800 1222·2022 ANNI

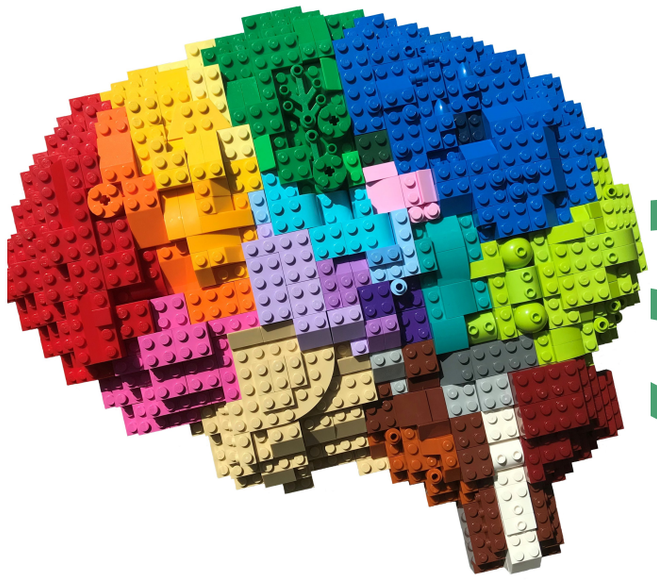
UNIVERSITÀ DEGLI STUDI DI PADOVA

CSAMA 2023 - BRIXEN/BRESSANONE

SINGLE-CELL RNA-SEQ

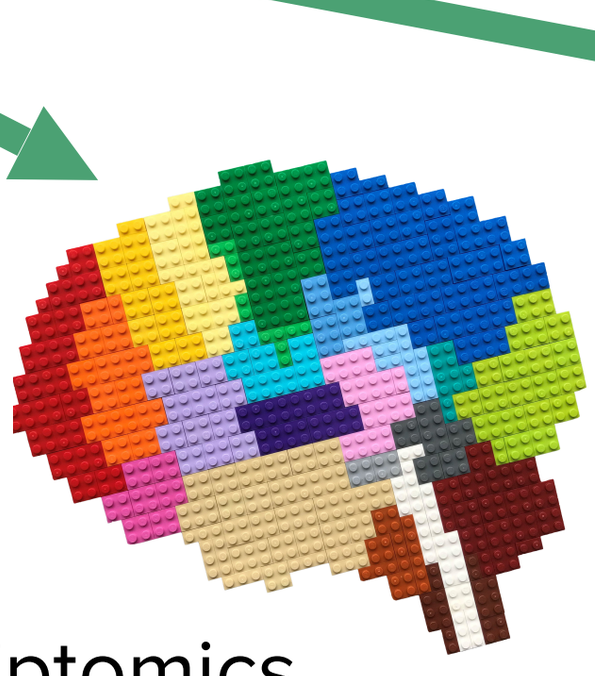
Davide Risso
@drisso1893
@drisso@genomic.social
@drisso

The evolution of gene expression measurements



Original organ

Bulk RNA-seq

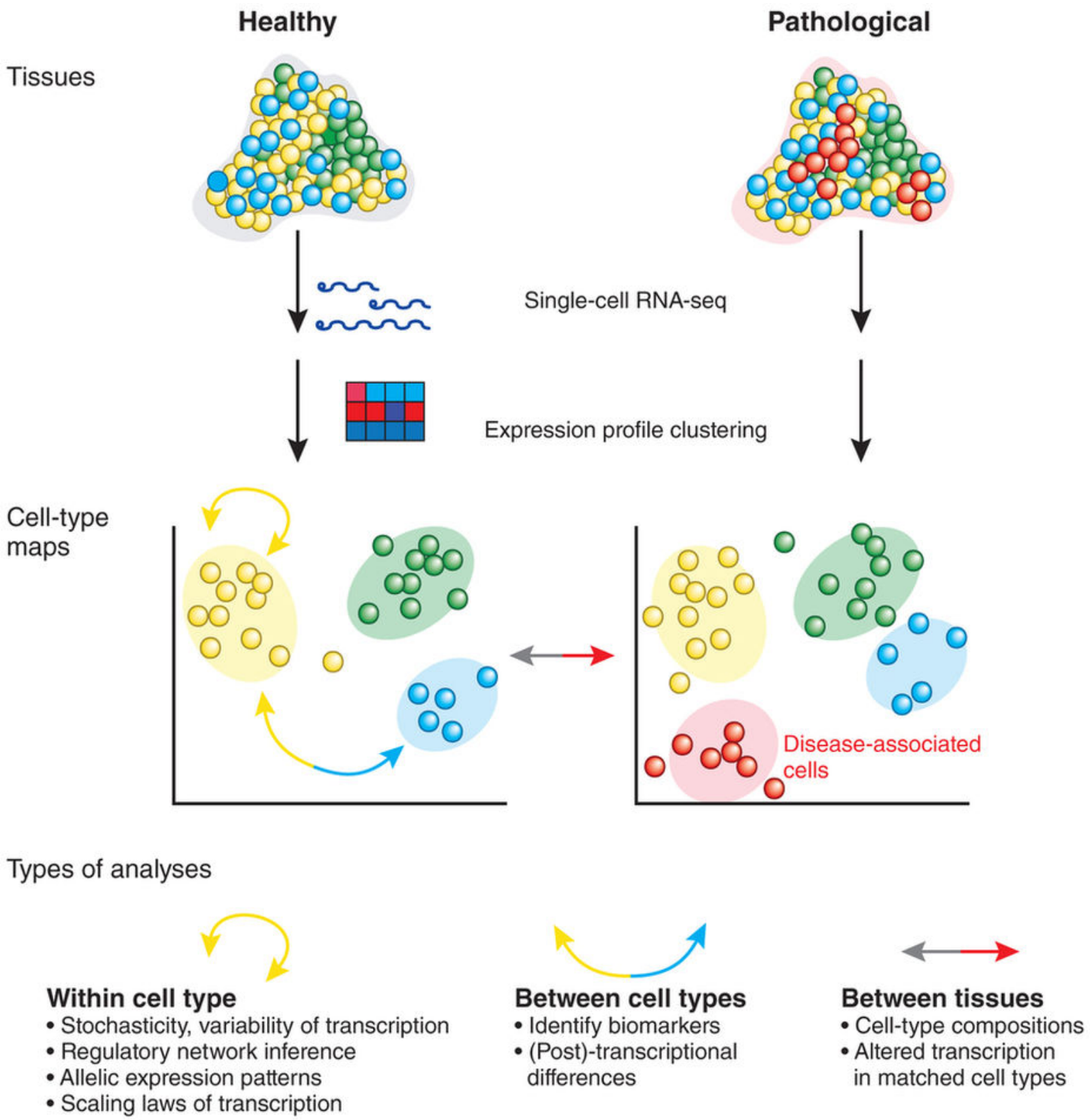


Spatial transcriptomics

Single-cell RNA-seq

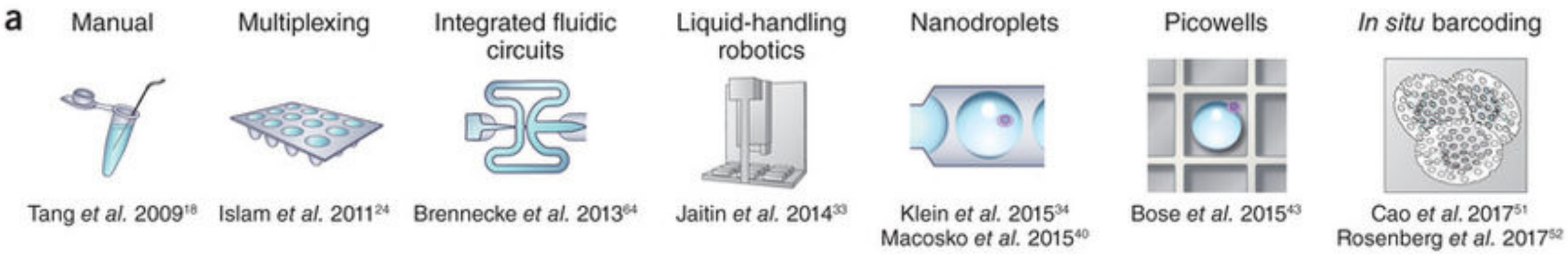
Image credit: @BoXia7

SINGLE-CELL RNA-SEQ



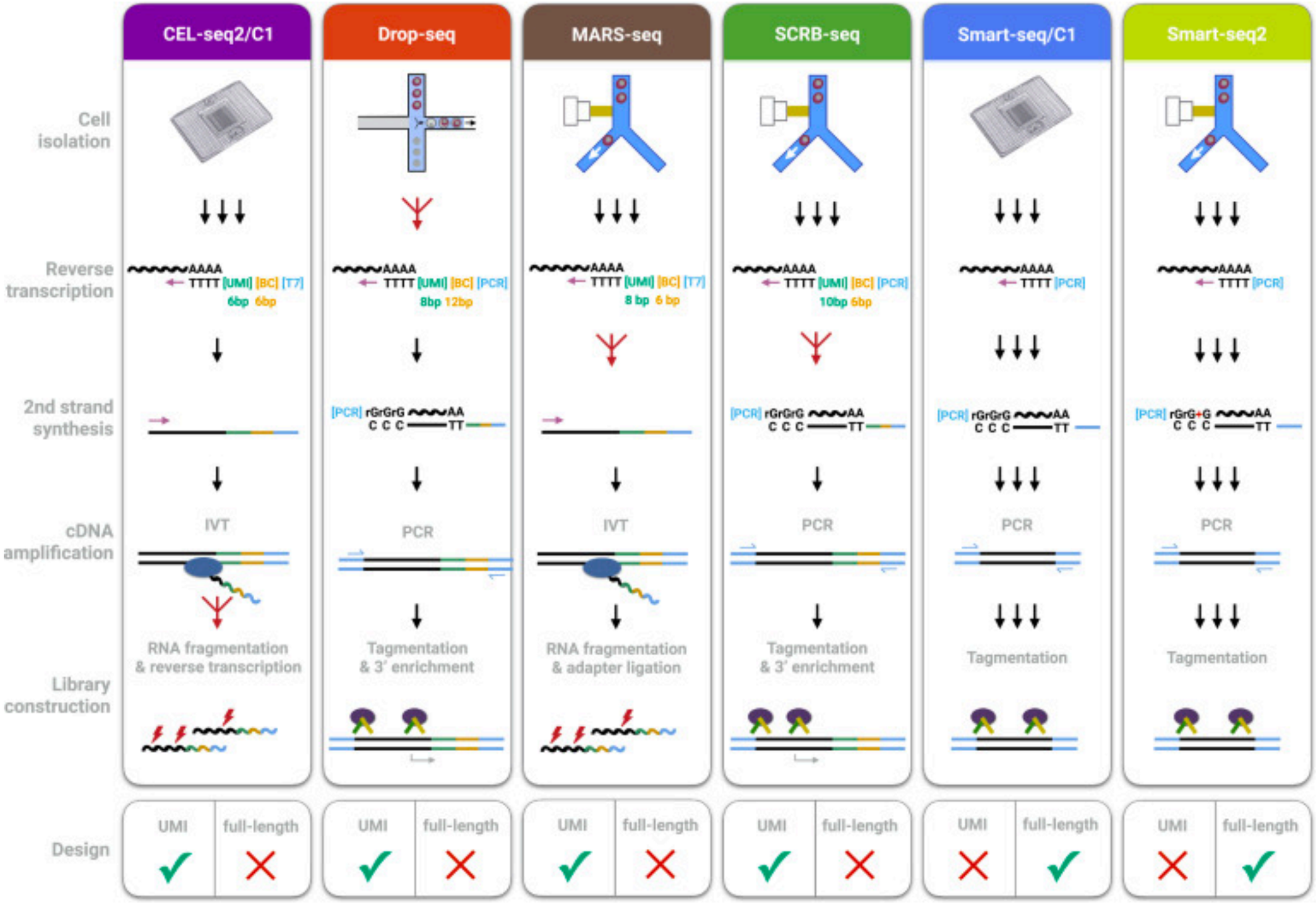
Sandberg (2014). *Nature Methods.*

SEVERAL PROTOCOLS AND PLATFORMS



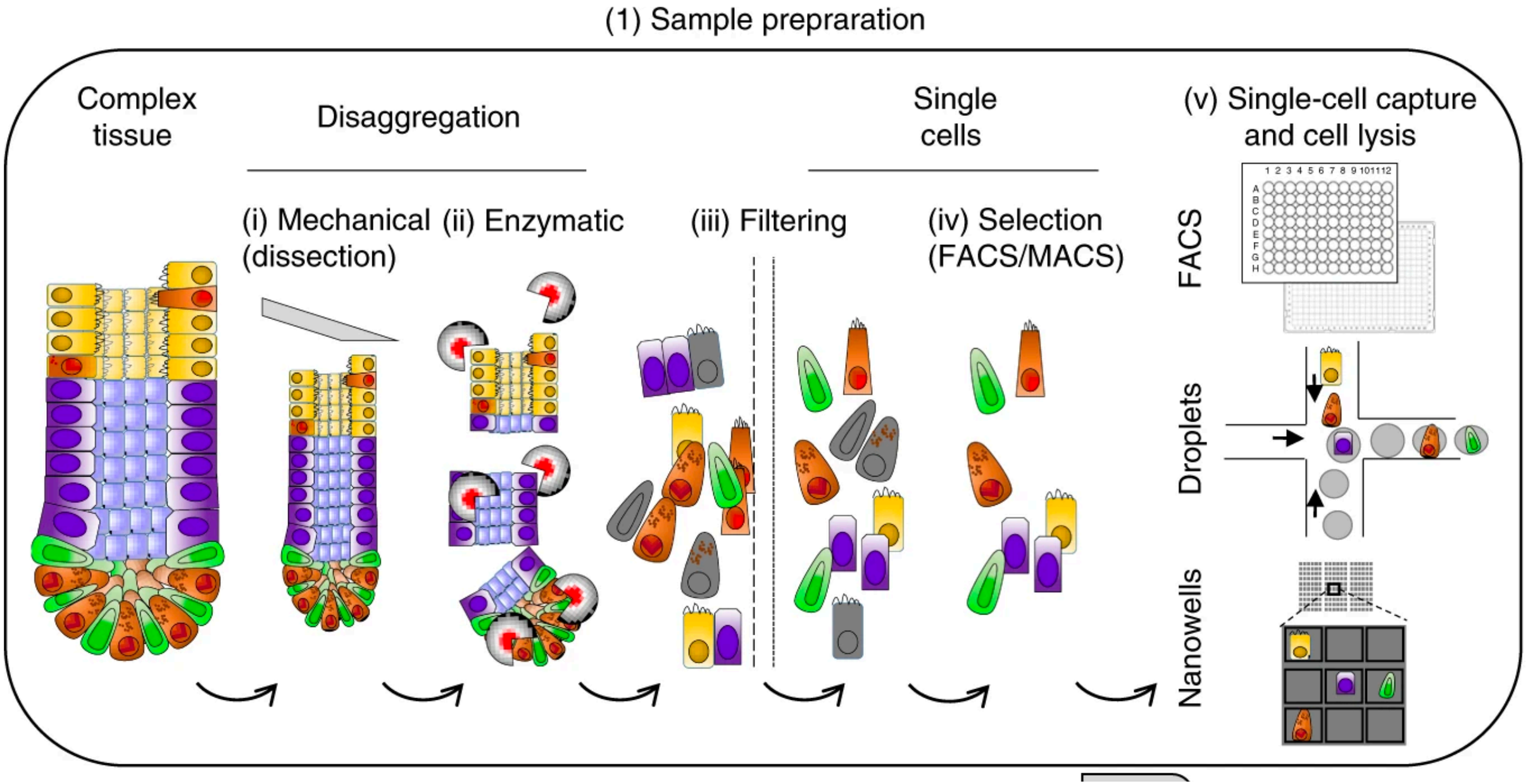
Svensson, Vento-Tormo, Teichmann (2018). *Nature Protocols.*

DIFFERENT PROTOCOLS HAVE DIFFERENT PROPERTIES



Ziegenhain et al. (2017). *Molecular Cell.*

SINGLE-CELL RNA-SEQ IN A NUTSHELL

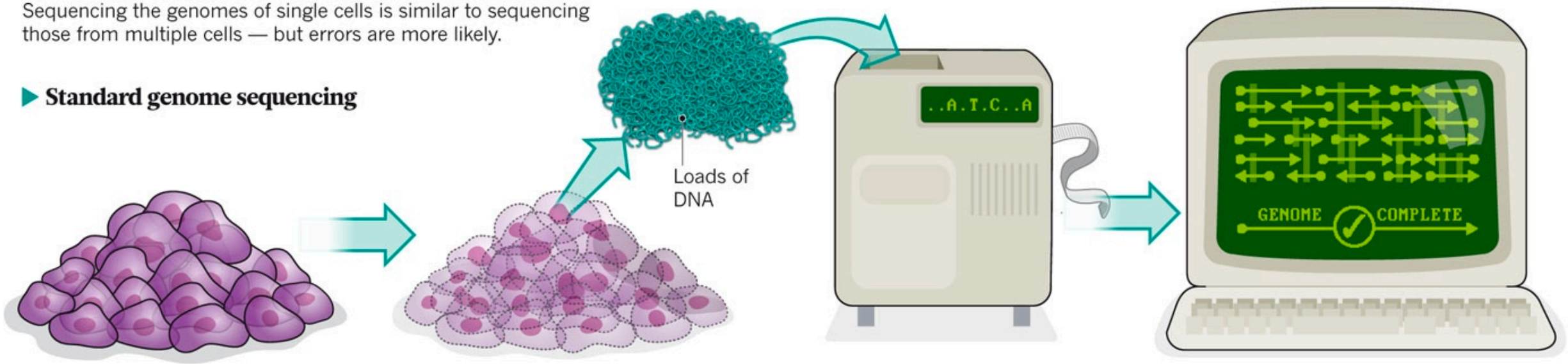


AMPLIFICATION BIAS LEADS TO…

ONE GENOME FROM MANY

Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

▶ Standard genome sequencing



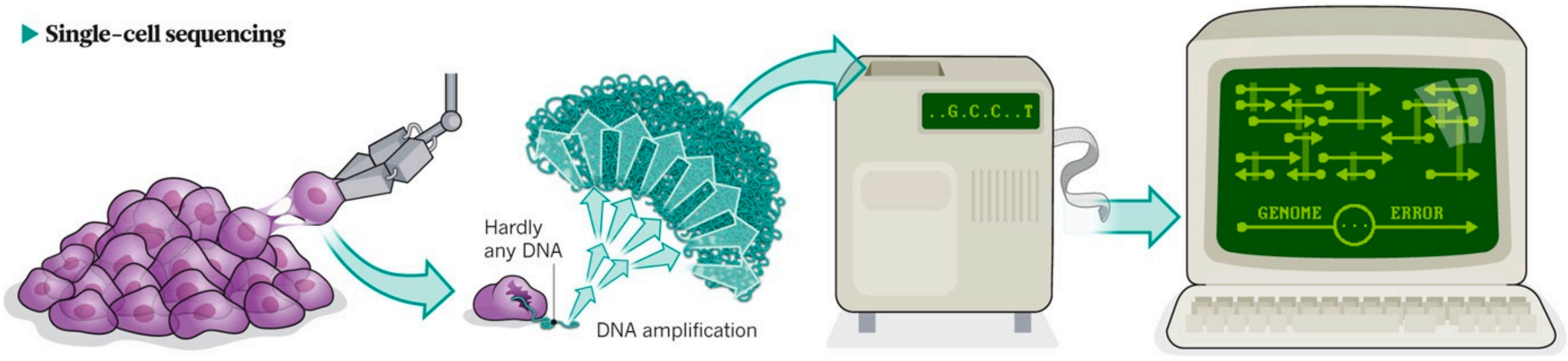
A sample containing thousands to millions of cells is isolated.

DNA is extracted from all the nuclei.

DNA is broken into fragments and then sequenced.

The sequences are assembled to give a common, 'consensus' sequence.

▶ Single-cell sequencing



A single cell is difficult to isolate, but it can be done mechanically or with an automated cell sorter.

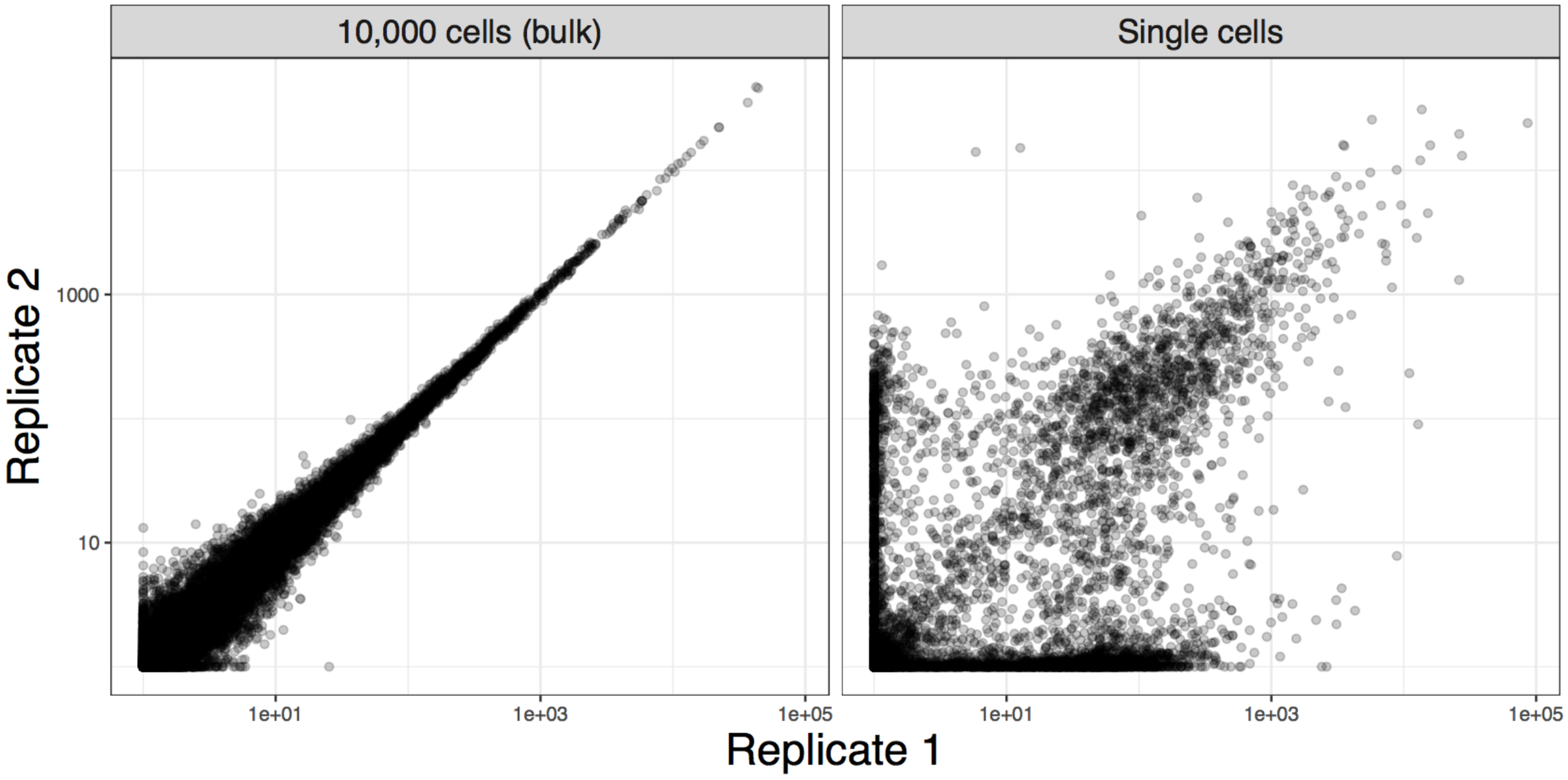
The DNA is extracted and amplified, during which errors can creep in.

Amplified DNA is sequenced.

