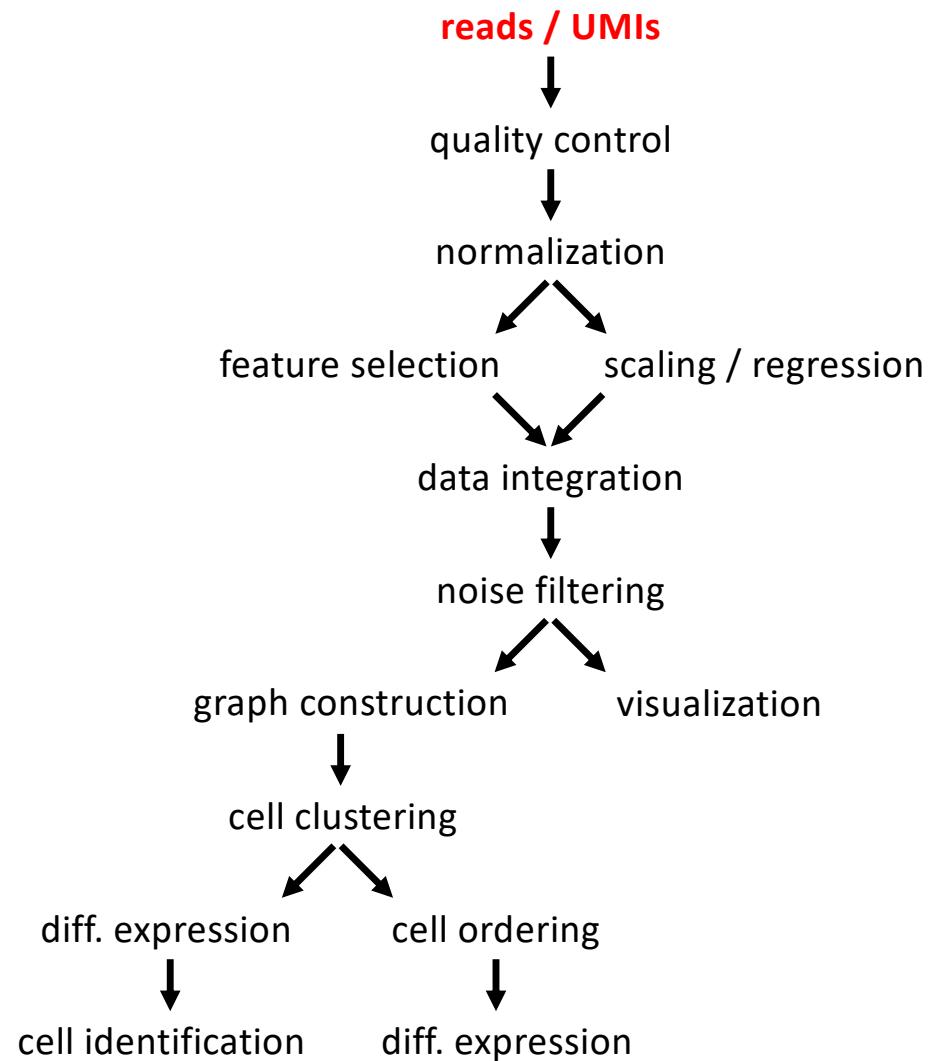


Deep
learning

scRNA-seq analysis workflow

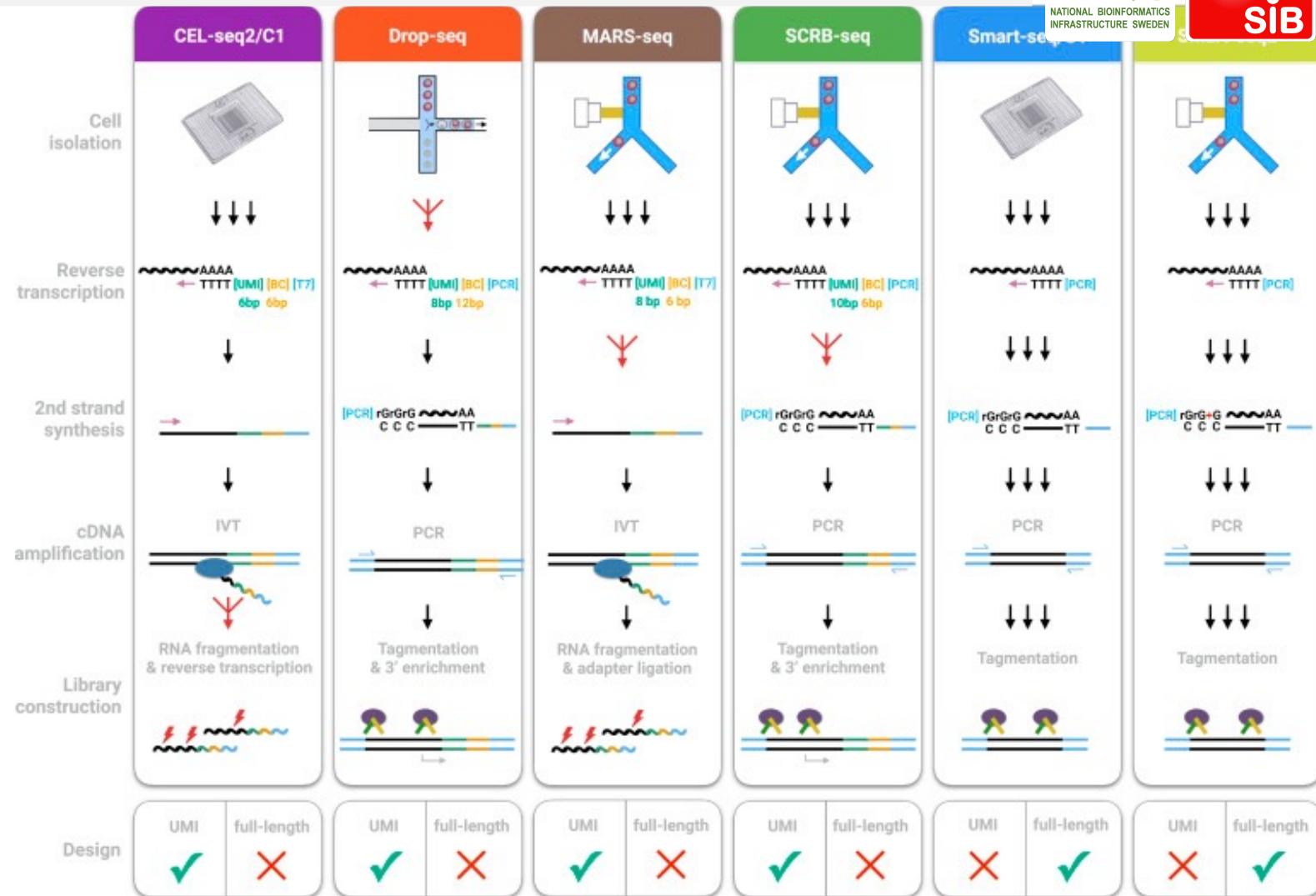
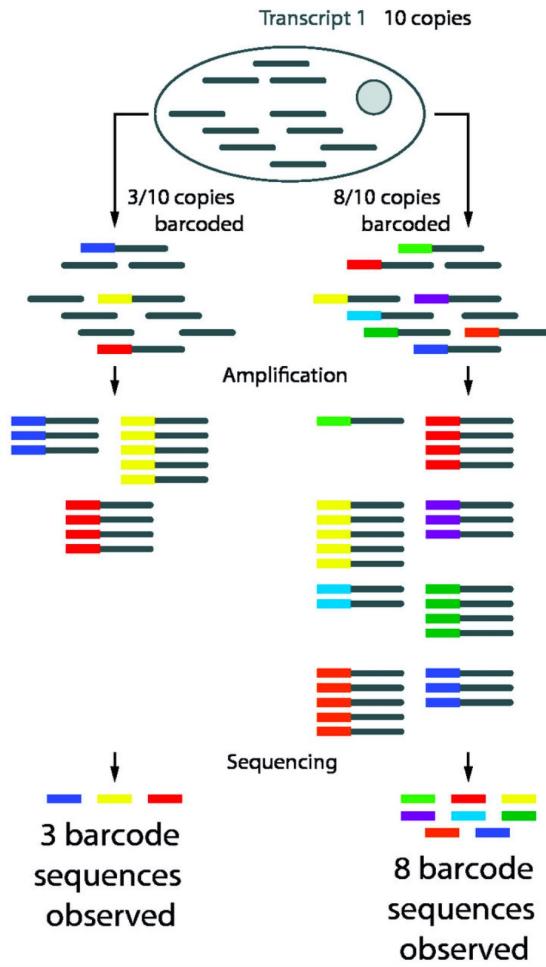


scRNA-seq analysis workflow



scRNA-seq technologies

UMI (unique molecular identifier)

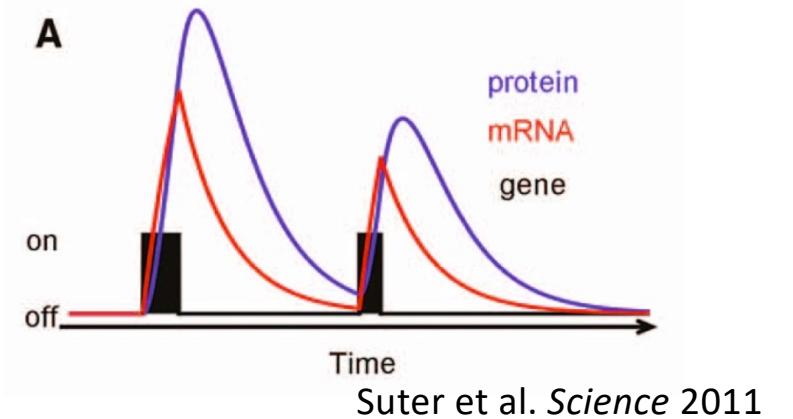


Zajac et al (2013) *Plos One*

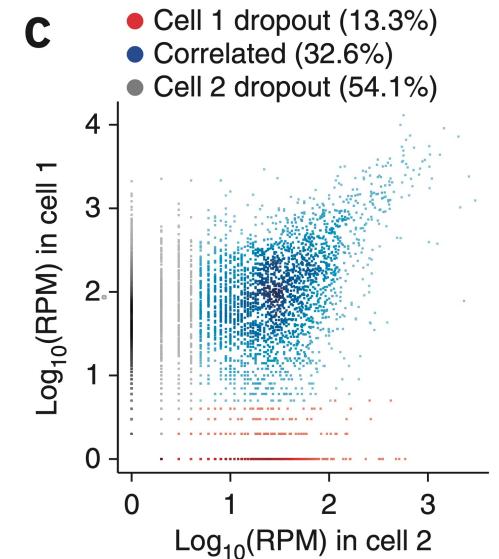
Ziegenhain et al (2015) *Molecular Cell*

