UNIVERSITÀ
DEGLI STUDI
DI PADOVA

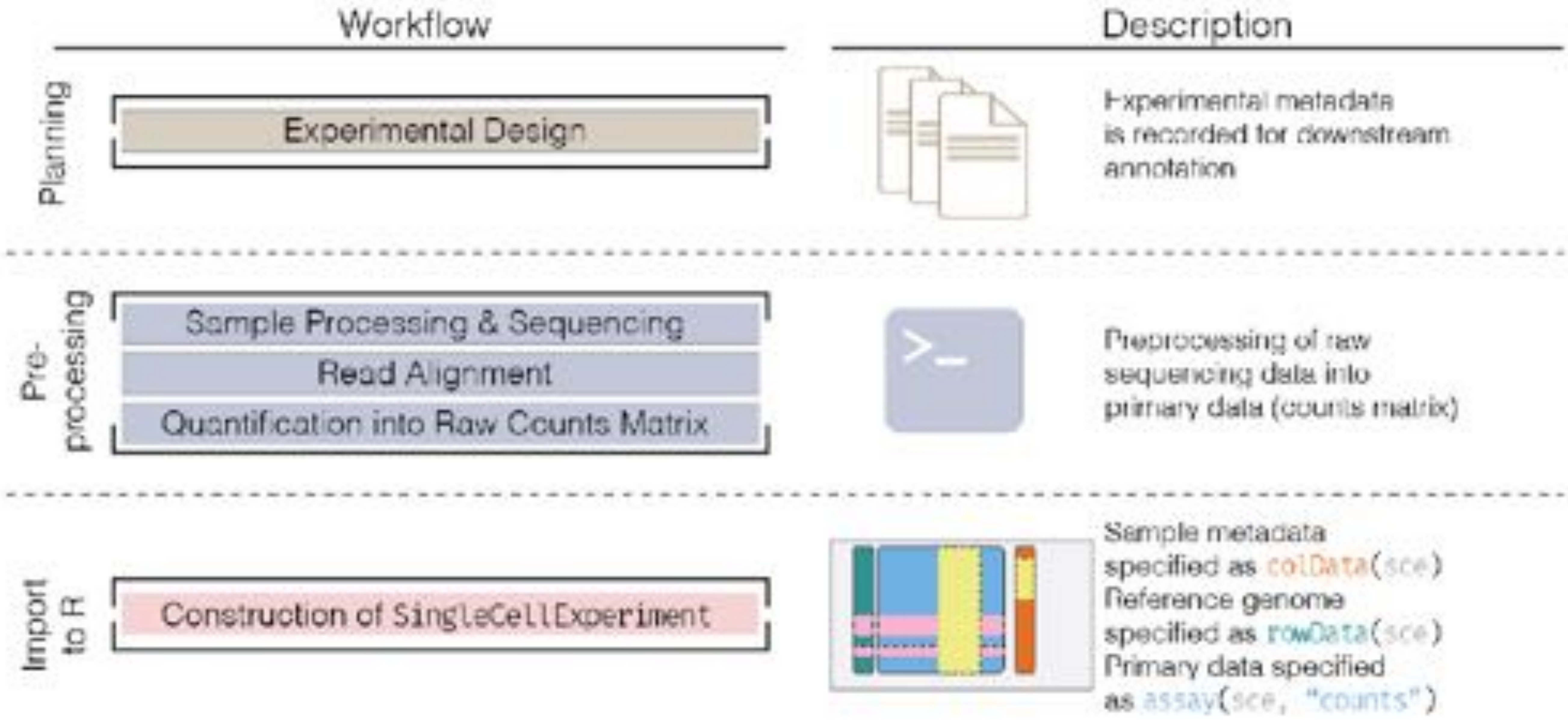
DAVIDE RISSO

QUANTIFICATION, QC & NORMALIZATION OF SCRNA-SEQ

OUTLINE

1. Quantification
2. Exploratory Data Analysis (EDA) & Quality Control (QC)
3. Normalization
4. Doublet detection

A TYPICAL ANALYSIS WORKFLOW



Workflow

Description

Planning

Experimental Design

Experimental metadata is recorded for downstream annotation

Pre-processing

Sample Processing & Sequencing

Read Alignment

Quantification into Raw Counts Matrix

Preprocessing of raw sequencing data into primary data (counts matrix)

Import to R

Construction of `SingleCellExperiment`

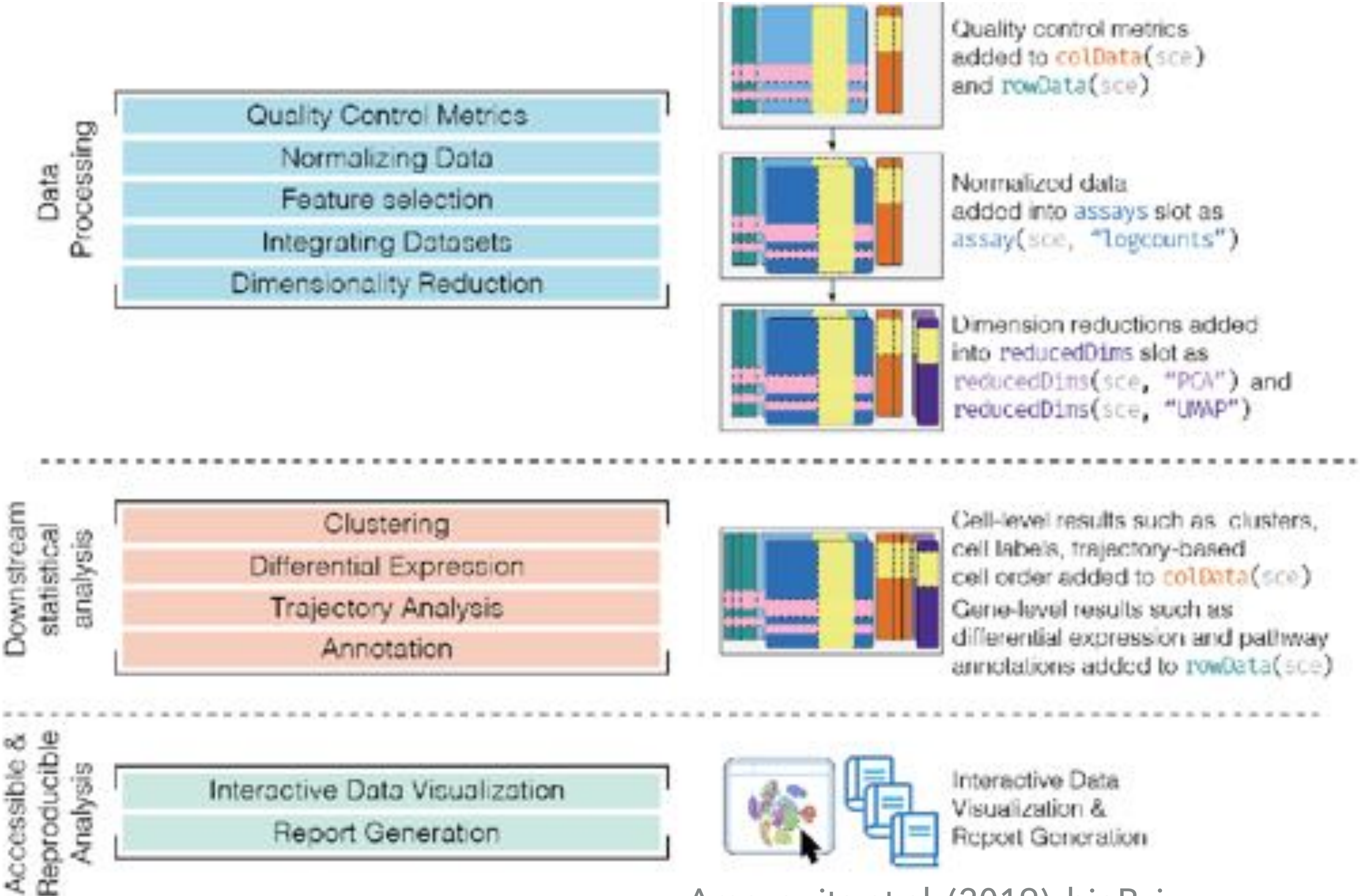
Sample metadata specified as `colData(sce)`

Reference genome specified as `rowData(sce)`

Primary data specified as `assay(sce, "counts")`

Amezquita et al. (2019). bioRxiv.

A TYPICAL ANALYSIS WORKFLOW



Data Processing

- Quality Control Metrics
- Normalizing Data
- Feature selection
- Integrating Datasets
- Dimensionality Reduction

Quality control metrics added to `colData(sce)` and `rowData(sce)`

Normalized data added into `assays` slot as `assay(sce, "logcounts")`

Dimension reductions added into `reducedDims` slot as `reducedDims(sce, "PCA")` and `reducedDims(sce, "UMAP")`

Downstream statistical analysis

- Clustering
- Differential Expression
- Trajectory Analysis
- Annotation

Cell-level results such as clusters, cell labels, trajectory-based cell order added to `colData(sce)`
Gene-level results such as differential expression and pathway annotations added to `rowData(sce)`

Accessible & Reproducible Analysis

- Interactive Data Visualization
- Report Generation

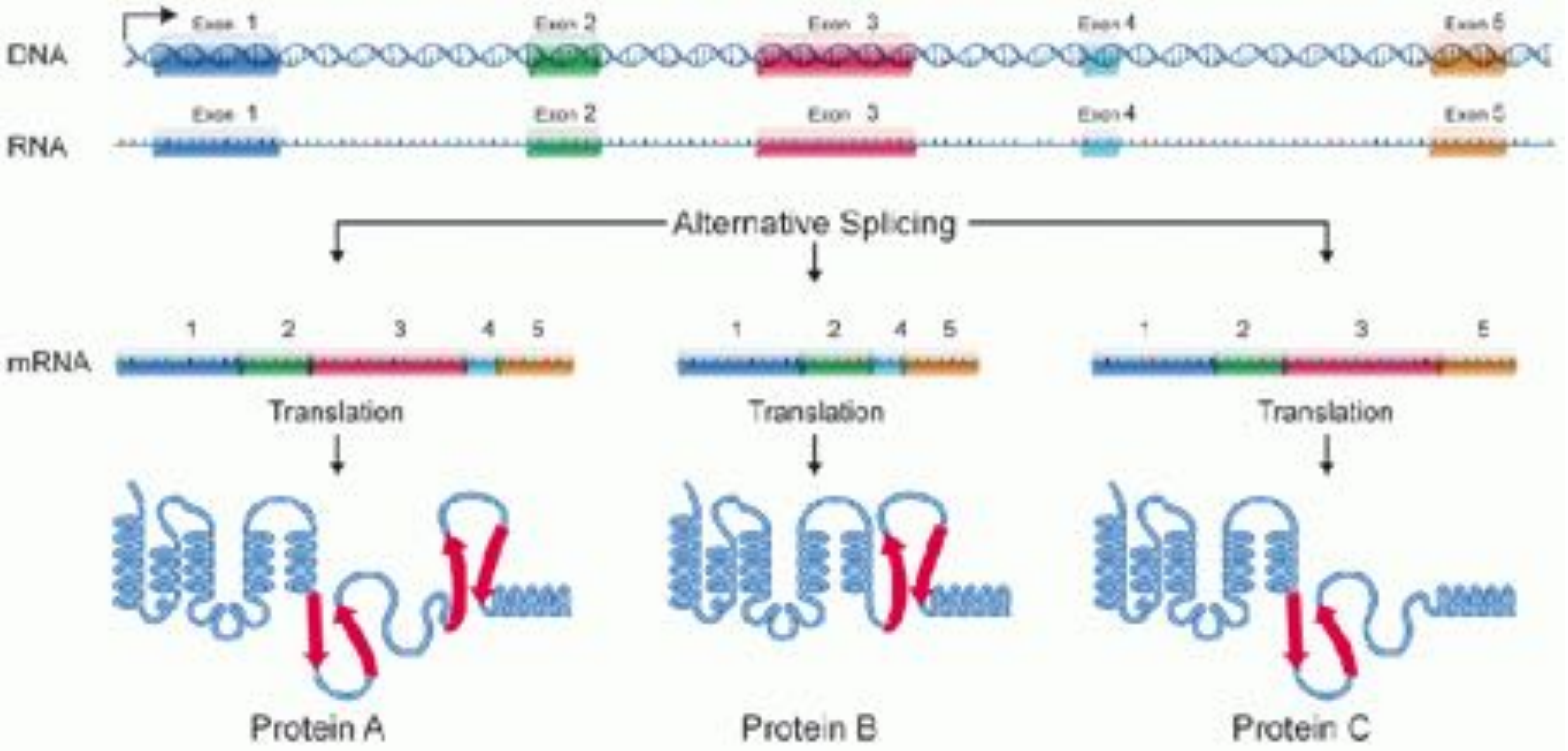
Interactive Data Visualization & Report Generation

Amezquita et al. (2019). bioRxiv.

Resources

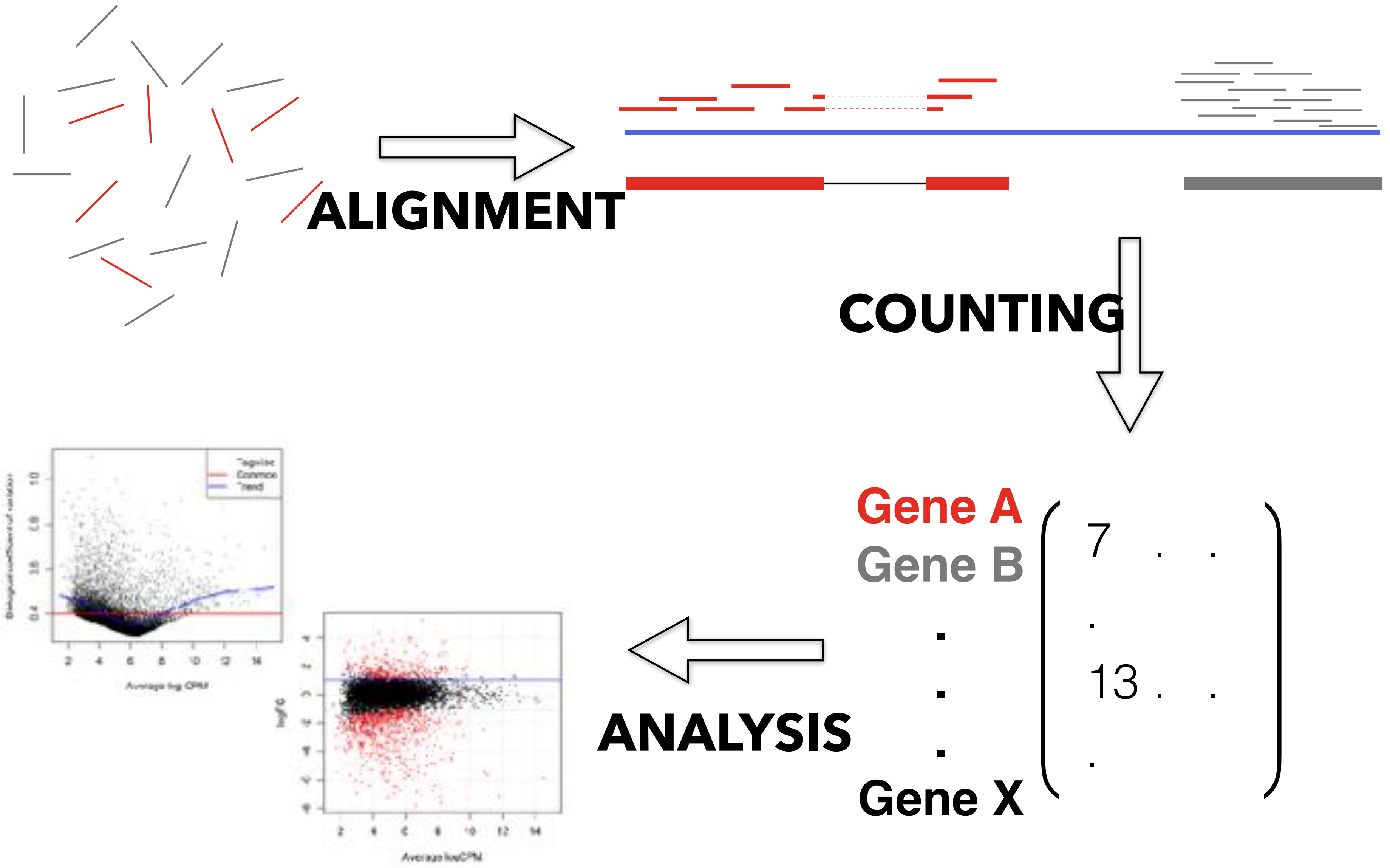
- A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor
    - https://f1000research.com/articles/5-2122/v2
- Bioconductor workflow for single-cell RNA sequencing
    - https://f1000research.com/articles/6-1158/v1
- github.com/seandavi/awesome-single-cell
- scrna-tools.org
- Seurat
    - https://satijalab.org/seurat/
- Bioconductor workshop materials
    - https://bioconductor.org/help/course-materials/
- Orchestrating Single Cell Analysis review
    - https://www.biorxiv.org/content/10.1101/590562v1.abstract
    - https://osca.bioconductor.org

QUANTIFICATION


DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1
2
3
4
5
1
2
4
5
1
2
3
5
Translation
Translation
Translation
Protein A
Protein B
Protein C

Wikipedia

Alignment-based RNA-seq workflow



ALIGNMENT

COUNTING

Gene A

Gene B

.