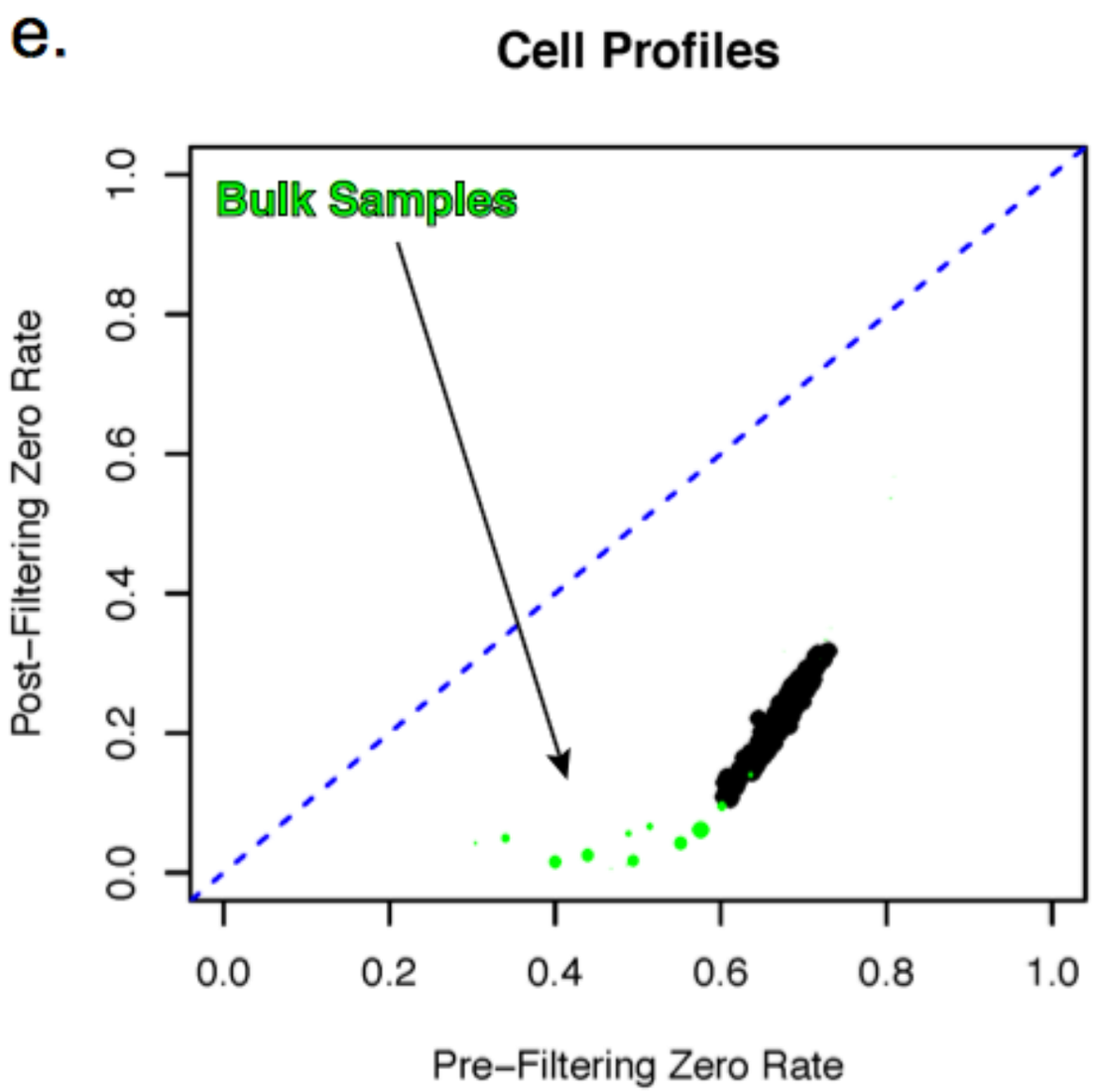
Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

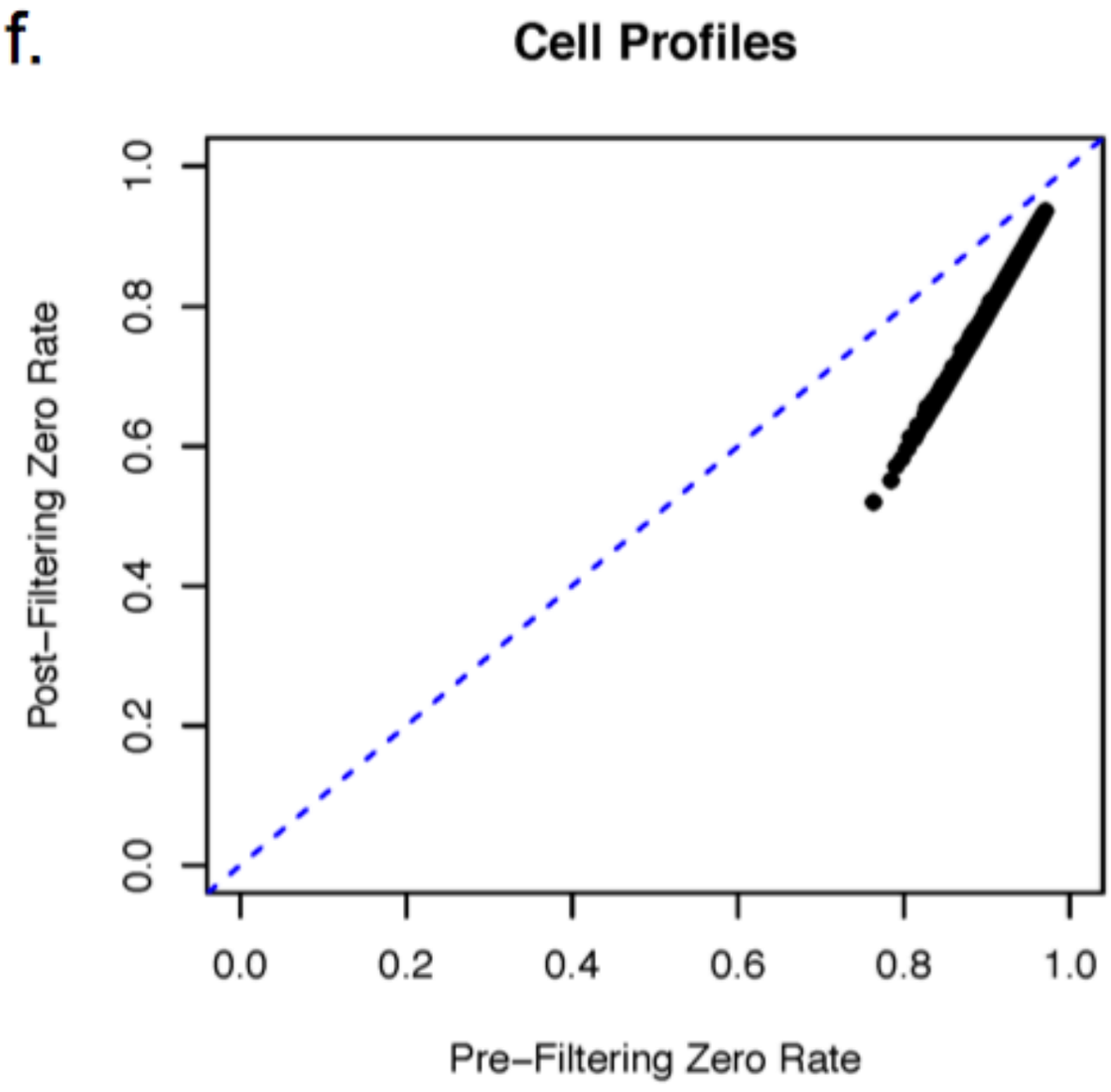
Owens (2012). *Nature*.

INCREASED VARIABILITY COMPARED TO "BULK" RNA-SEQ



EXCESS OF ZERO COUNTS





Cole et al. (2019). *Cell Systems*.

UNIQUE MOLECULAR IDENTIFIERS

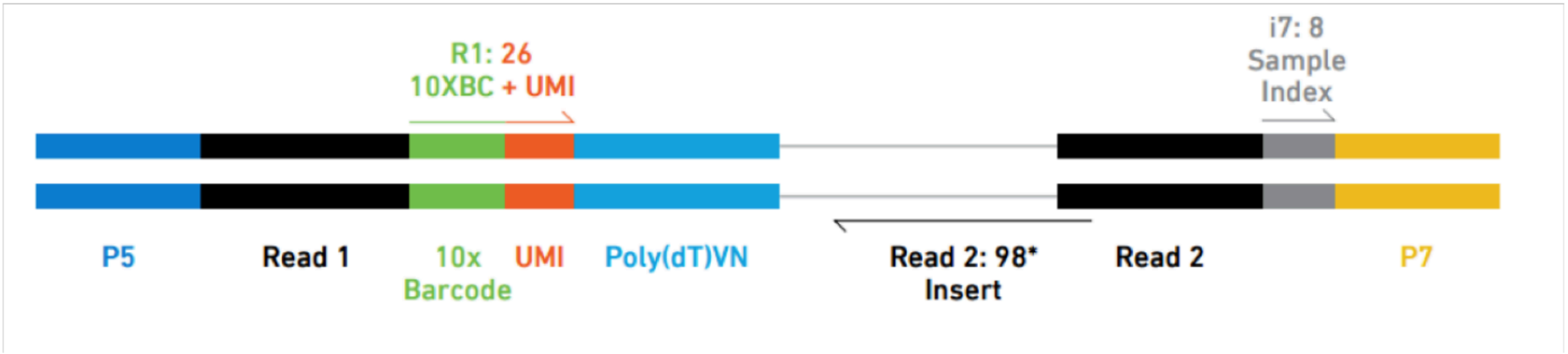
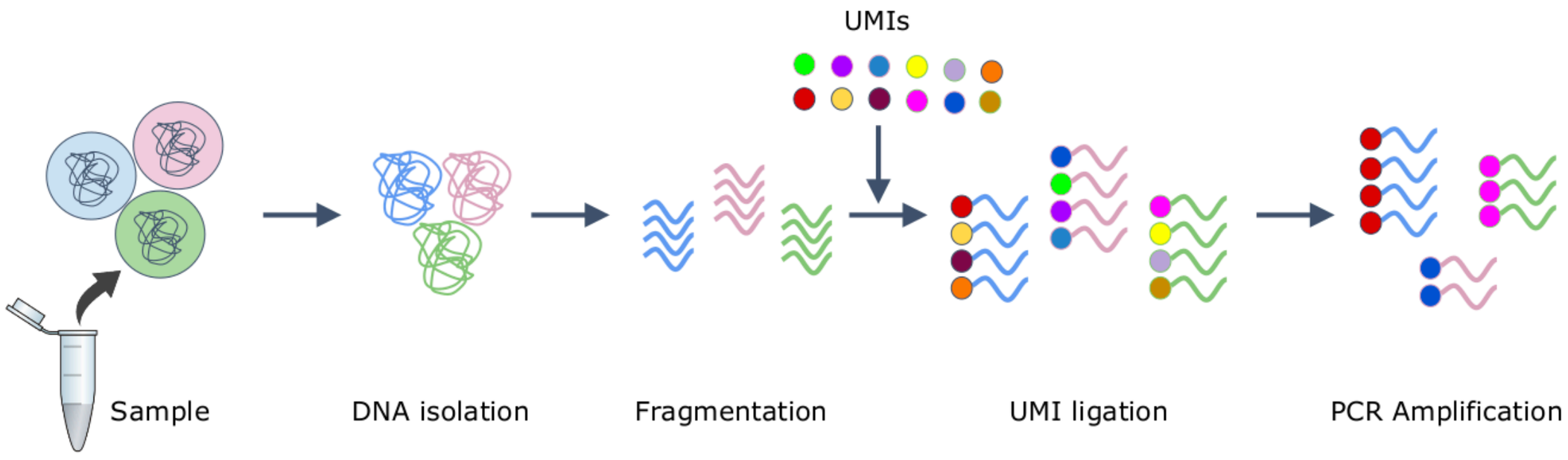


Fig. 2. Schematic of a fragment from a final Chromium™ Single Cell 3' v2 library. *Can be adjusted.



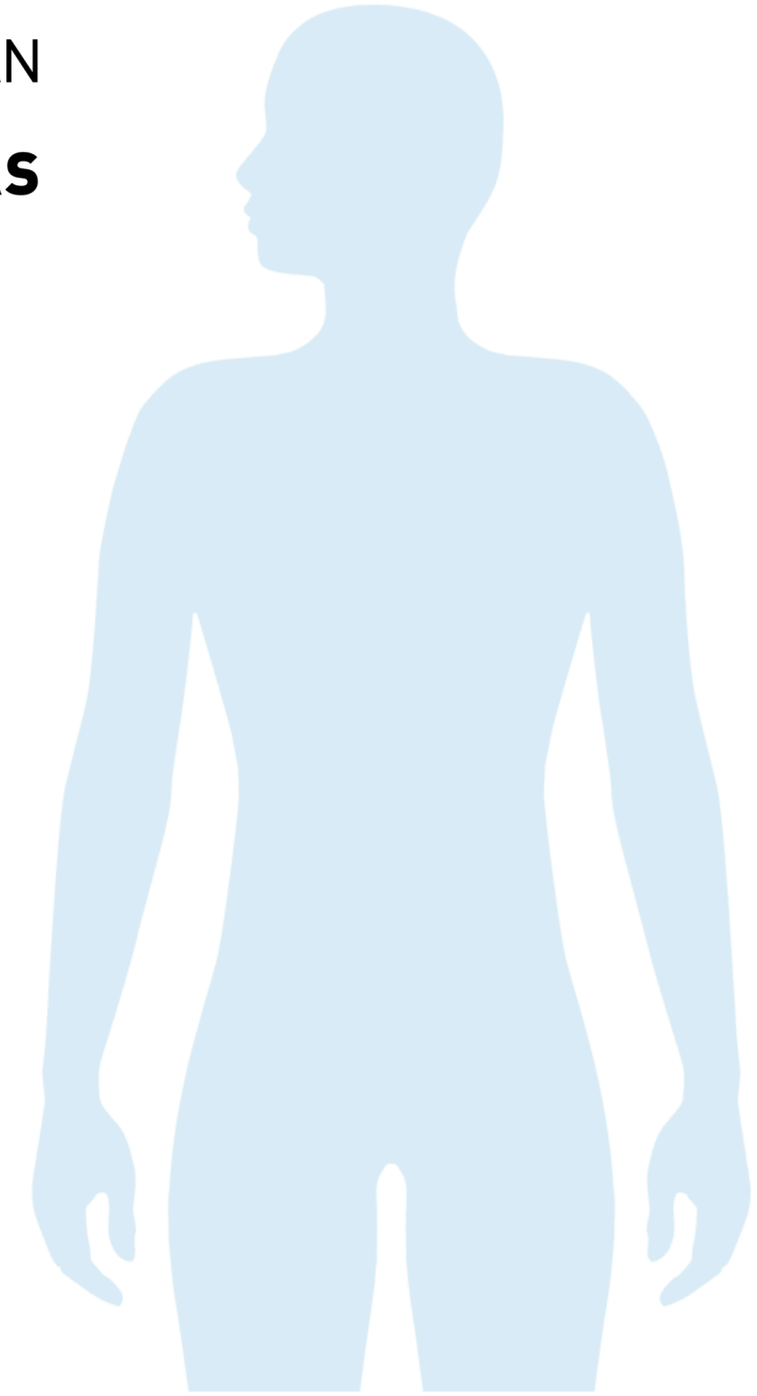
HIGH DIMENSIONALITY AND SAMPLE SIZE

44.5M Cells

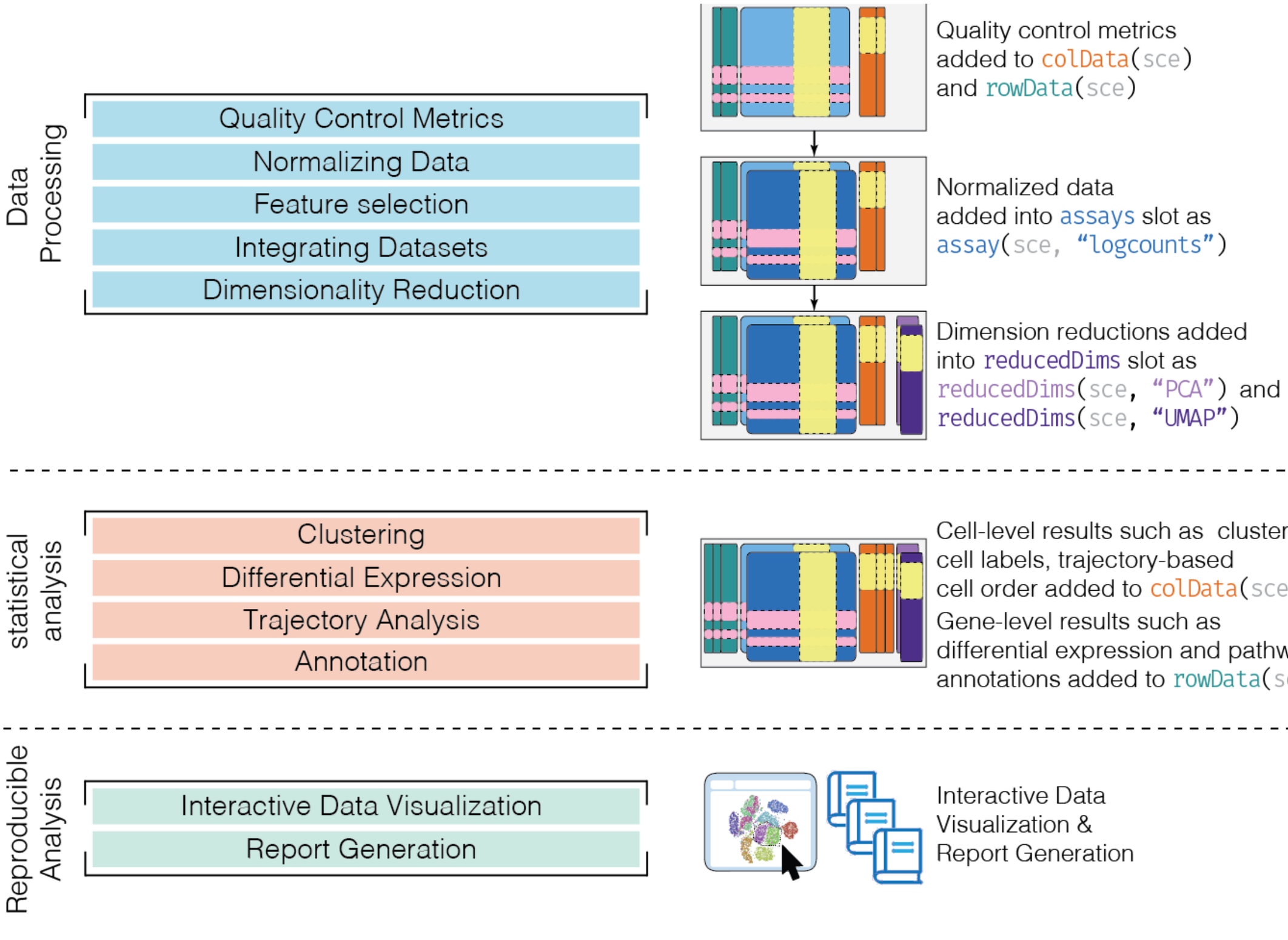
ALL CELLS

- Blood
- Kidney
- Bone
- Liver
- Brain
- Lung
- Pancreas
- Heart
- Immune System
- Skin

HUMAN CELL ATLAS

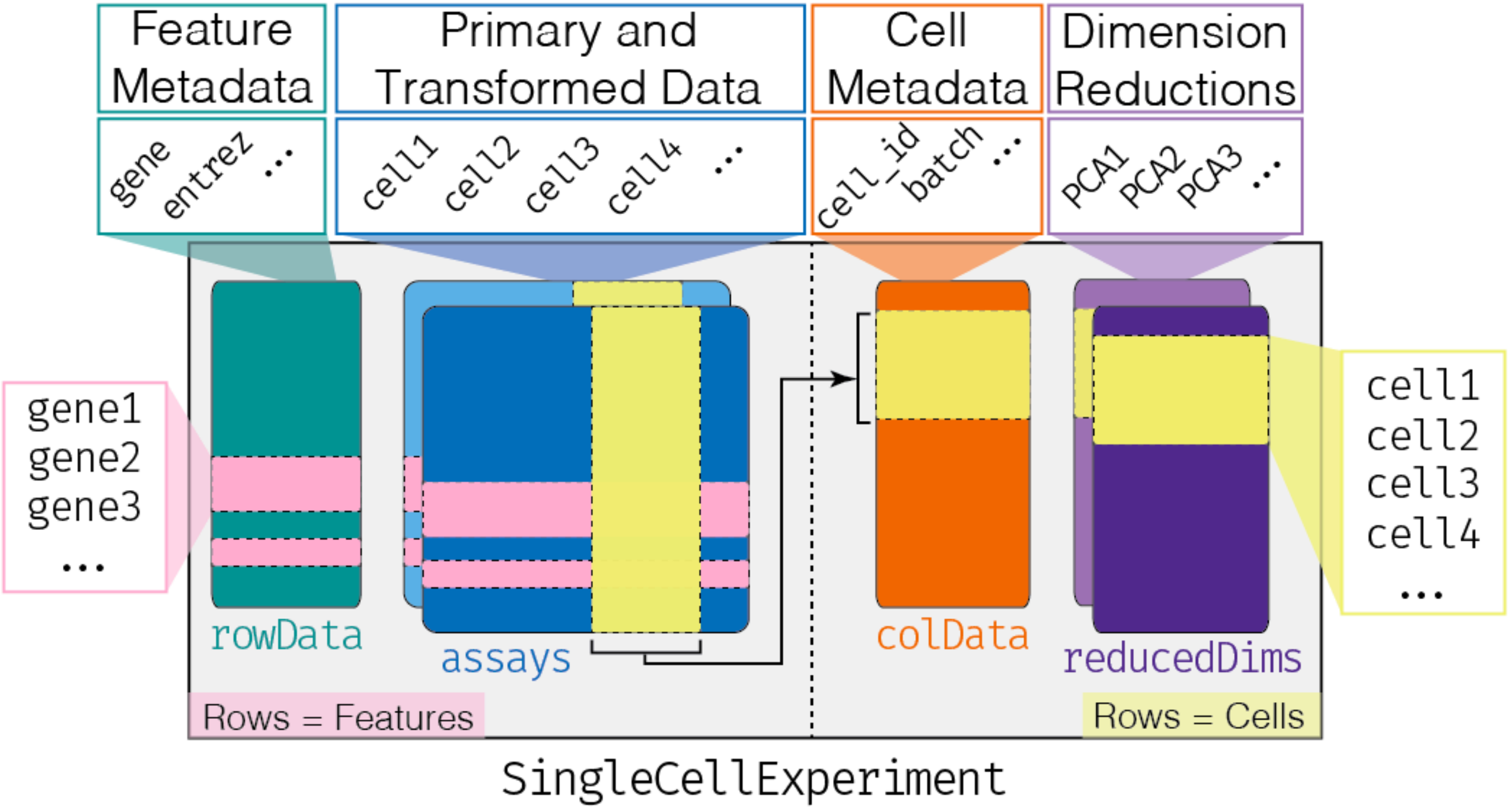


A TYPICAL WORKFLOW



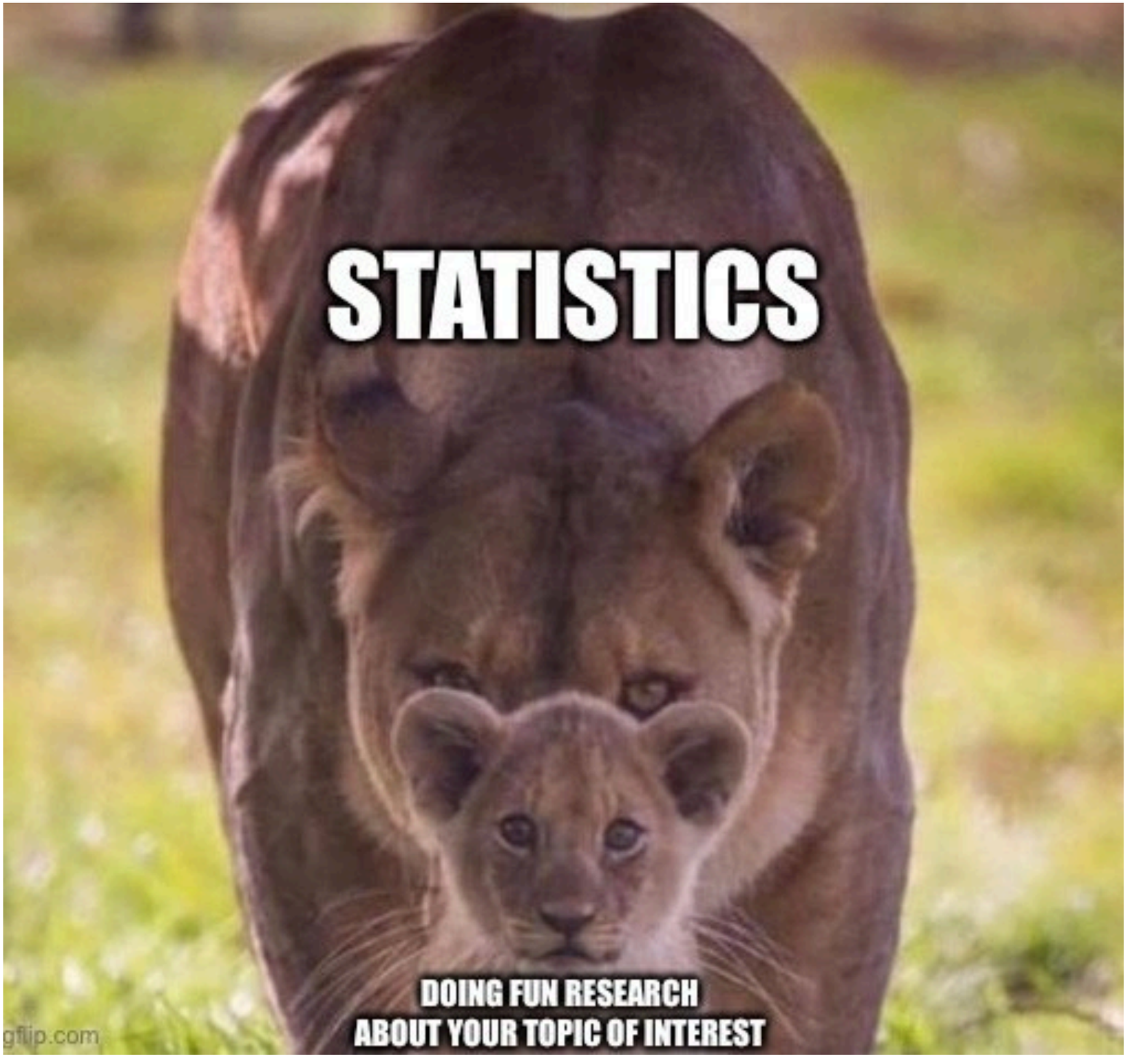
Amezquita et al. (2020). *Nature Methods.*