scRNA-seq biases

- 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

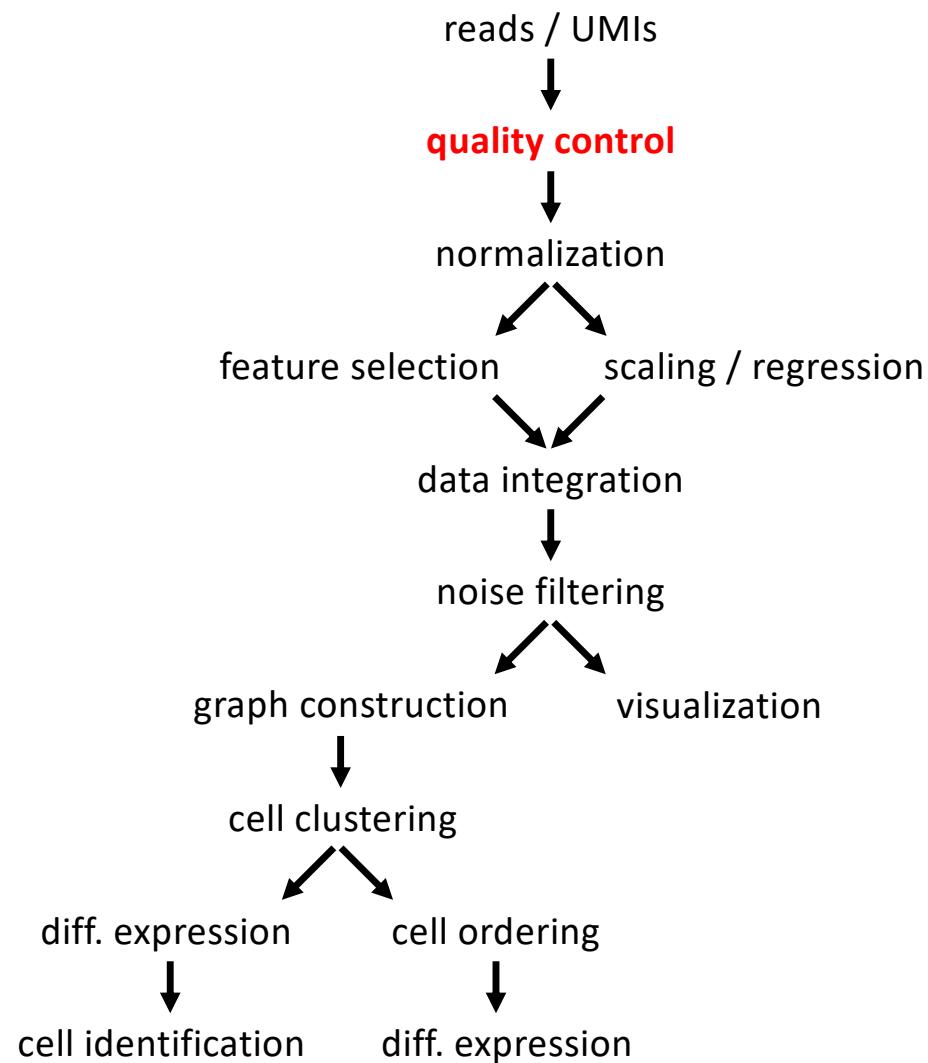


Suter et al. *Science* 2011



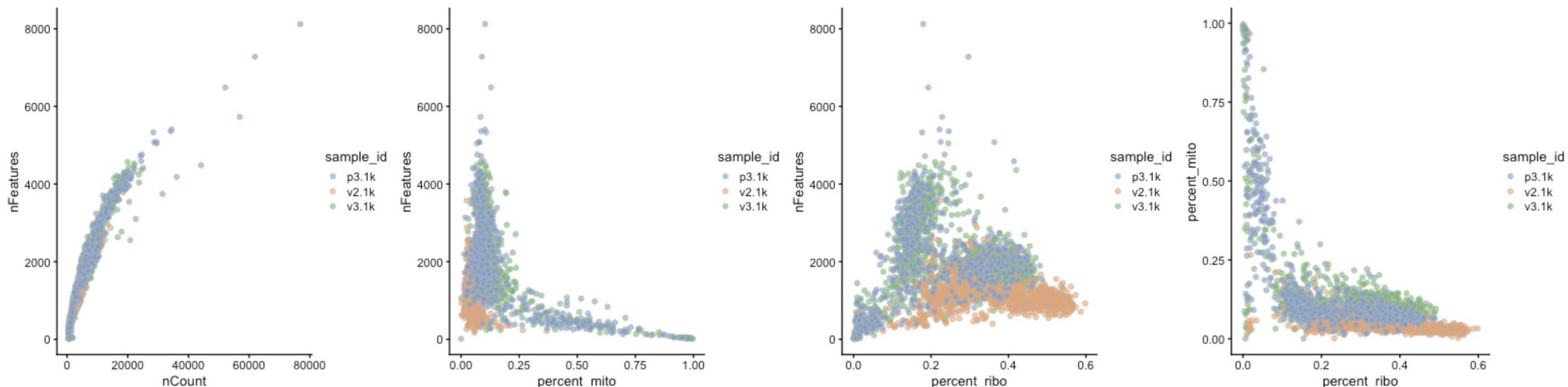
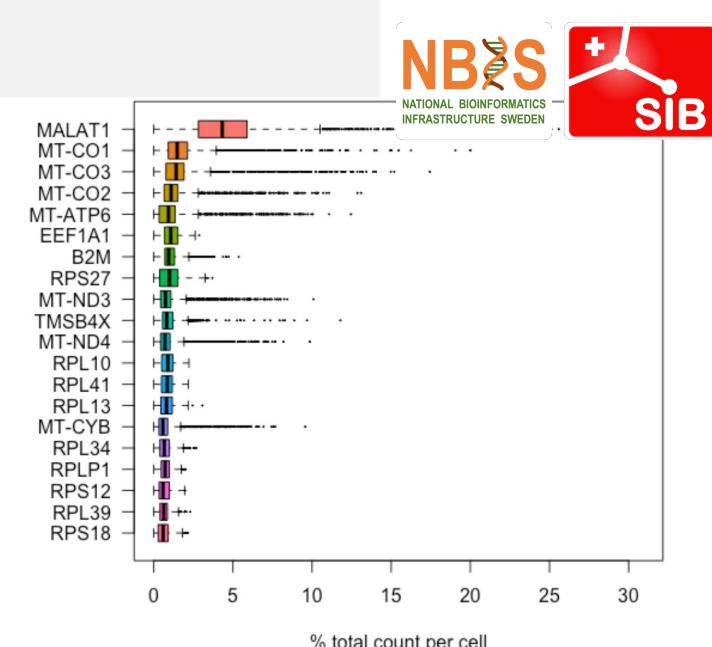
Karchenko et al. *Nature Methods* 2014

scRNA-seq analysis workflow

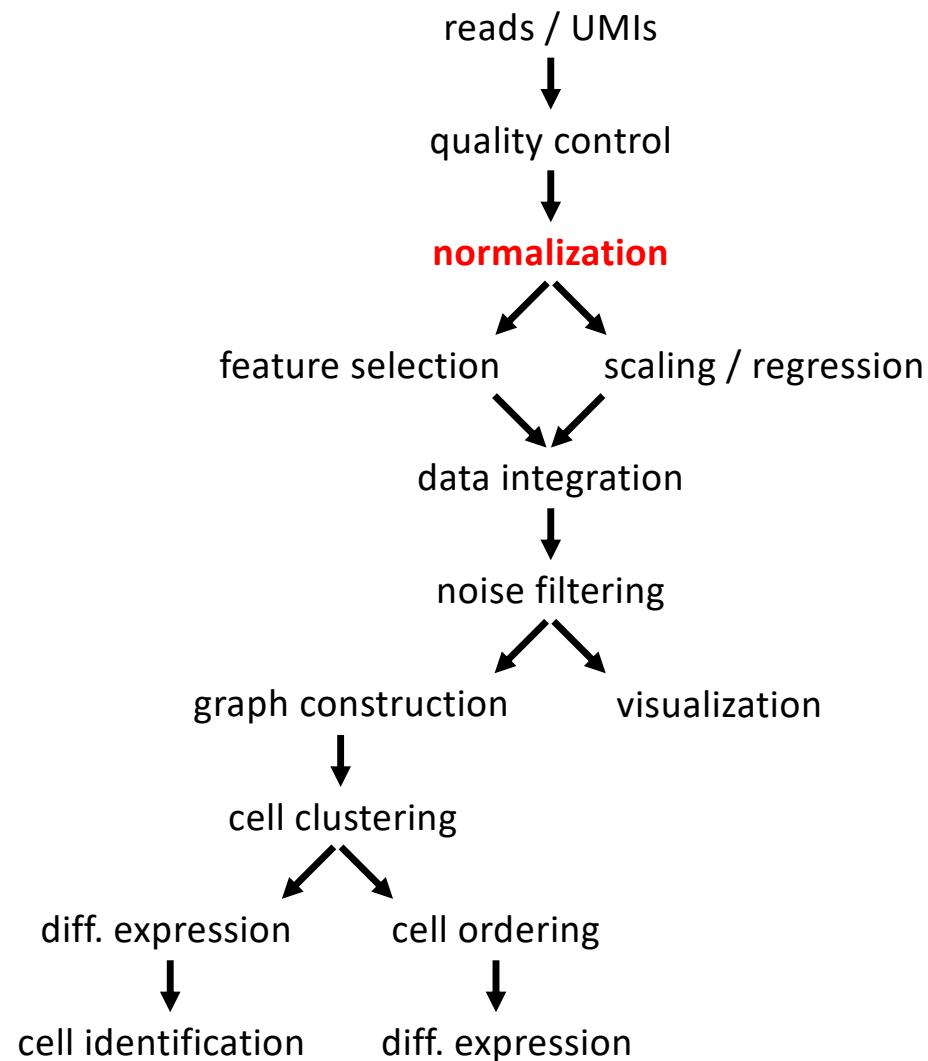


scRNA-seq quality control

- 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 - $\log[CP10K+1]$

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

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

single-cell

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

Most common for UMI data / fast

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

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

scRNA-seq analysis workflow



Count normalization (UMI and read counts)

for uneven sequencing depth

- CPM - $\log[CP10K+1]$

Gene length normalization (read counts)

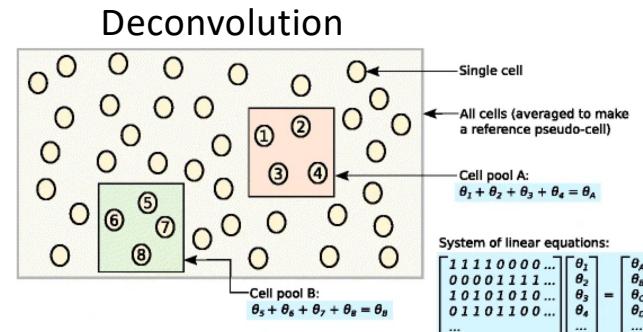
for differences in gene detection due to gene length

- TPM (closer to UMI counts)
- FPKM

Drop-out rate normalization (UMI and read counts)

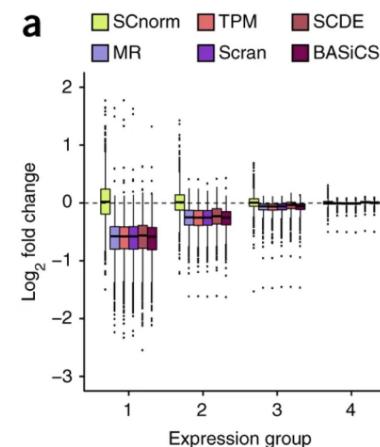
for differences in RNA content / drop-out rates