.

.

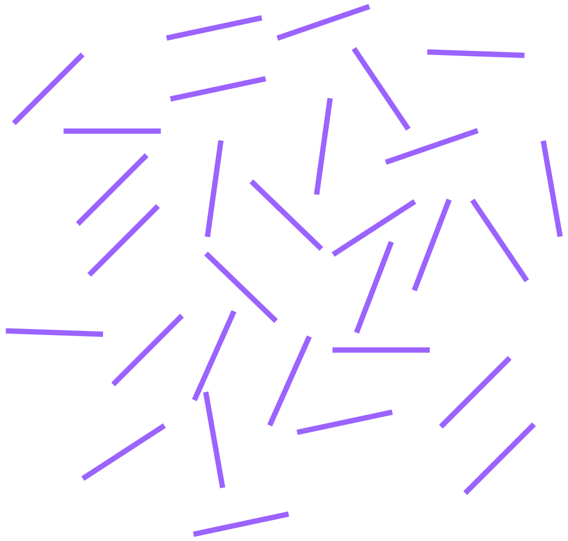
Gene X

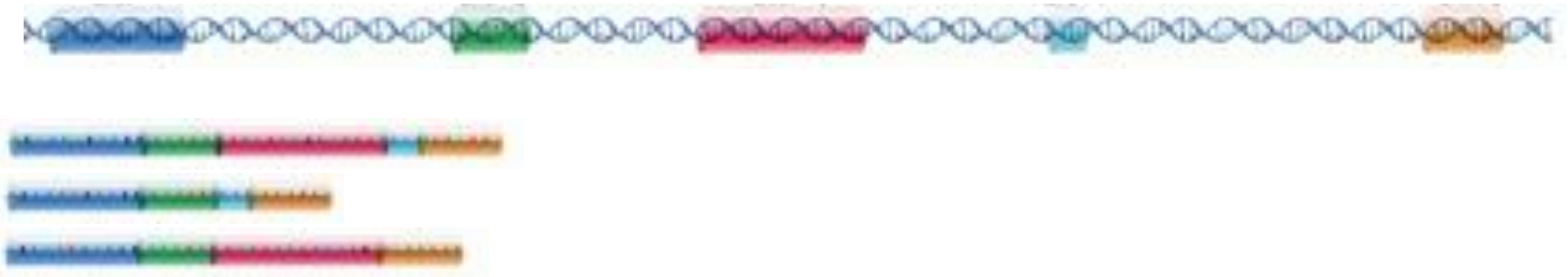
$$\begin{pmatrix} 7 & . & . \\ . & & \\ 13 & . & . \\ . & & \end{pmatrix}$$

ANALYSIS

Charlotte Soneson

Abundance quantification

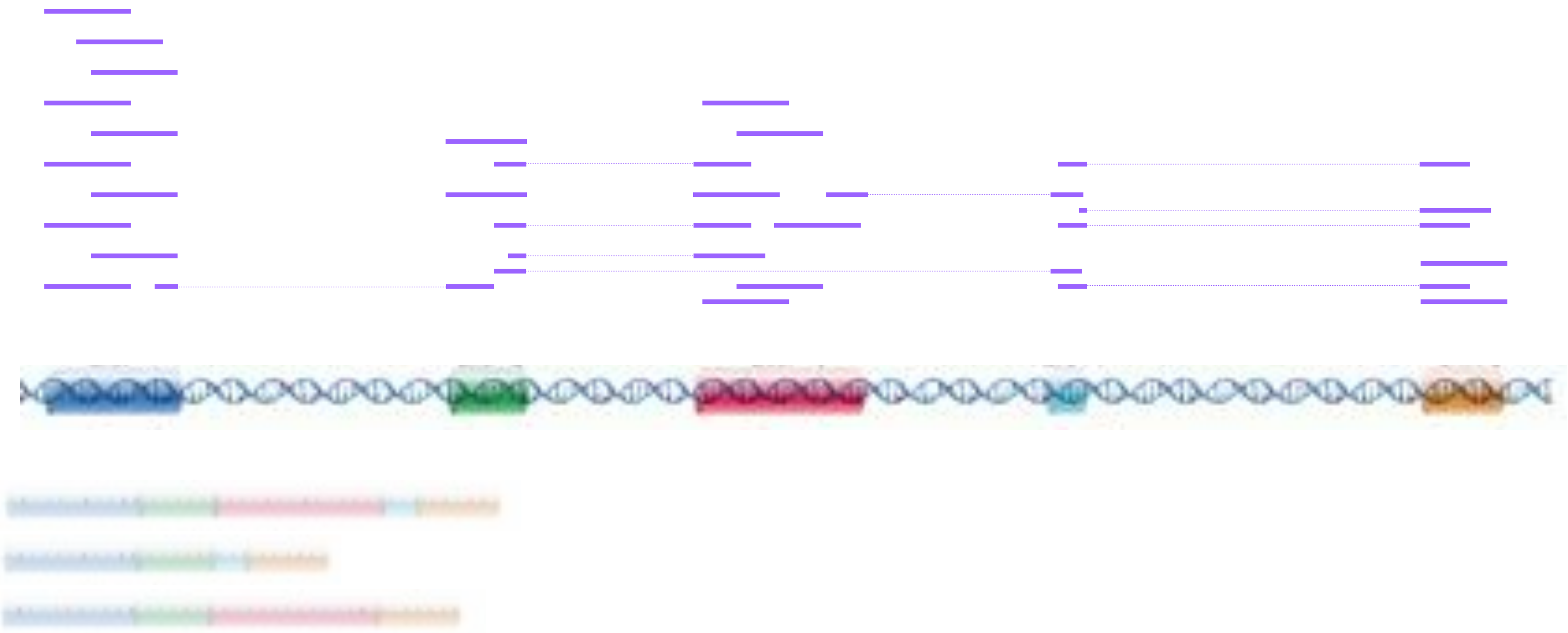




Charlotte Soneson

Abundance quantification

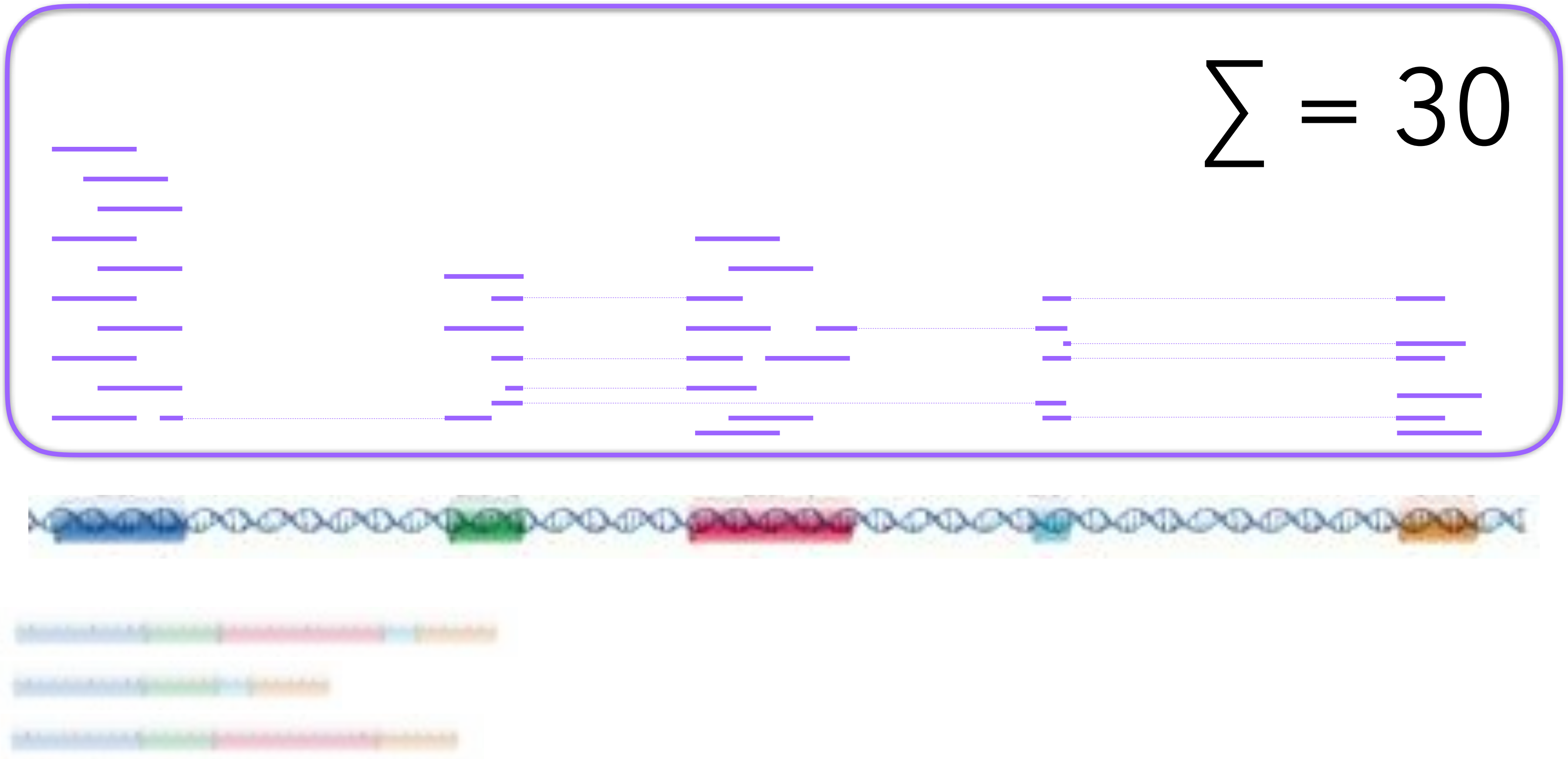
Gene-level counts, often obtained by genome alignment + overlap counting



Charlotte Soneson

Abundance quantification

Gene-level counts, often obtained by genome alignment + overlap counting

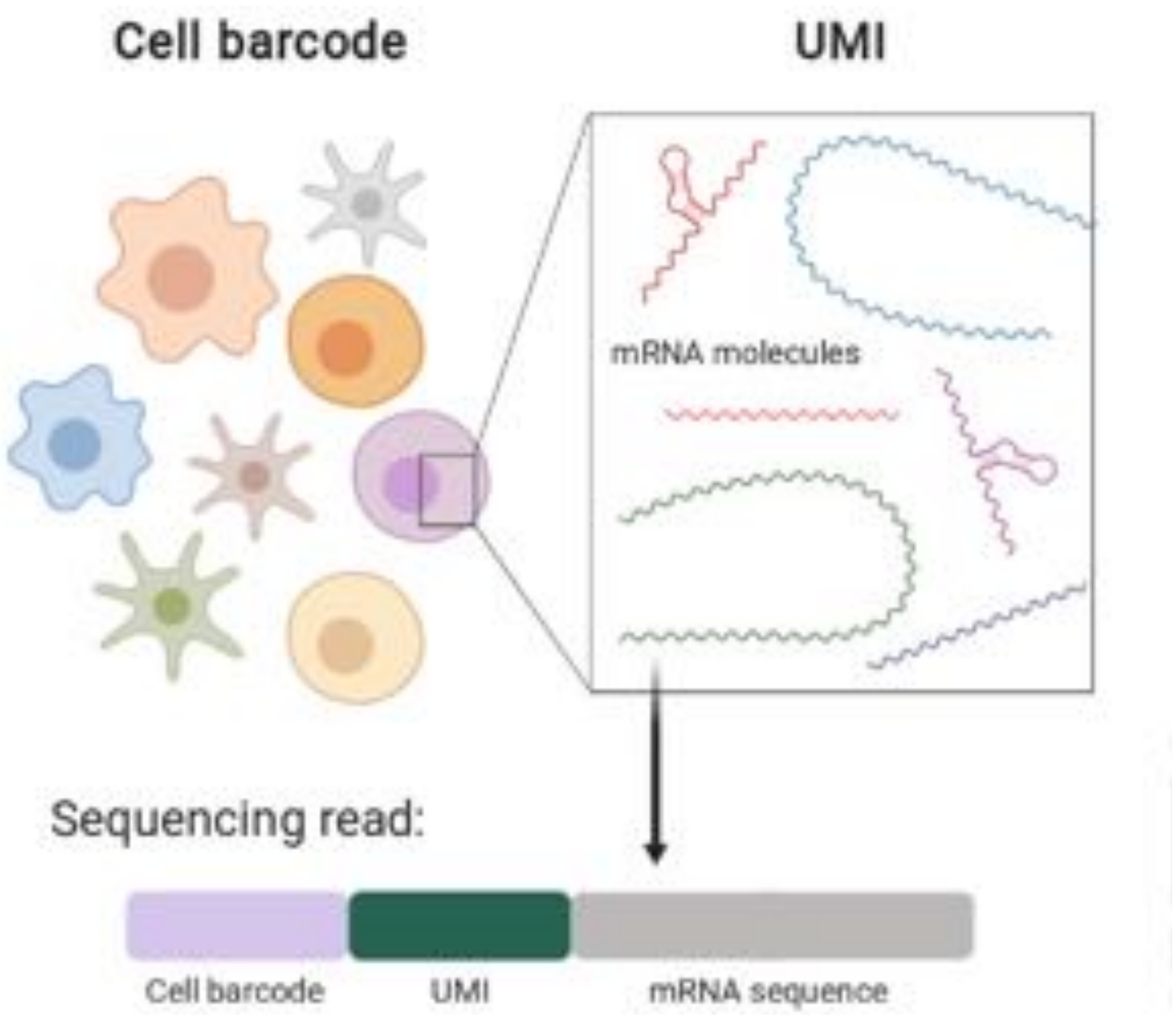


Charlotte Soneson

Cell barcode and unique molecular identifier (UMI)