THE SINGLECELLEXPERIMENT CLASS

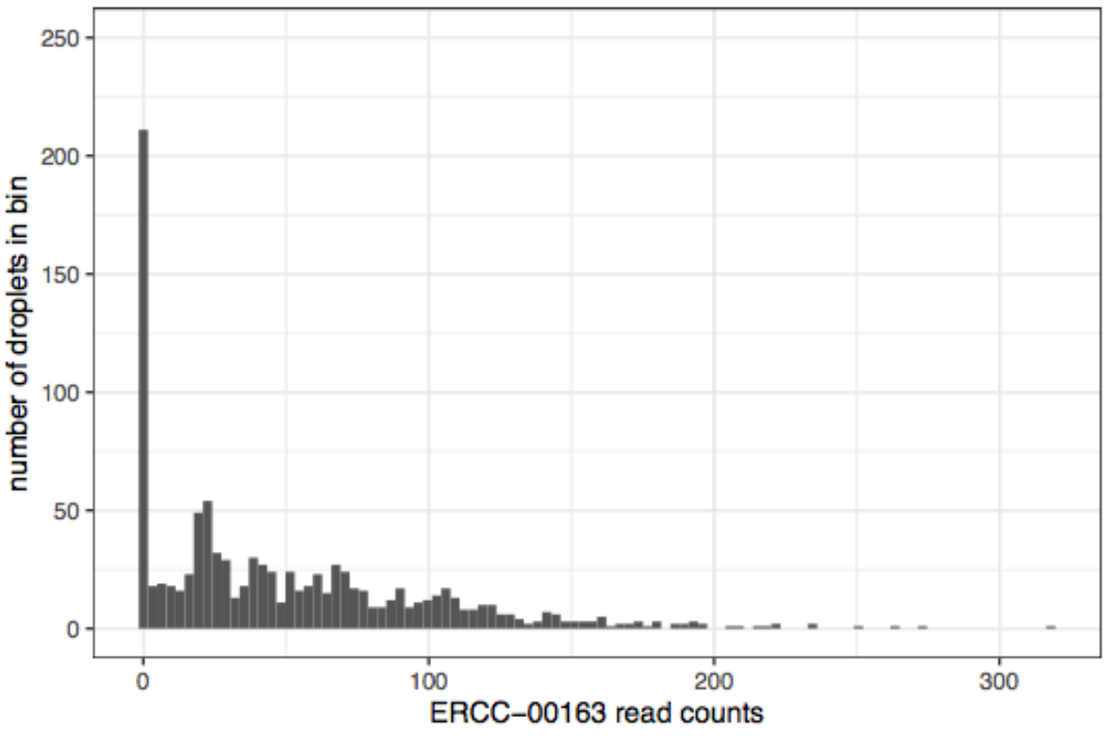


Amezquita et al. (2020). *Nature Methods.*

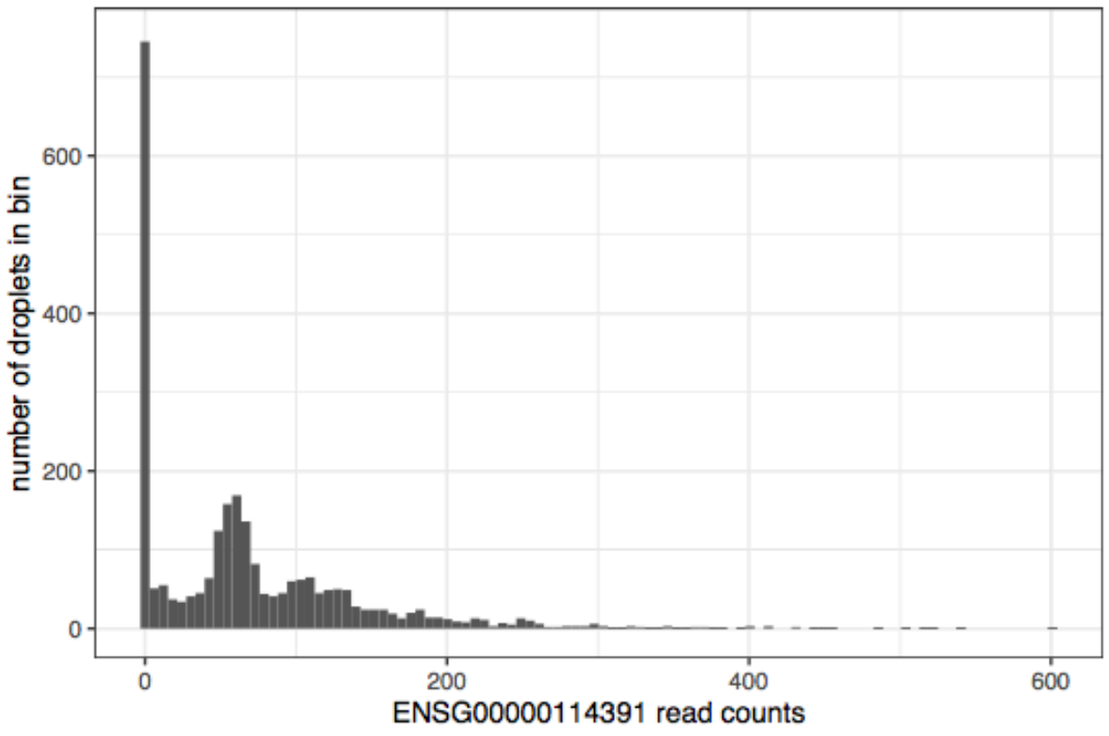
DATA PROPERTIES



READ COUNT DISTRIBUTION



(a) Read counts- technical replicates



(b) Read counts- biological replicates

Townes et al. (2019). *Genome Biology.*

THE ZERO-INFLATED NEGATIVE BINOMIAL

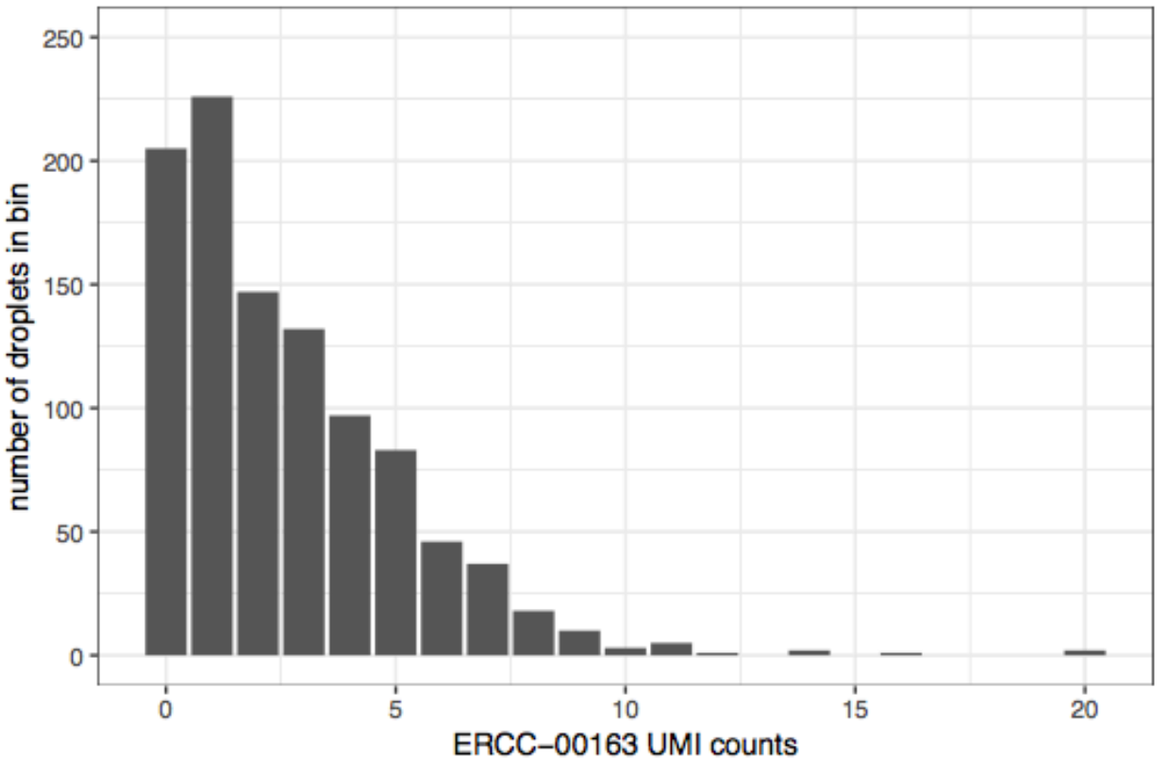
For any $\pi \in [0, 1]$, the PMF of the zero-inflated negative binomial (ZINB) distribution is given by

$$f_{ZINB}(y; \mu, \theta, \pi) = \pi\delta_0(y) + (1 - \pi)f_{NB}(y; \mu, \theta), \quad \forall y \in \mathbb{N},$$

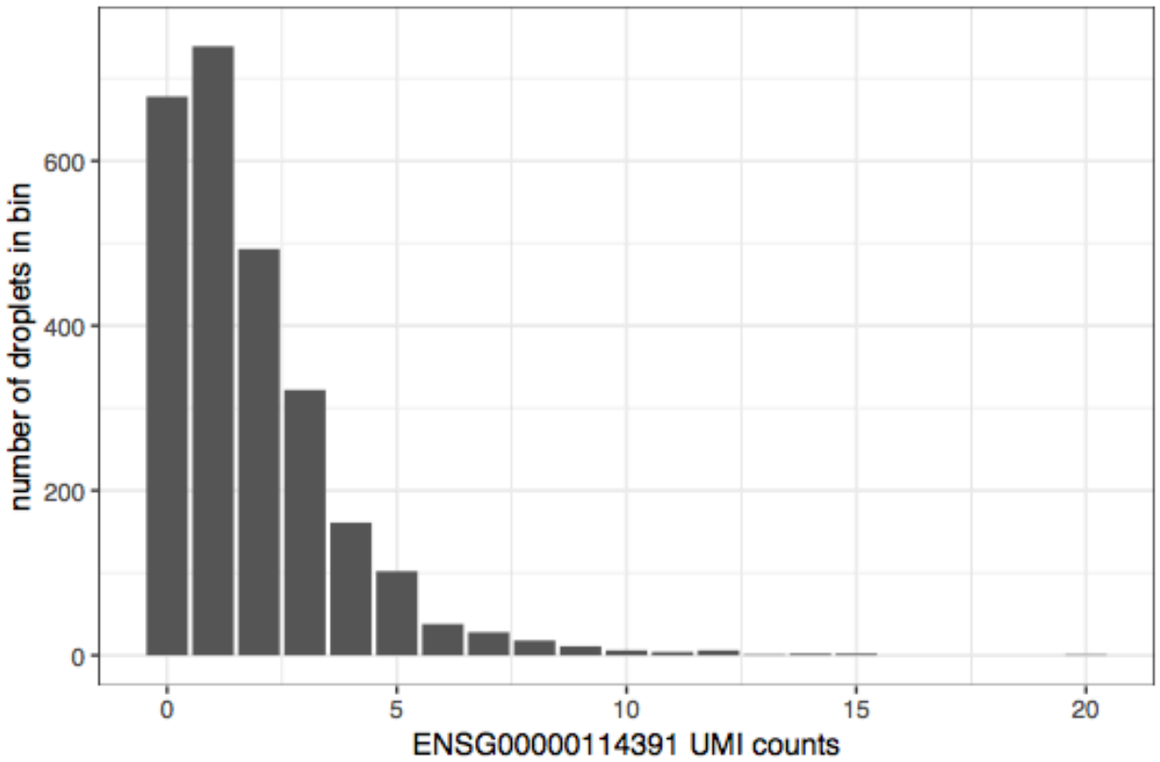
where $\delta_0(\cdot)$ is the Dirac function.

Here, π can be interpreted as the probability that a 0 is observed instead of the actual count, resulting in an inflation of zeros compared to the NB distribution, hence the name ZINB.

UMI COUNT DISTRIBUTION



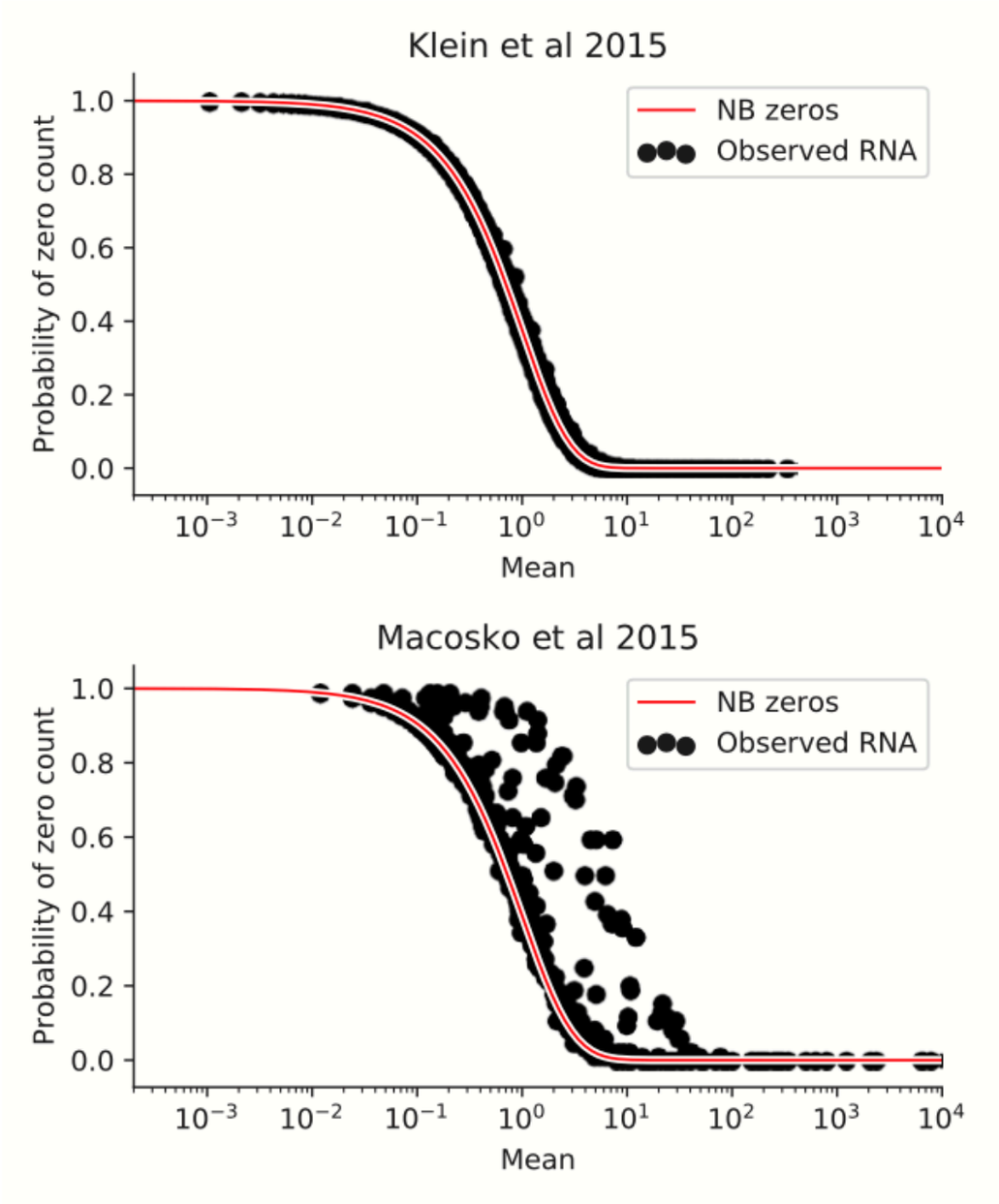
(c) UMI counts- technical replicates

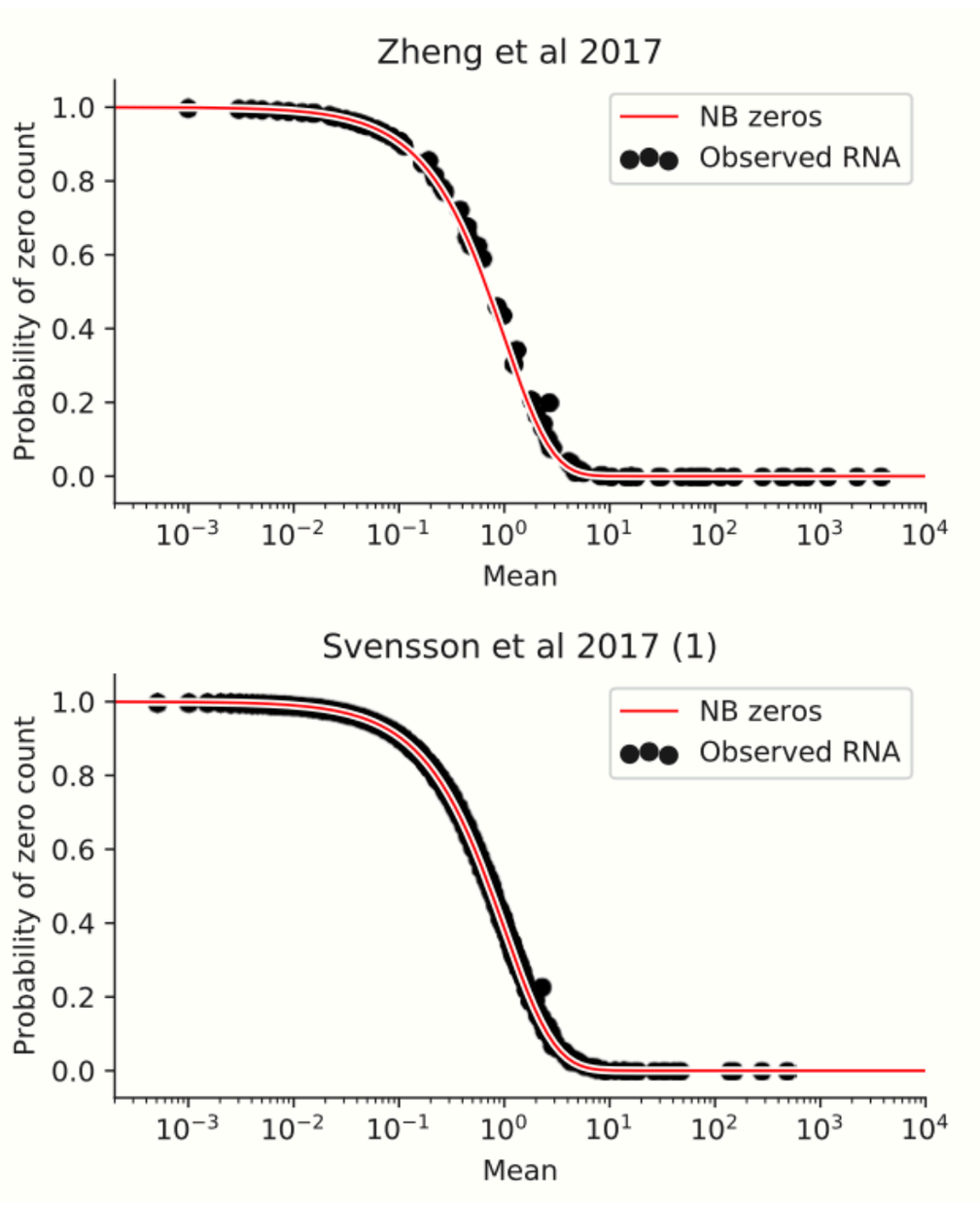


(d) UMI counts- biological replicates

Townes et al. (2019). *Genome Biology.*

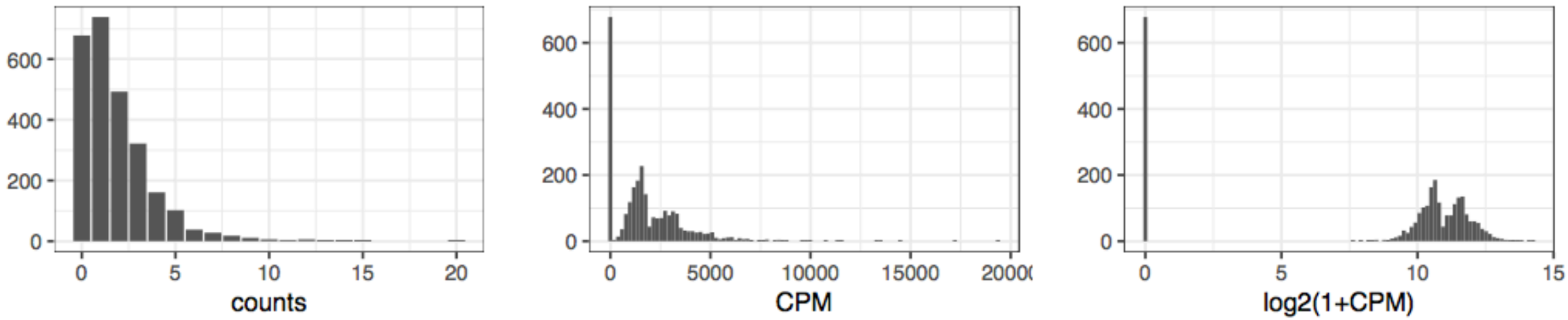
SHOULD WE MODEL ZERO INFLATION?





Svensson (2020). *Nature Biotechnology.*

LOGCPM TRANSFORMATION DOES NOT HELP!



(a) UMI counts (b) counts per million (CPM) (c) log of CPM

Townes et al. (2019). *Genome Biology.*

SHOULD WE MODEL ZERO INFLATION?

See also:
Vieth et al. (2017)
Jiang et al. (2022)
Nguyen et al. (2023)

- Non-UMI data: very likely.
- UMI data: probably not.

EXPLORATORY DATA ANALYSIS!

- Measurement vs expression models (Sarkar & Stephens 2021)

Table 1 | Single-gene models for scRNA-seq data

Expression model	Observation model	Method[a]
Point mass (no variation)	Poisson	Analytic
Gamma	Negative binomial	MASS[41], edgeR[42], DESeq2 (ref. [43]), SAVER[20], BASICS[44]
Point-Gamma	ZINB	PSCL[45]
Unimodal (nonparametric)	Unimodal	ashr[24,46]
Point-exponential family	Flexible	DESCEND[4]
Fully nonparametric[47]	Flexible	ashr

Different expression models, when combined with the Poisson measurement model, yield different observation models. [a]Previously published methods and software packages that use the corresponding observation model to analyze data.

Table 2 | Multigene models for scRNA-seq data

Link function	Noise distribution	Method[a]
Identity	None	NMF[48], scHPF[49]
Identity	Gamma	NBMF[50]
log	None	GLM-PCA[19]
log	Gamma	scNBMF[51], GLM-PCA[19]
log	Point-Gamma	ZINB-WaVE[52]
Neural network	Point-Gamma	scVI[29], DCA[21]

Multigene models partition variation in true expression into structured and stochastic components. The link function describes a transformation and the noise distribution indicates an assumption about the stochastic component. [a]Previously published methods and software packages that use the corresponding observation model to analyze data.

QUALITY CONTROL AND FILTERING

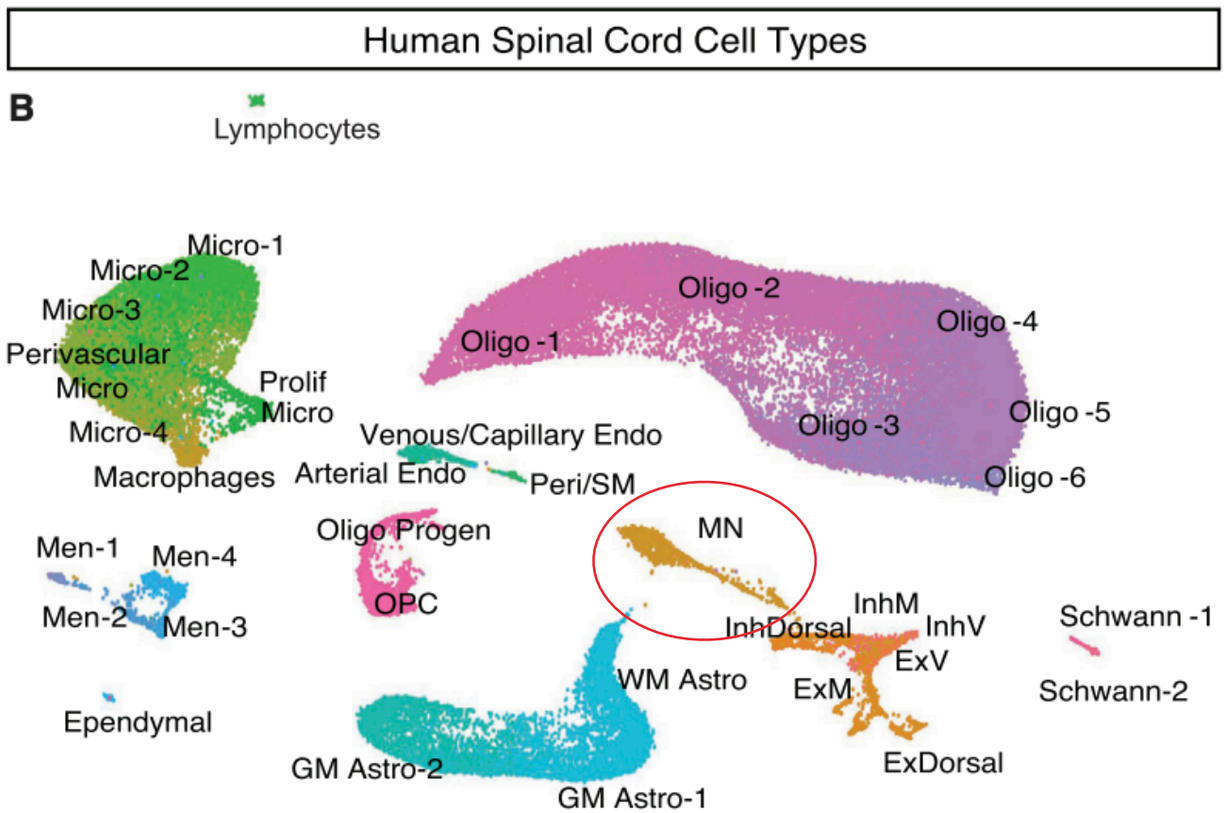
QUALITY CONTROL

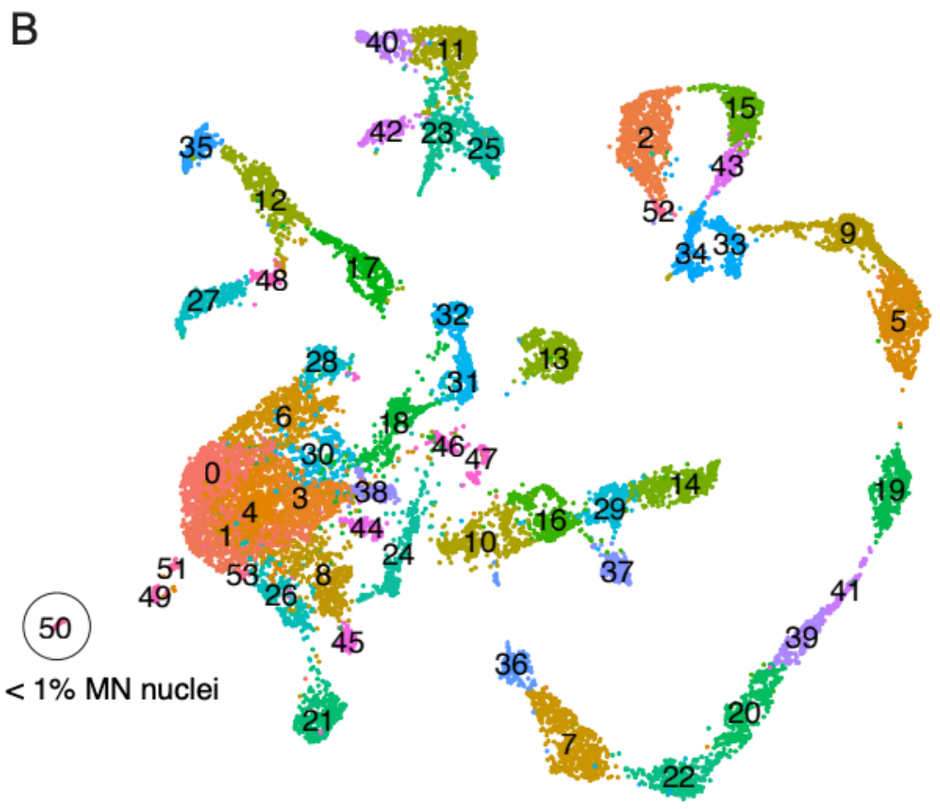
Unlike with “bulk” data, in single-cell datasets we can afford to remove some low-quality cells.