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



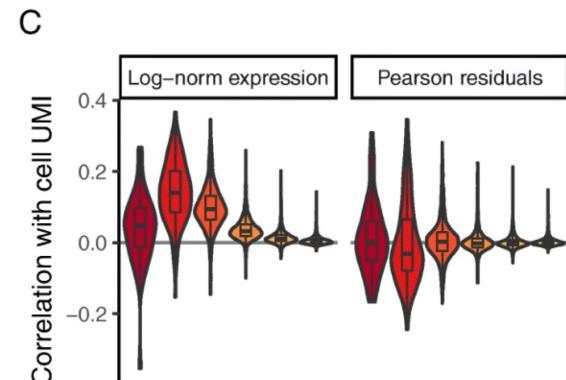
Lun et al. Genome Biol. 2016

SCnorm



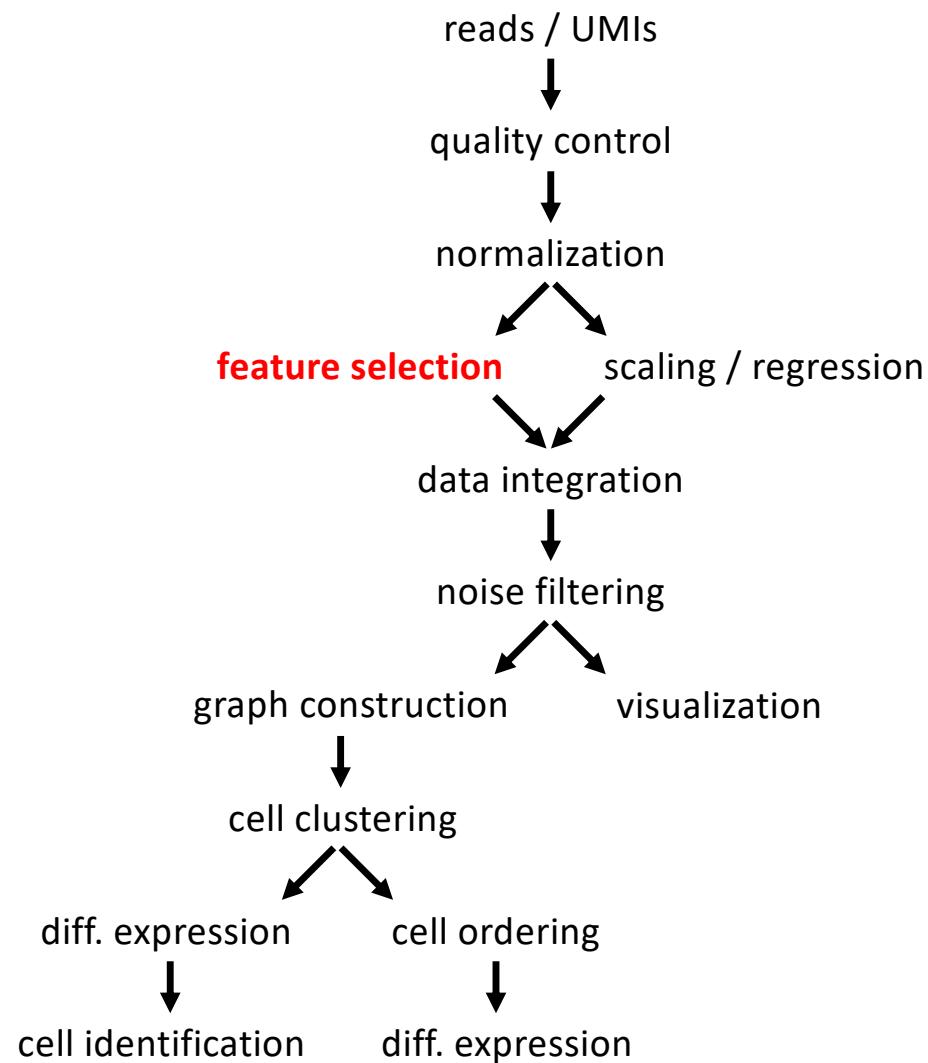
Bacher et al. Nature Methods 2017

SCTransform



Hafmeister &
Satija Genome Biology 2019

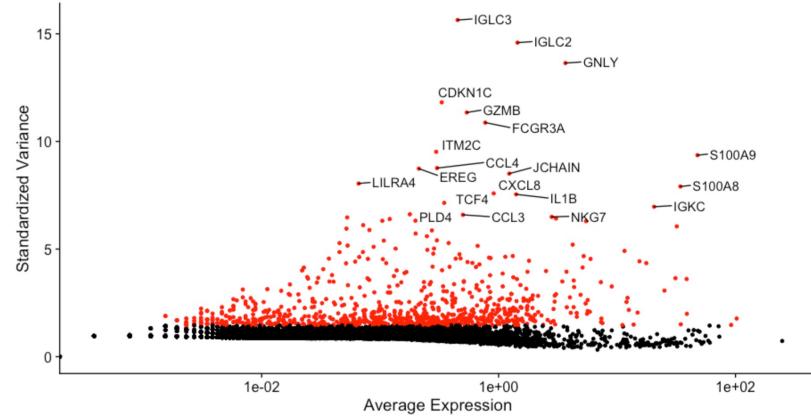
scRNA-seq analysis workflow



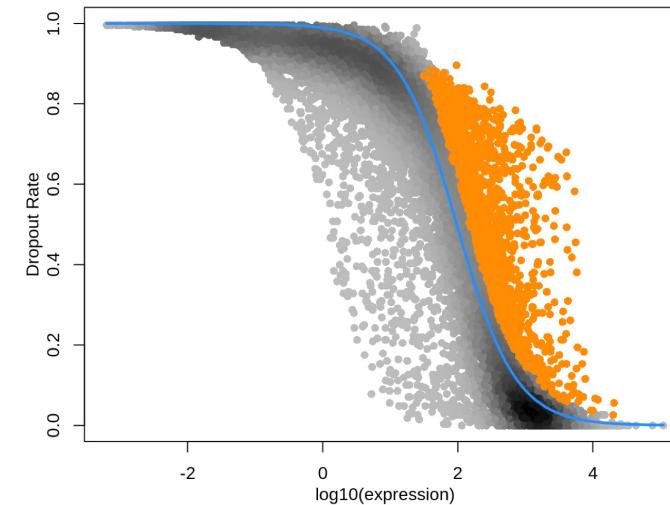
scRNA-seq feature selection

Not all genes are important to define your cell types

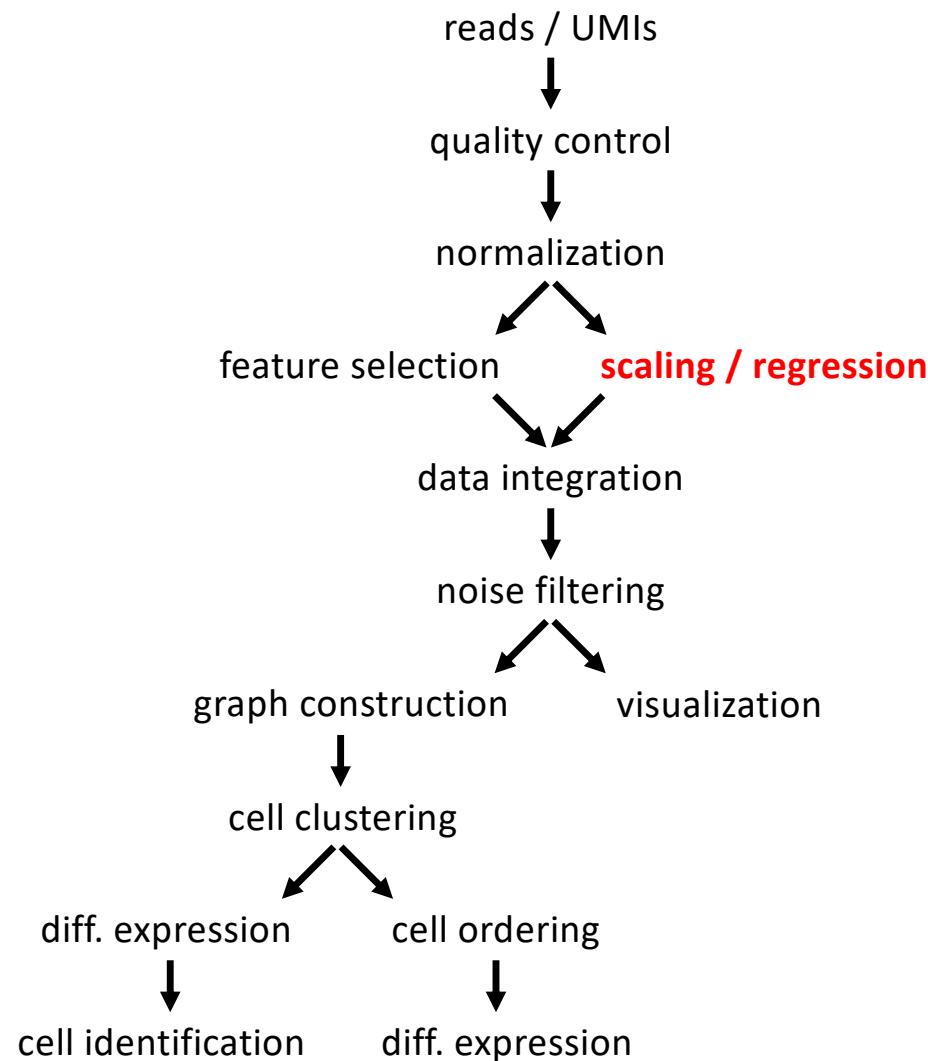
$$HVG = \frac{variance}{\log(meanExpression)}$$



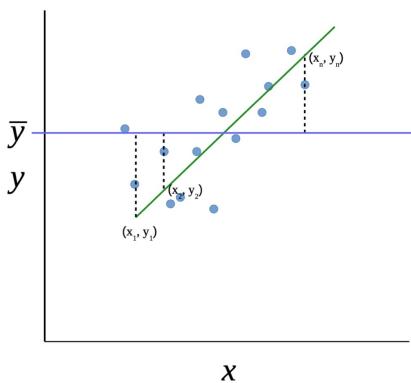
$$HVG = \frac{\log(meanExpression)}{dropout_{rate}}$$



scRNA-seq analysis workflow



scRNA-seq scaling and regression of biases



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

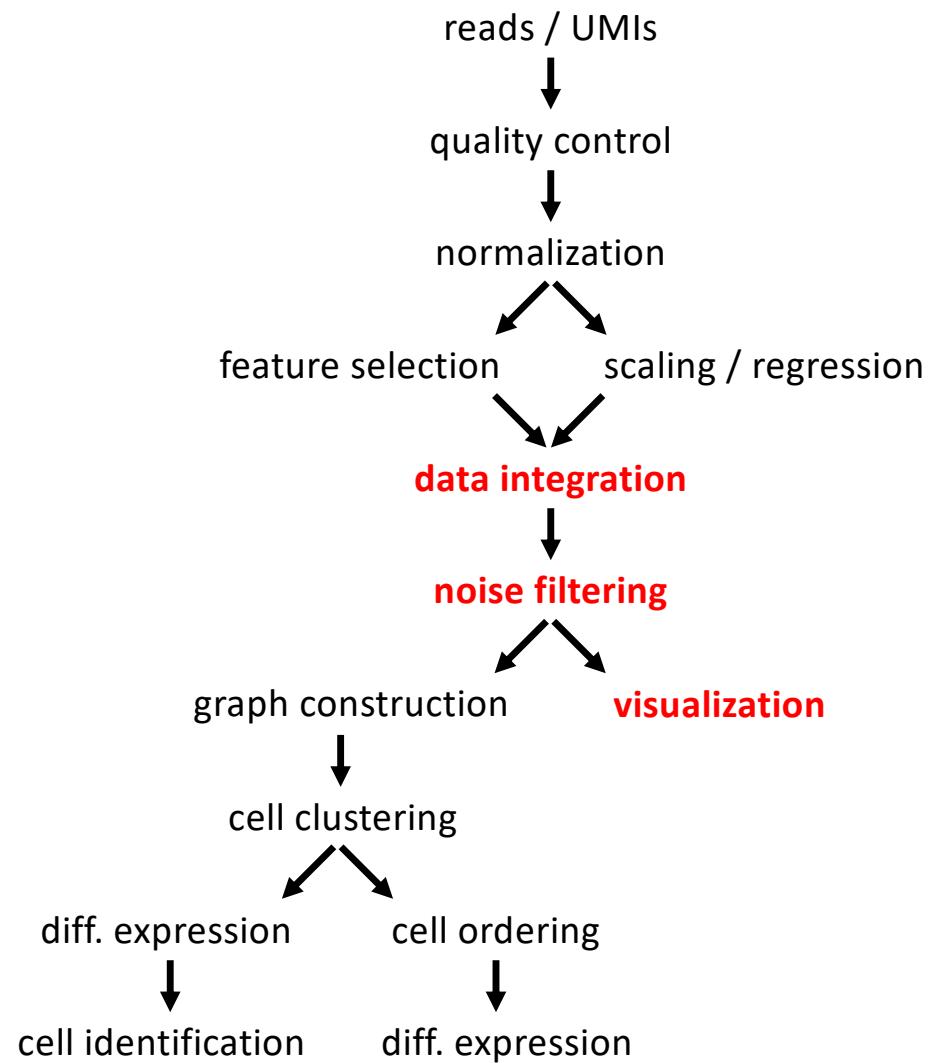
- Fit a line to the gene expression *vs* variable of interest
- Calculate residuals
- Remove variance explained by the variable of interest by taking the residuals.
- Linear / Negative Binomial / Poisson distributions