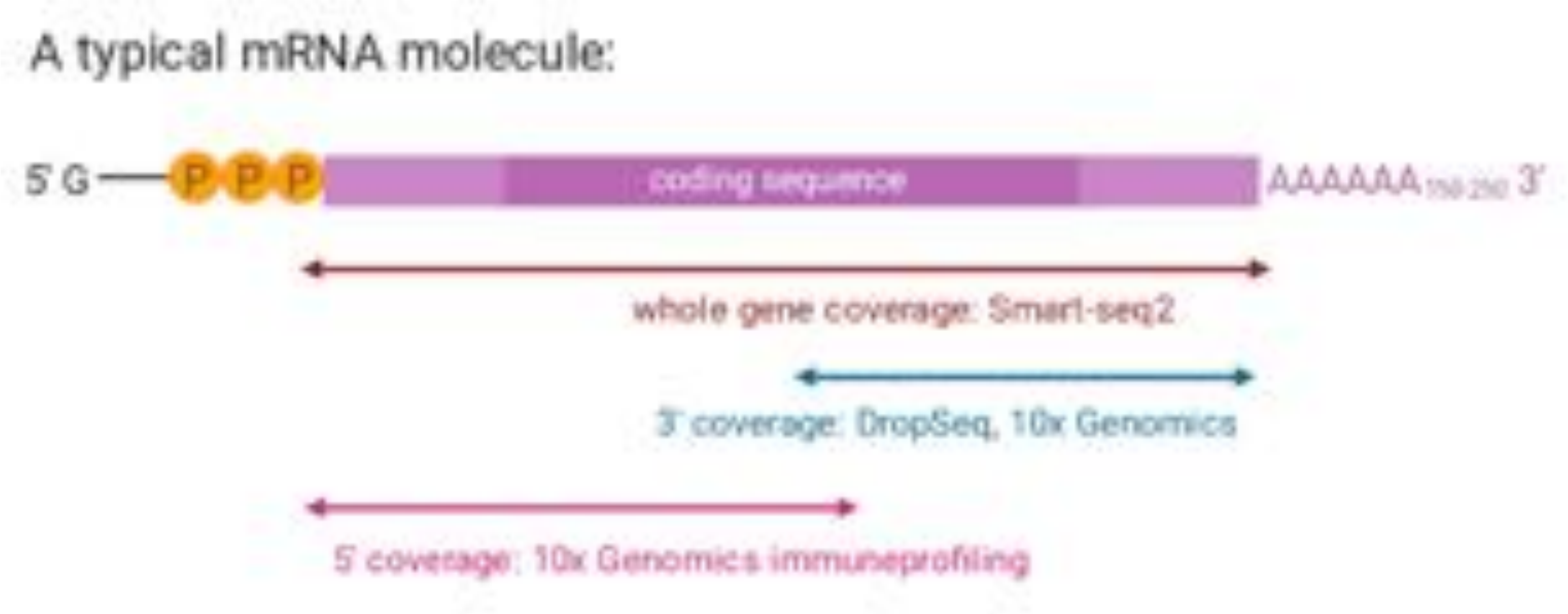
Sequencing data preserves information:

- Which cell did the sequenced transcript belong to? → **cell barcode**
- How many times did one transcript get sequenced? → **UMI**



Whole gene vs. 3’ or 5’ sequencing

Depending on the library preparation and sequencing protocols that you are using, you will get different coverage of mRNA molecules.



A typical mRNA molecule:
5' G
P P P
coding sequence
AAAAAA 150-250 3'
whole gene coverage: Smart-seq2
3' coverage: DropSeq, 10x Genomics
5' coverage: 10x Genomics immuneprofiling

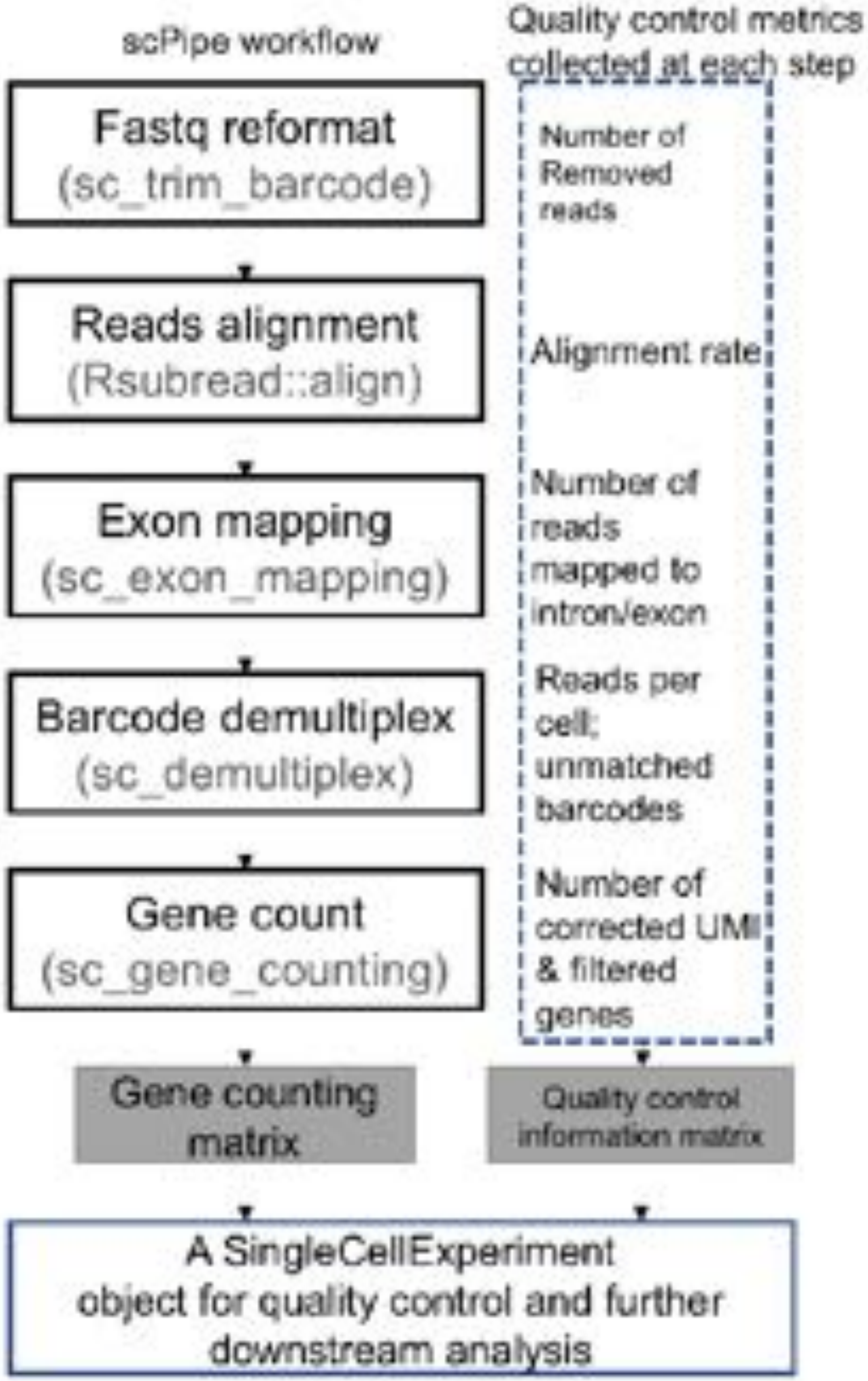
SINGLE-CELL SPECIFIC PROBLEMS FOR QUANTIFICATION

- Correctly detect barcode sequences
- Assign reads to the right barcode (cell)
- Identify empty droplets and barcode swapping
- UMI quantification, starting from read alignments (UMI deduplication)

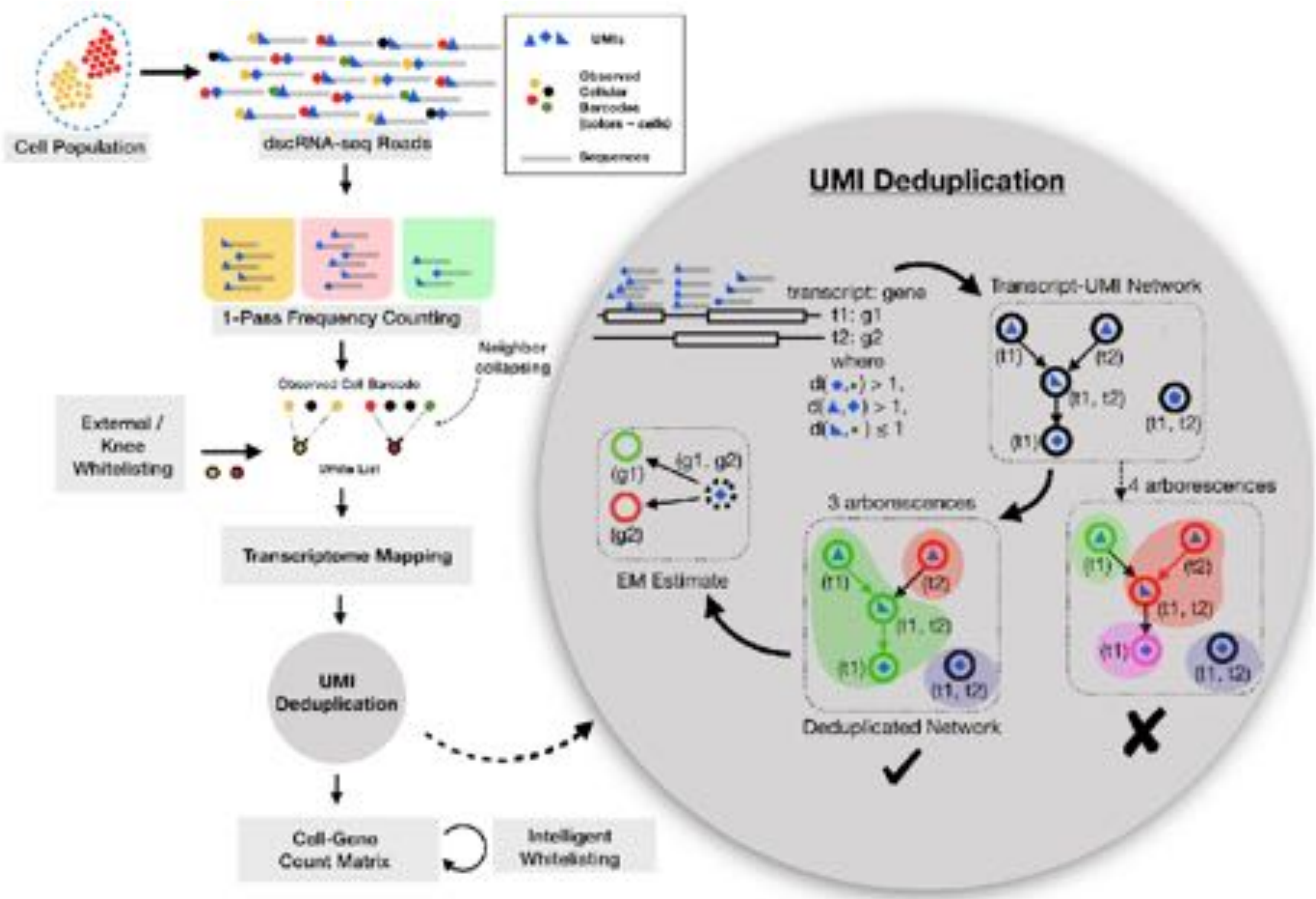
USEFUL TOOLS

- CellRanger (for 10X Genomics data)
- Alevin (salmon)
- Kallisto | bustools
- scPipe (Rsubread)
- Scruff (CEL-seq and CEL-seq2 data)

SCPIPE


scPipe workflow
Quality control metrics collected at each step
Fastq reformat (sc_trim_barcode)
Reads alignment (Rsubread::align)
Exon mapping (sc_exon_mapping)
Barcode demultiplex (sc_demultiplex)
Gene count (sc_gene_counting)
Number of Removed reads
Alignment rate
Number of reads mapped to intron/exon
Reads per cell; unmatched barcodes
Number of corrected UMI & filtered genes
Gene counting matrix
Quality control information matrix
A SingleCellExperiment object for quality control and further downstream analysis

ALEVIN



Srivastava et al. (2019). Genome Biology.

UMI DEDUPLICATION

- Each RNA molecule is tagged with a UMI.
  - Obviously, the reads with the same UMI should map to the same gene.
- Naive approach is to discard reads that map to more than one gene (ambiguous reads).
  - 15-20% of input reads in 3′-end methods.
- Discarding reads can bias gene expression estimates.

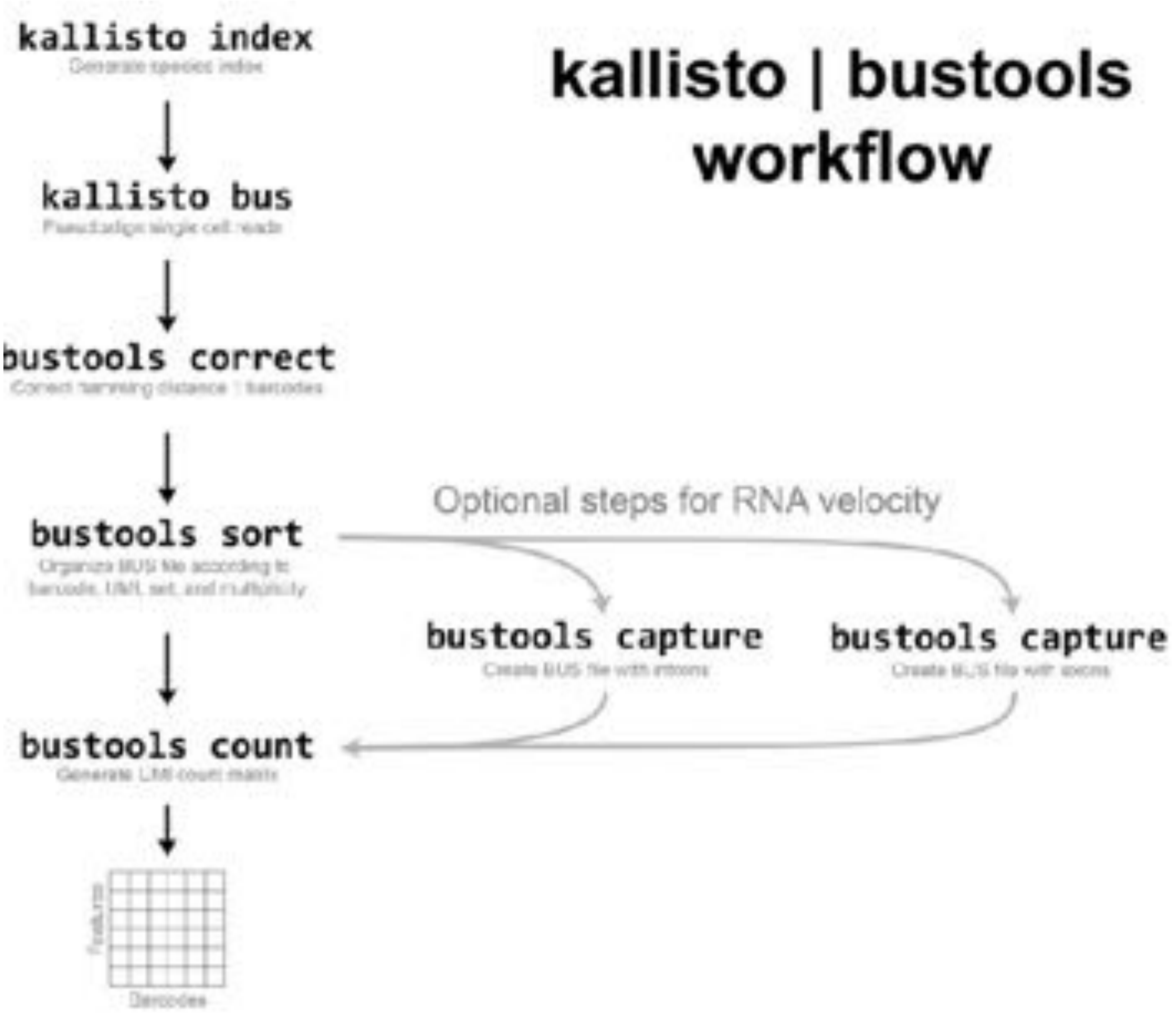
Srivastava et al. (2019). Genome Biology.