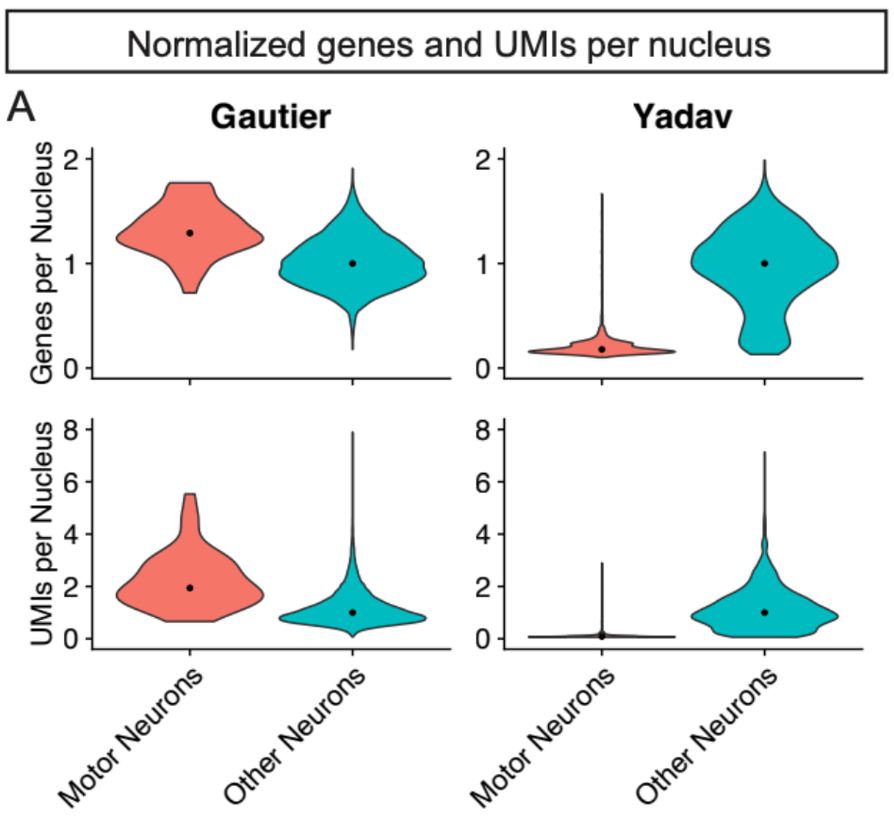
We typically want to remove low quality samples and outliers.

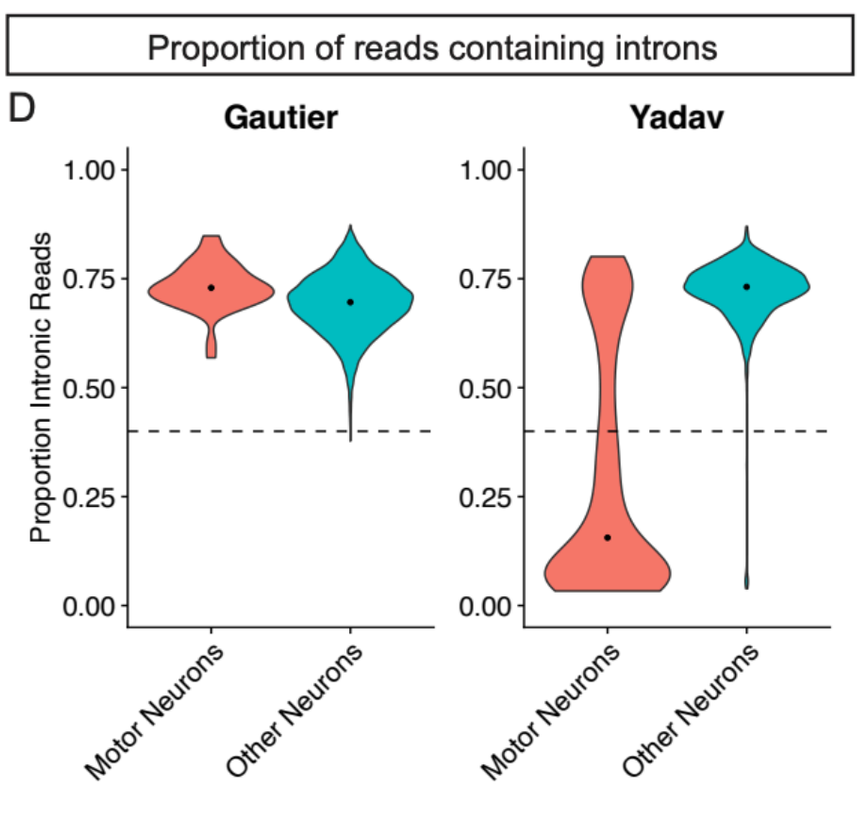
In addition, it is important to identify and remove empty droplets and doublets.

WHY IT MATTERS

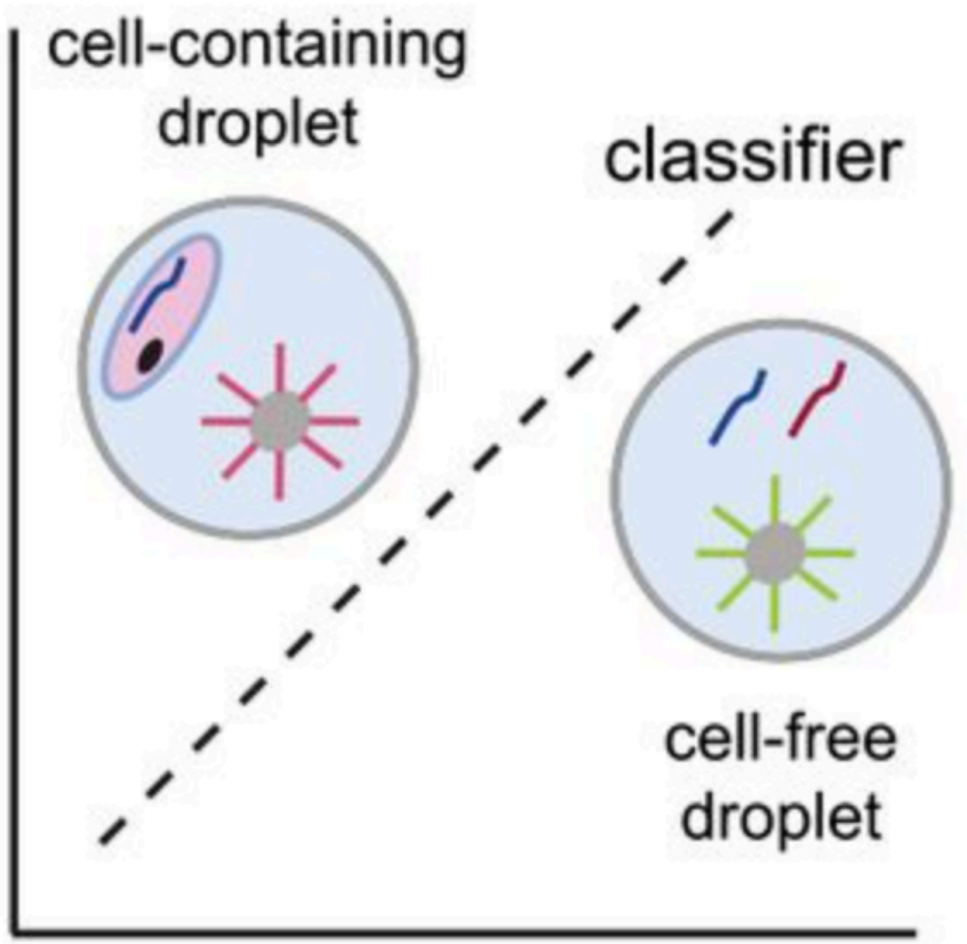








EMPTY DROPLETS

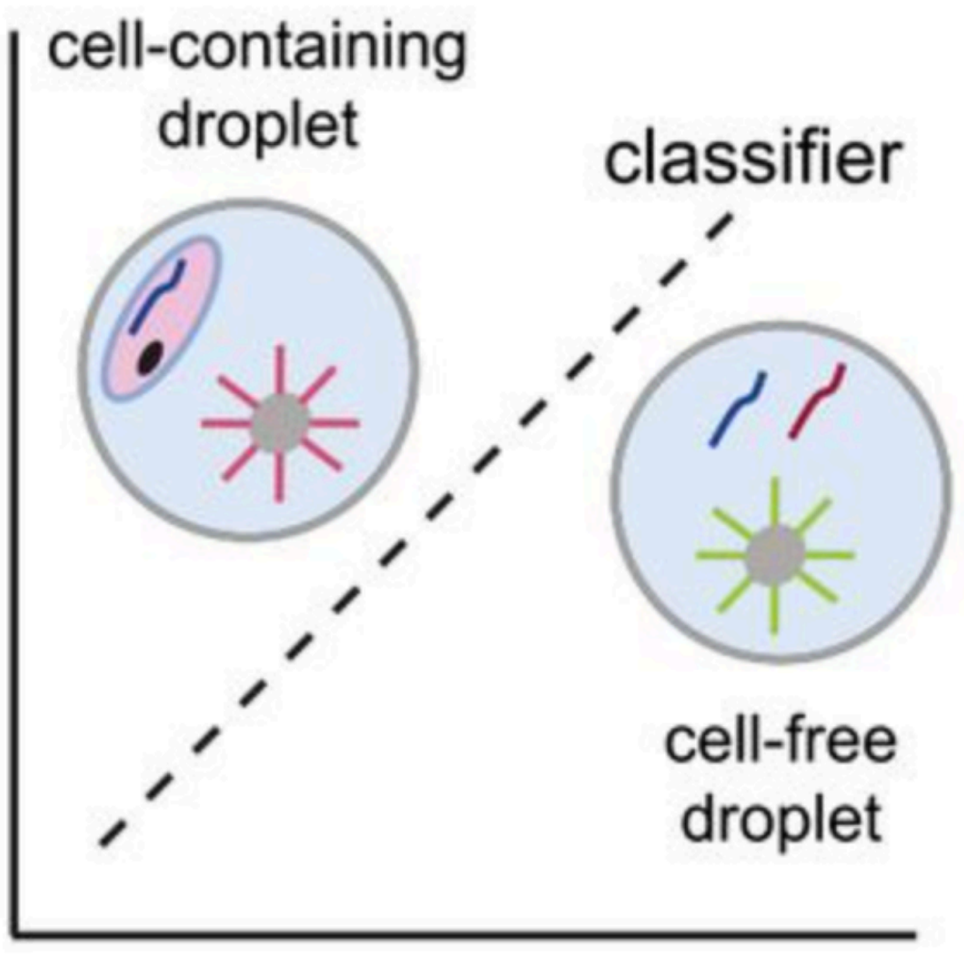


In droplet-based methods, we sequence the content of each droplet, *independently on whether they contain live cells.*

We may have contaminations due to cell raptures and ambient RNA.

We need statistical methods to identify and remove empty droplets.

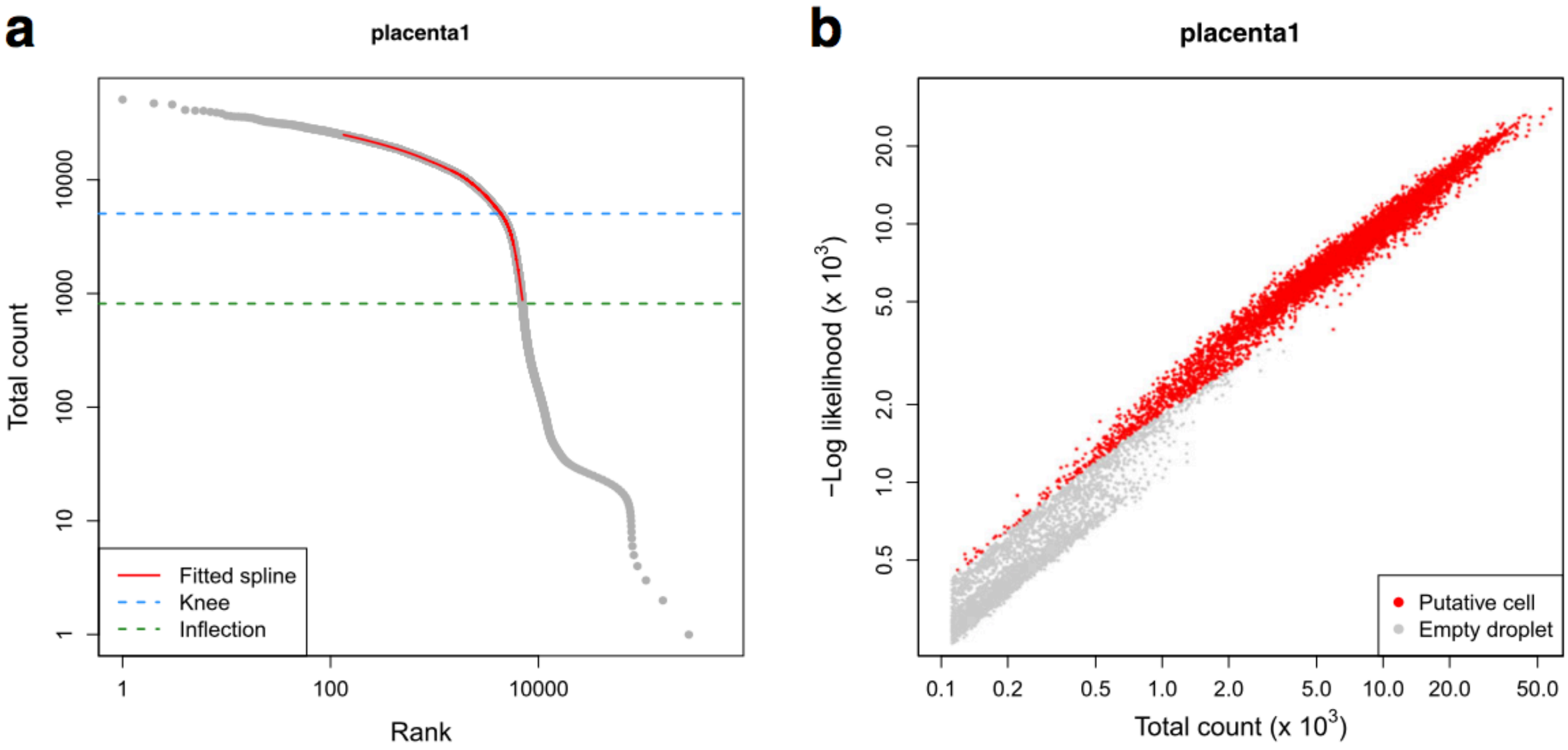
EMPTY DROPLETS



We can use an algorithm to:

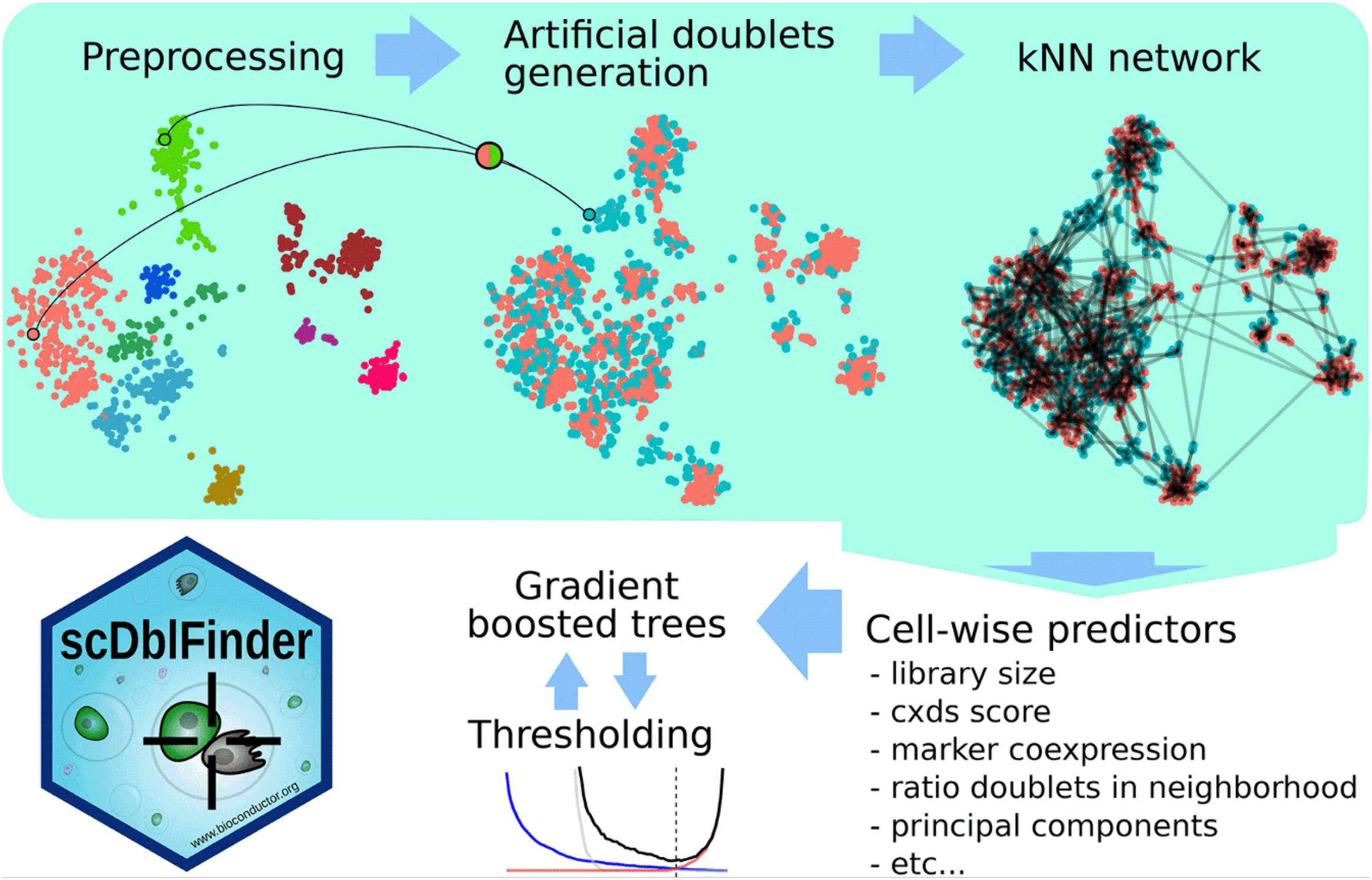
- Estimate ambient RNA profile using droplets that contain few UMI counts
- For each droplet, test the null hypothesis that its UMI count distribution is compatible with such profile

DROPLETUTILS BIOCONDUCTOR PACKAGE



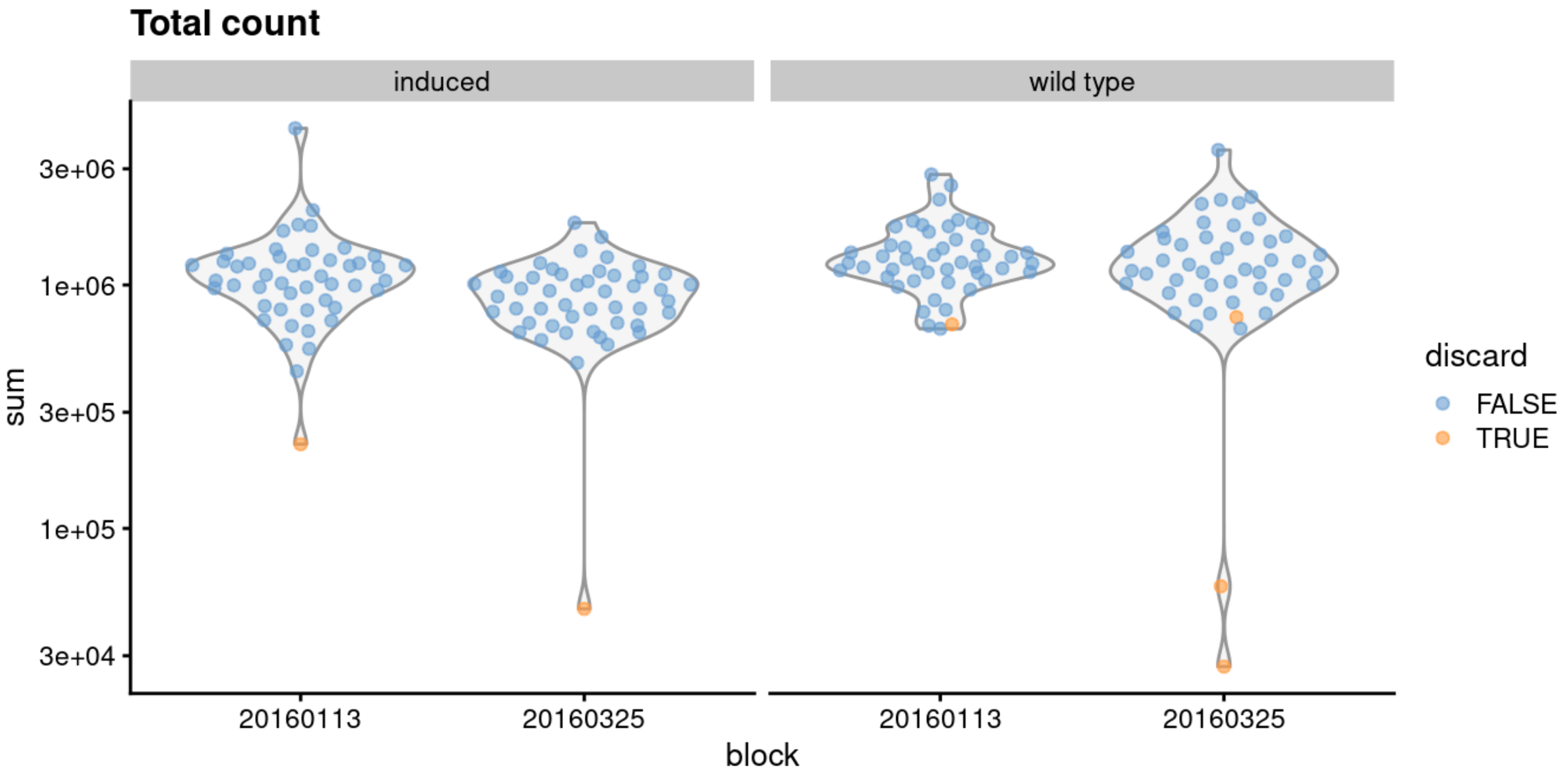
Lun et al. (2019). *Genome Biology.*

DOUBLET IDENTIFICATION

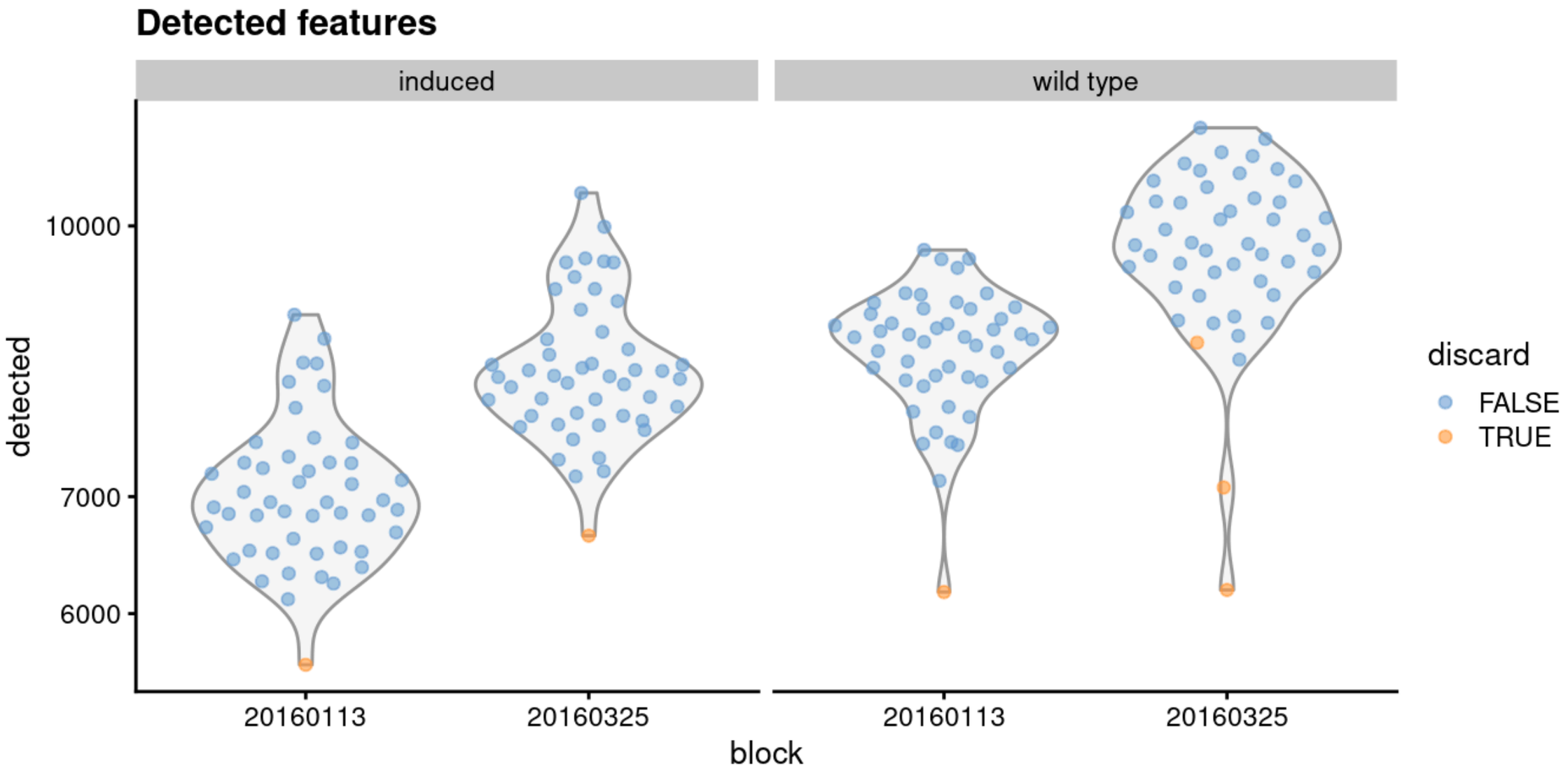


Germain et al. (2021)