logNormalized counts follows a log-linear distribution

↓
fast

↓
slower (but more accurate)
Regressing counts directly is better with count-based distributions

scRNA-seq analysis workflow

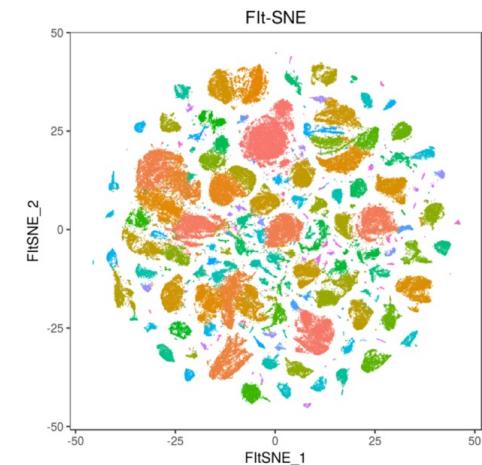
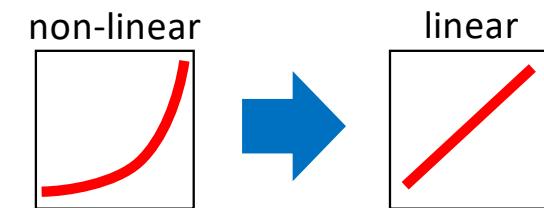
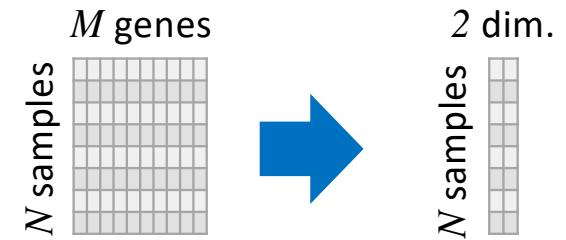


scRNA-seq dimensionality reduction

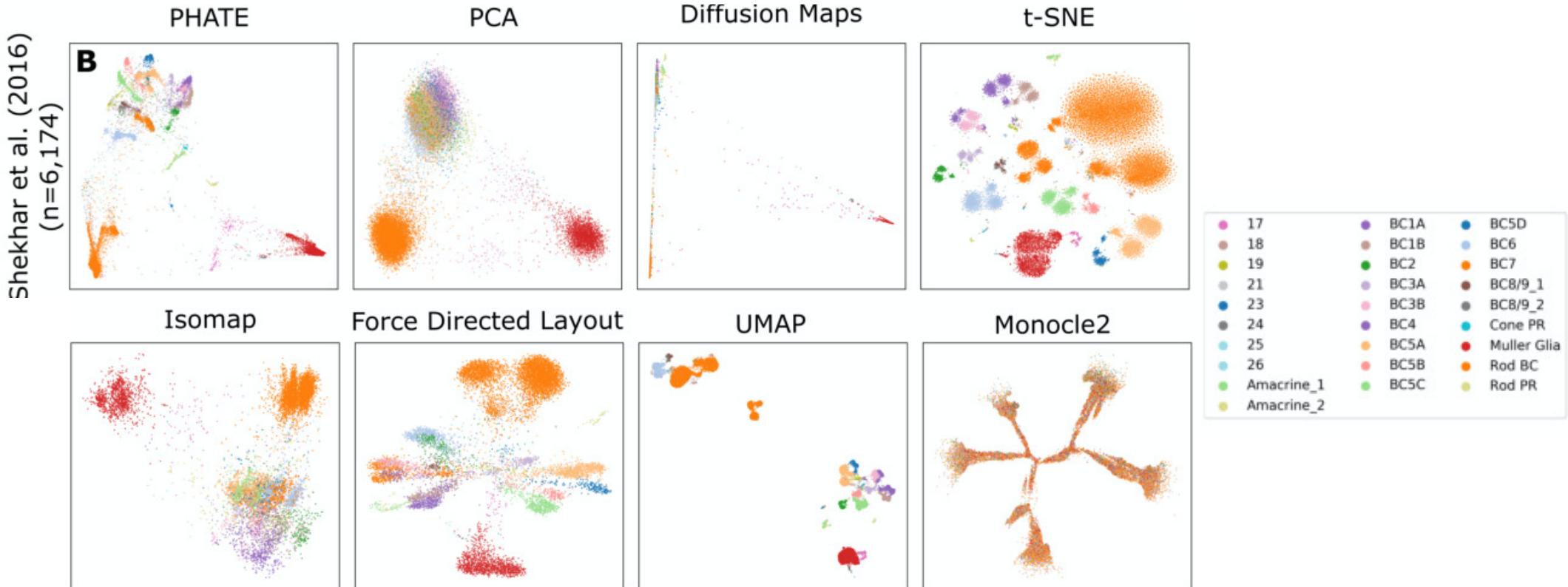


- Simplify complexity, so it becomes easier to work with.
Reduce number of features (genes)
In some: Transform non-linear relationships to linear
- “Remove” redundancies in the data
- Identify the most relevant information (find and filter noise)
- Reduce computational time for downstream procedures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

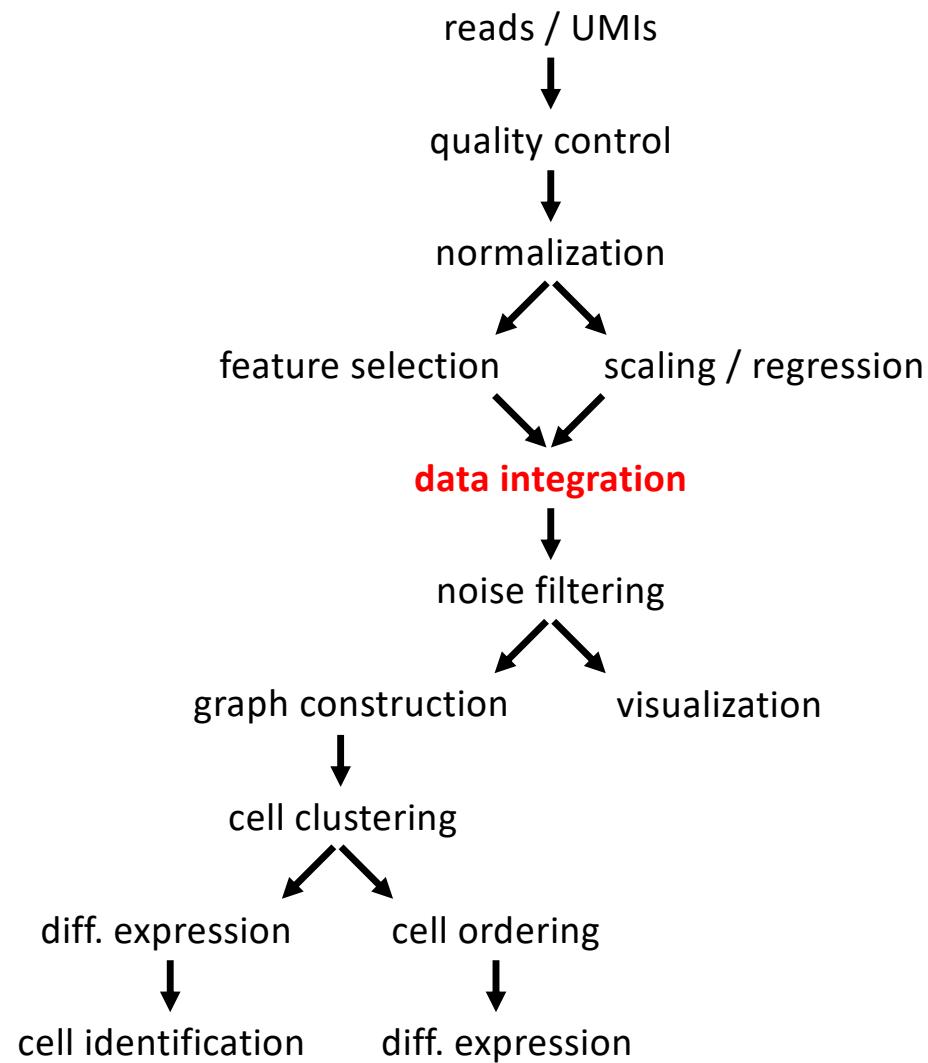
... and more ...



scRNA-seq analysis workflow



scRNA-seq analysis workflow



scRNA-seq data integration

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

Goal:

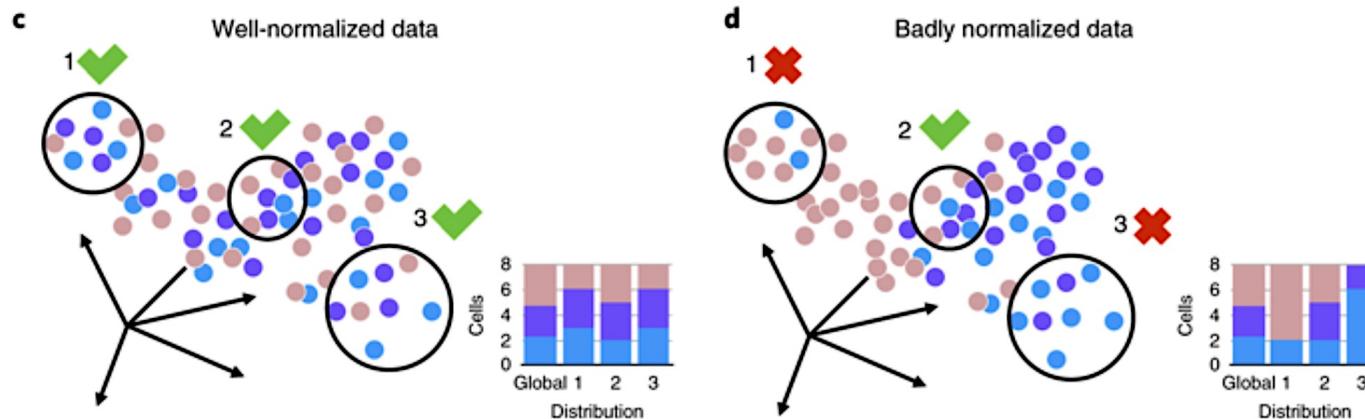
1. The batch-originating variance is erased
2. Meaningful heterogeneity is preserved
3. No artefactual variance is introduced