KALLISTO | BUSTOOLS

- Uses pseudo-alignment and a new format called BUS (Barcode, UMI, Set) to efficiently produce UMI count matrices.
- It can correct barcode sequencing errors and “collisions”, but empirically only a negligible fraction of UMIs are affected.
- Automatically generates spliced and unspliced RNA matrices for fast RNA velocity estimates.

Melsted et al. (2019). bioRxiv.

KALLISTO | BUSTOOLS



kallisto | bustools workflow

kallisto index

Generate species index

kallisto bus

Pseudoalign single cell reads

bustools correct

Correct hamming distance 1 barcodes

bustools sort

Organize BUS file according to barcode, UMI, set, and multiplicity

Optional steps for RNA velocity

bustools capture

Create BUS file with introns

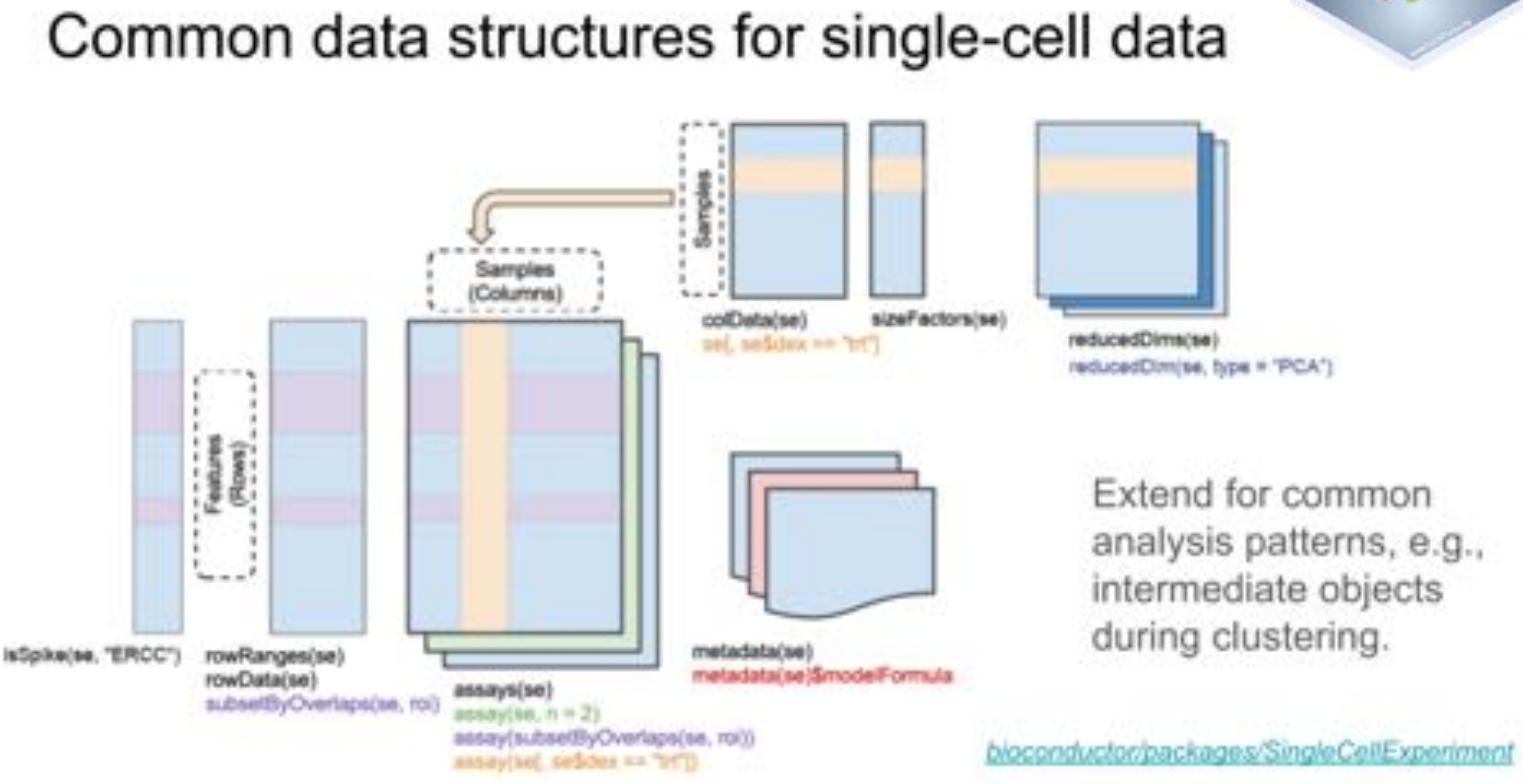
bustools capture

Create BUS file with exons

bustools count

Generate UMI count matrix

THE SINGLECELLEXPERIMENT CLASS



Common data structures for single-cell data

Extend for common analysis patterns, e.g., intermediate objects during clustering.

bioconductor/packages/SingleCellExperiment

The SingleCellExperiment class

```
sce
```

```
## class: SingleCellExperiment
## dim: 3079 1000
## metadata(1): log.exprs.offset
## assays(2): counts logcounts
## rownames(3079): ENSG00000188976 ENSG00000187608 ...
##   ENSG00000198727 ENSG00000220023
## rowData names(12): ENSEMBL_ID Symbol_TENx ... total_counts
##   log10_total_counts
## colnames(1000): Cell1 Cell2 ... Cell999 Cell1000
## colData names(56): Sample Barcode ...
##   pct_counts_in_top_200_features_mito
##   pct_counts_in_top_500_features_mito
## reducedDimNames(2): PCA zinbwave
## spikeNames(0):
```

QUALITY CONTROL

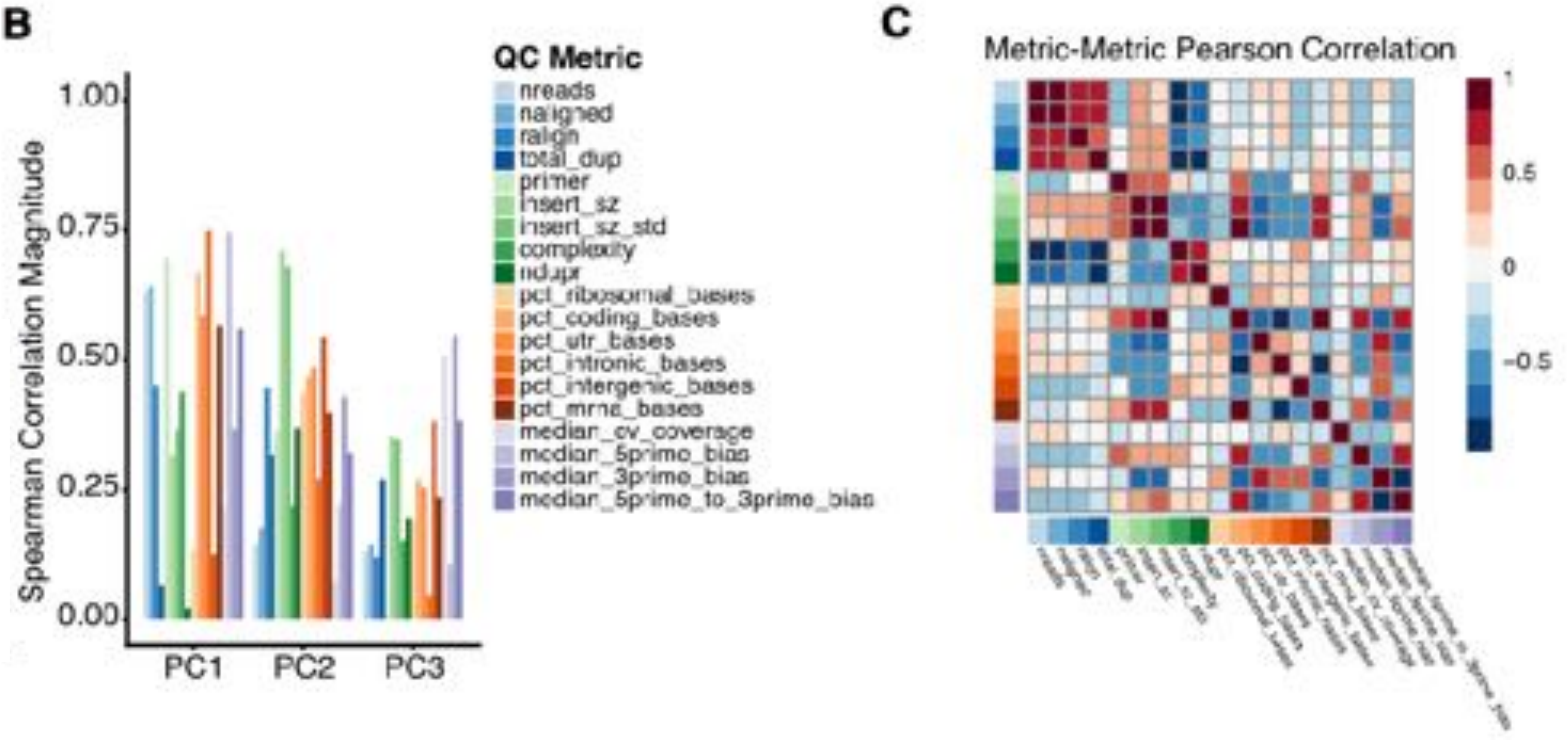
QUALITY CONTROL AND FILTERING

- Exploratory data analysis (EDA) and quality control (QC) are of utmost importance in genomics.
- With single cell data we have the luxury of having a large number of samples, hence we can filter out low quality cells as well as lowly expressed genes.
- There are some simple metrics that we can compute as a proxy of the quality of the samples.

Computing QC metrics

```r
sce <- TENxPBMCData::TENxPBMCData("pbmc4k")
sce <- scater::calculateQCMetrics(sce)
```

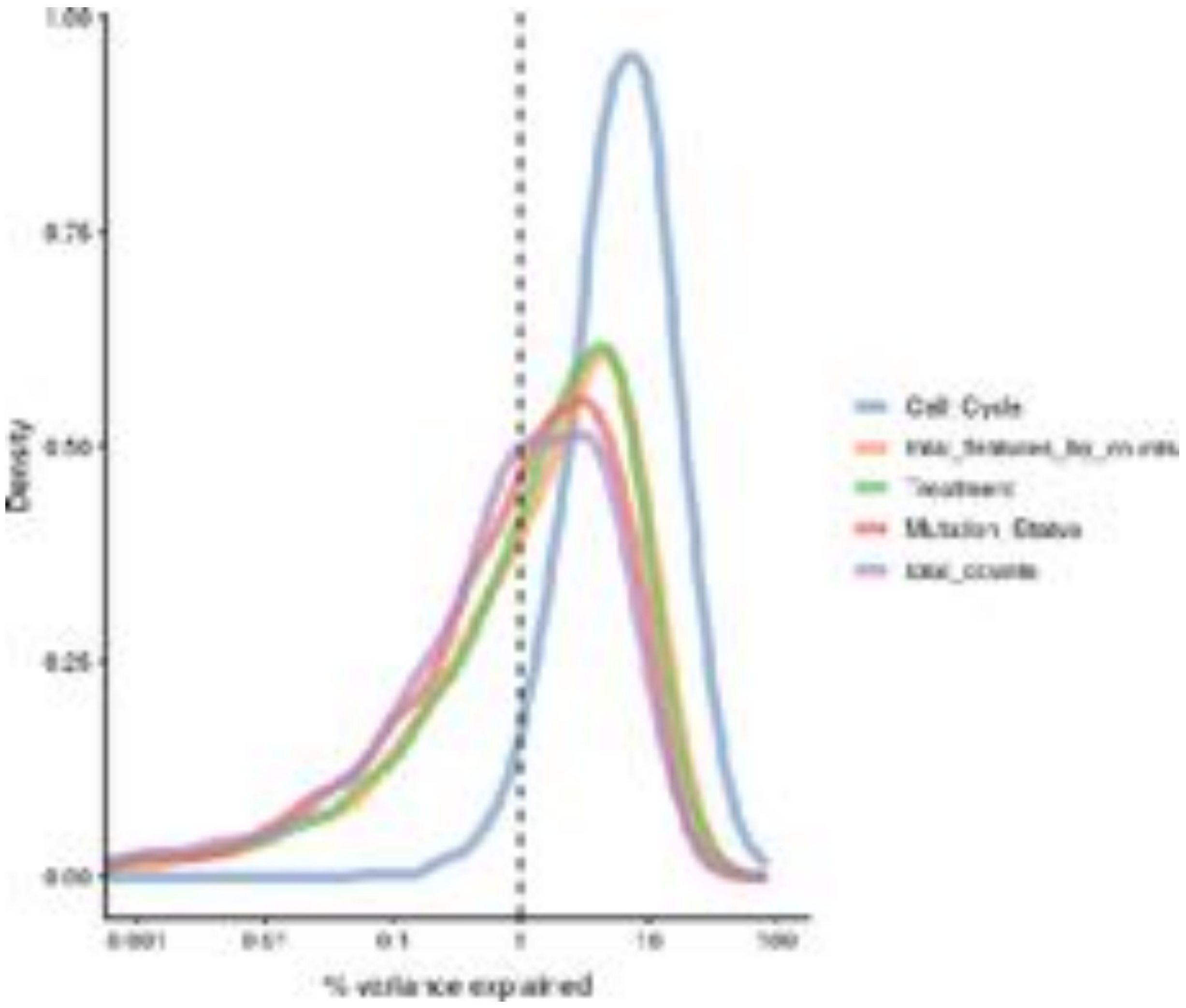
QC METRICS

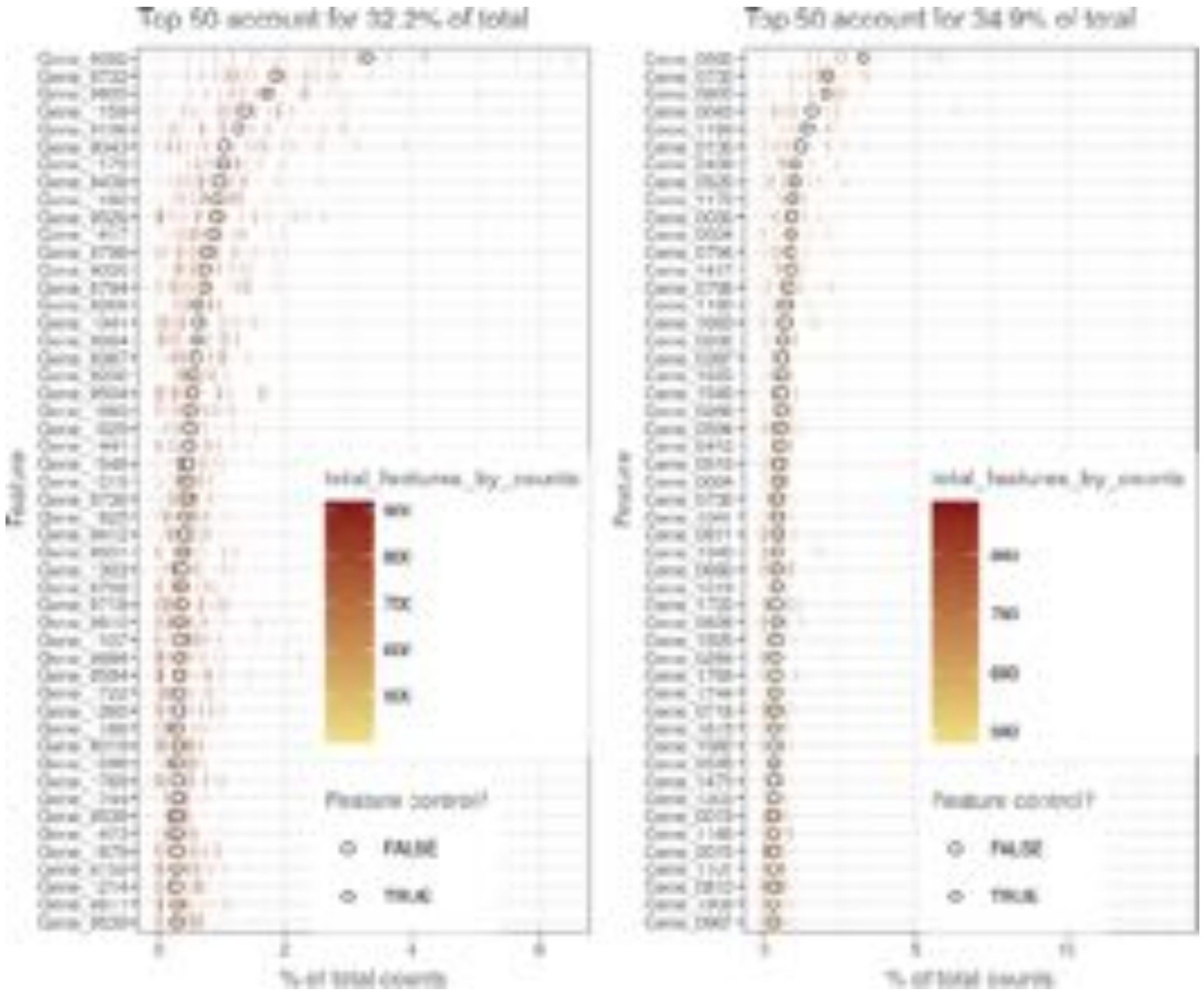


Cole et al. (2019). Cell Systems.