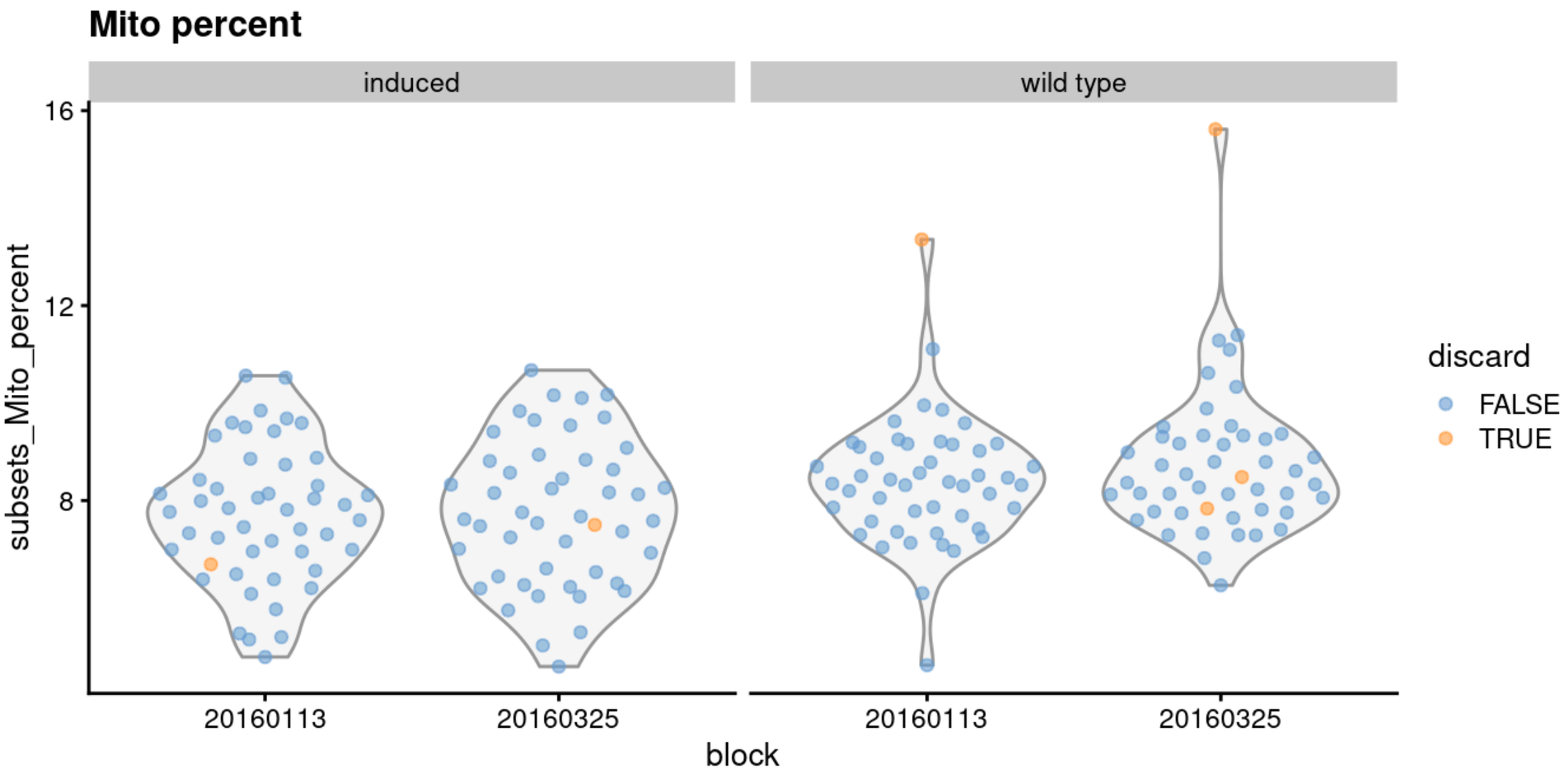
OUTLIER IDENTIFICATION



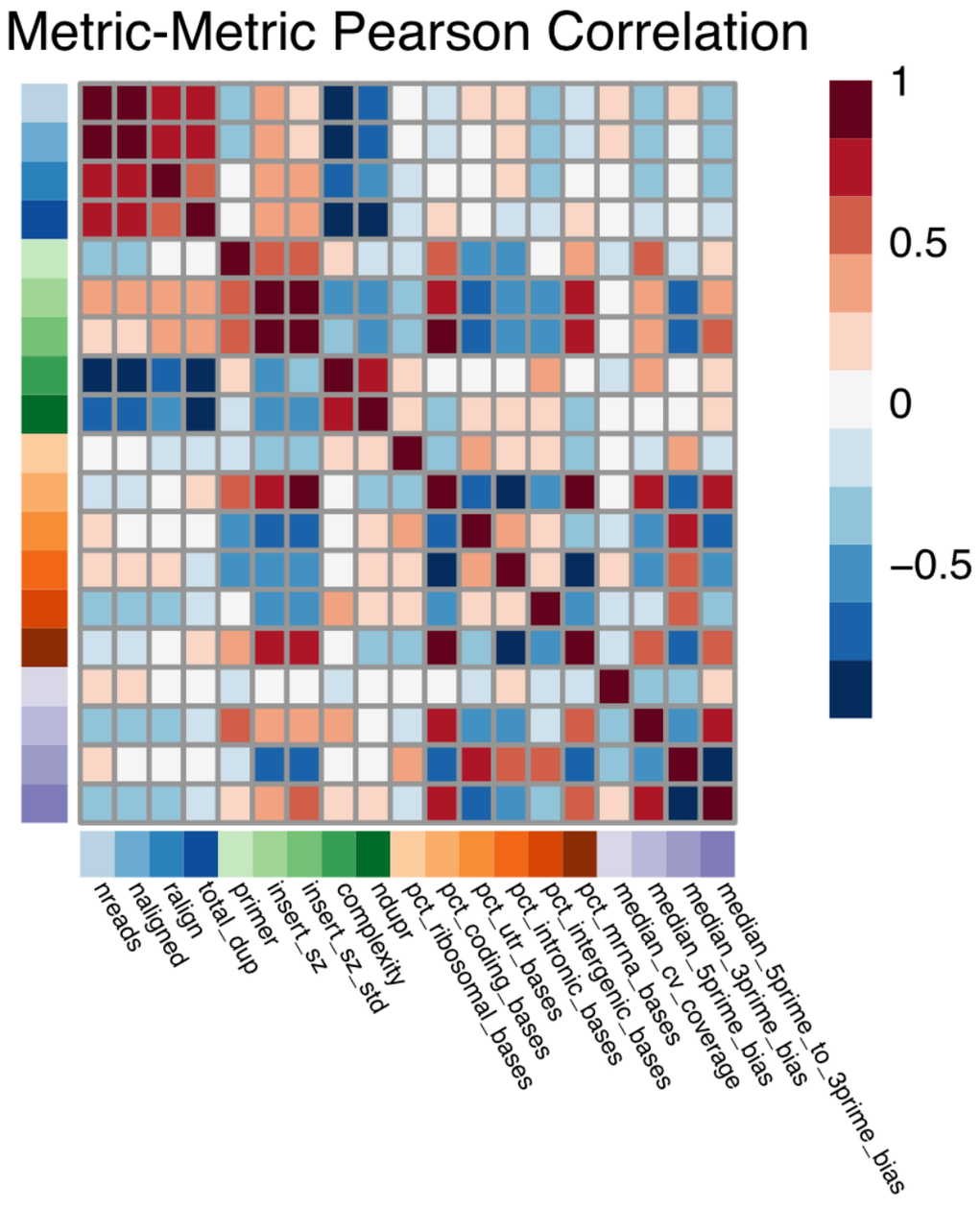
OUTLIER IDENTIFICATION

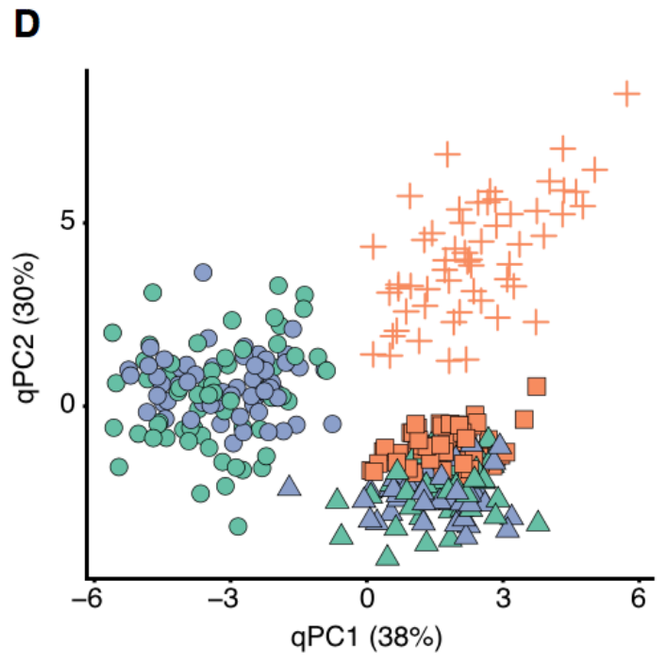


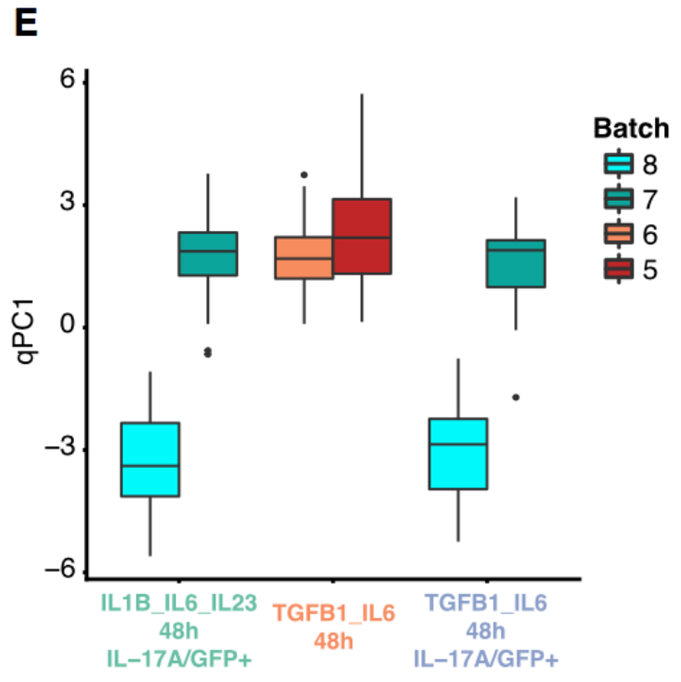
OUTLIER IDENTIFICATION



MANY CORRELATED METRICS



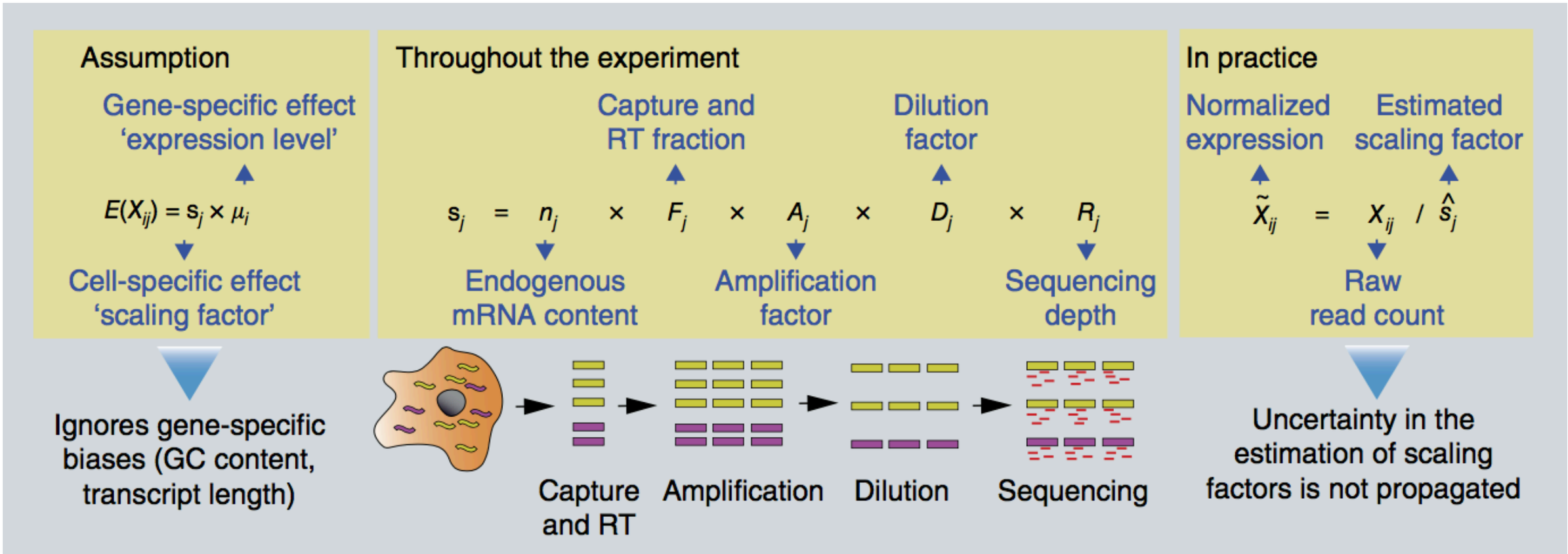




Cole et al. (2019). *Cell Systems*.

NORMALIZATION

As for bulk RNA-seq, we need to account for sequencing depth and other cell-specific technical effects.



Vallejos et al. (2017). *Nature Methods*.

EXAMPLE: TABULA MURIS DATA

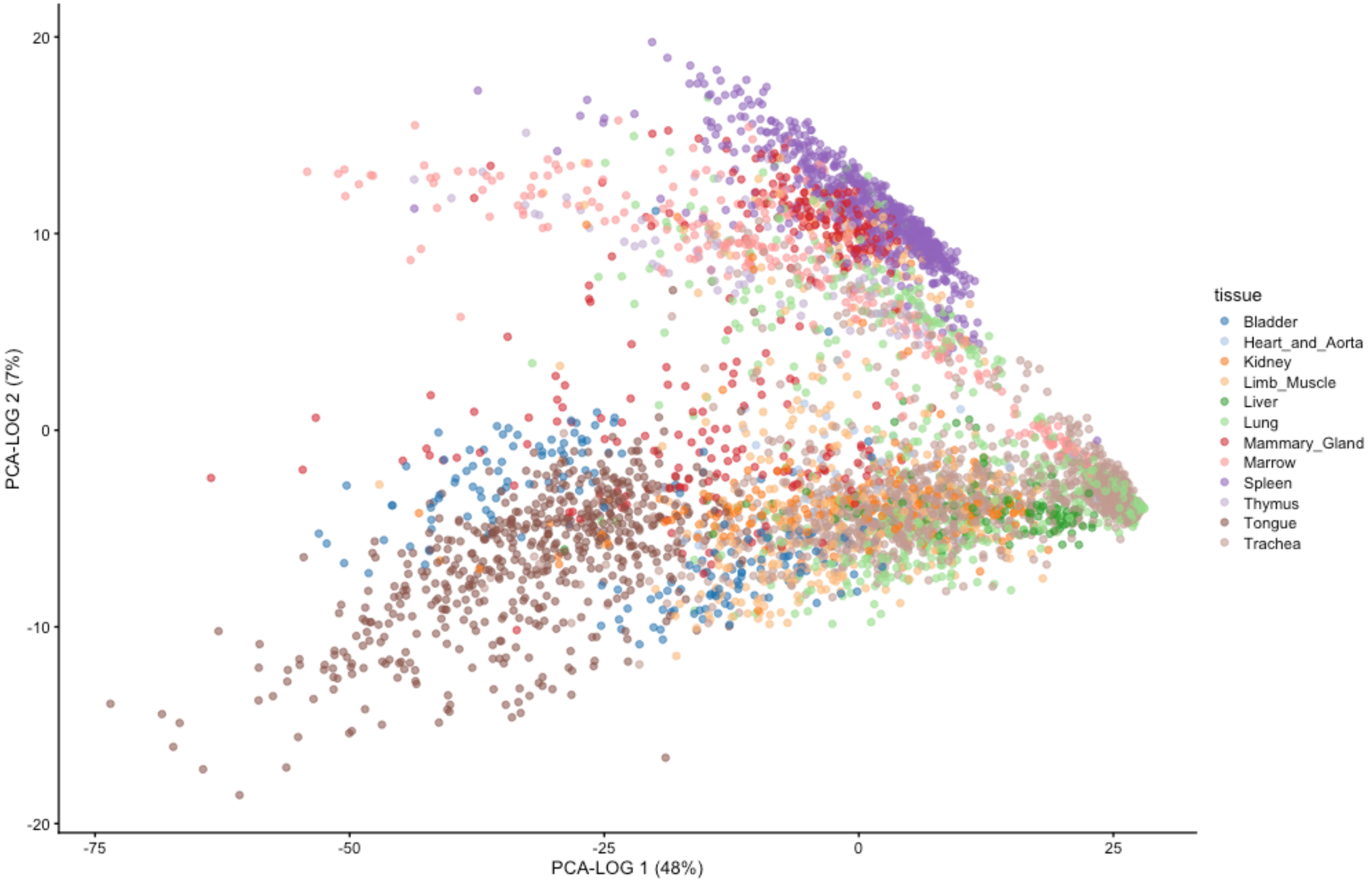
Tabula Muris is a project aimed at characterizing all cell types in the mouse.

The droplet dataset comprised 70,000 cells from 12 tissues.

We see here a random subset of 5,000 cells, limiting the dataset to the 1,000 most variable genes.

TabulaMurisData Bioconductor package.

TABULA MURIS: PCA (LOG SCALE)



TABULA MURIS: PCA (LOG SCALE)

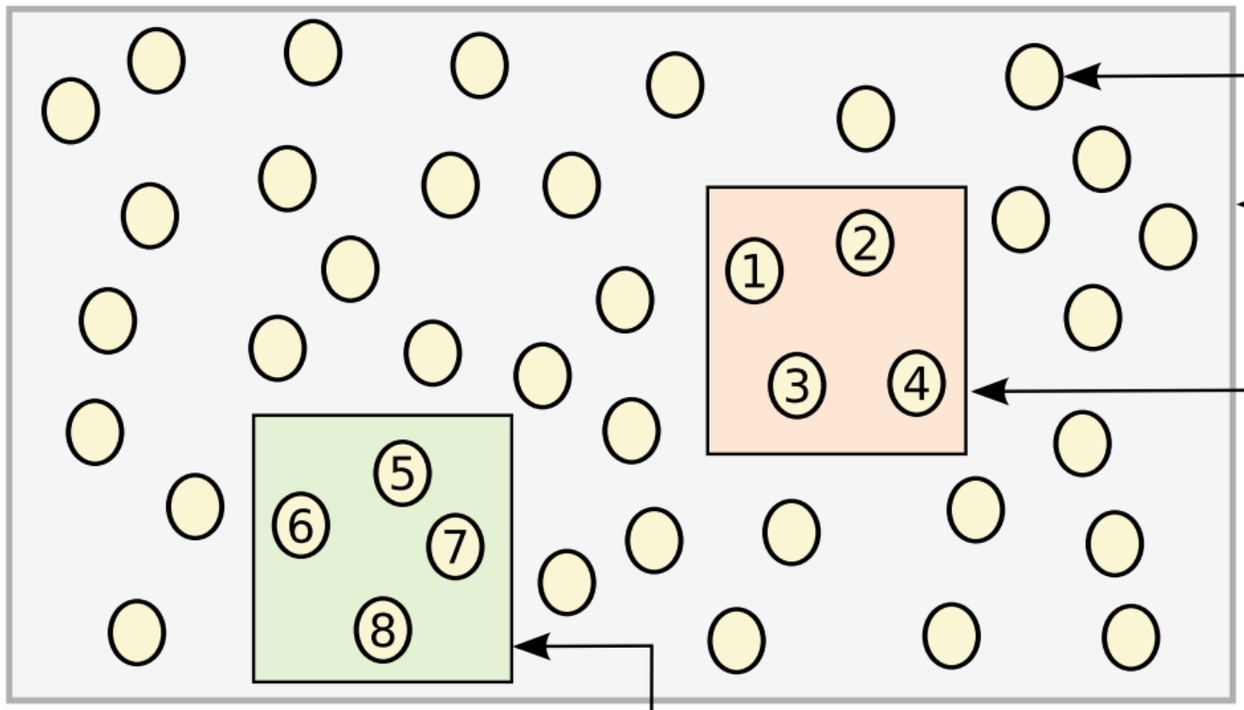
```
Call:
lm(formula = pc1 ~ subset$sum + subset$detected)

Residuals:
    Min      1Q  Median      3Q     Max
-15.201  -3.817   0.037   3.710  34.968

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.360e+01  1.546e-01  152.69   <2e-16 ***
subset$sum      -2.791e-04  2.036e-05  -13.71   <2e-16 ***
subset$detected -1.369e-02  1.435e-04  -95.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.747 on 4997 degrees of freedom
Multiple R-squared:  0.9057,    Adjusted R-squared:  0.9057
F-statistic: 2.4e+04 on 2 and 4997 DF,  p-value: < 2.2e-16
```

SCRAN NORMALIZATION



Single cell

All cells (averaged to make a reference pseudo-cell)

Cell pool A:

$\theta_1 + \theta_2 + \theta_3 + \theta_4 = \theta_A$

Cell pool B:

$\theta_5 + \theta_6 + \theta_7 + \theta_8 = \theta_B$

System of linear equations:

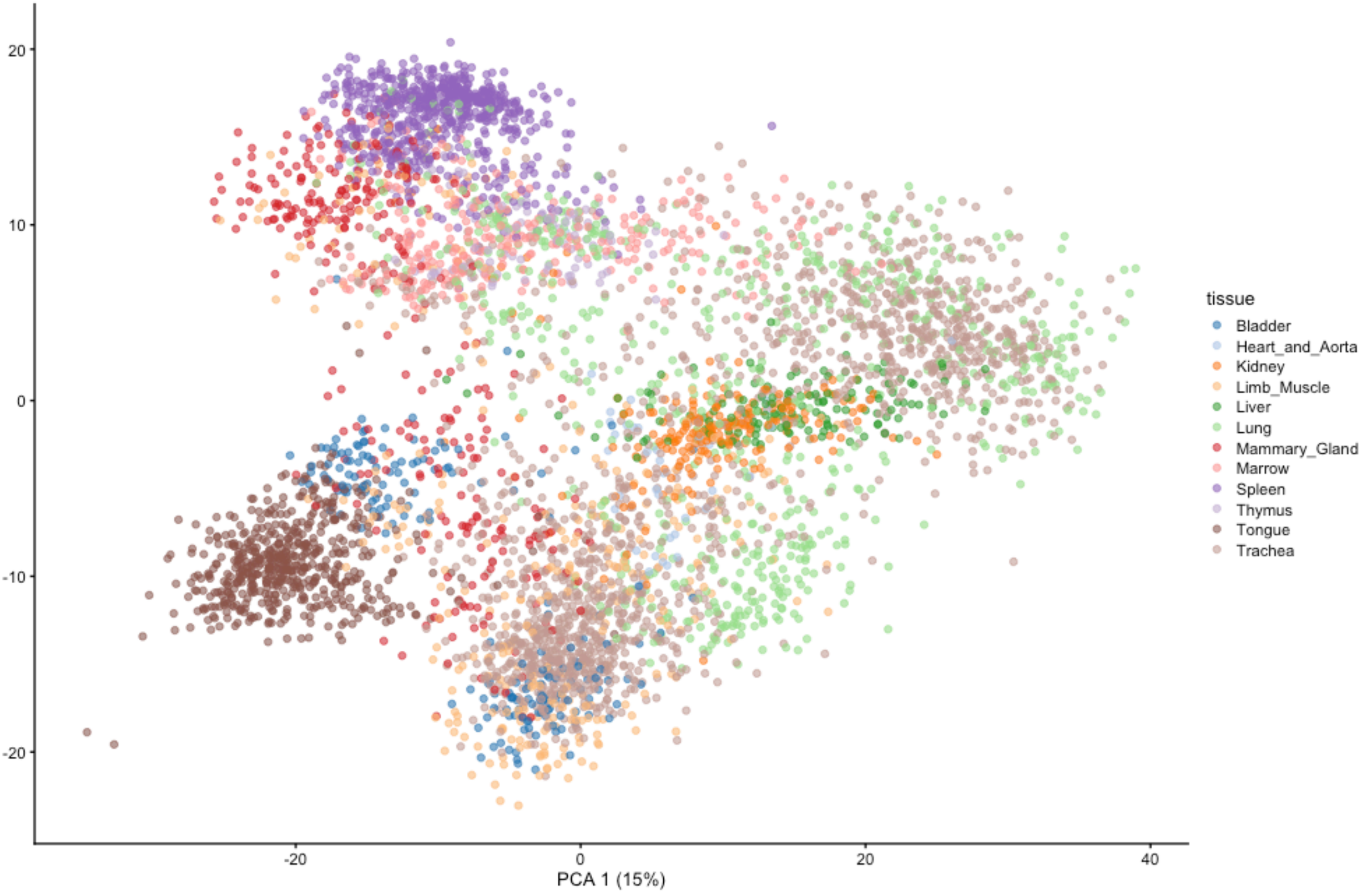
$$\begin{bmatrix} 1\,1\,1\,1\,0\,0\,0\,0\,\ldots \\ 0\,0\,0\,0\,1\,1\,1\,1\,\ldots \\ 1\,0\,1\,0\,1\,0\,1\,0\,\ldots \\ 0\,1\,1\,0\,1\,1\,0\,0\,\ldots \\ \ldots \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \ldots \end{bmatrix} = \begin{bmatrix} \theta_A \\ \theta_B \\ \theta_C \\ \theta_D \\ \ldots \end{bmatrix}$$

The computation of normalization factors may be difficult due to the many zero counts.