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*

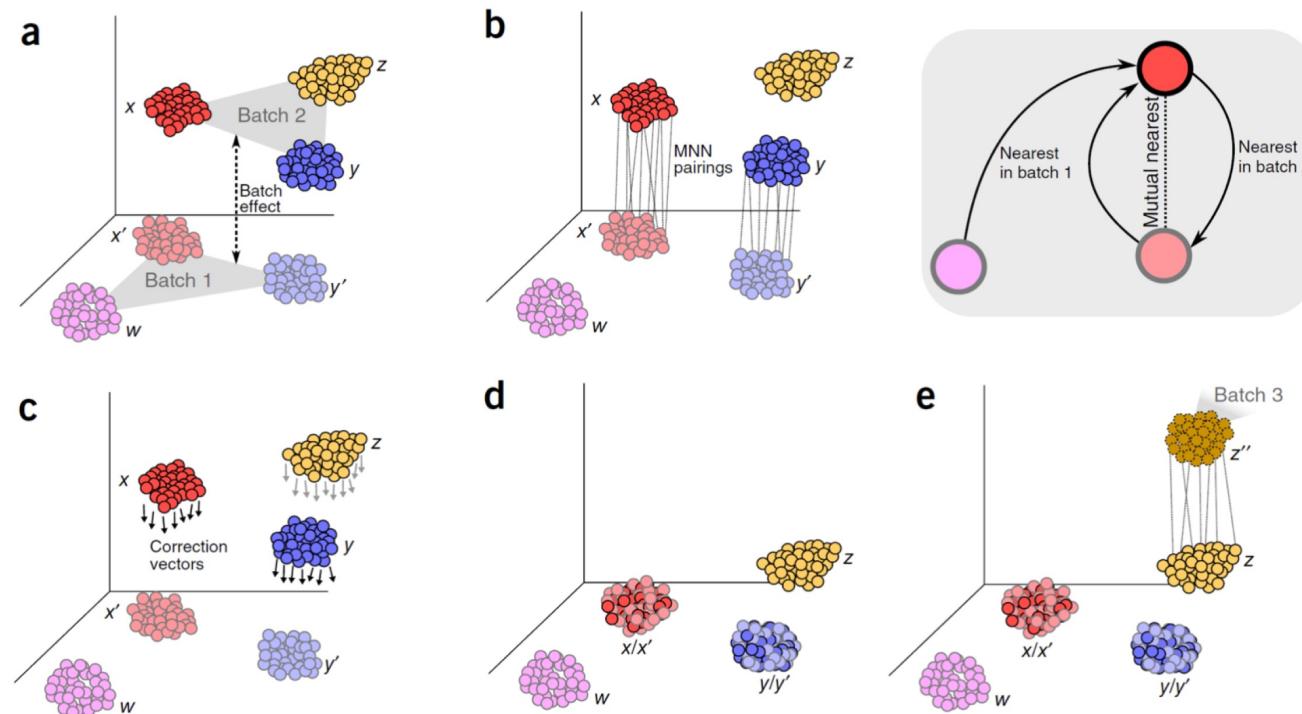
scRNA-seq analysis workflow



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

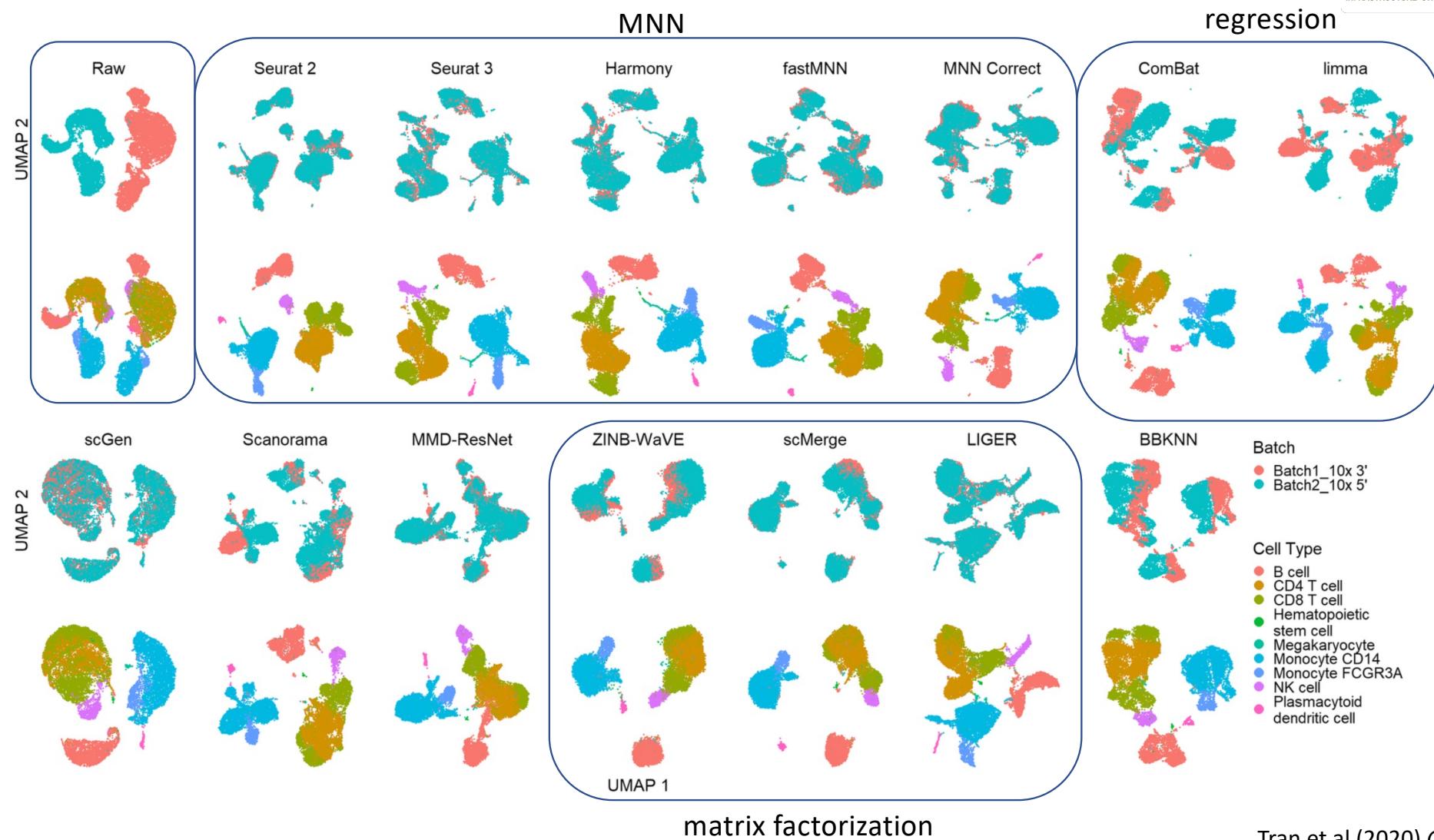
Modern data integration methods are based on the same principle:

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

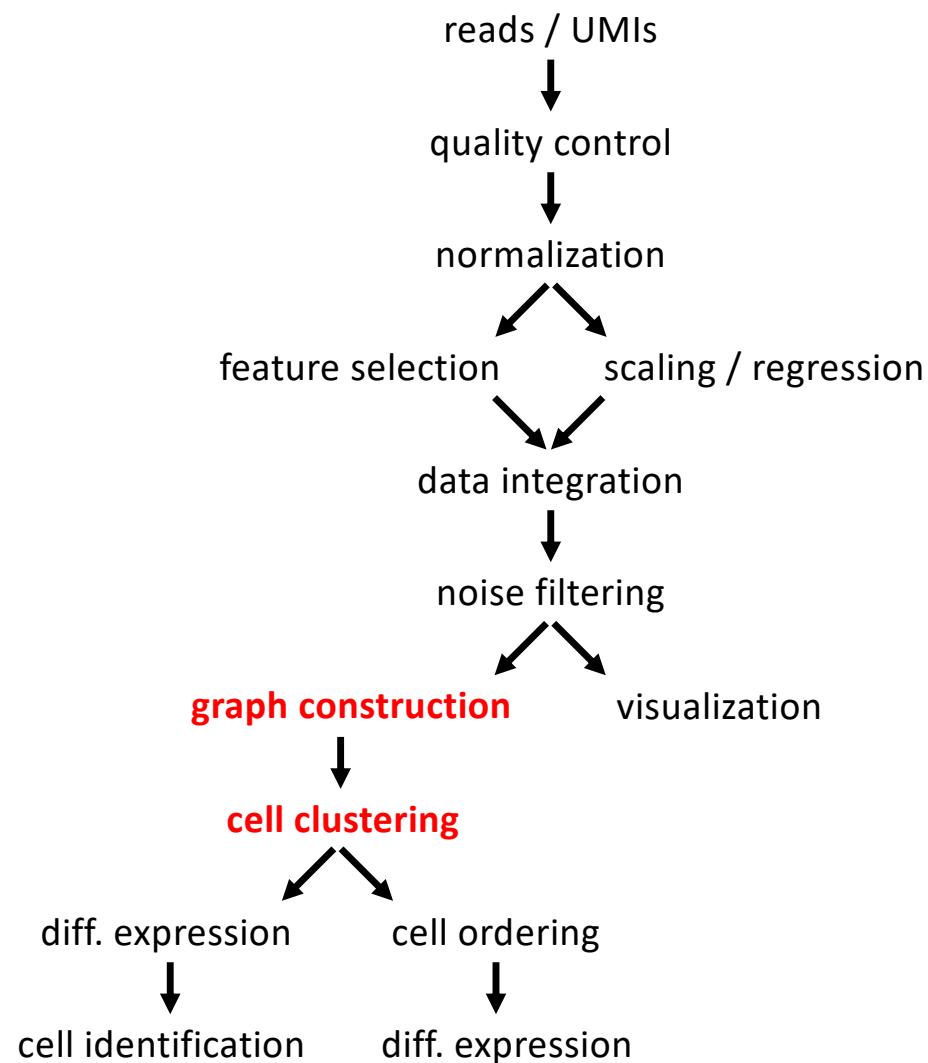


Haghverdi et al (2017) Nat Biotechnology

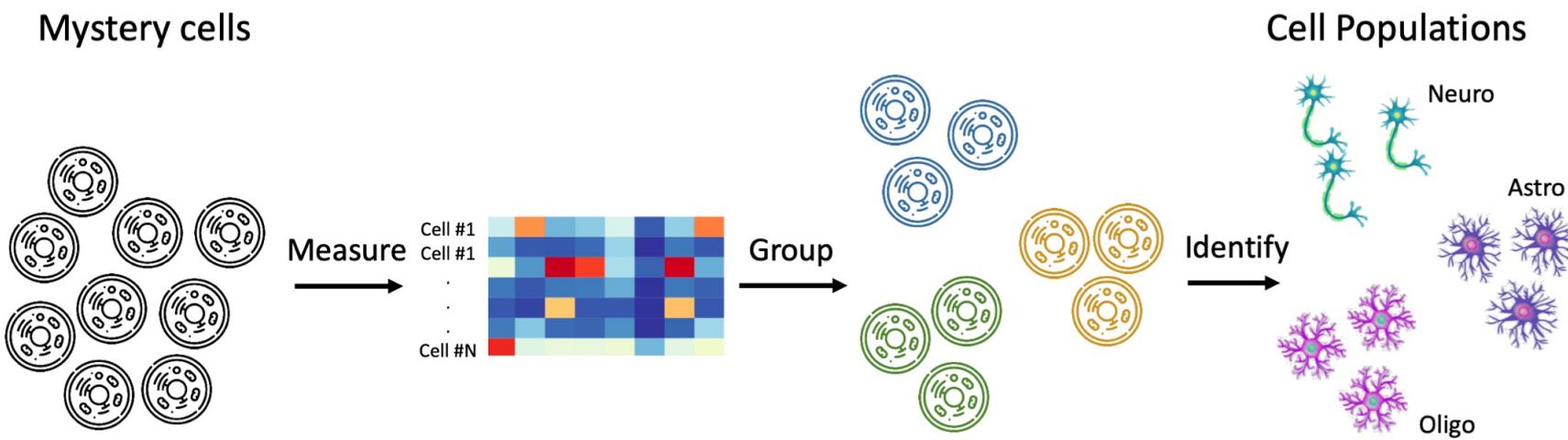
scRNA-seq analysis workflow



scRNA-seq analysis workflow

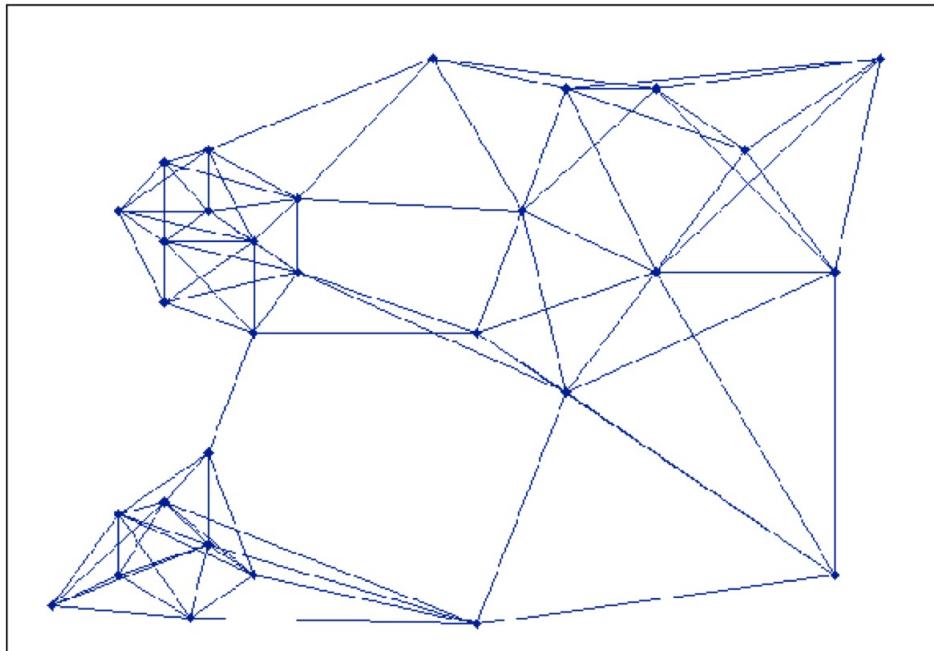


scRNA-seq clustering

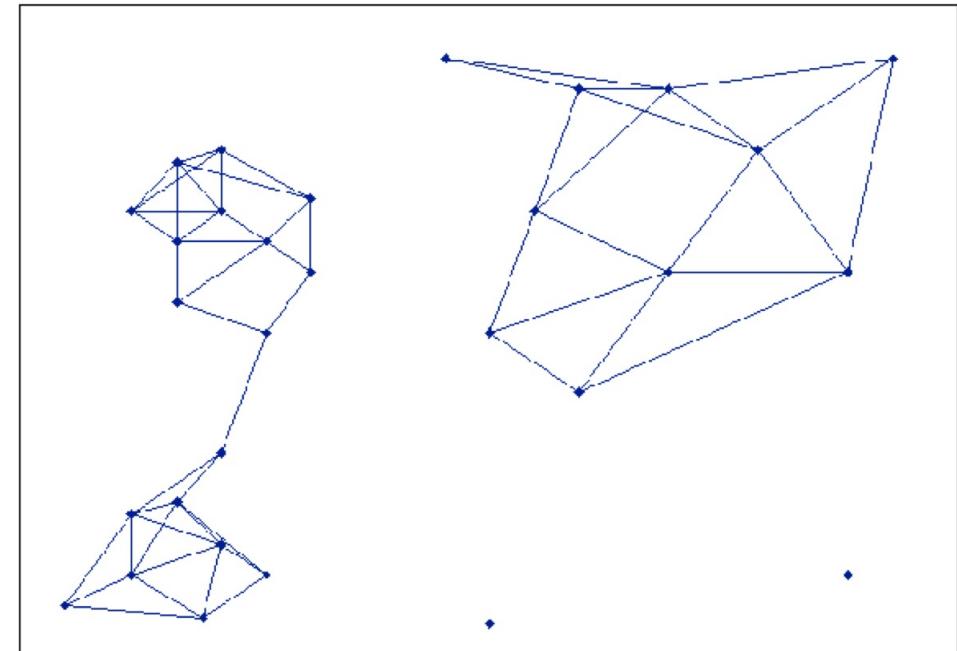


scRNA-seq graph construction

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 .

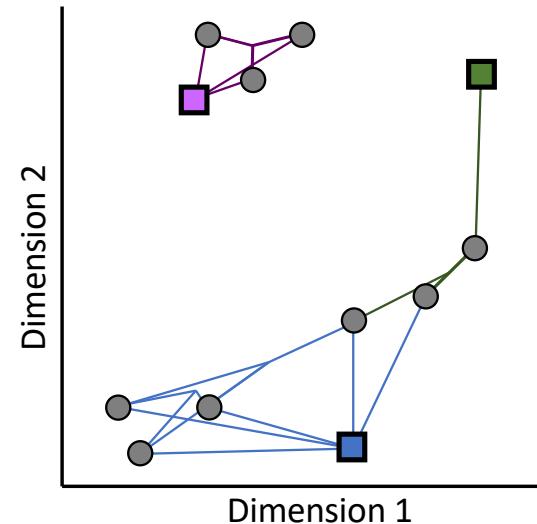


The **Shared Nearest Neighbor (SNN)** graph has weights that defines proximity, or similarity between two edges in terms of the number of neighbors (i.e., directly connected vertices) they have in common.

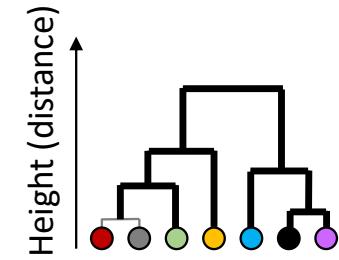
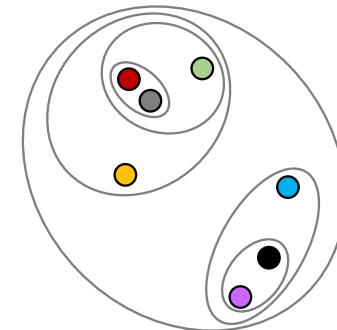