scone Bioconductor Package

EXPLORING DATA QUALITY

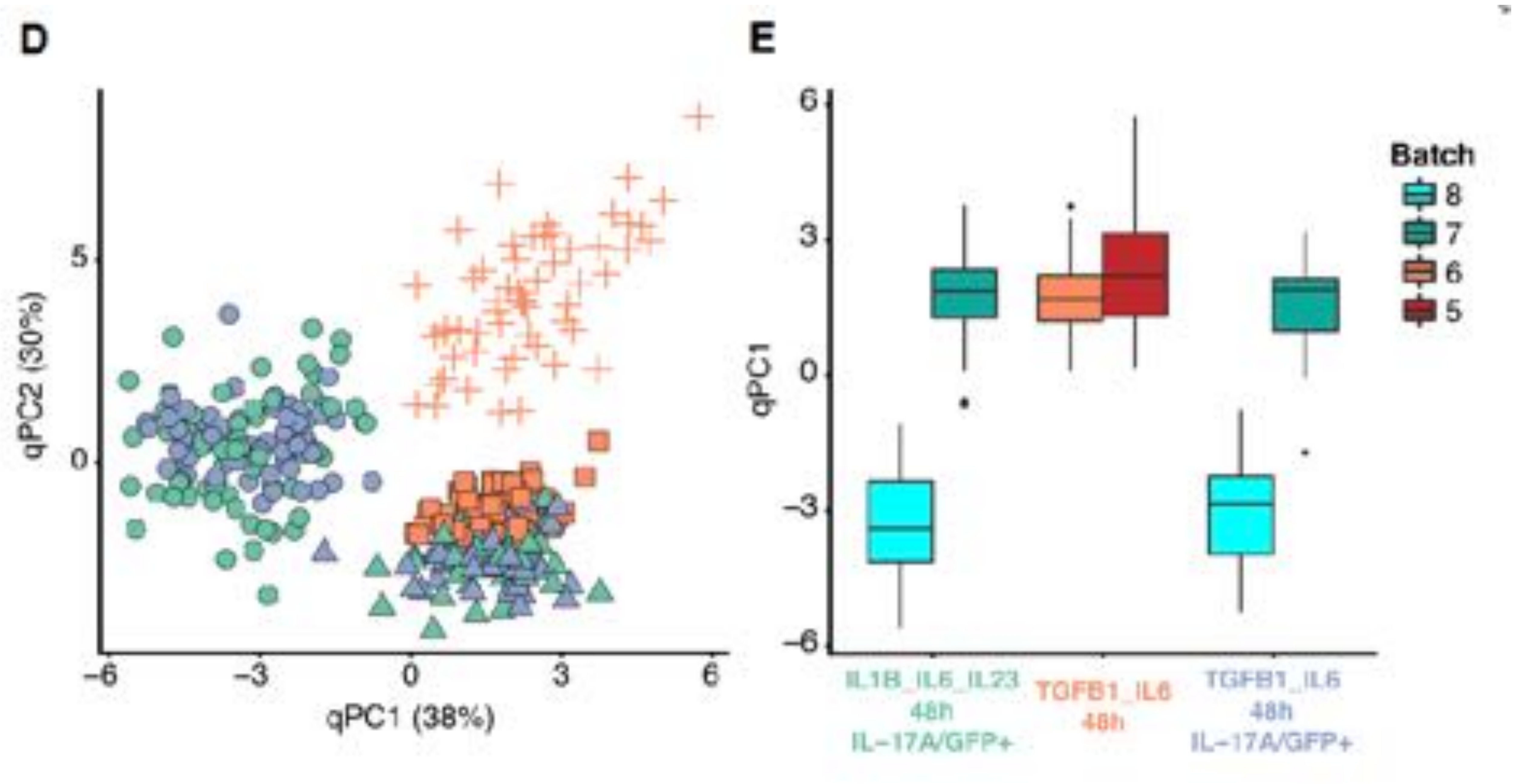




McCarthy et al. (2019). Bioinformatics.

scater Bioconductor Package

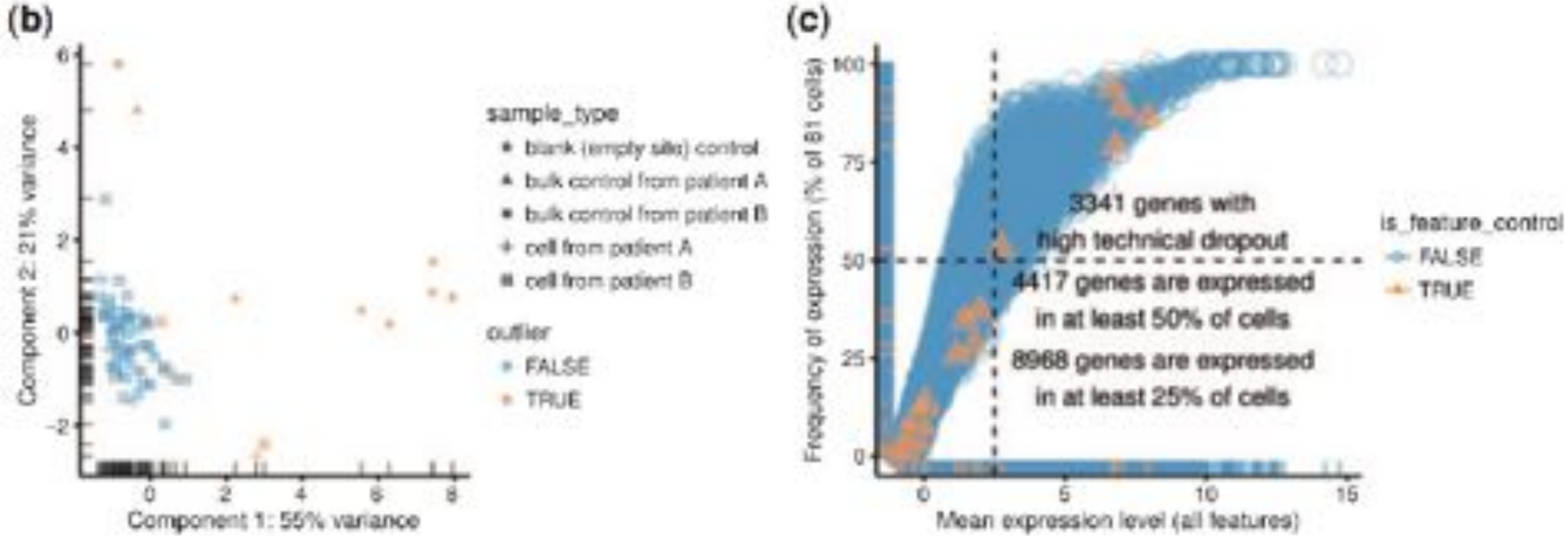
EXPLORING DATA QUALITY



Cole et al. (2019). Cell Systems.

scone Bioconductor Package

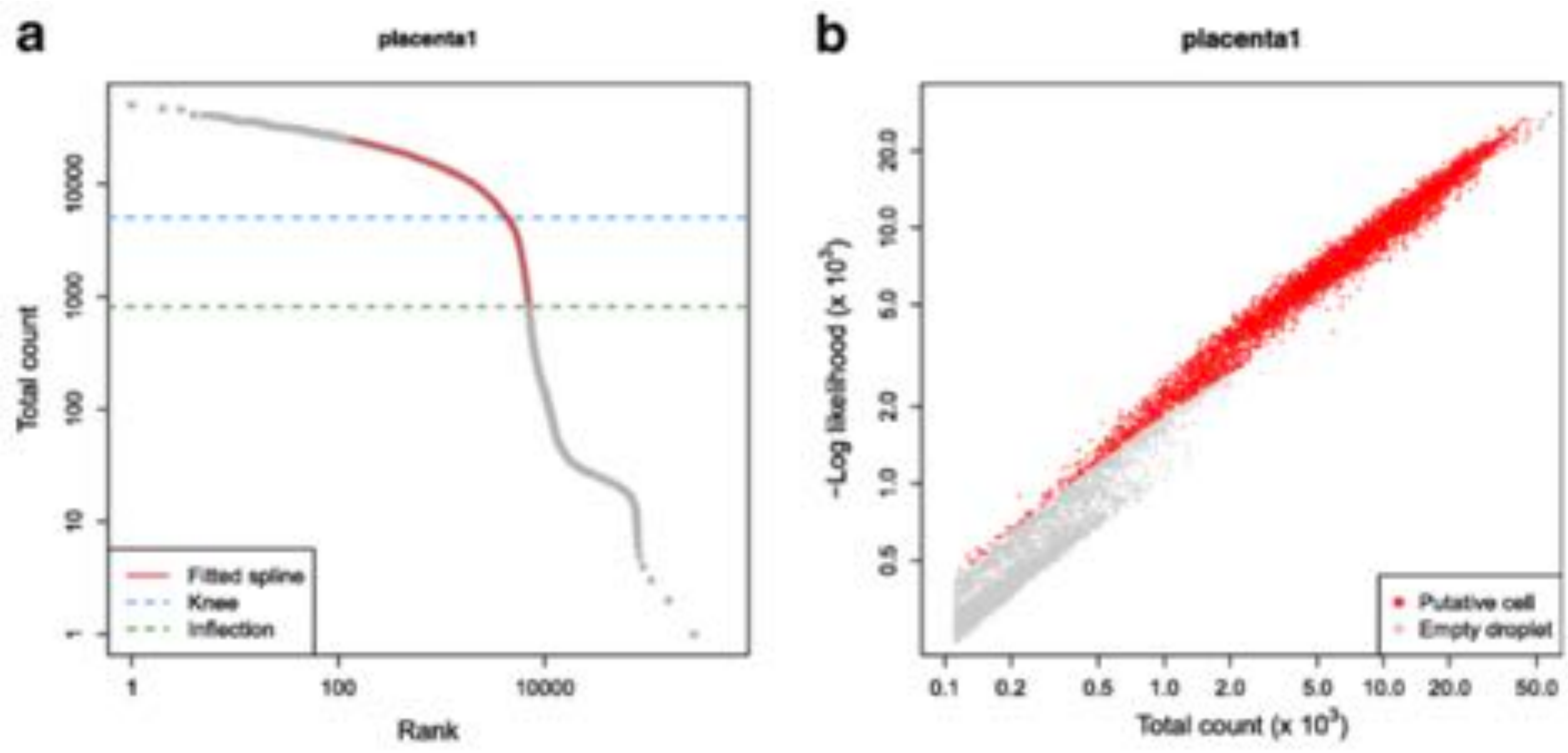
FILTERING GENES AND SAMPLES



McCarthy et al. (2019). Bioinformatics.

scater Bioconductor Package

EMPTY DROPLETS VS CELLS



Lun et al. (2019). Genome Biology.

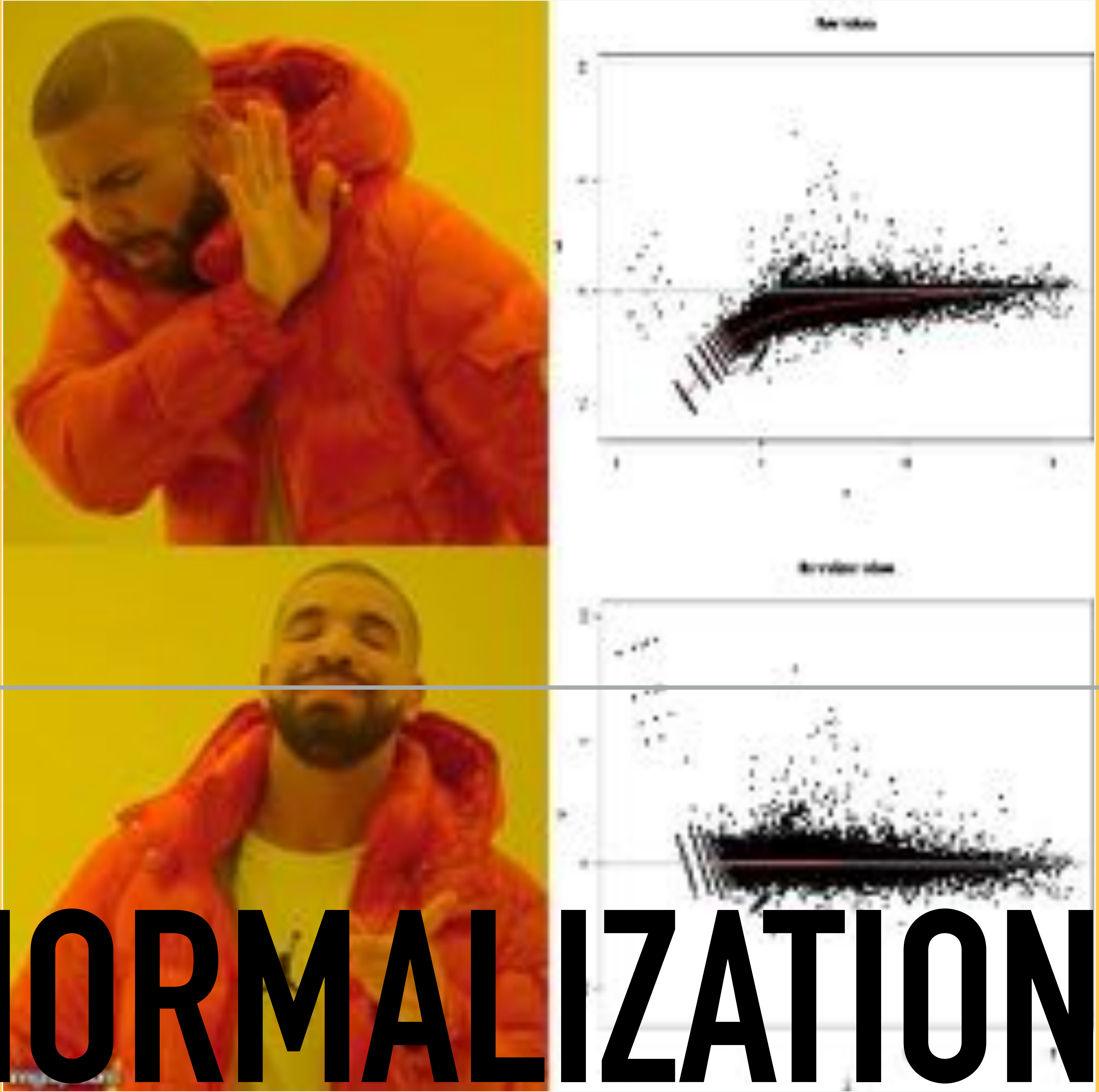
DropletUtils Bioconductor Package

THE EMPTYDROPS METHOD

- Estimate the expression profile of ambient RNA from the droplets with less than *T* total UMI counts

- Test deviation from this profile using a Dirichlet-multinomial model to identify non-empty (i.e., cell containing) droplets.