One approach is to pool together cells to compute the normalization factors and then deconvolve them back at single-cell resolutions.

Lun et al. (2016). *Genome Biology*.

TABULA MURIS: PCA AFTER SCRAN NORMALIZATION (LOG SCALE)



TABULA MURIS: PCA AFTER SCRAN NORMALIZATION (LOG SCALE)

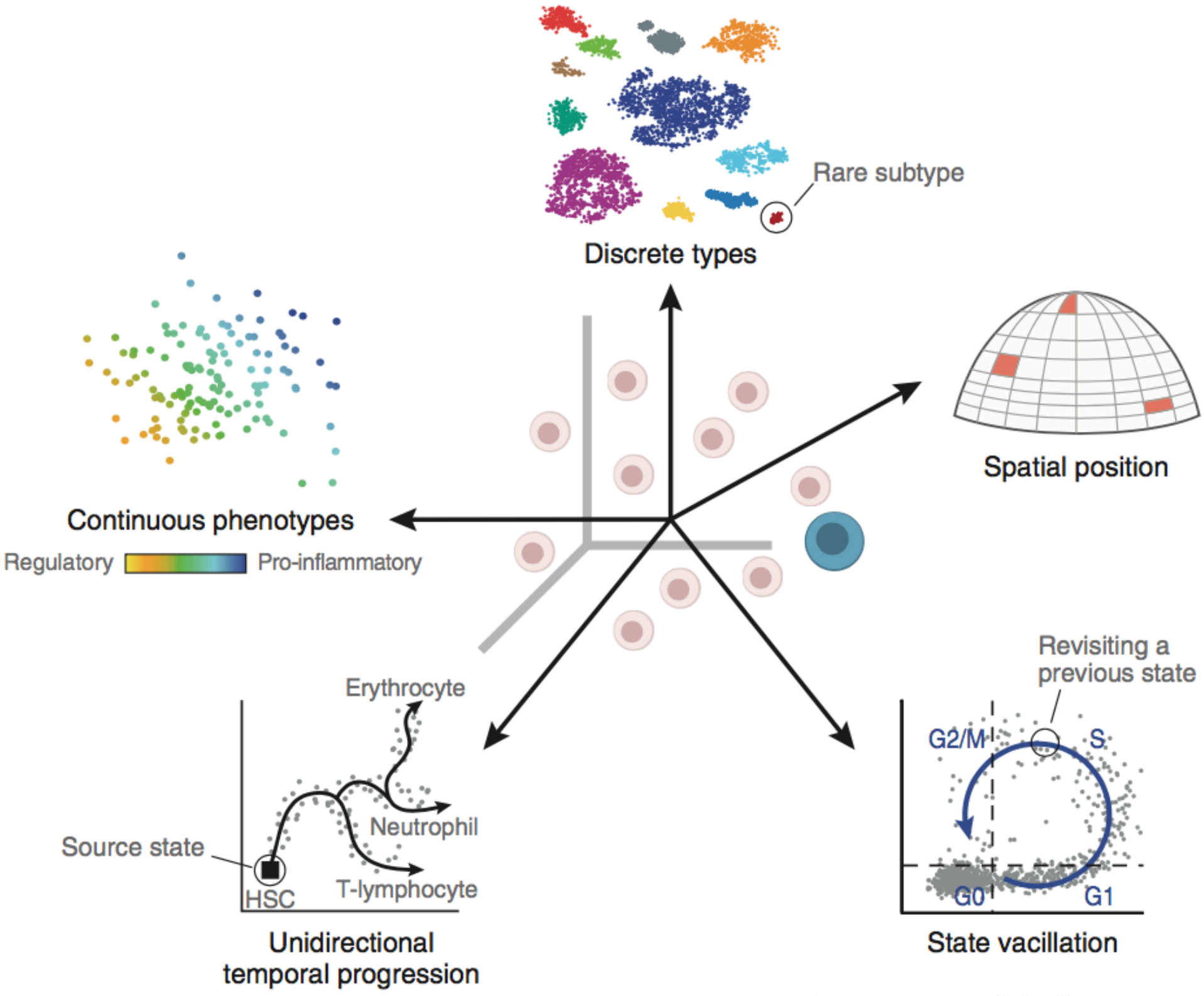
```
Call:
lm(formula = pc1ls ~ subset$sum + subset$detected)

Residuals:
    Min      1Q  Median      3Q     Max
-30.900  -8.646  -0.446   8.452  45.700

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.298e+01  3.078e-01  42.179   <2e-16 ***
subset$sum       6.363e-05  4.055e-05   1.569    0.117
subset$detected -8.403e-03  2.857e-04 -29.407   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.44 on 4997 degrees of freedom
Multiple R-squared:  0.3953,    Adjusted R-squared:  0.3951
F-statistic:  1633 on 2 and 4997 DF,  p-value: < 2.2e-16
```

DIMENSIONALITY REDUCTION



Wagner, Regev, Yosef (2016). *Nature Biotechnology.*

DIMENSIONALITY REDUCTION

We talk about "dimensionality reduction" when referring to two different goals:

1. **Visualize** high-dimensional data
    - Usually 2-3 dimensions
    - Non-linear, local techniques
2. **Infer** low-rank signal from high-dimensional data
    - Usually 10-50 dimensions
    - Factor analysis models

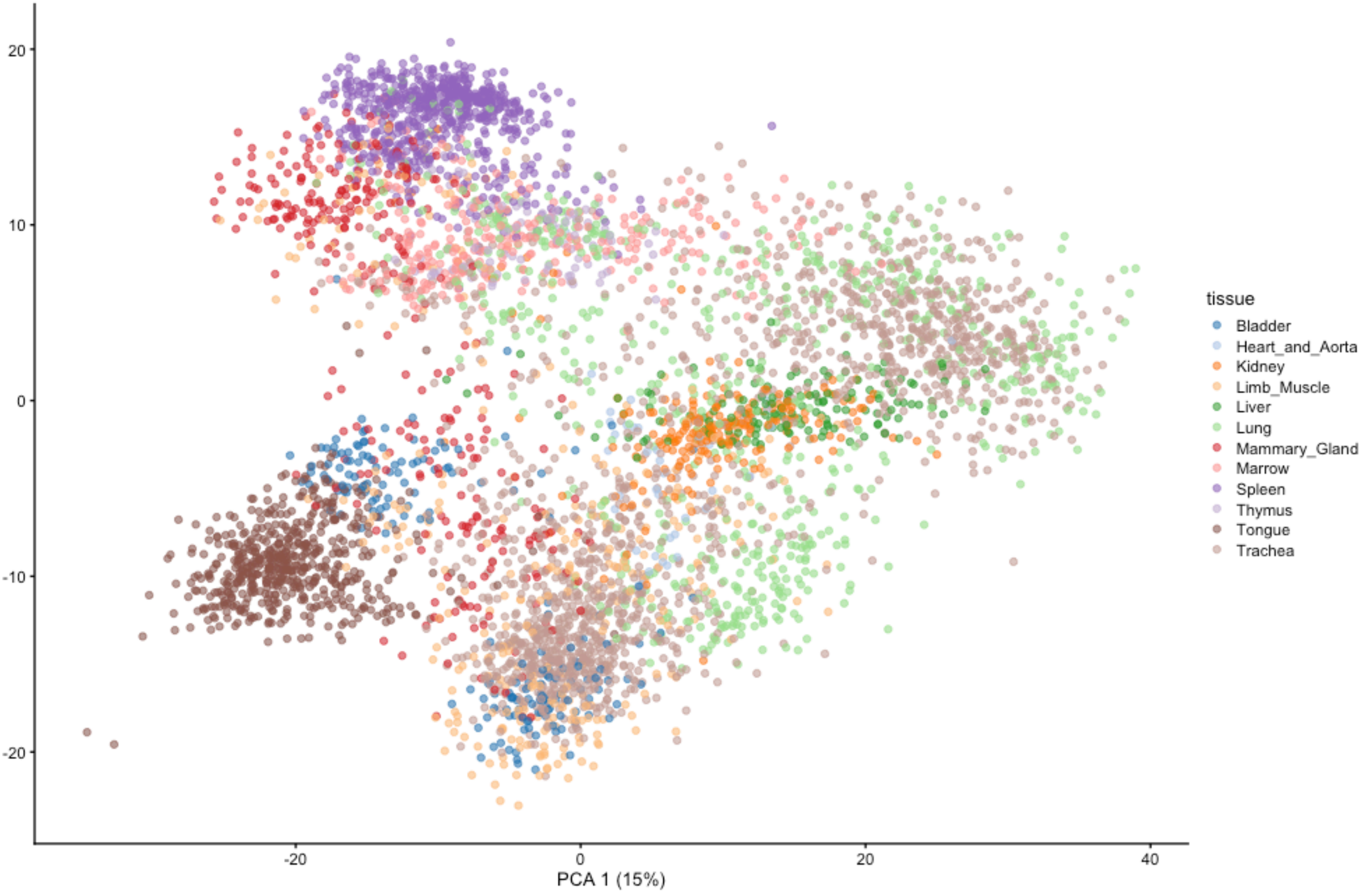
PRINCIPAL COMPONENT ANALYSIS (PCA)

- PCA is the starting point and baseline approach for both types of analysis.
- PCA can be used to visualize high-dimensional data in 2-3 dimensions.
- PCA can be seen as a solution of a factor analysis model for Gaussian data.

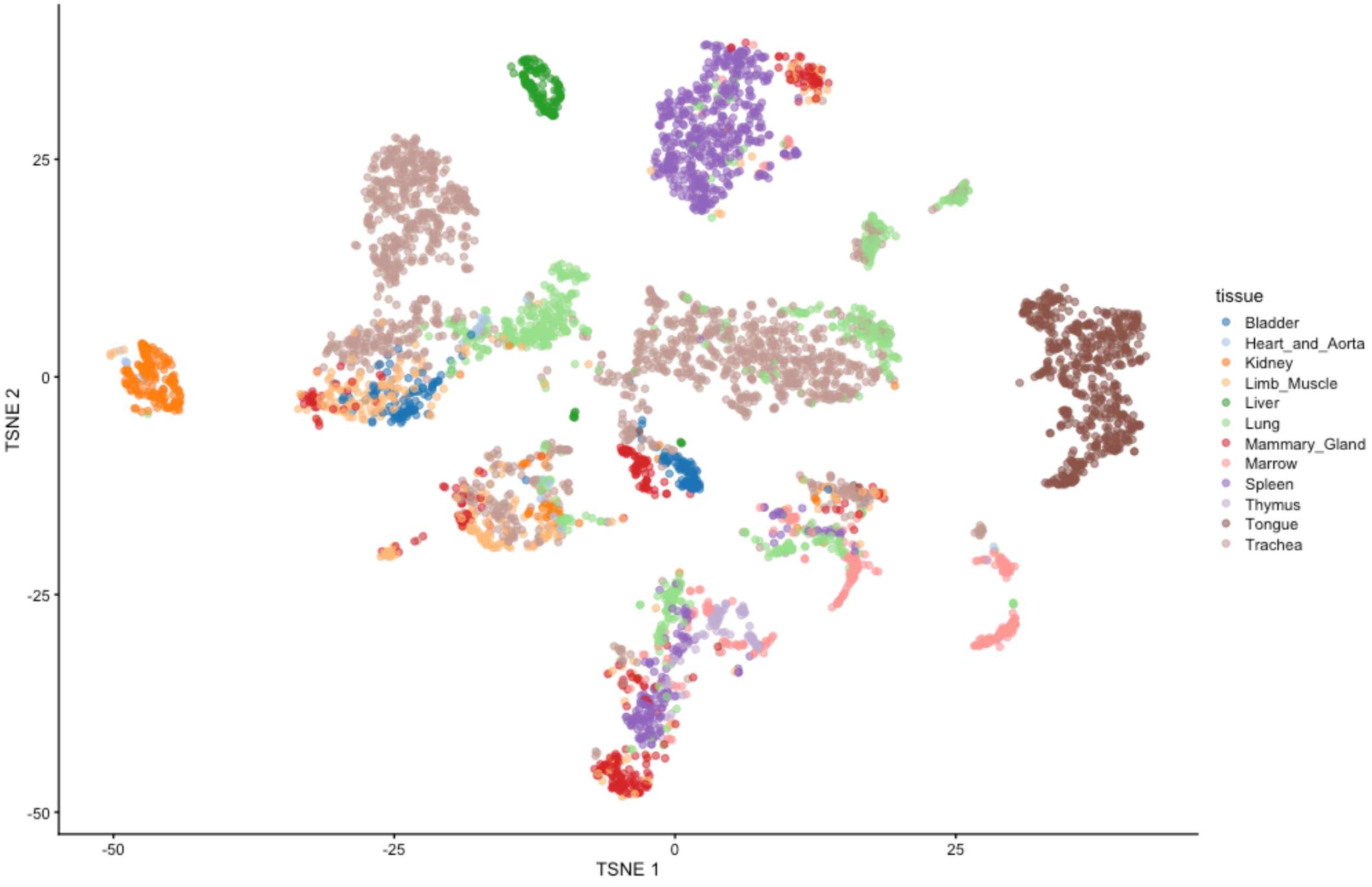
DESIRED PROPERTIES OF DIMENSIONALITY REDUCTION MODELS

- Accounting for the count nature of the data, over-dispersion, and possibly zero inflation.
- General and flexible.
- Extract low-dimensional signal from the data.
- Adjust for complex effects (batch effects, sample quality).
- Scalable.

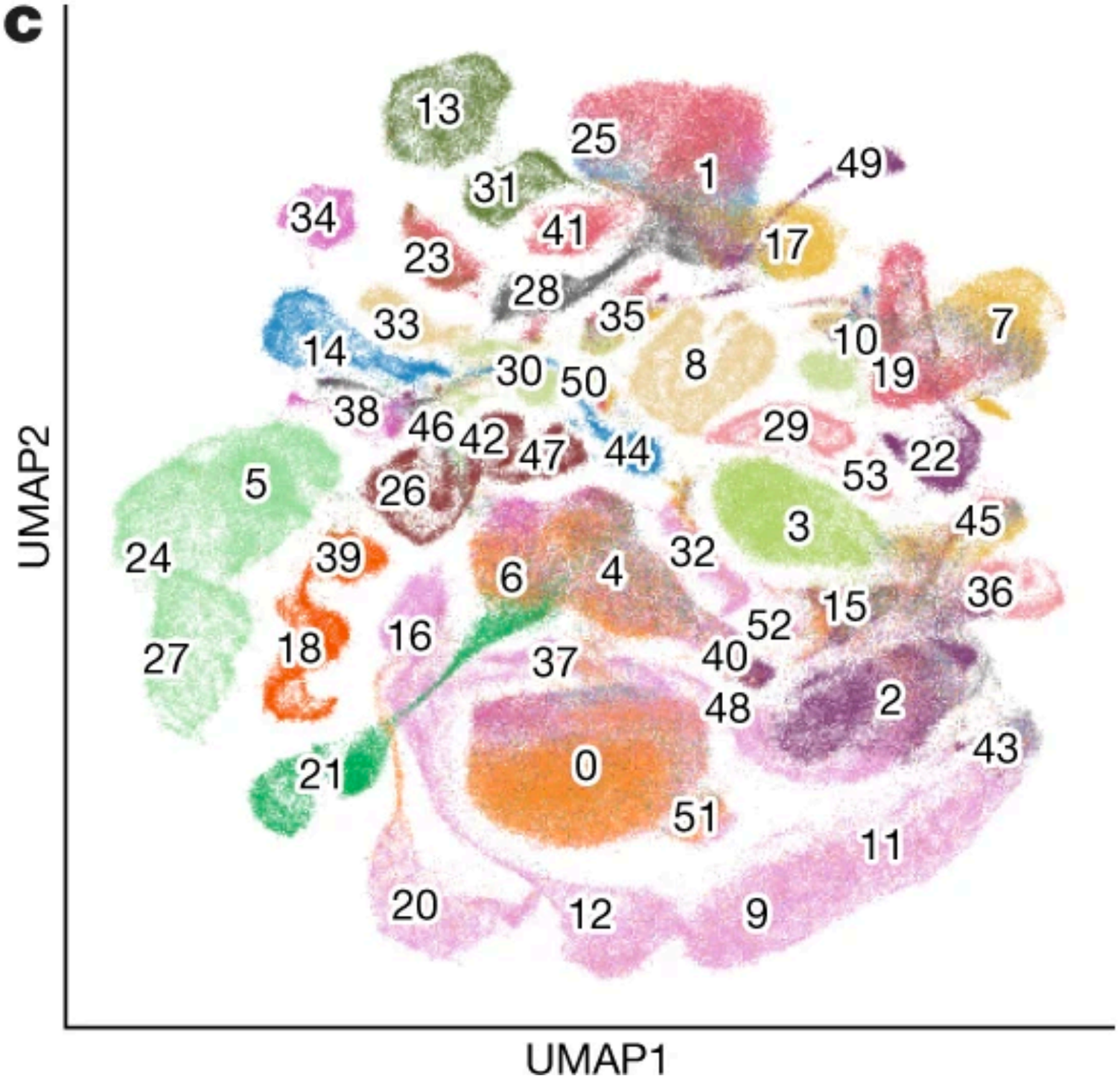
TABULA MURIS: PCA AFTER SCRAN NORMALIZATION (LOG SCALE)



TABULA MURIS: T-SNE



TABULA MURIS: UMAP (ALL CELLS)



PCA IS A LINEAR METHOD

- One way to define the first principal component is: the **linear combination** of the original variables that explain the most variability in the data.
- Similarly, subsequent PCs are linear combinations of the original variables that are orthogonal to the first and explain the most variance among the orthogonal combinations.
- Are we limiting ourselves by only looking at **linear** combinations?
- Would a non-linear method have more flexibility in explaining our data?

T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (T-SNE)

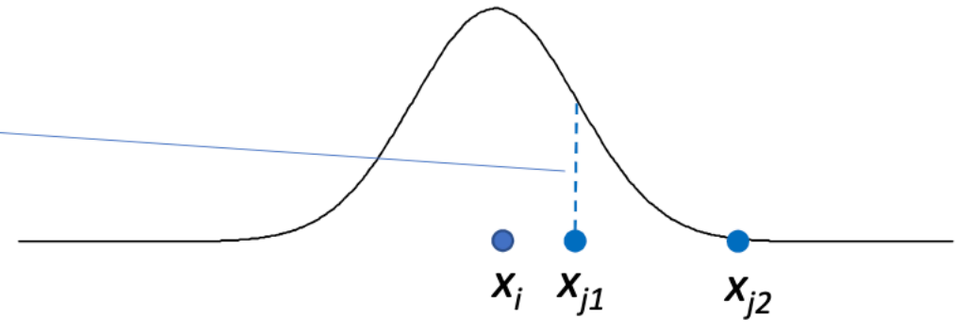
- One option, very popular in single-cell genomics, is t-distributed Stochastic Neighbor Embedding (t-SNE).
- Briefly, the problem that we want to solve is to represent in a 2-3 dimensional map (*embedding*) the observations from a high-dimensional space preserving as much as possible the distance between points.

STOCHASTIC EMBEDDING: PROBABILISTIC REPRESENTATION OF DISTANCES

- Similarity between two points, x_i and x_j in the ***original high-dimensional space*** is given by

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

The denominator scales the sum of all the scores to 1



- Essentially, the probability that x_i would pick x_j as its neighbor.

- We set $p_{i|i} = 0$ and actually use a symmetrized version that ensures $p_{ij} = p_{ji}$.

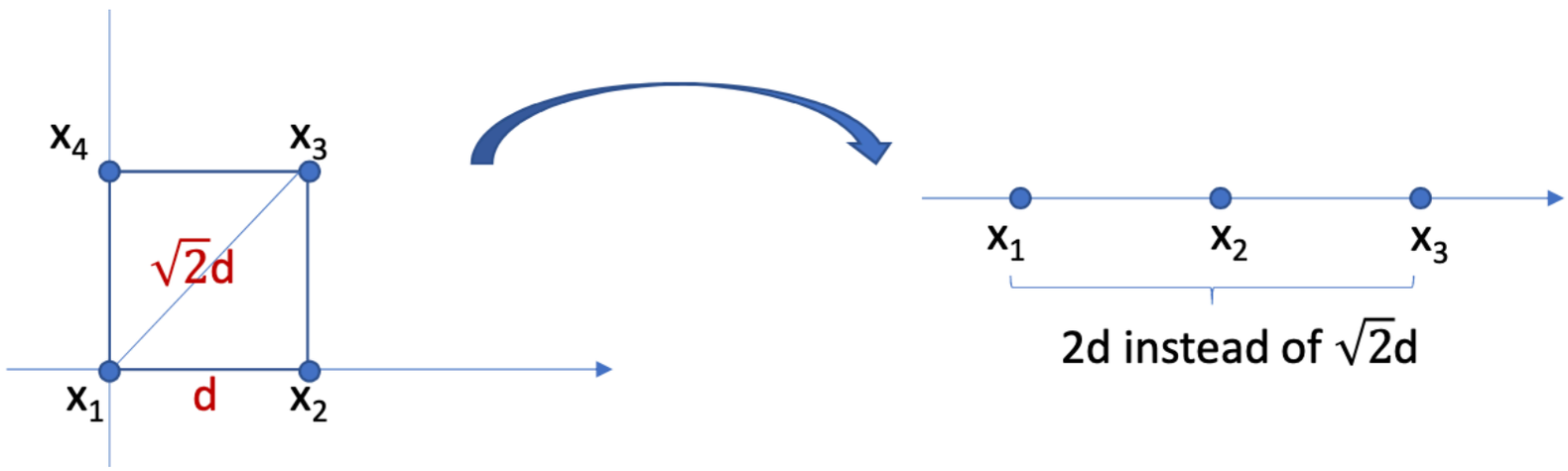
Hinton and Roweis (2002)

STOCHASTIC EMBEDDING: PROBABILISTIC REPRESENTATION OF DISTANCES

- We could define a similar density in the ***low-dimensional space***, but we use a t-distribution instead of a Gaussian kernel

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq l} \exp\left(-\|y_k - y_l\|^2\right)}, \qquad q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}.$$

- The *t* distribution has heavier tails and partially account for the *crowding problem*.

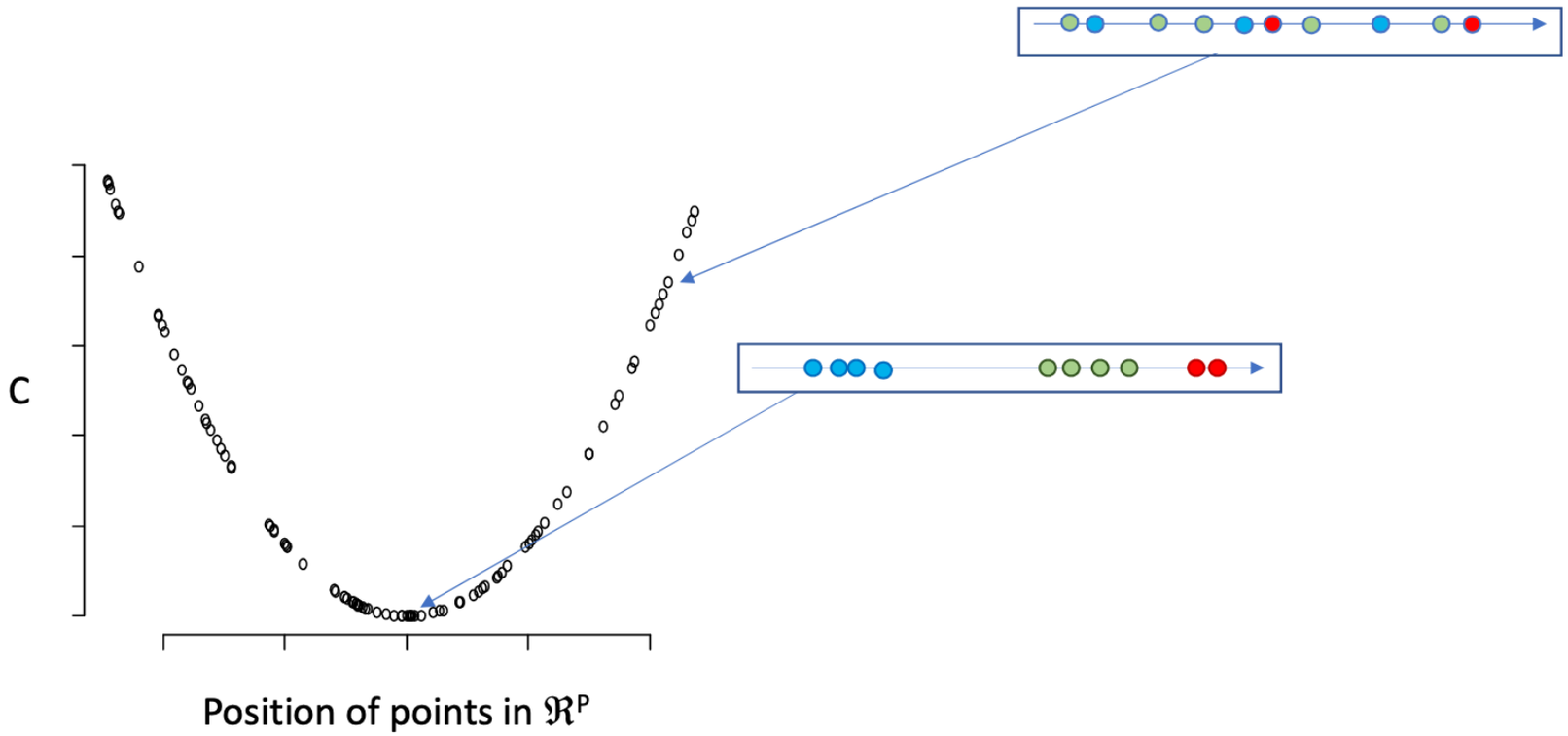


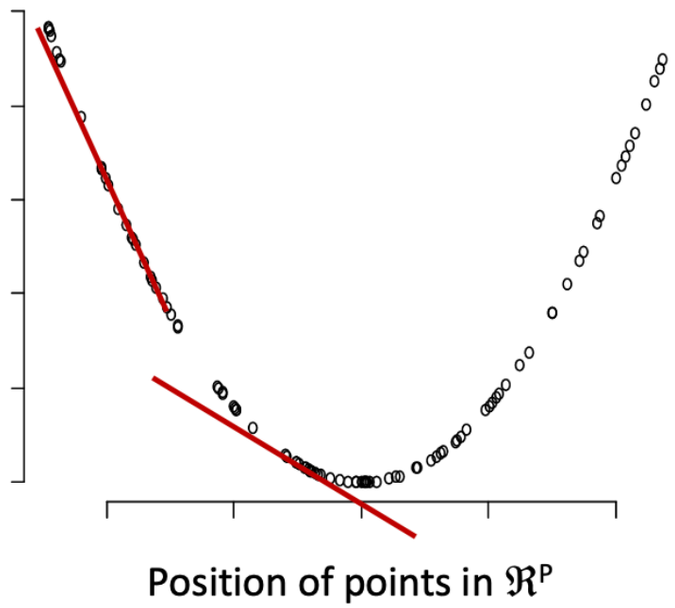
T-SNE ALGORITHM

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq l} \exp\left(-\|y_k - y_l\|^2\right)}, \qquad p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\|x_k - x_l\|^2 / 2\sigma^2\right)},$$

- We minimize the Kullback-Leibler (KL) divergence between the two distributions with *gradient descent*.