scRNA-seq clustering

K-means $K = 3$



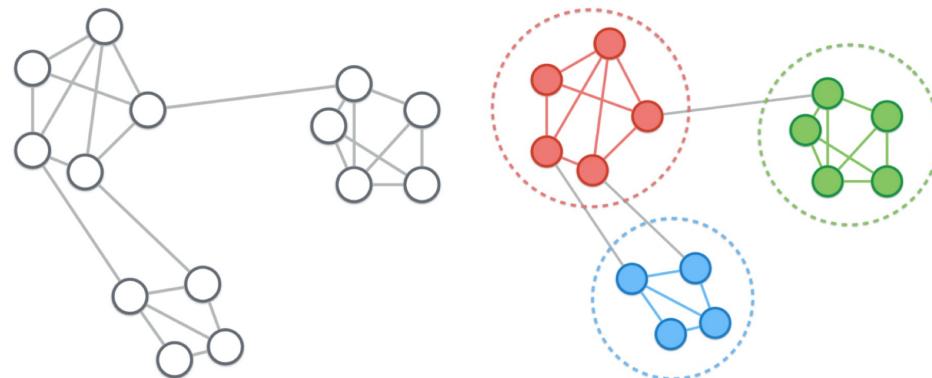
Hierarchical Clustering



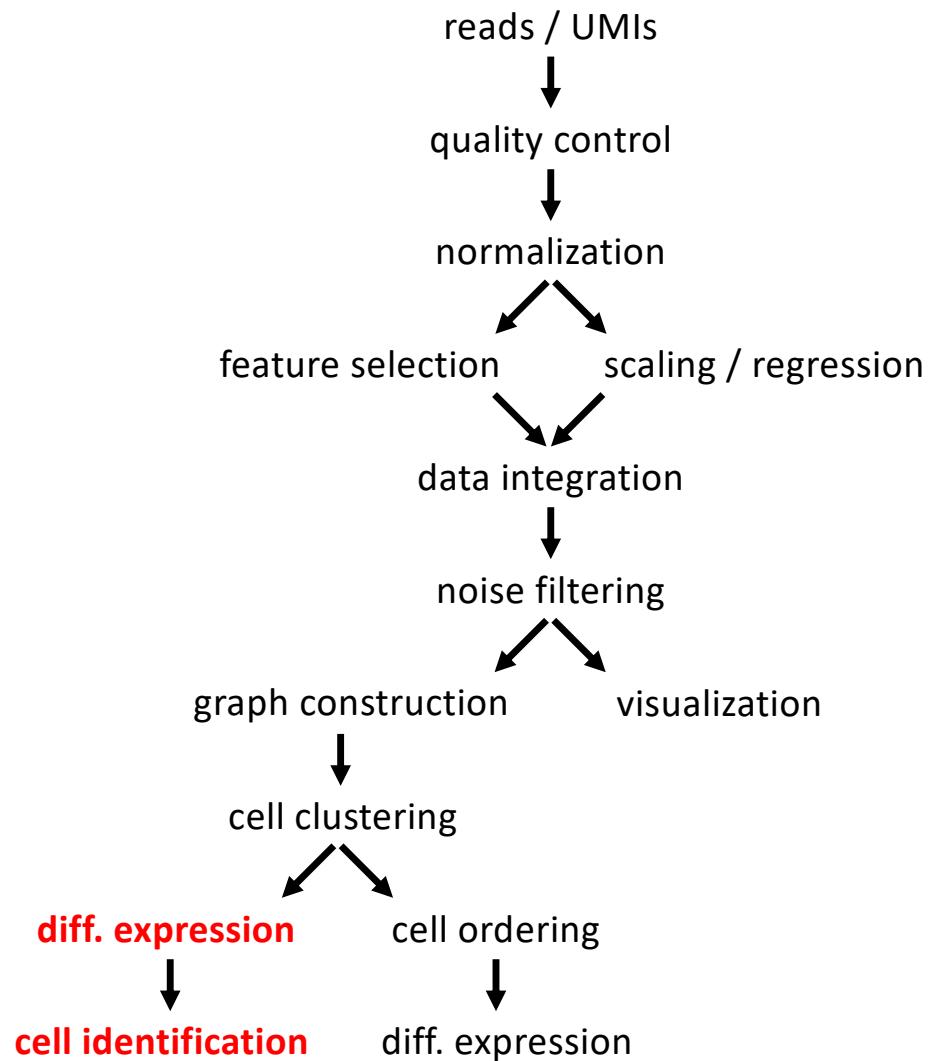
GRAPH

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



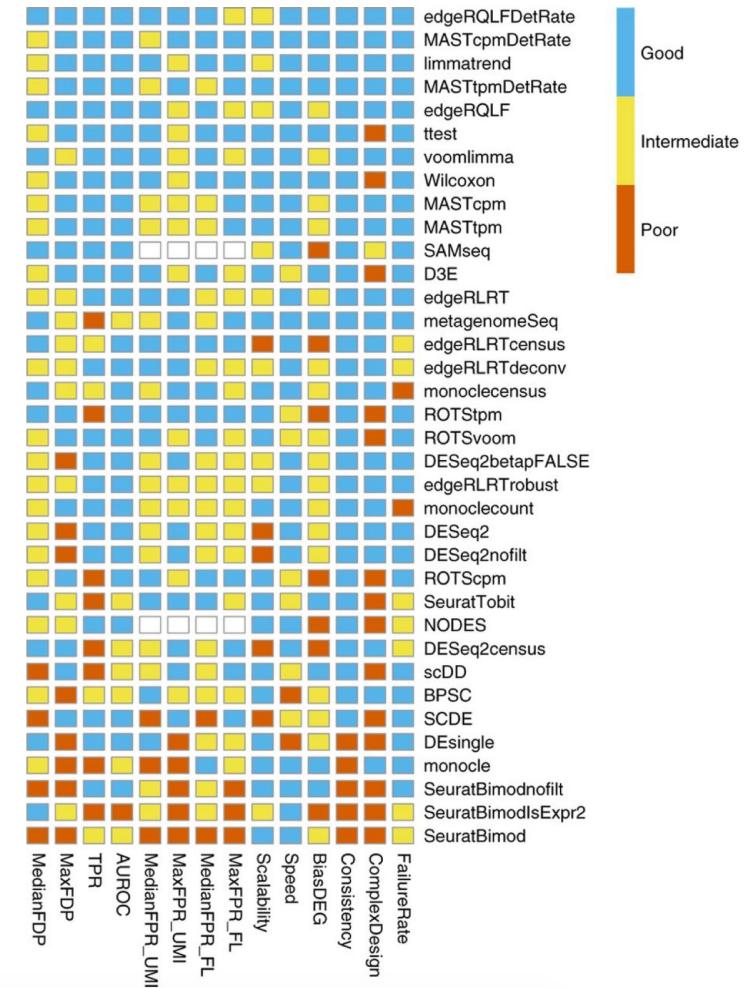
scRNA-seq differential gene expression

Typically we have more than two clusters in a data set

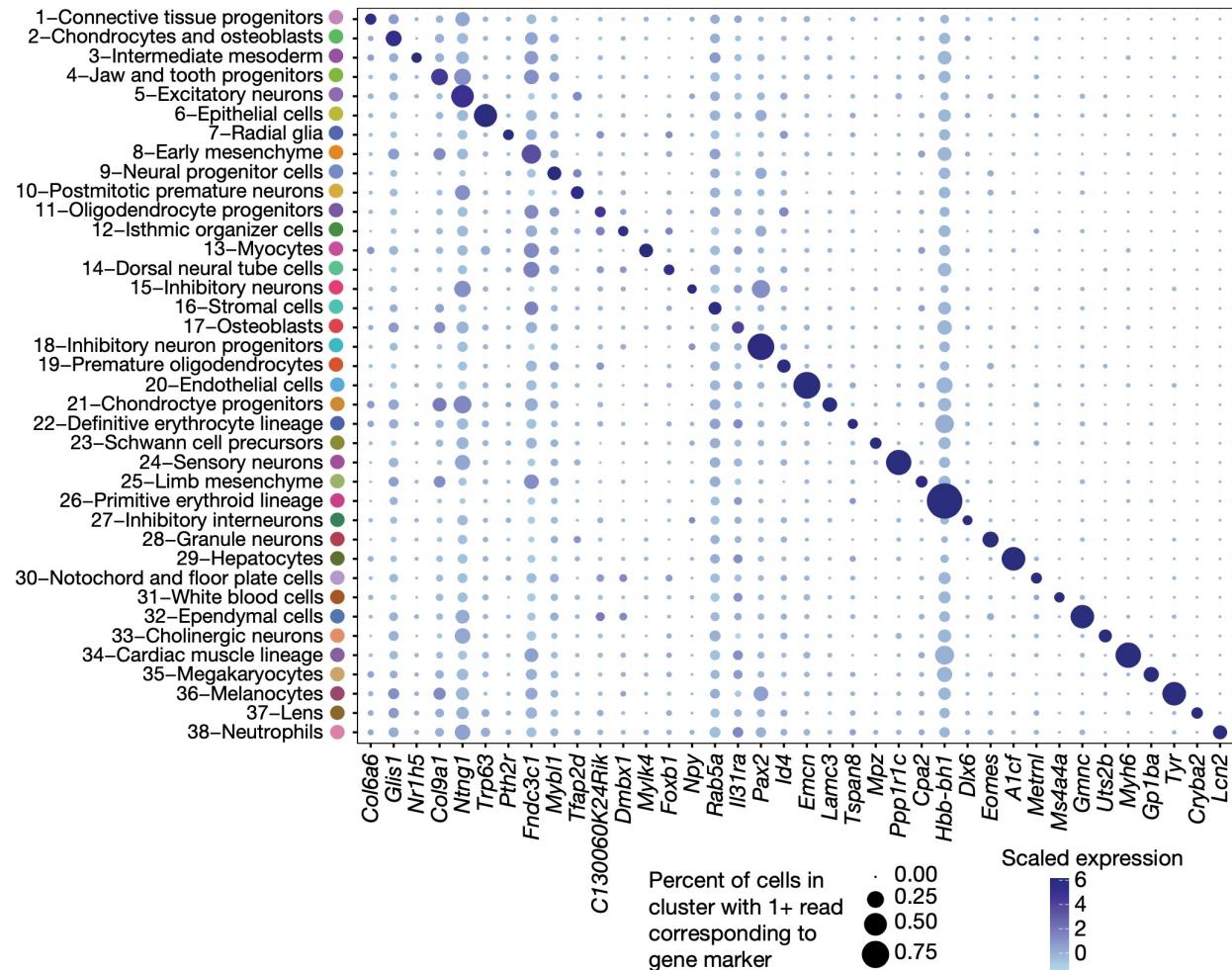
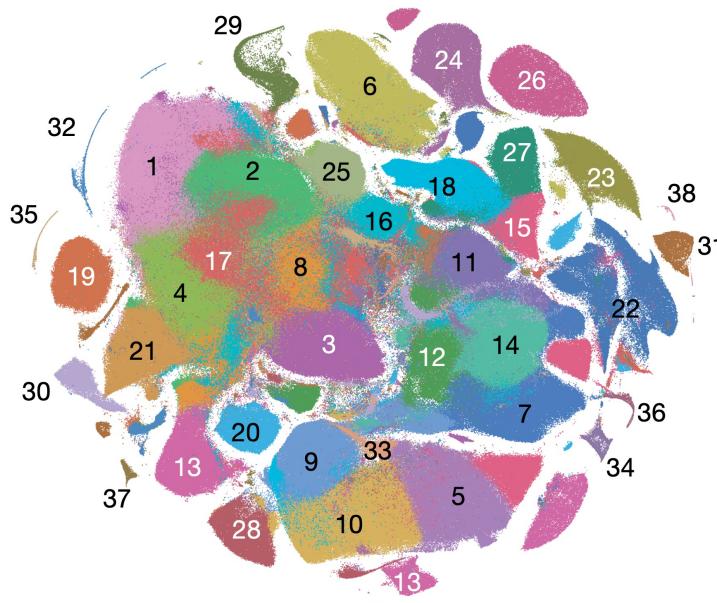
For a given cluster, are we interested in “marker genes” that are:

- **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

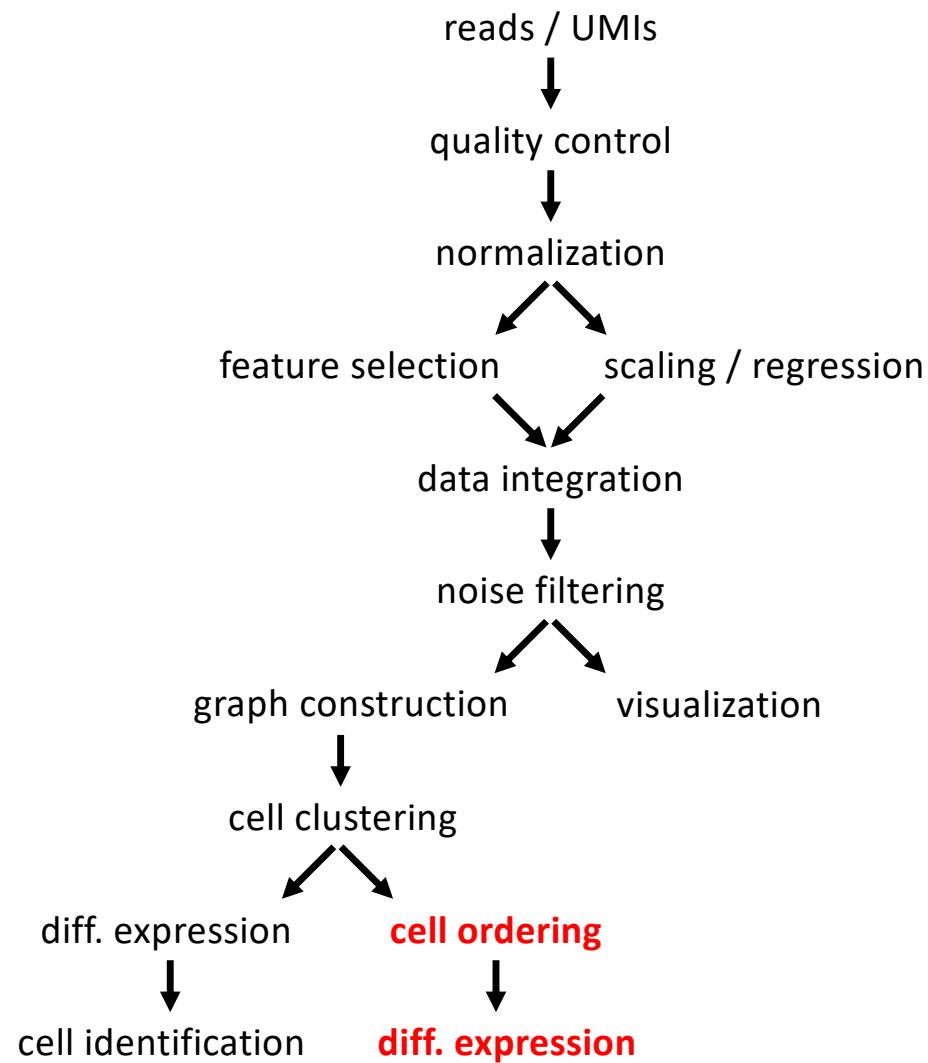
Even the t-test and the Wilcoxon test work well (assuming that you have at least a few dozen cells to compare)



scRNA-seq differential gene expression



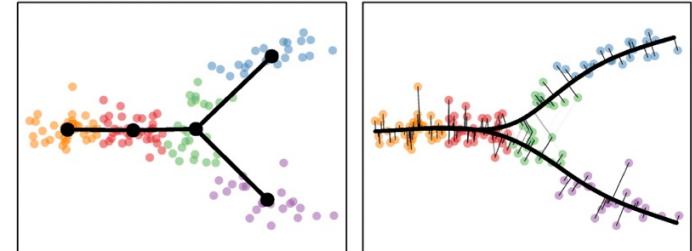
scRNA-seq analysis workflow



scRNA-seq trajectory inference



Are you sure that you have a developmental trajectory?