- To avoid incorrectly calling ambient-like cells as empty droplets, a “knee point” is identified by fitting a spline and cells with total count greater than the knee point are always retained.



NORMALIZATION

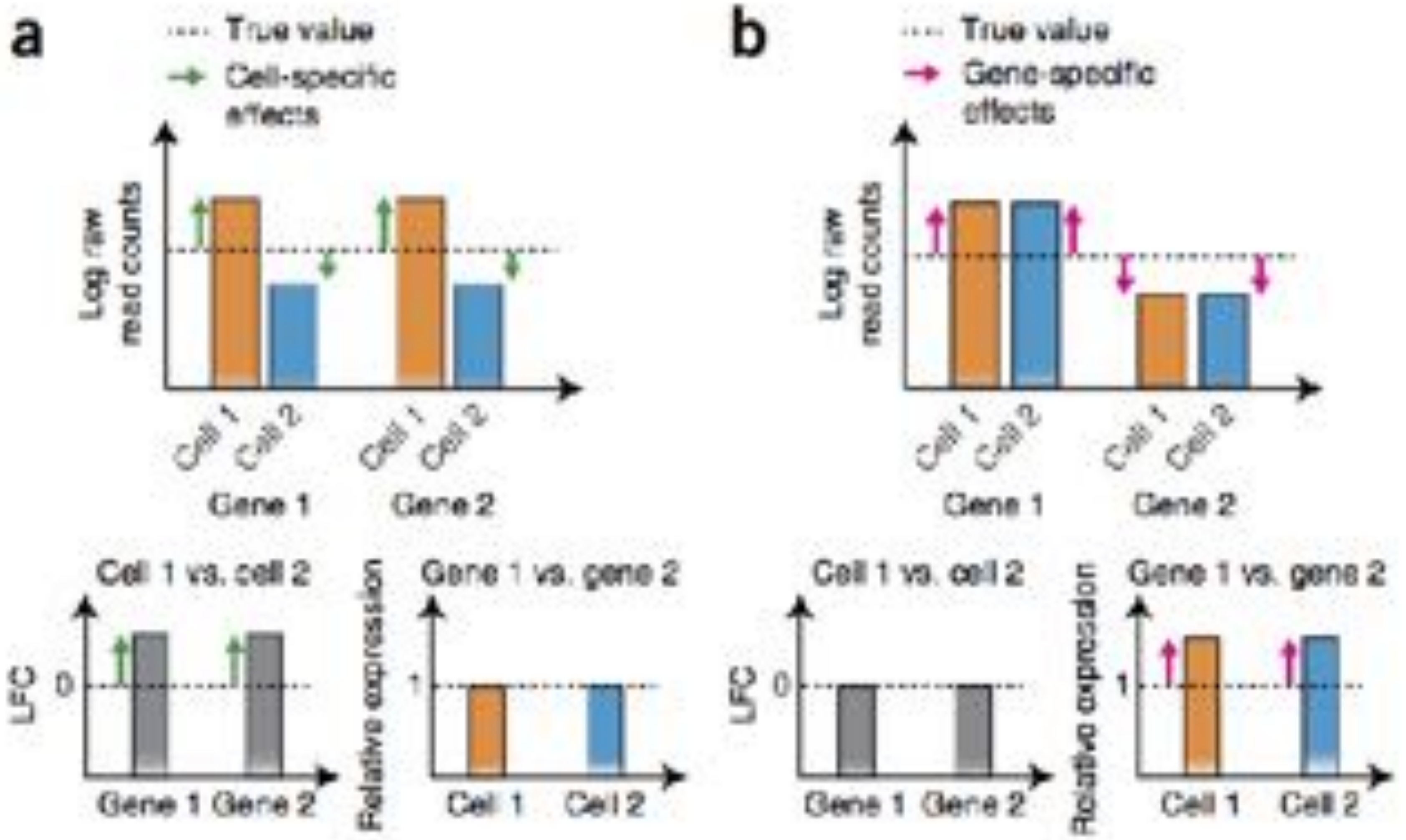
NORMALIZATION

- As with bulk RNA-seq, it is important to account for differences in sequencing depth and other biases that may affect the expression levels.

- Usually, it is a preprocessing step prior to other analyses.

- Some methods, such as **MAST**, **ZINB-WaVE**, and **BASiCS**, include normalization factors as part of the models and estimate them along with the other parameters.

NORMALIZATION

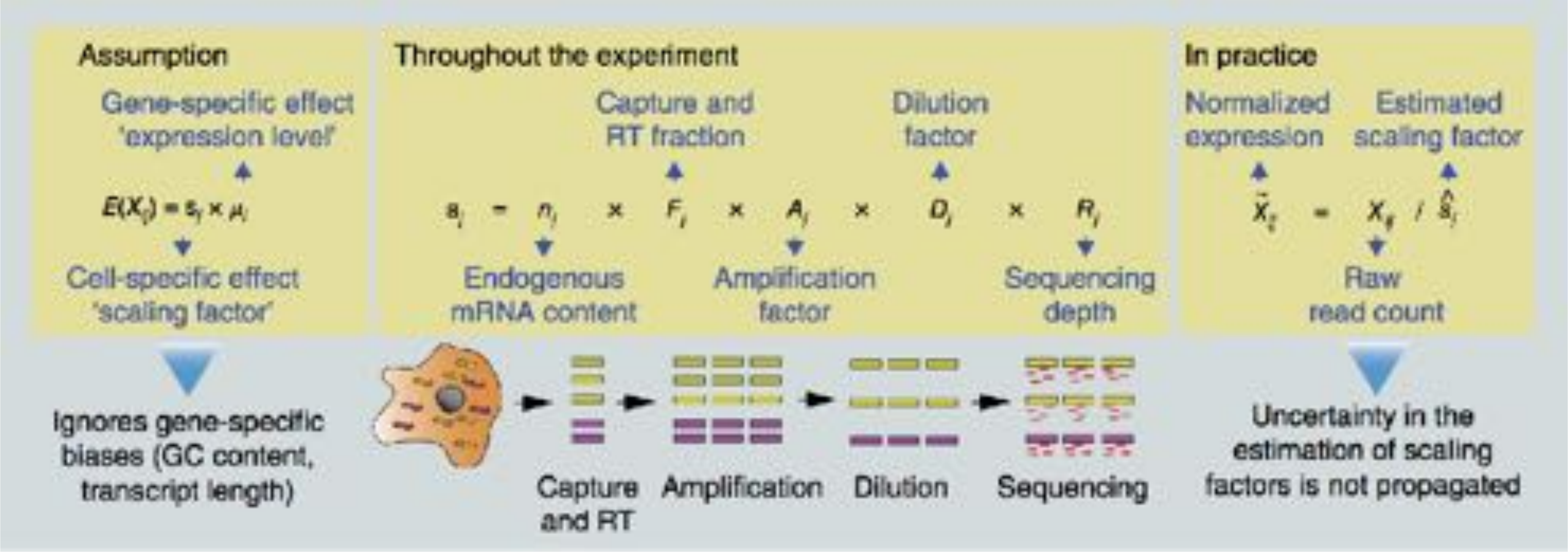


c

	Cell-specific effects	Gene-specific effects	Not removed by UMIs
Sequencing depth	✓		✓
Amplification	✓	✓	
Capture and RT efficiency	✓	✓	✓
Gene length		✓	
GC content	✓	✓	✓
mRNA content	✓		✓

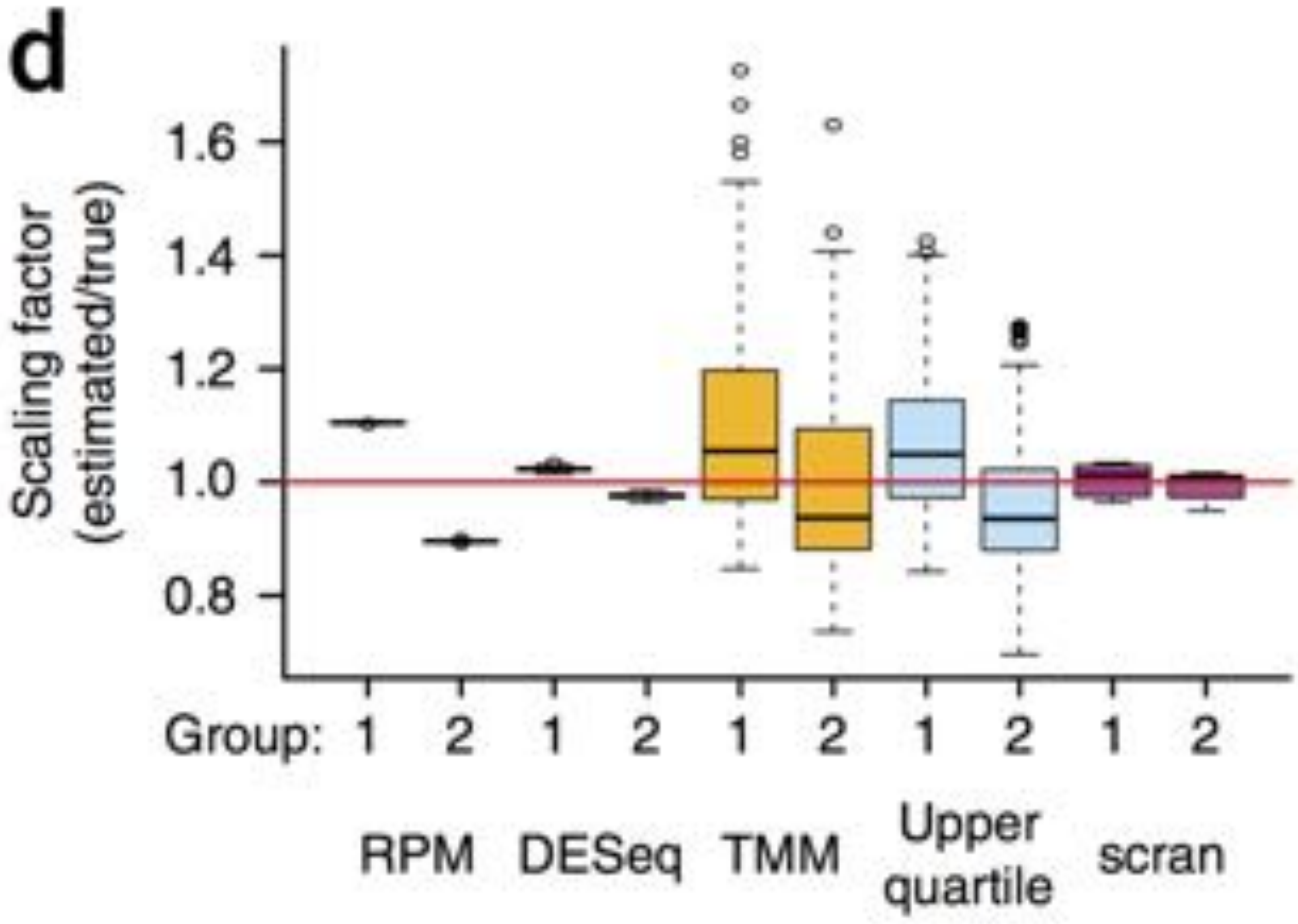
Vallejos et al. (2017). Nat Methods.

NORMALIZATION


Assumption
Gene-specific effect 'expression level'
$E(X_{ij}) = s_j \times \mu_i$
Cell-specific effect 'scaling factor'
Throughout the experiment
Capture and RT fraction
Dilution factor
$s_j = n_j \times F_j \times A_j \times D_j \times R_j$
Endogenous mRNA content
Amplification factor
Sequencing depth
In practice
Normalized expression
Estimated scaling factor
$\tilde{X}_{ij} = X_{ij} / \hat{s}_j$
Raw read count
Ignores gene-specific biases (GC content, transcript length)
Capture and RT
Amplification
Dilution
Sequencing
Uncertainty in the estimation of scaling factors is not propagated

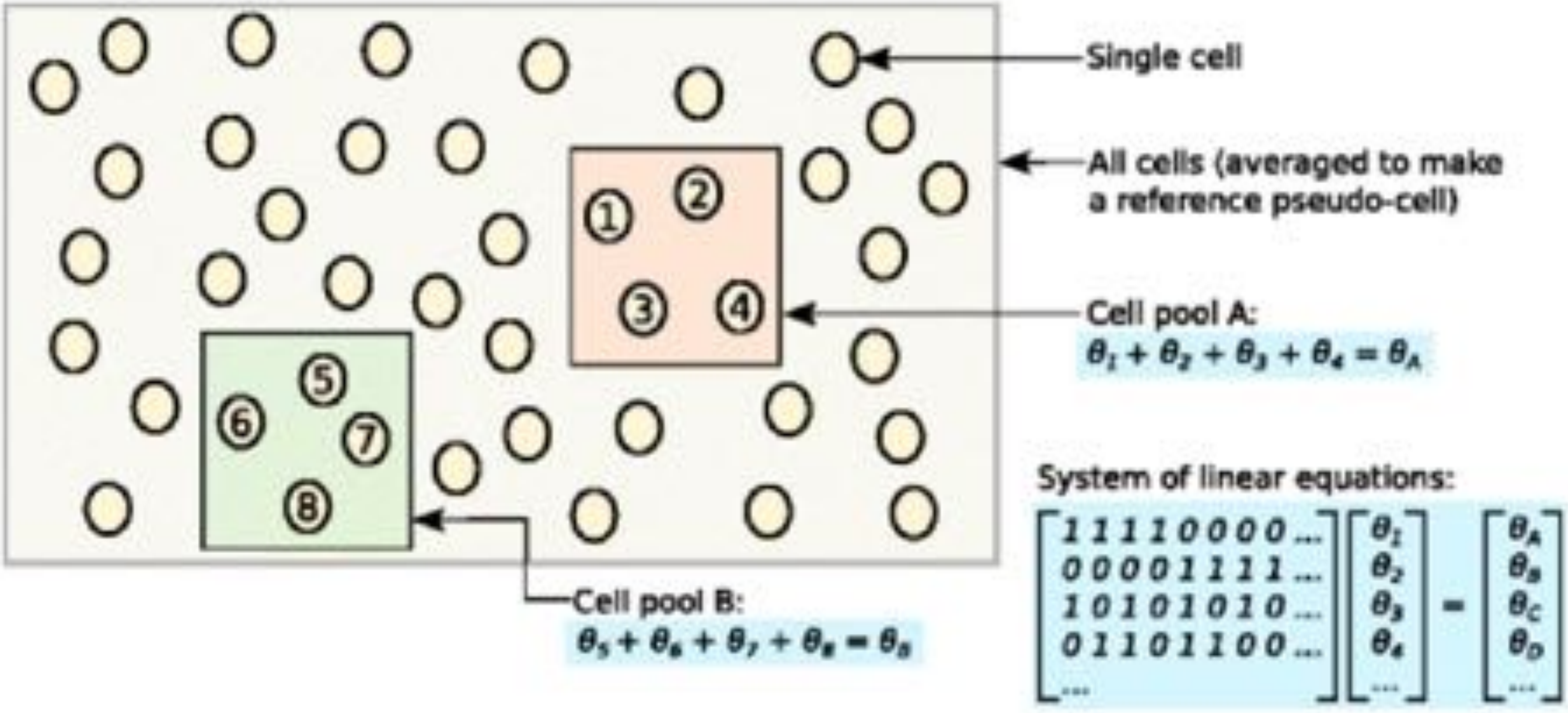
Vallejos et al. (2017). Nat Methods.