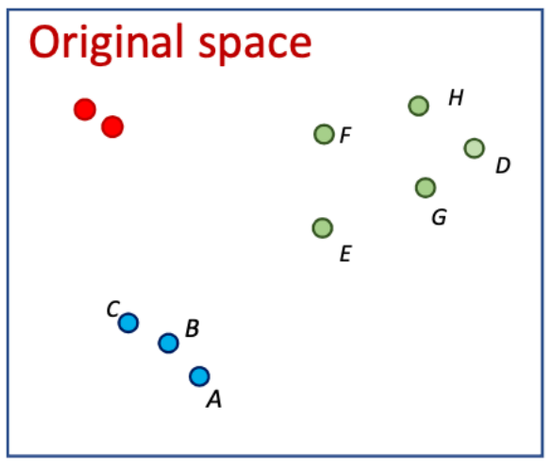


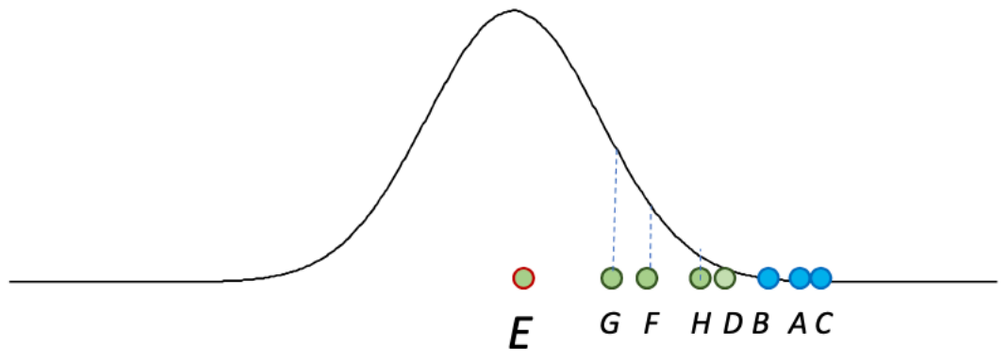


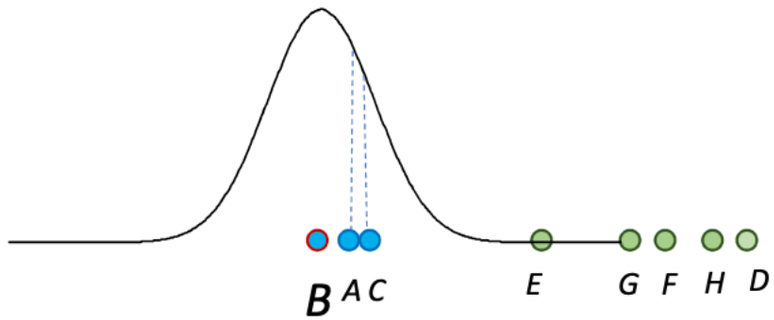
van der Maaten et al. (2008)

CHOICE OF σ^2

- It's not appropriate to have a single value of σ^2 as you need a smaller value in more dense regions.
- The user controls it through a parameter called perplexity
- Perplexity can have a big impact on the result!





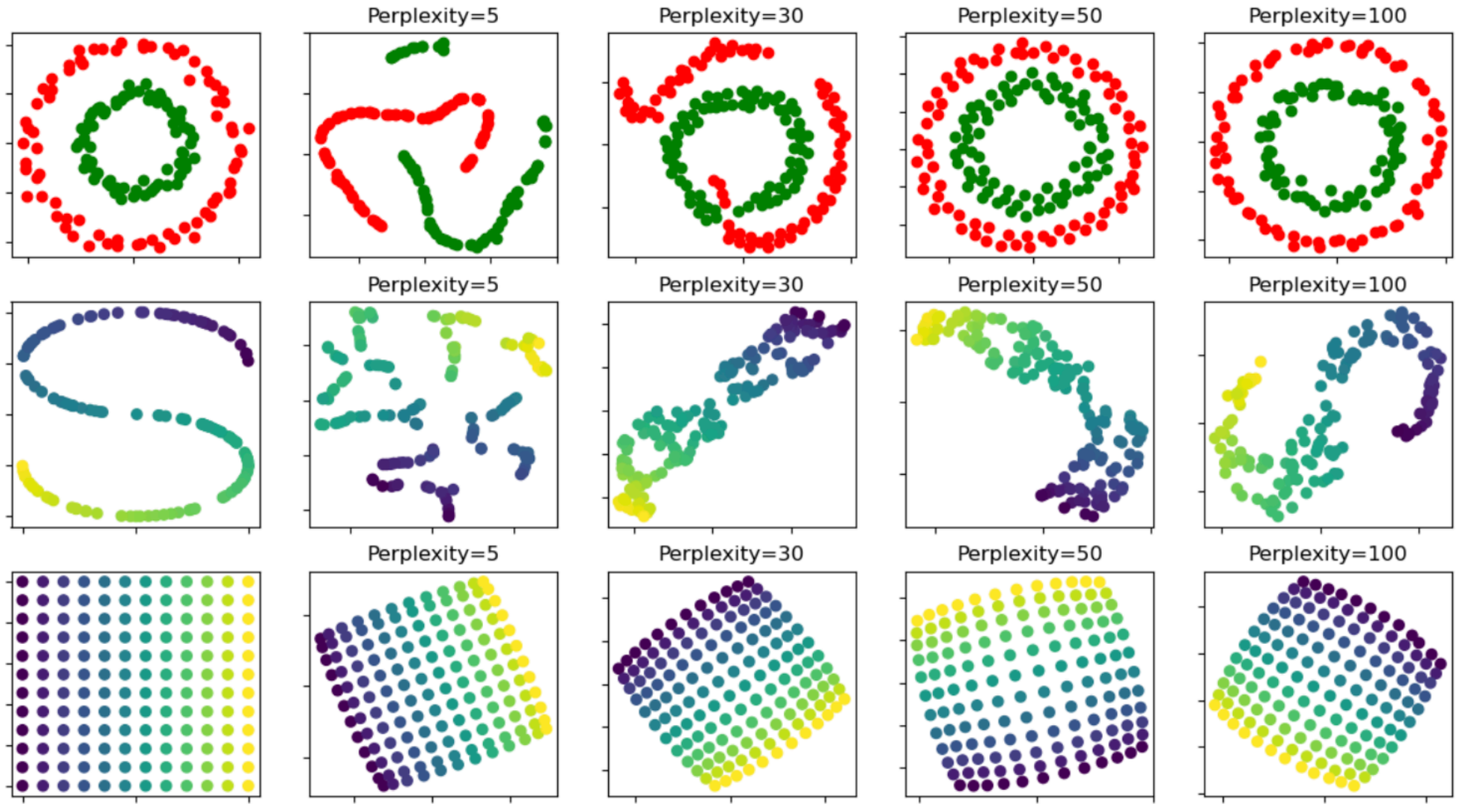


$$Perp(P_i) = 2^{H(P_i)}, \qquad H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}.$$

H(*Pi*) is the Shannon entropy of *Pi* measured in bits

T-SNE ART, OR THE CHOICE OF THE PERPLEXITY PARAMETER

https://distill.pub/2016/misread-tsne/

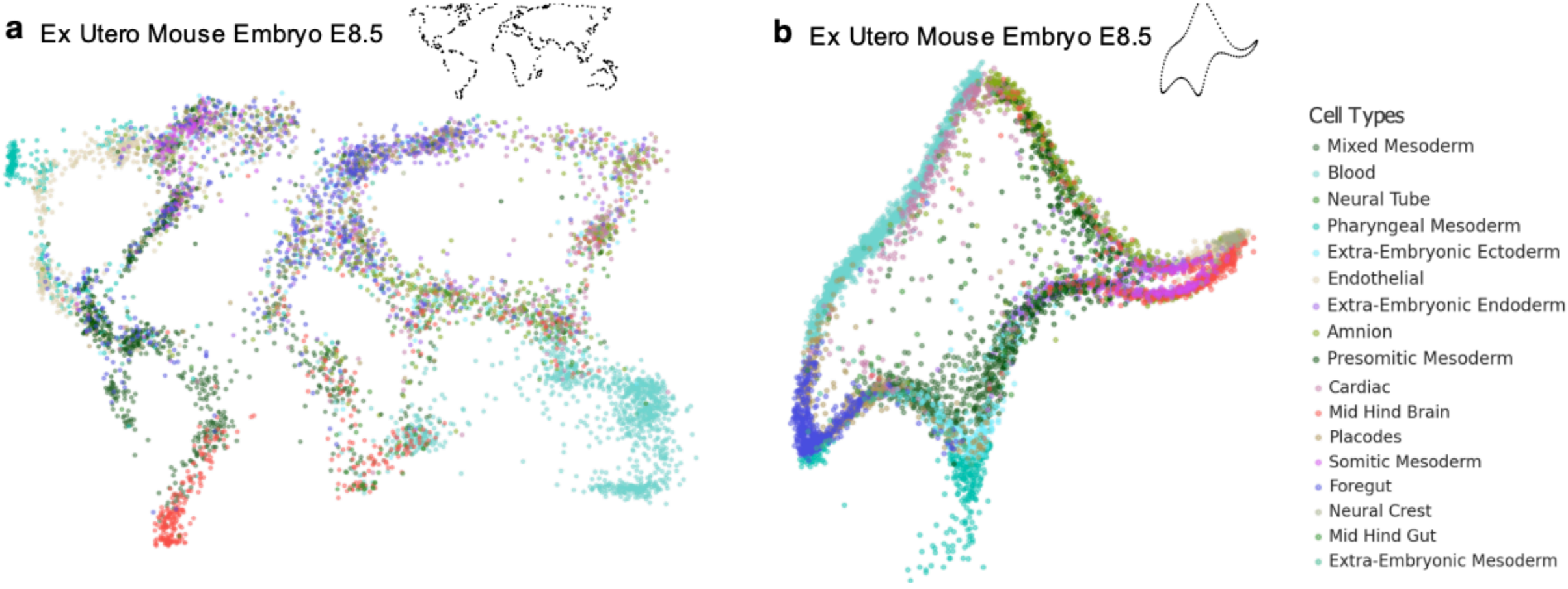


LIMITATIONS OF T-SNE (AND UMAP)

- Unlike PCA, we do not have a simple interpretation for our low-dimensional embedding (the axes have "no meaning").
- t-SNE preserves only the local structure (who is neighbor of whom) but not the global structure.
- There is no guarantee of convergence to the global minimum (non-convex problem), hence different runs will lead to different embeddings.
- Some argue that t-SNE and UMAP do not even preserve the local structure or the neighbors (Chari et al. 2021).

LIMITATIONS OF T-SNE (AND UMAP)

- The “shape” of the data in the embedding is arbitrary.

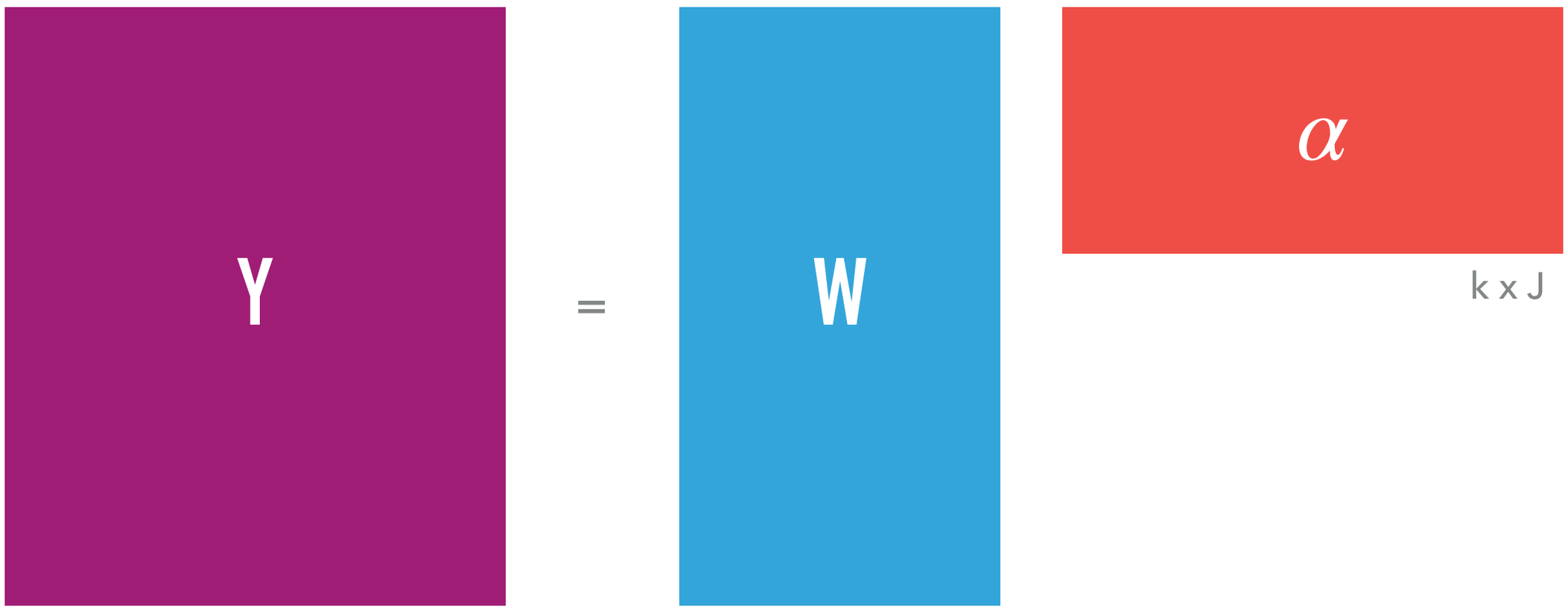


Chari et al (2021). *bioRxiv.*

FACTOR ANALYSIS

From a statistical perspective, we can state the problem using the following model

$$Y = W\alpha + \varepsilon$$



FACTOR ANALYSIS

The goal is to find $k \ll J$ *factors* that describe, with the minimum possible loss of information, the J original variables (genes).

We can show that if ε (or equivalently Y) is Gaussian, a solution of the model is PCA.

ADVANTAGES OF PCA

In one word: **interpretability**!

- The first principal component is the direction of greater variability in the data.
- It is easy to compute the variance explain by the first *m* principal components.
- Very computationally efficient.

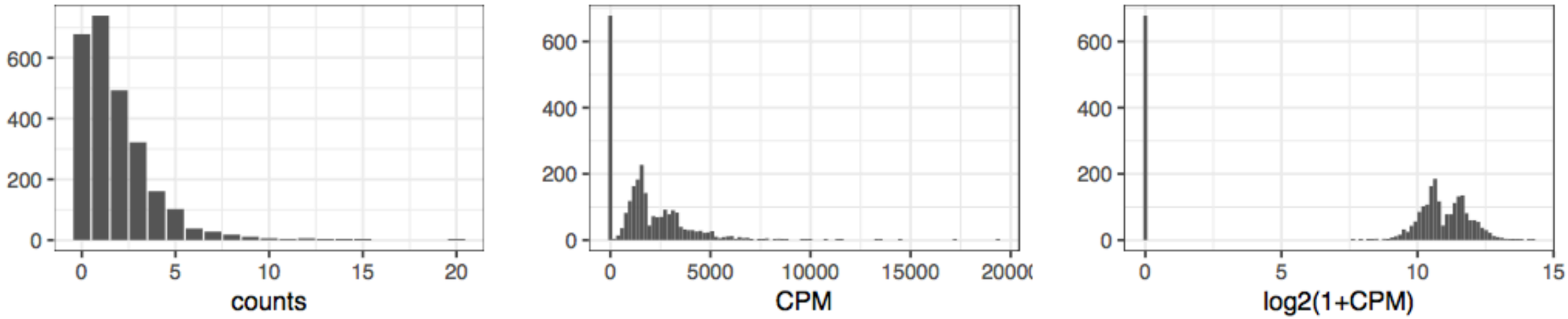
SO… WHY NOT PCA?

The main issue of PCA for single-cell data is that the data are non-negative integer counts, which exhibit skewed distributions and are not well fit by a Gaussian model.

A simple, and somewhat effective, solution is to transform the data, e.g., by $\log(x+1)$, but this is not always straightforward:

- Which transformation to use?
- Do we need to normalize the data for sequencing depth and other cell-specific effects?
- Zero counts complicate the analysis.

REMEMBER THE EFFECT OF LOGCPM...



(a) UMI counts (b) counts per million (CPM) (c) log of CPM

Townes et al. (2019). *Genome Biology.*

GLM-PCA

One alternative to transforming the data, is to generalize our model to non-Gaussian data.

This can be done by defining a set of models, known as GLM-PCA that extend the framework to a set of well behaving distributions (exponential family) similar to how GLM extends the linear model.

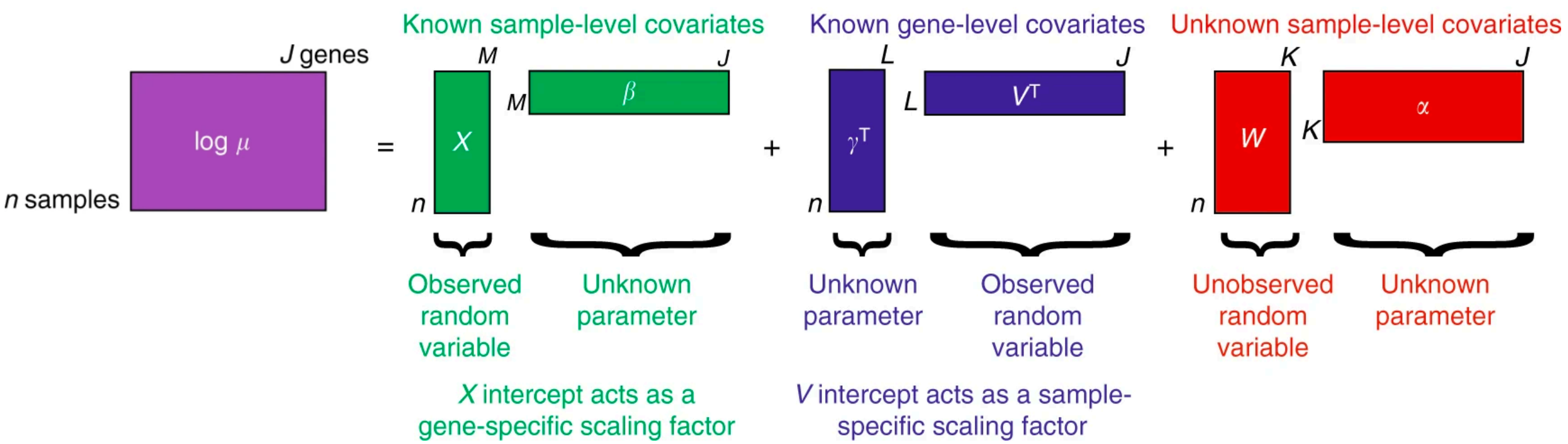
In particular, since we have count data, we can use the Poisson or negative binomial model, which has a log link function.

$$E[Y \mid W] = \mu, \qquad \log \mu = W\alpha$$

Townes et al. (2019). *Genome Biology.*
scry package

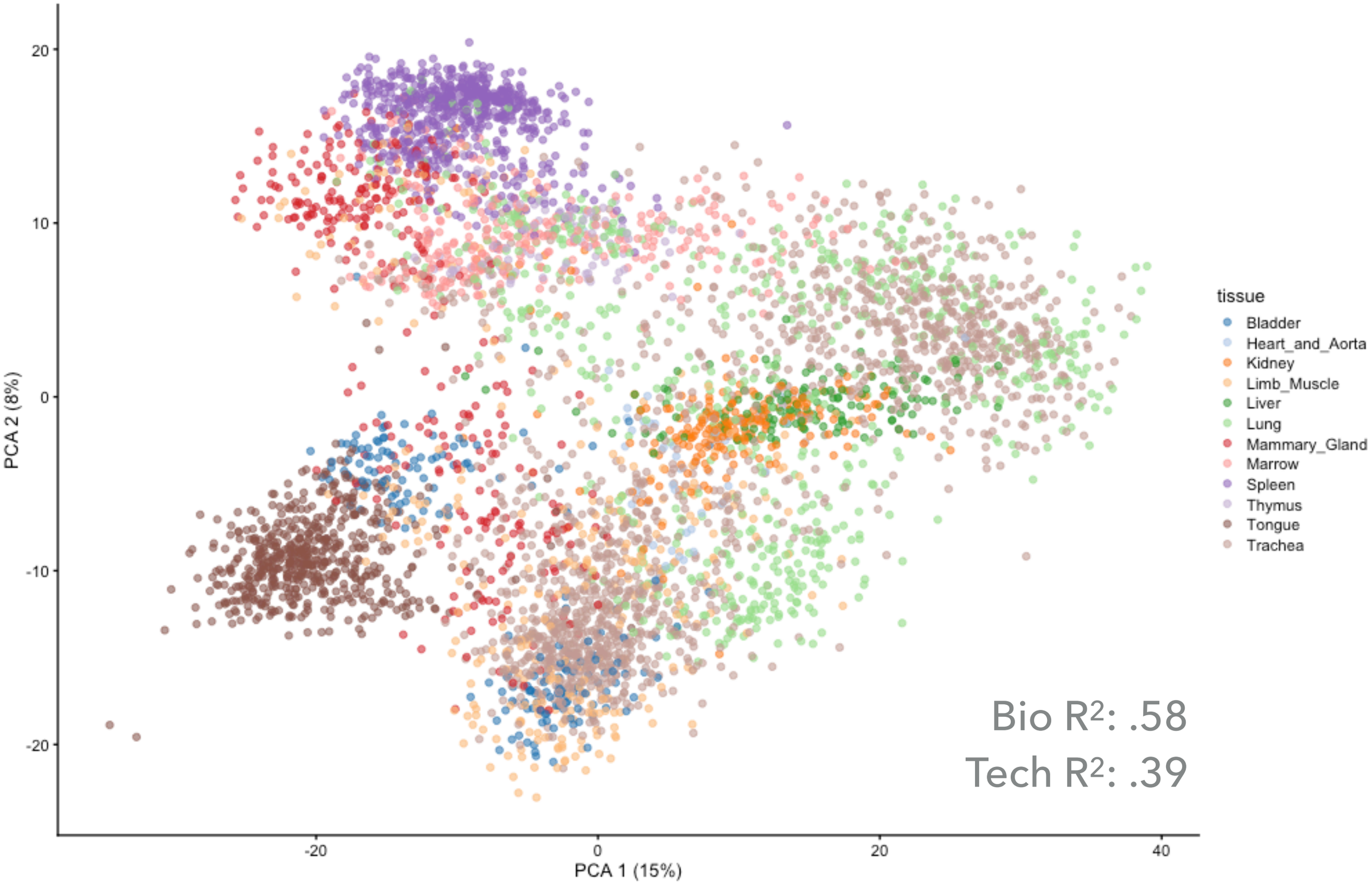
GLM-PCA / WANTED VARIATION EXTRACTION

A further generalization allows us to include *observed covariates* in the model. These can be covariates at the cell and gene level and it is useful for normalization and batch effect correction.