Street et al (2018) *BMC Genomics*

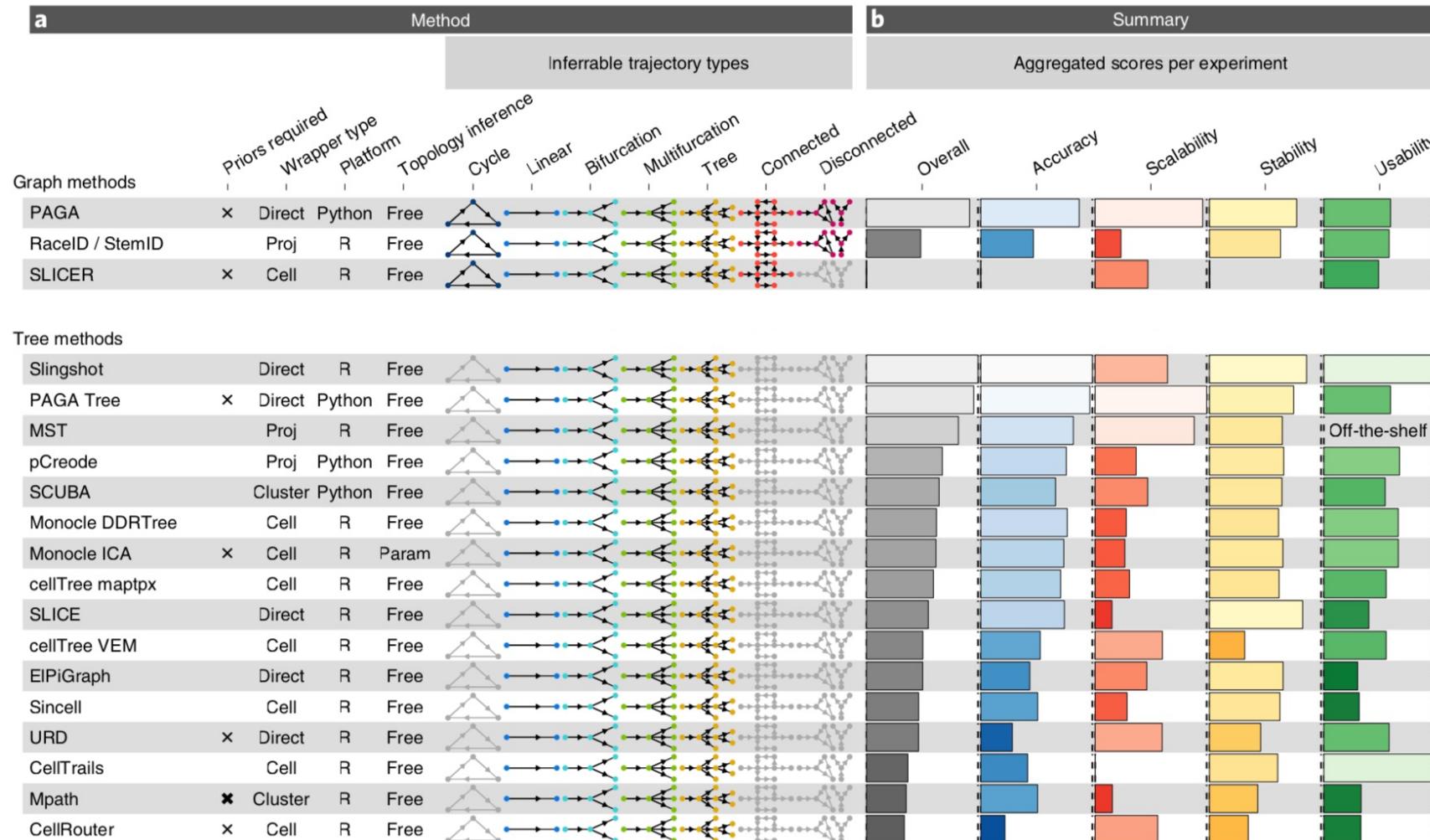
Do you have intermediate states?

Do you believe that you have branching in your trajectory?

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

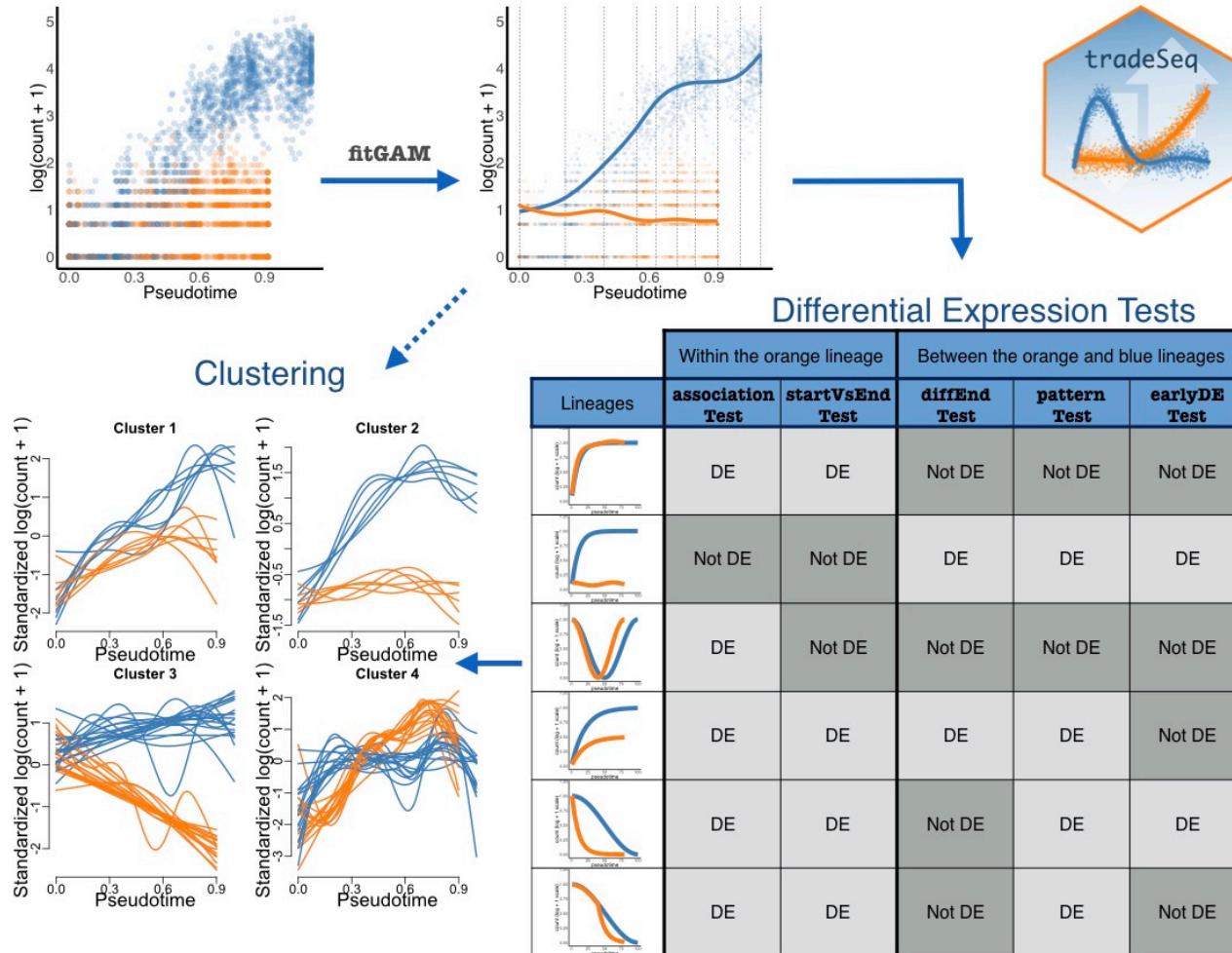
! First make sure that gene set and dimensionality reduction captures what you expect.

scRNA-seq trajectory inference

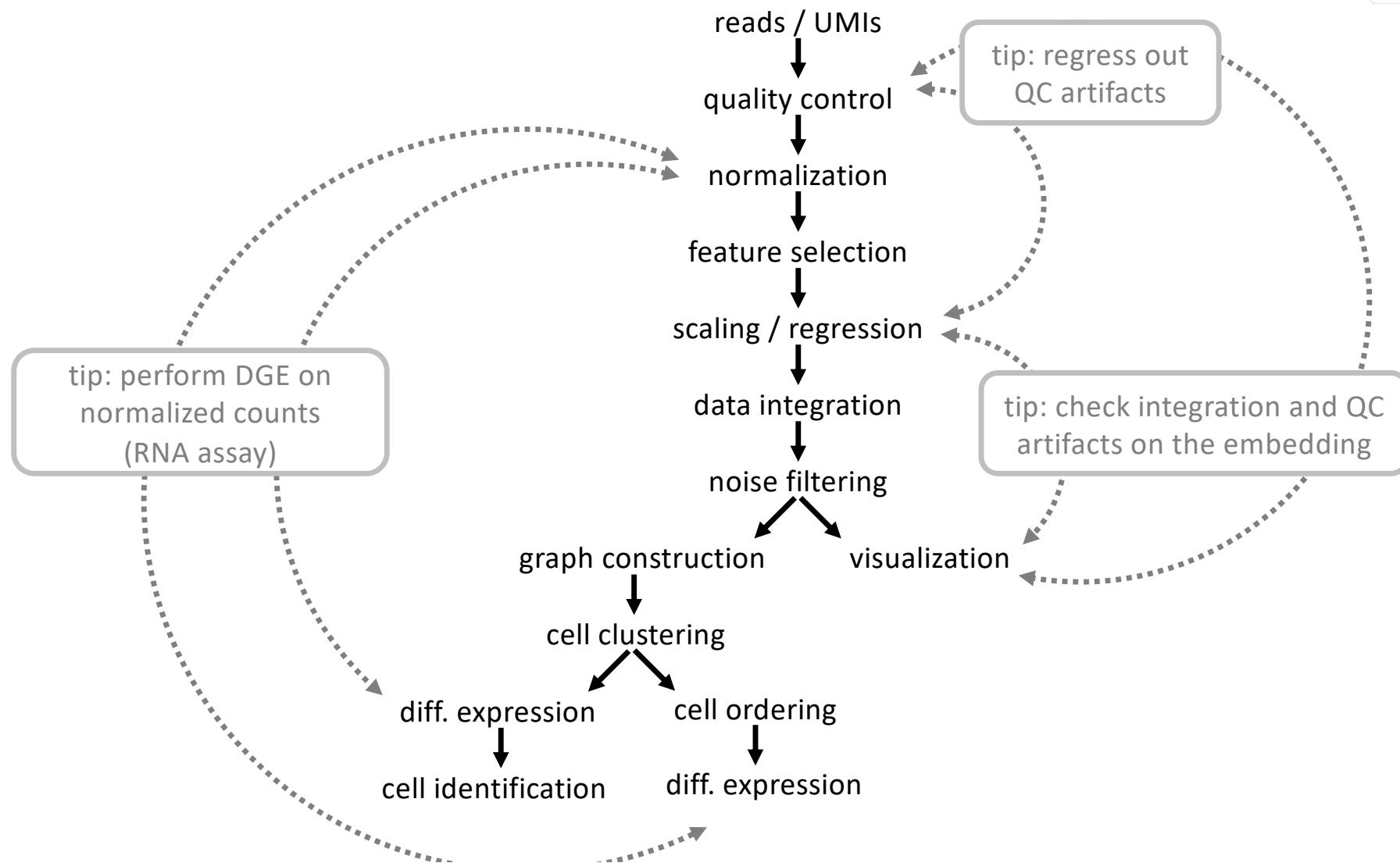


Saelens et al 2019 *Nat Biotechnology*

scRNA-seq trajectory inference



scRNA-seq analysis workflow



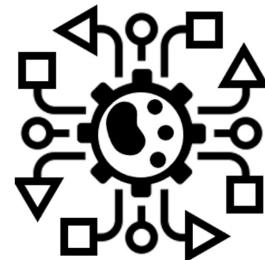
scRNA-seq analysis workflow



Spatial
transcriptomics



RNA
velocity



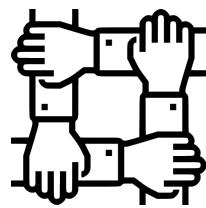
Omics
integration



Deep
learning



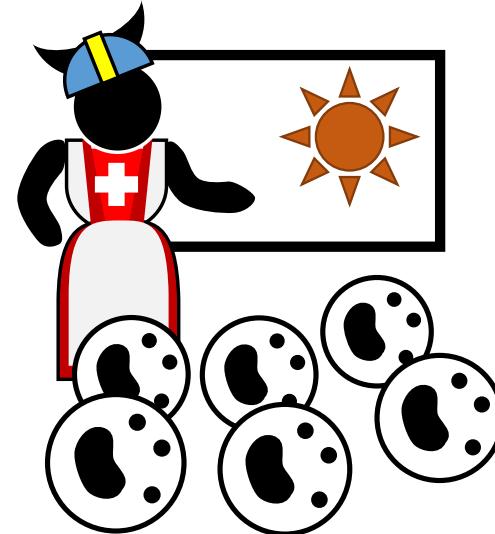
Learning Strategy



Working in Groups



Tips for a good dynamic



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