NORMALIZATION



Vallejos et al. (2017). Nat Methods.

POOLING ACROSS CELLS HELPS

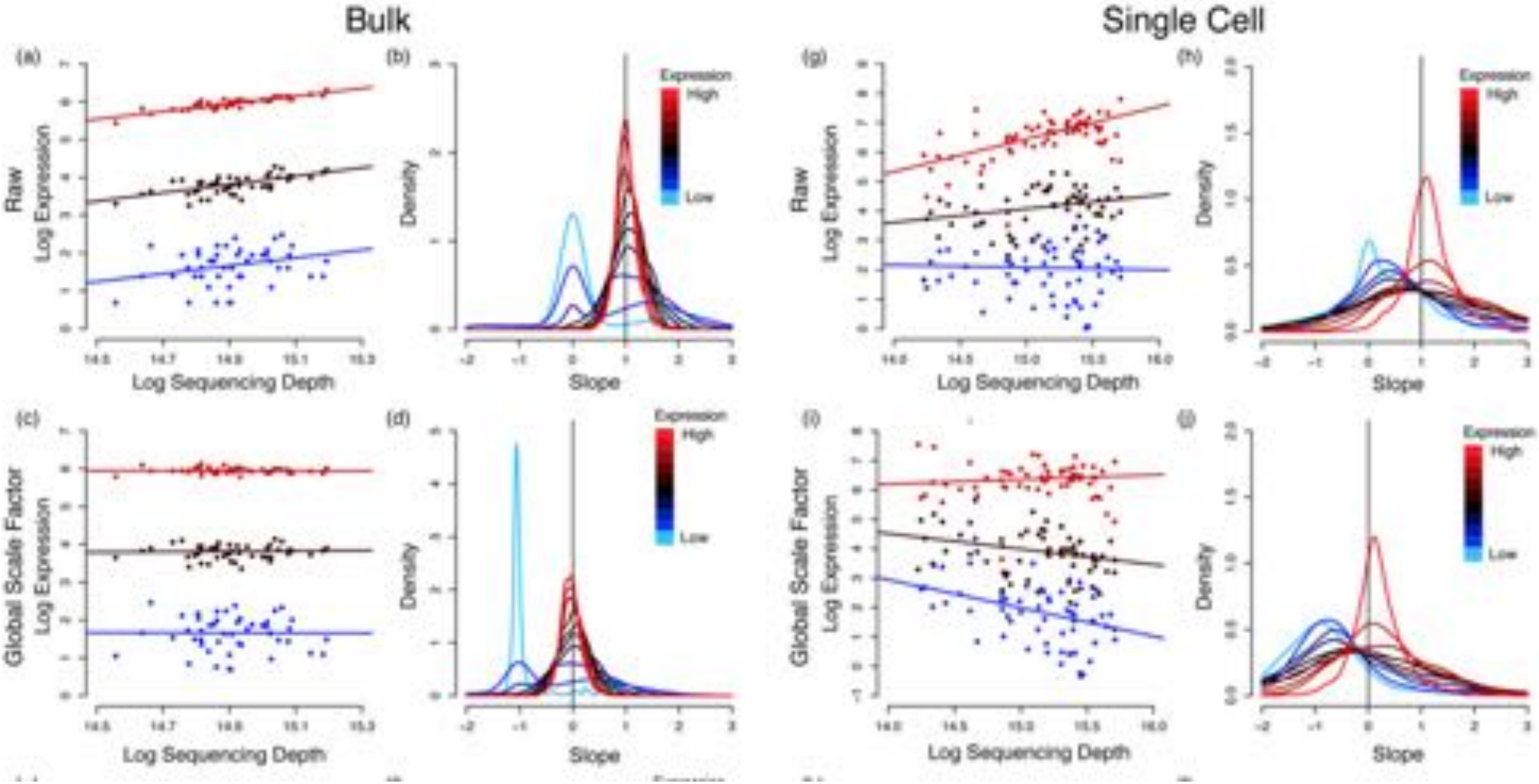

1
2
3
4
5
6
7
8
Single cell
All cells (averaged to make a reference pseudo-cell)
Cell pool A:
$\theta_1 + \theta_2 + \theta_3 + \theta_4 = \theta_A$
Cell pool B:
$\theta_5 + \theta_6 + \theta_7 + \theta_8 = \theta_B$
System of linear equations:

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

Lun et al. (2016). Genome Biology.

scran Bioconductor Package

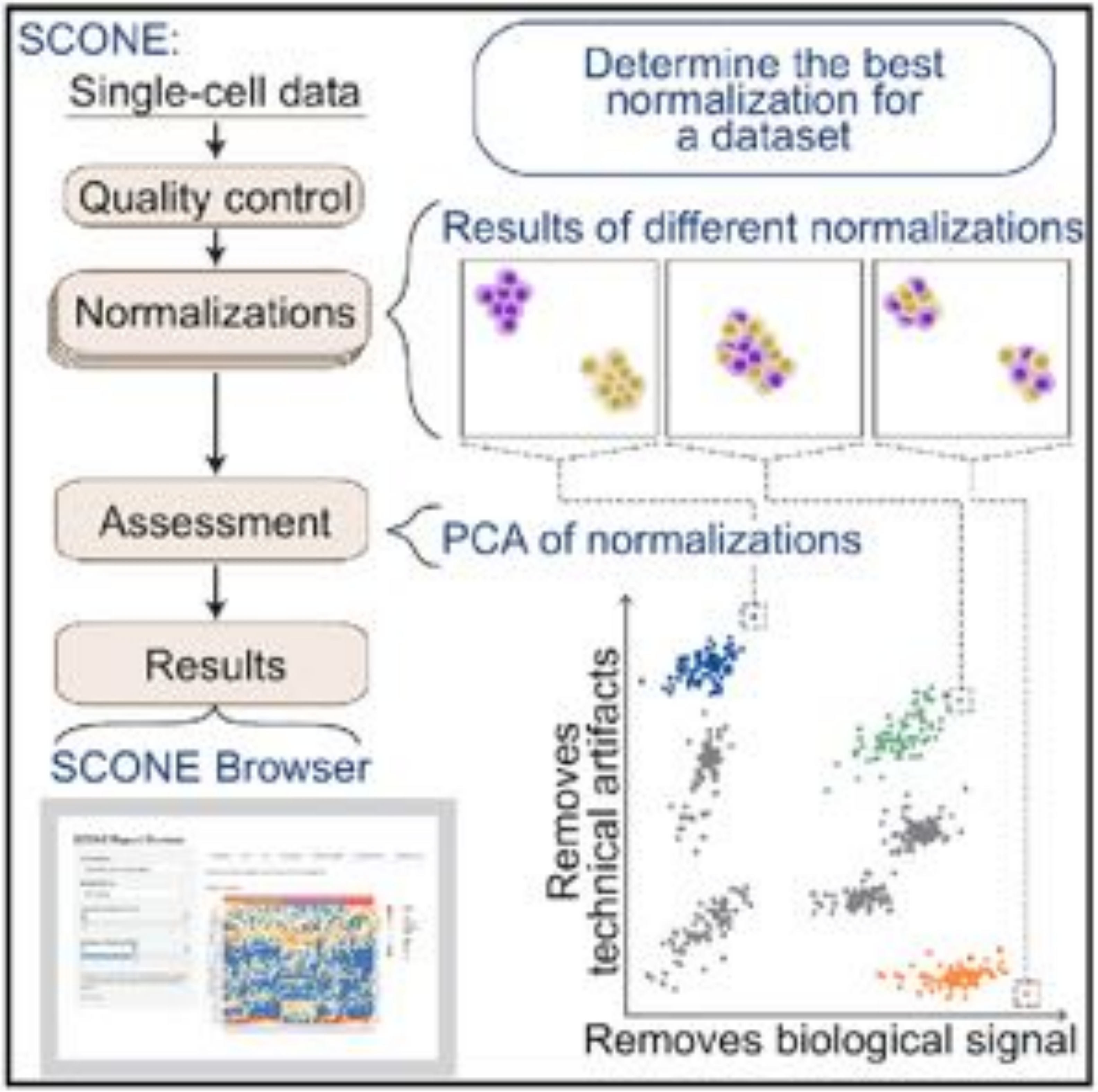
NON-LINEAR NORMALIZATION



Bacher et al. (2017). Nat Methods.

SCnorm Bioconductor Package

RANKING NORMALIZATION BY PERFORMANCE


SCONE:
Single-cell data
Quality control
Normalizations
Assessment
Results
SCONE Browser
Determine the best normalization for a dataset
Results of different normalizations
PCA of normalizations
Removes technical artifacts
Removes biological signal

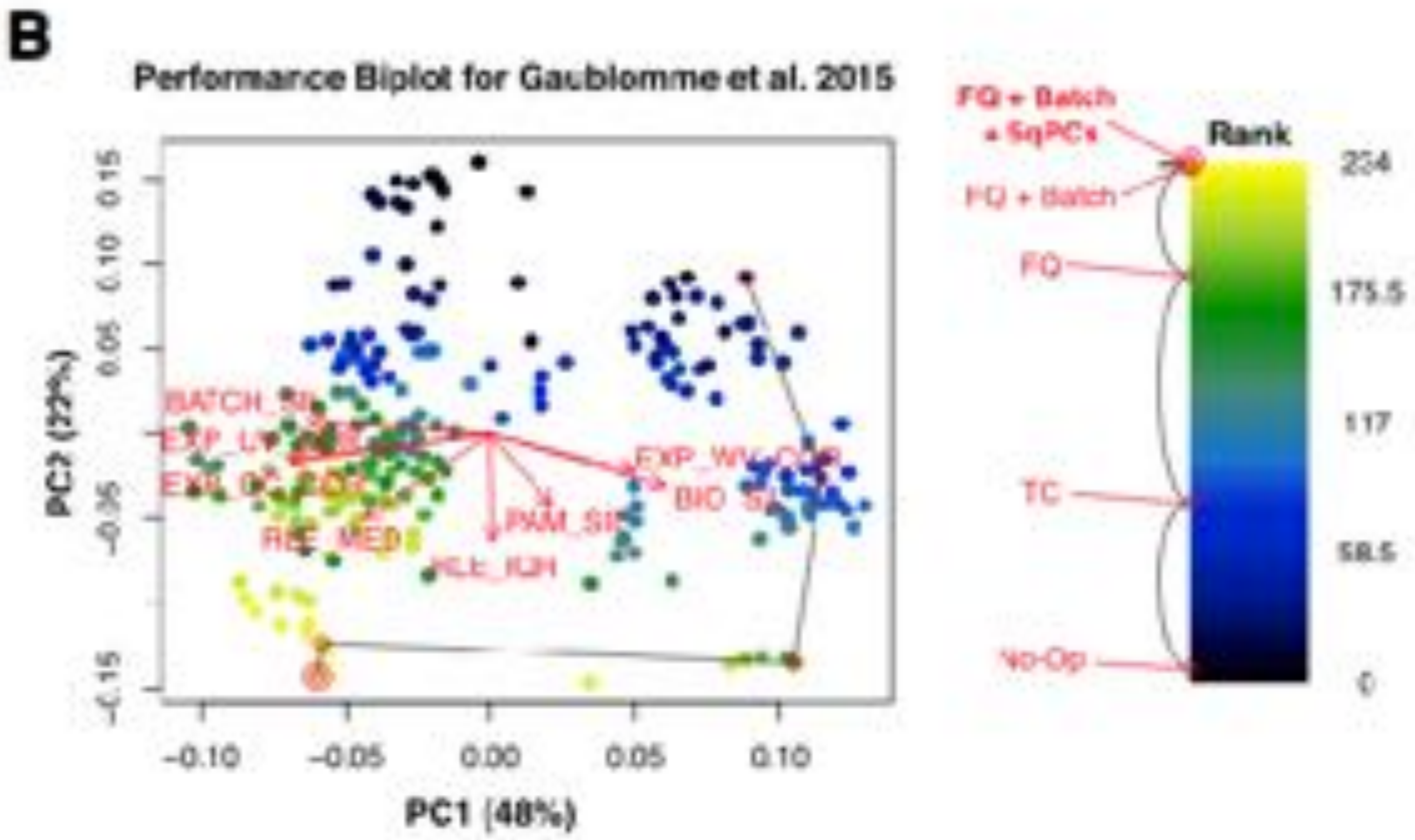
Cole et al. (2019). Cell Systems.