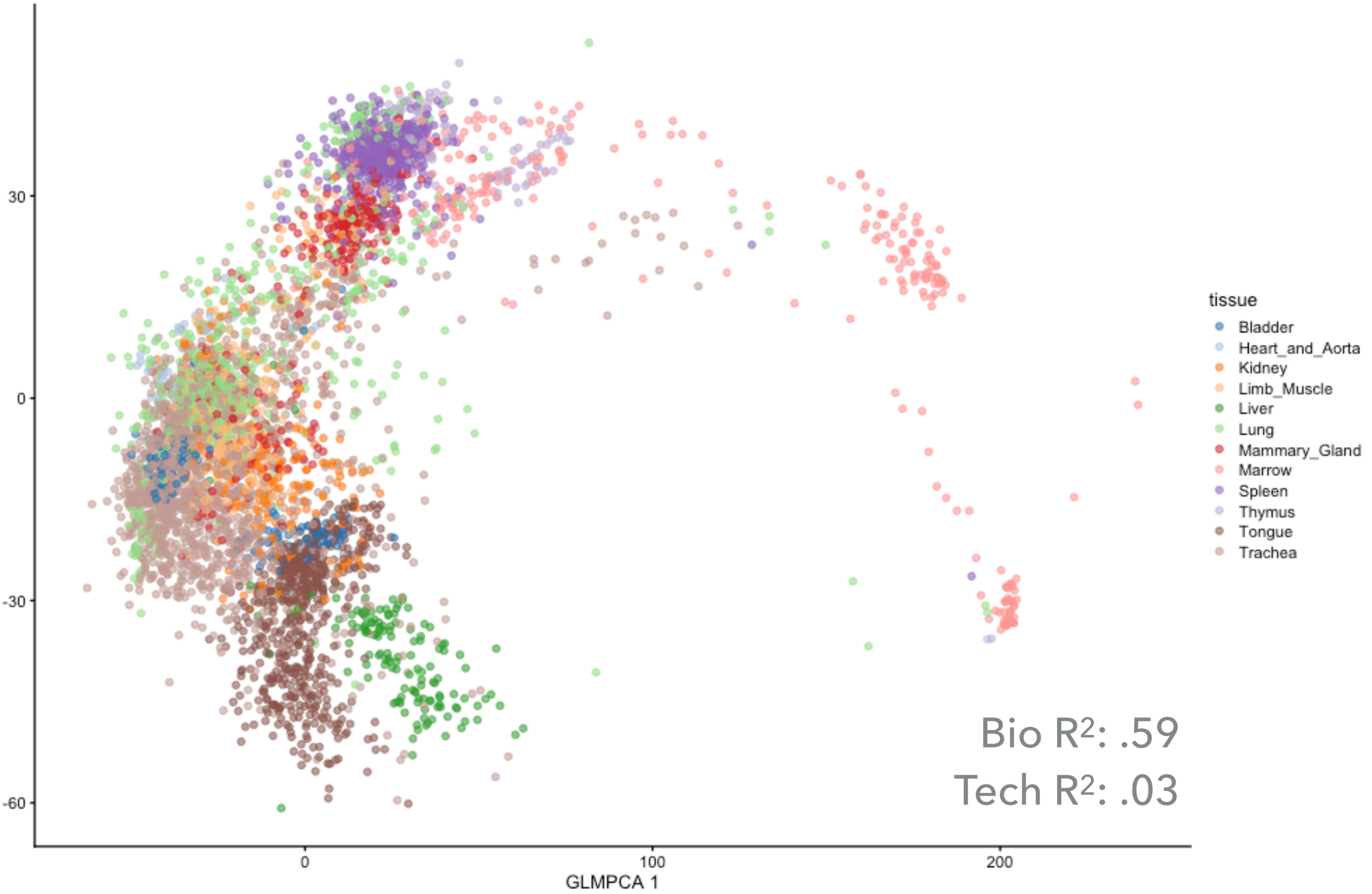


Risso et al. (2018). *Nature Communications*.
Agostinis et al. (2022). *Bioinformatics.*
NewWave package

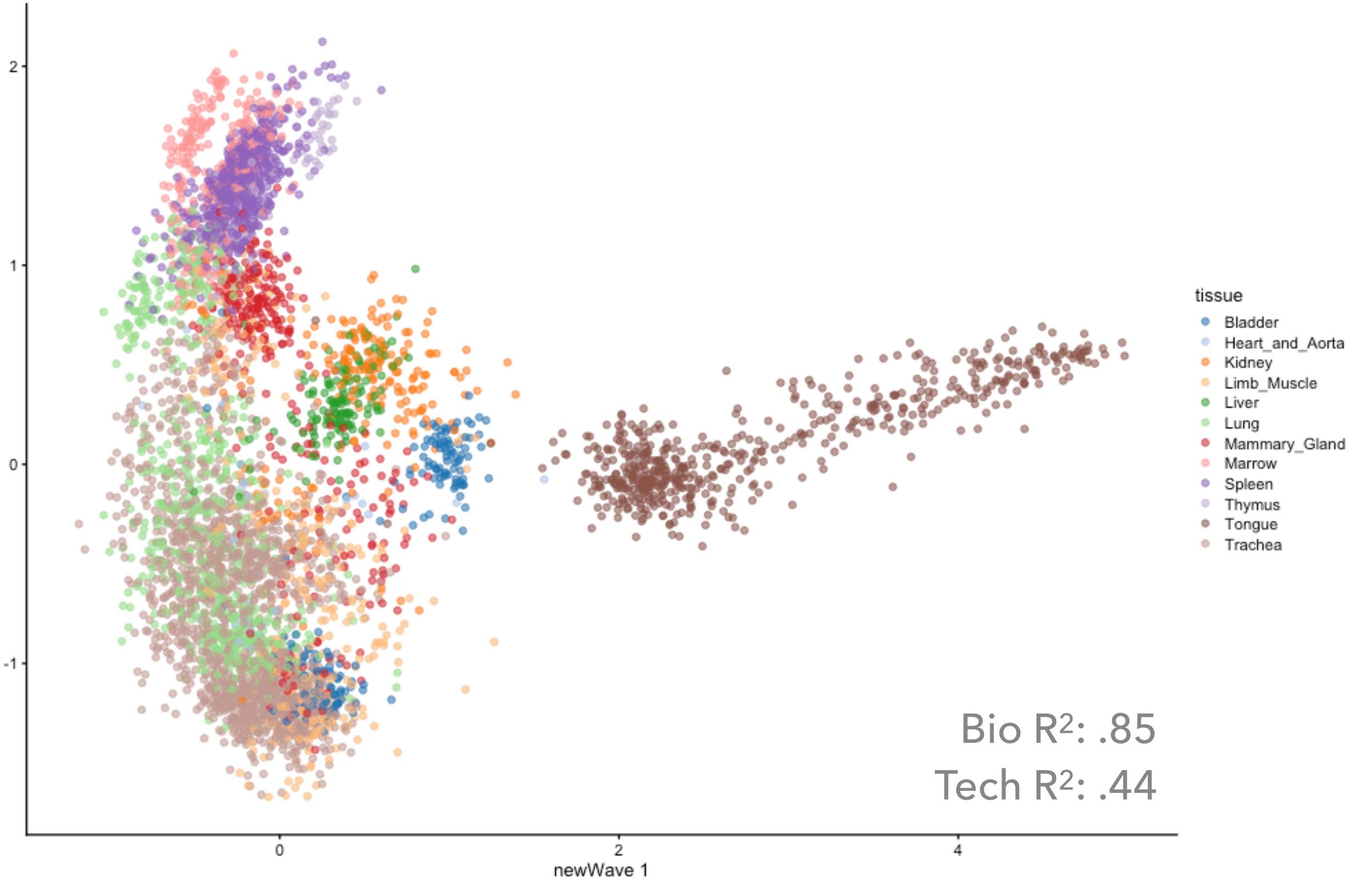
TABULA MURIS: PCA AFTER SCRAN NORMALIZATION (LOG SCALE)



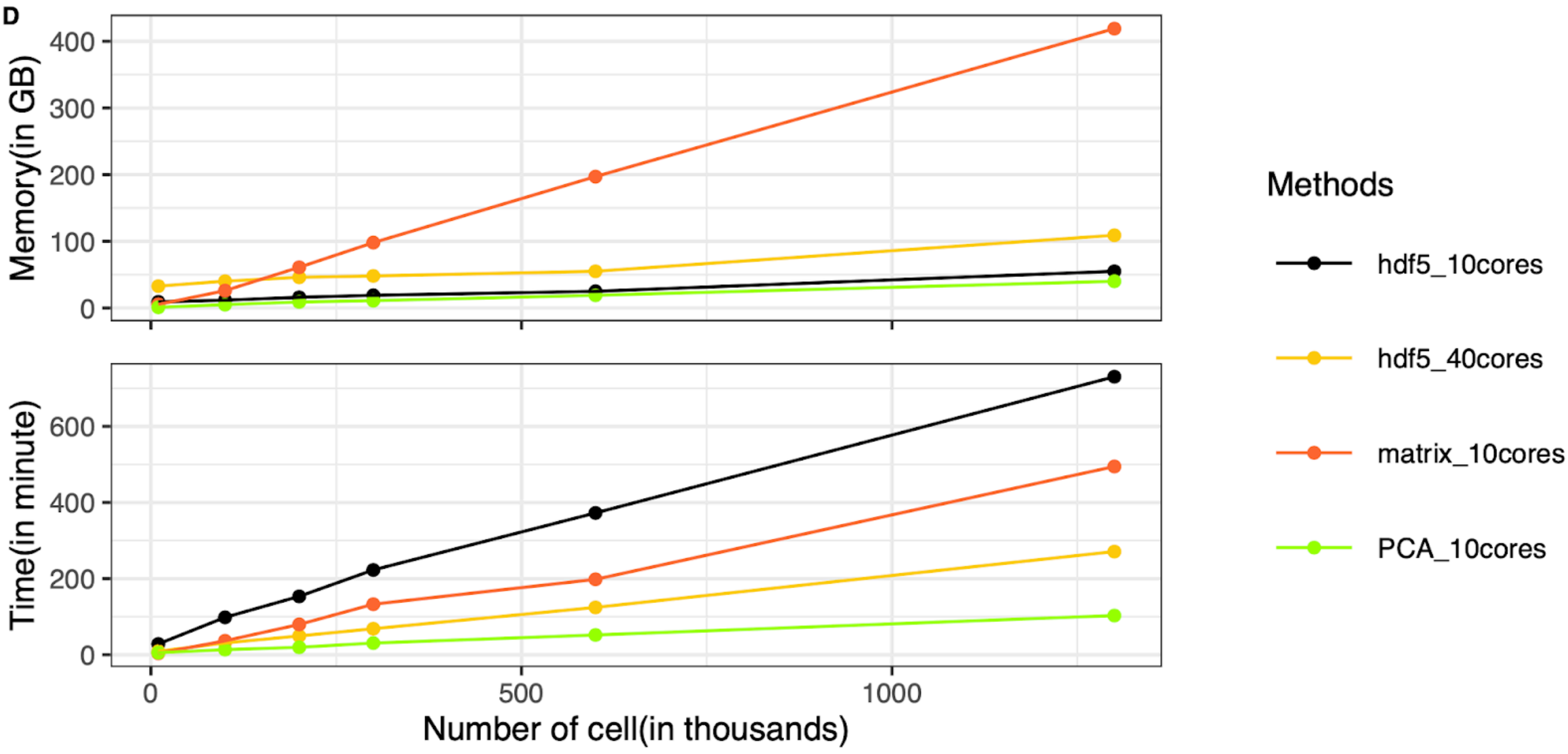
TABULA MURIS: GLM-PCA (POISSON)



TABULA MURIS: NEWAVE (NEGATIVE BINOMIAL)



SCALABILITY

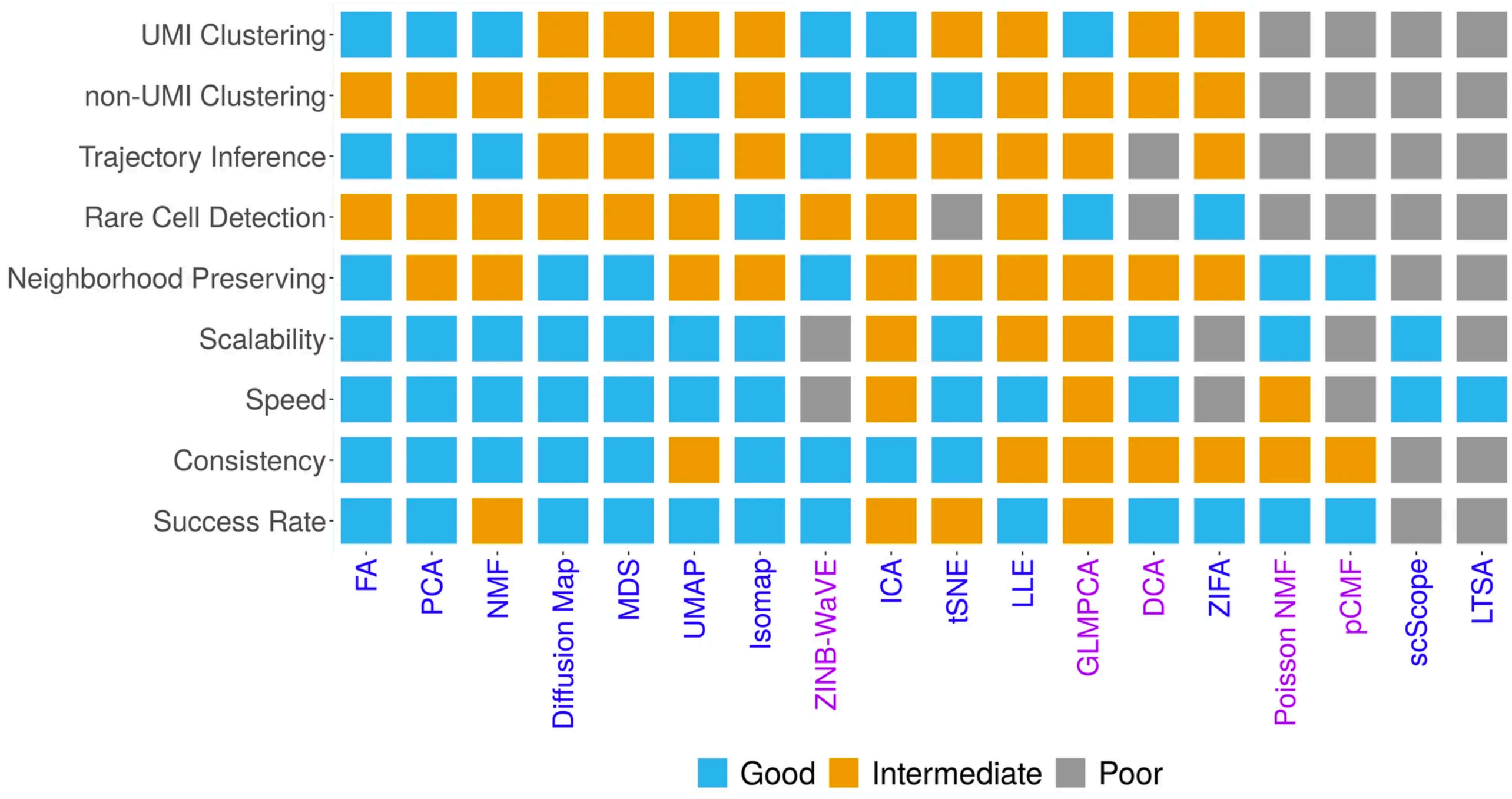


Agostinis et al. (2022). *Bioinformatics*.

SCALABILITY

- Townes et al. (2019) propose an *approximate approach* to speed up computations.
- Essentially, they compute Pearson or deviance residuals of a *GLM* fit on each gene independently, and then compute PCA of the rediduals.
- A similar approach, *correspondence analysis*, uses chi-squared Pearson residuals + PCA/SVD.
- These methods are implemented in the scry and corral Bioconductor packages, respectively.

WHICH SHOULD I USE?



Sun et al. (2019). *Genome Biology*.

WHICH ONE SHOULD I USE?

EXPLORATORY DATA ANALYSIS!

Advantages of (Generalised) Linear Factor Analysis Models:

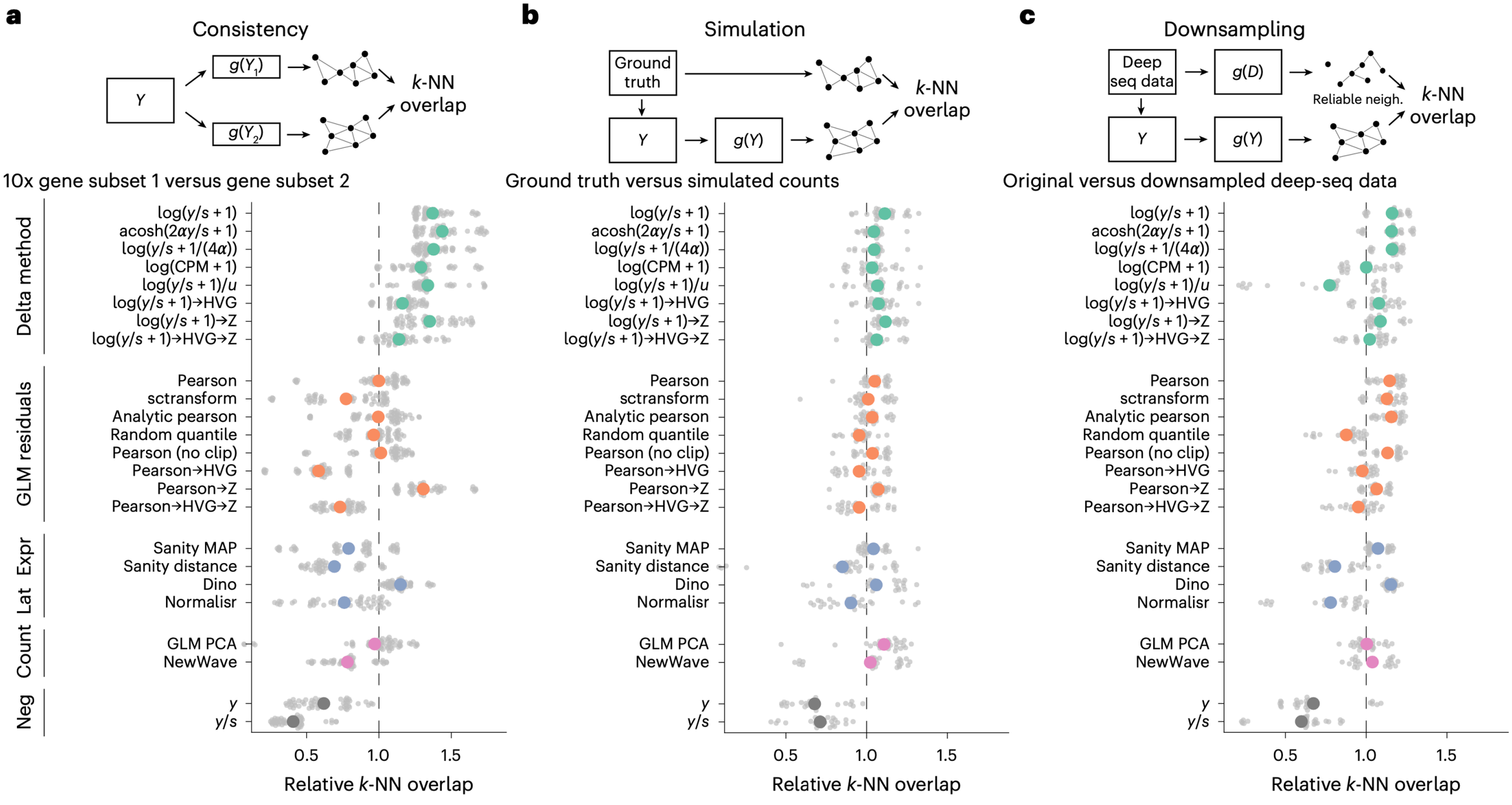
- Allow to reduce the dimensionality of the problem
- Ability to control for batch effects and other confounders
- Interpretation of the inferred factors

Importance of simple models and interpretability of the solutions.

MORE QUESTIONS THAN ANSWERS

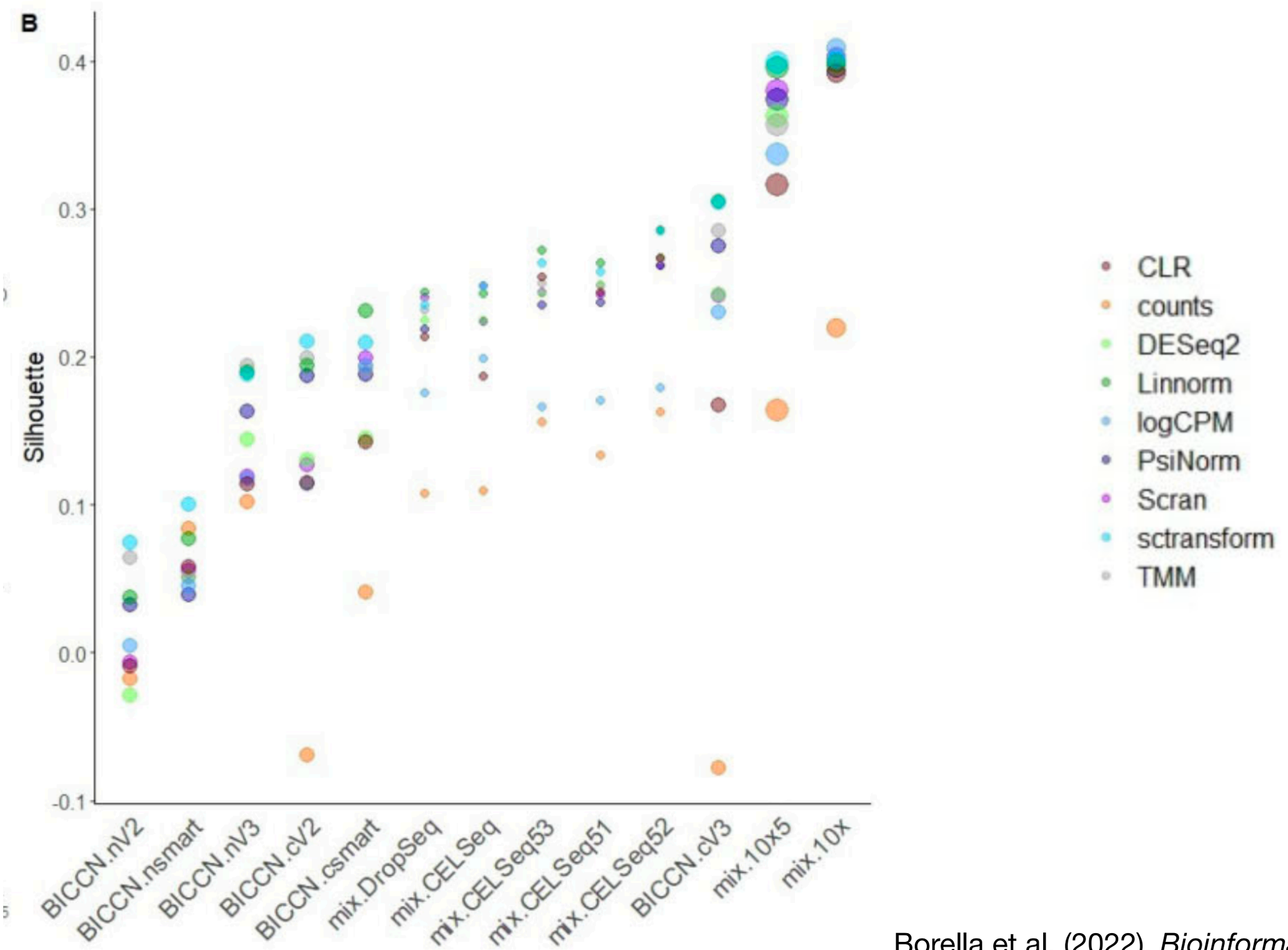
- How many factors should I estimate?
- Should I include covariates? Which ones?
- If PCA, should I scale the data? (Covariance or correlation?)
- Which data transformations should I use?
- Which normalization should I use?
- Why not deep neural networks?
- Importance of simple models and interpretability of the solutions.

SOME ANSWERS…



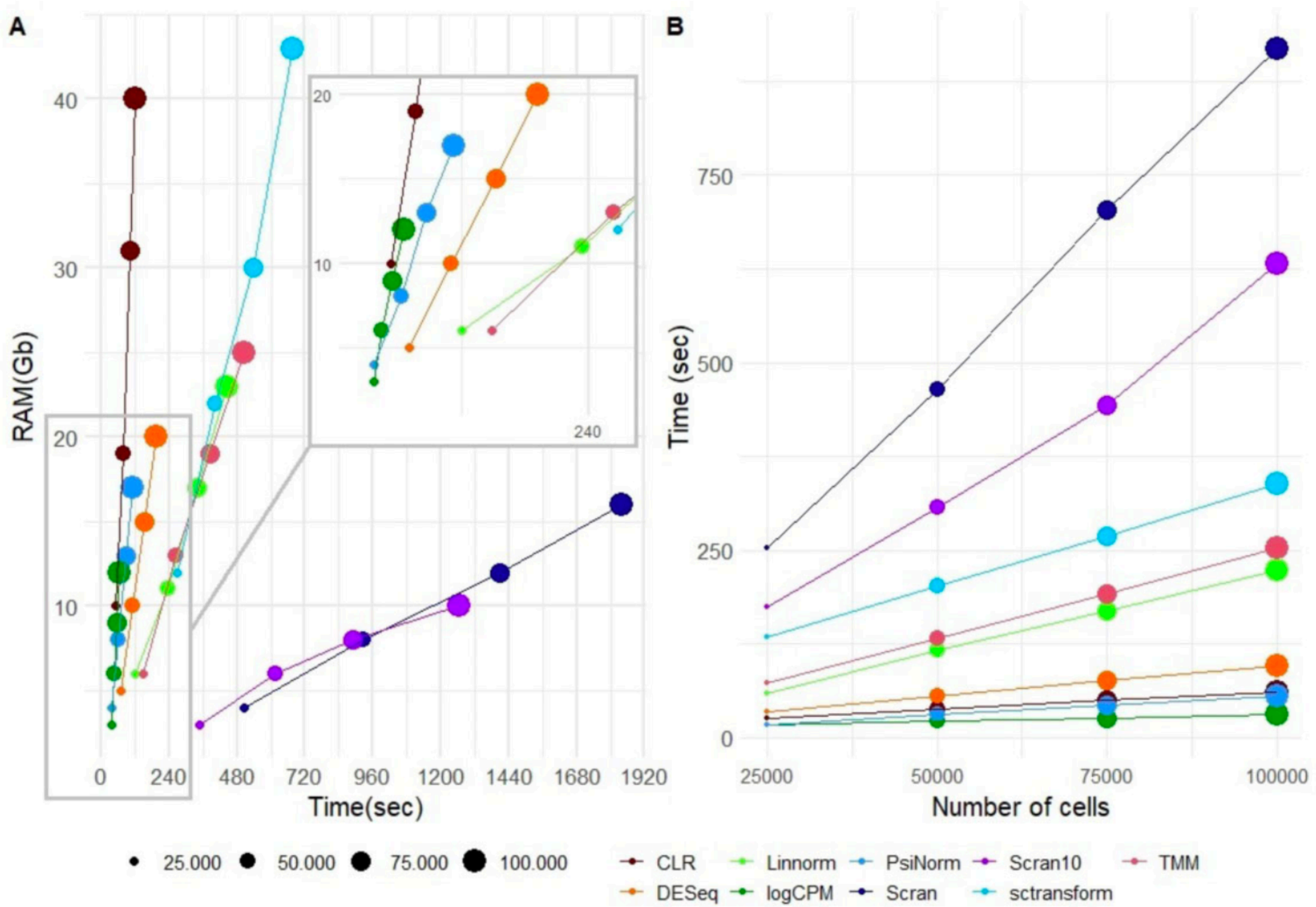
Ahlmann-Eltze & Huber (2023). *Nature Methods*.

SOME ANSWERS…



Borella et al. (2022). *Bioinformatics*.

SOME ANSWERS…



Borella et al. (2022). *Bioinformatics*.

TAKE-HOME MESSAGE

- t-SNE / UMAP are fine for visualization
- Do not use them for inference (e.g., clustering)
- Linear/more interpretable techniques should be preferred

THANKS FOR YOUR ATTENTION!