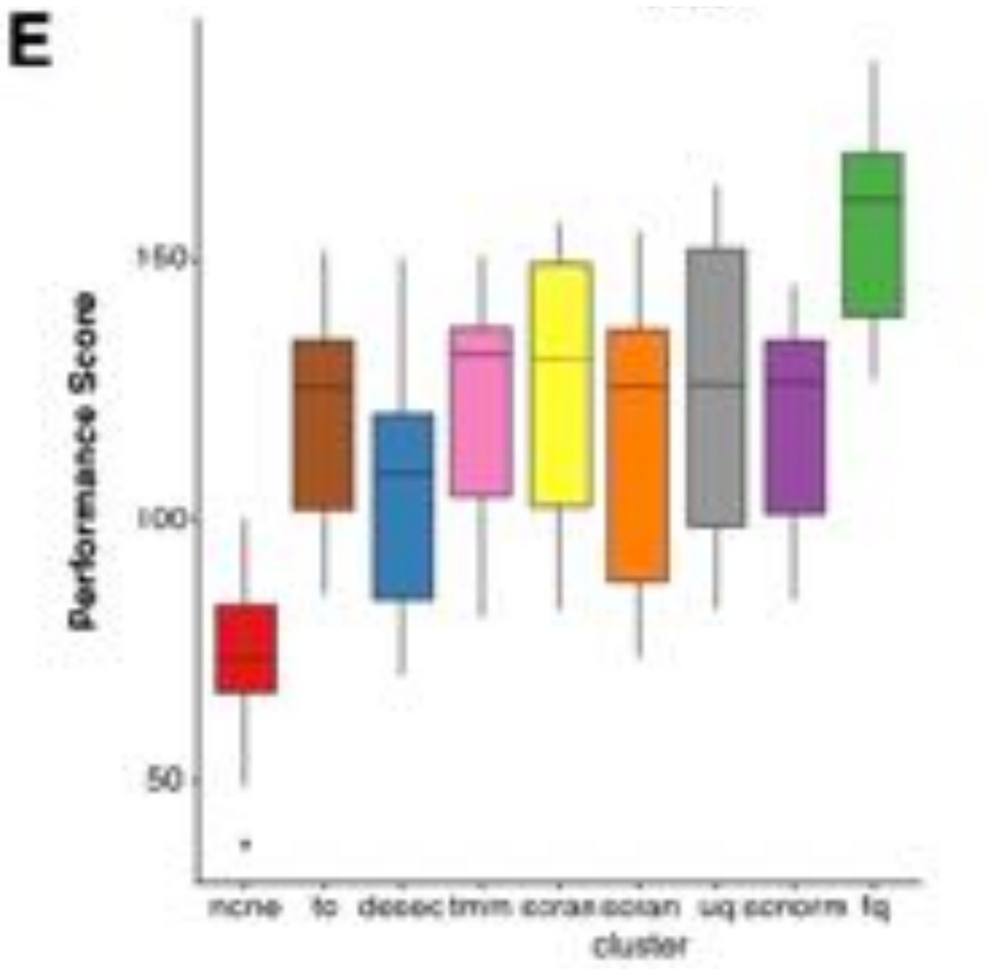
scone Bioconductor Package

SCONE PERFORMANCE METRICS

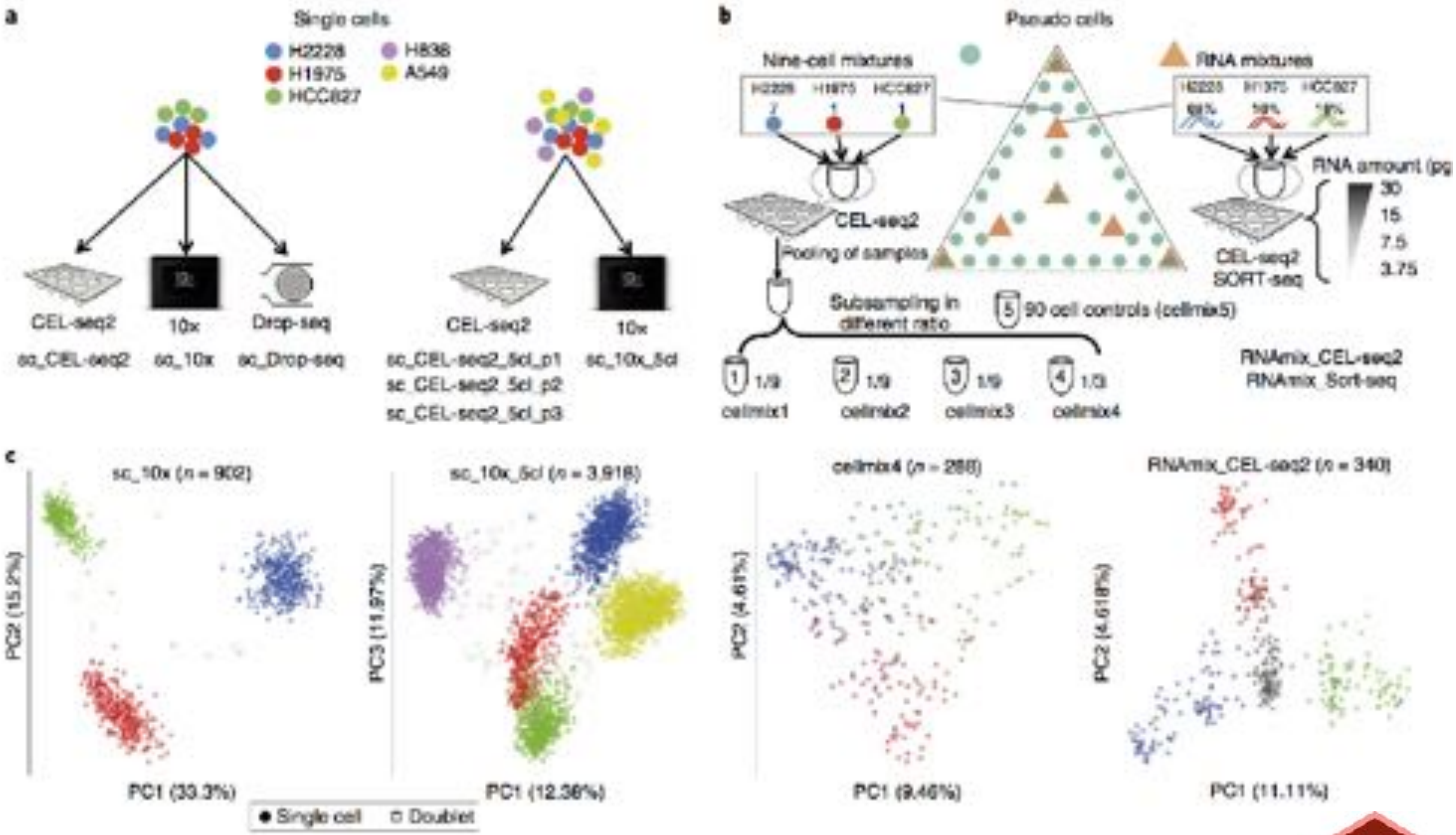
1. Clustering of samples according to factors of wanted and unwanted variation.
    - Average silhouette width, with samples grouped by cell type, batch.
2. Association of expression with factors of wanted and unwanted variation.
    - Correlation with QC measures, positive and negative controls.
3. Between-sample distributional properties of the expression measures.
    - Relative-log-expression (RLE).

RANKING NORMALIZATION USING SCONE





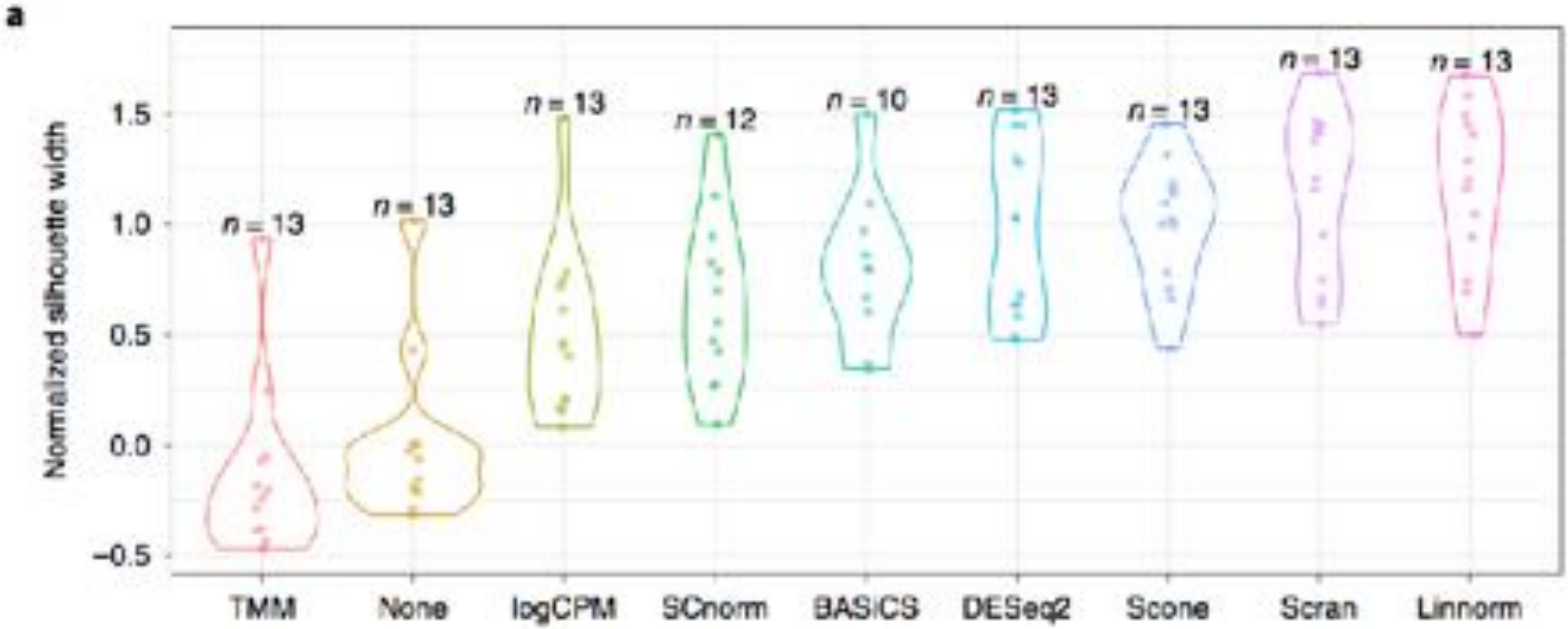
BENCHMARKING USING EXPERIMENTAL MIXTURES



Tian et al. (2019). Nat Methods.

CellBench Bioconductor Package

BENCHMARKING USING EXPERIMENTAL MIXTURES



Tian et al. (2019). Nat Methods.

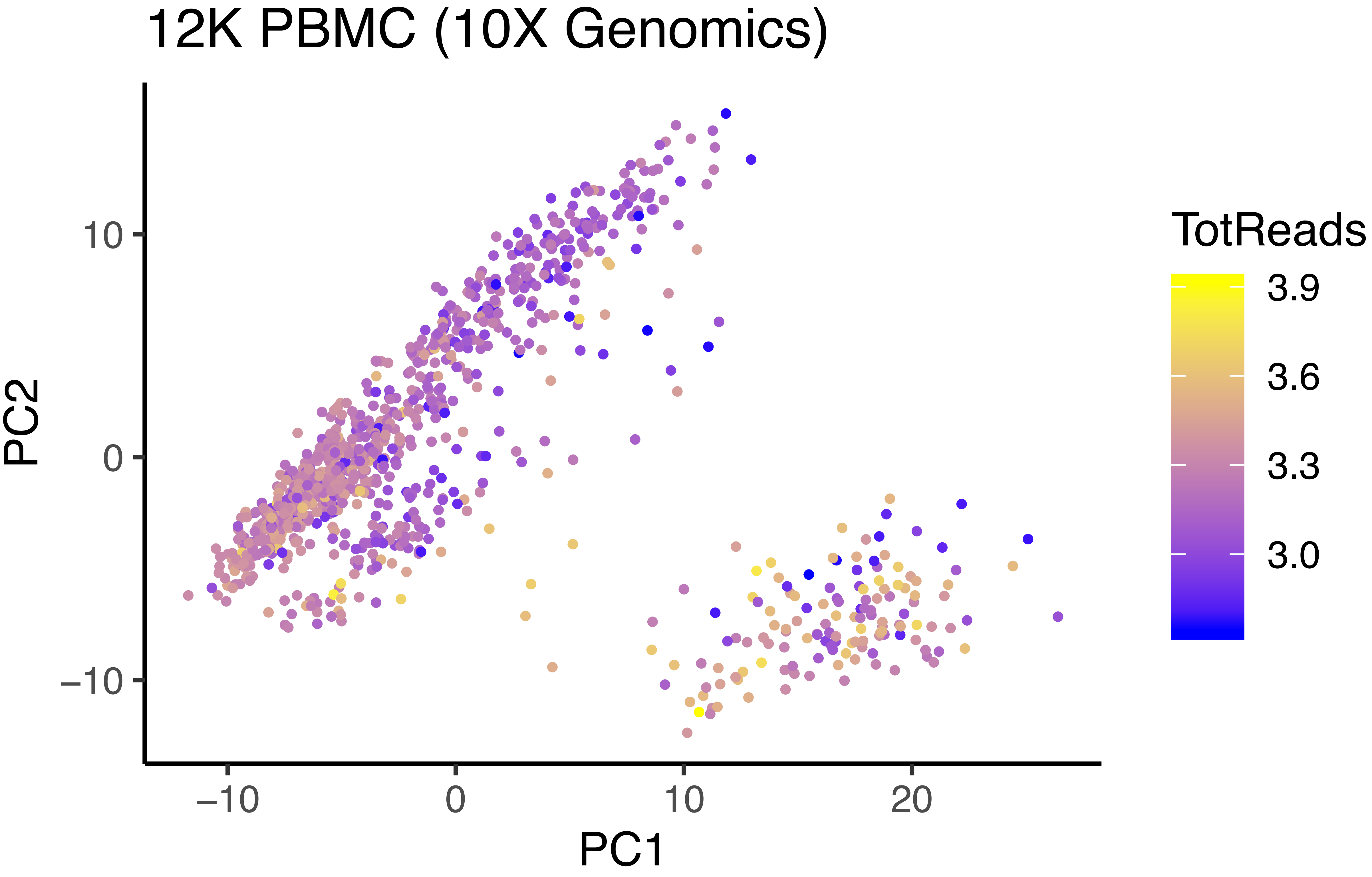
CellBench Bioconductor Package

CellBench

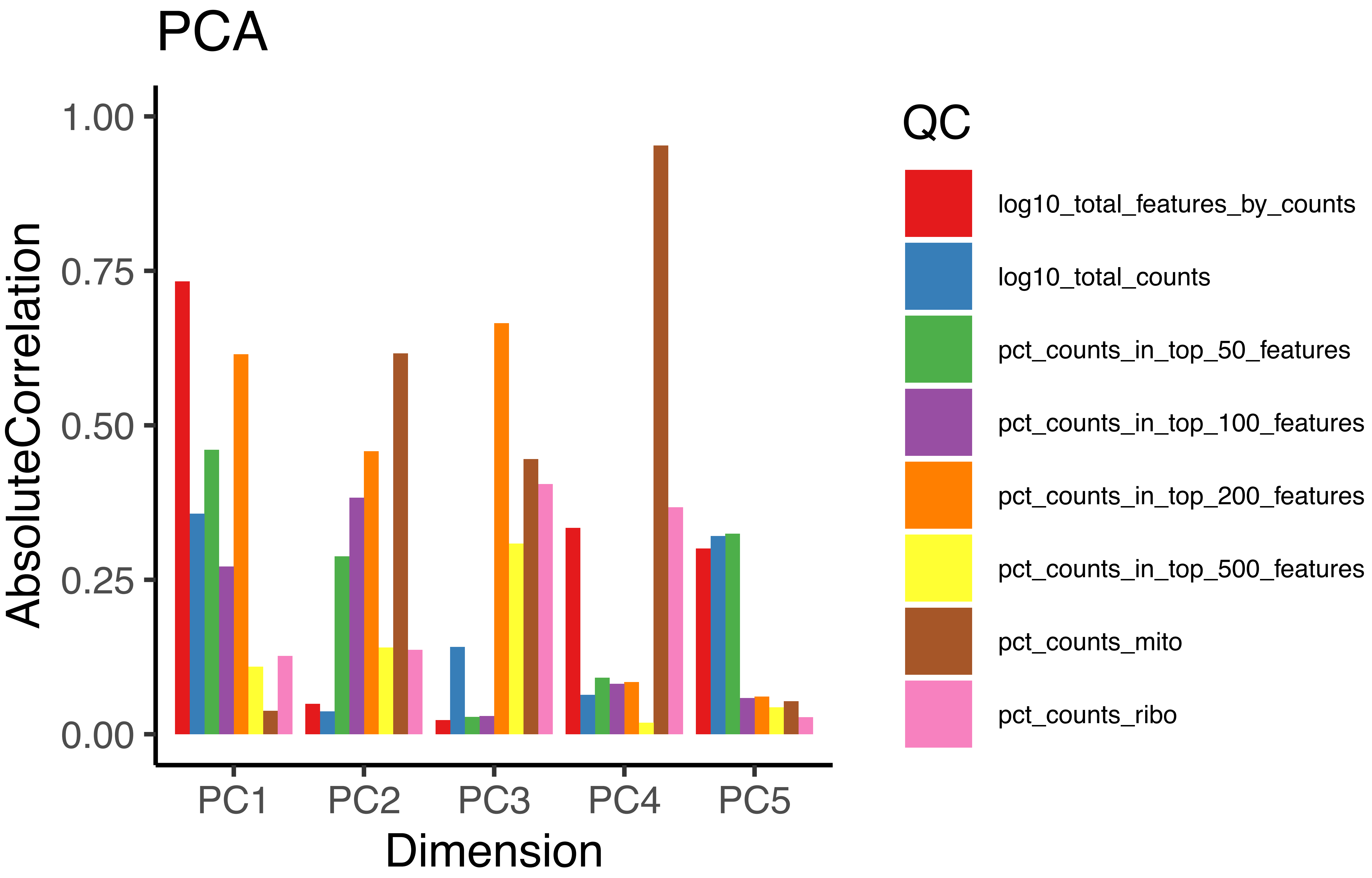
DIRECTLY ACCOUNTING FOR QUALITY

- The normalization methods seen so far are global scaling methods.
- An alternative is to account for the quality of the samples (and batch effects) directly in the statistical model.
- Several methods do that
  - MAST and BASiCS for differential expression.
  - ZINB-WaVE, scVI, and GLM-PCA for dimensionality reduction.
- We will see ZINB-WaVE as an example.

Sample quality affects PCA

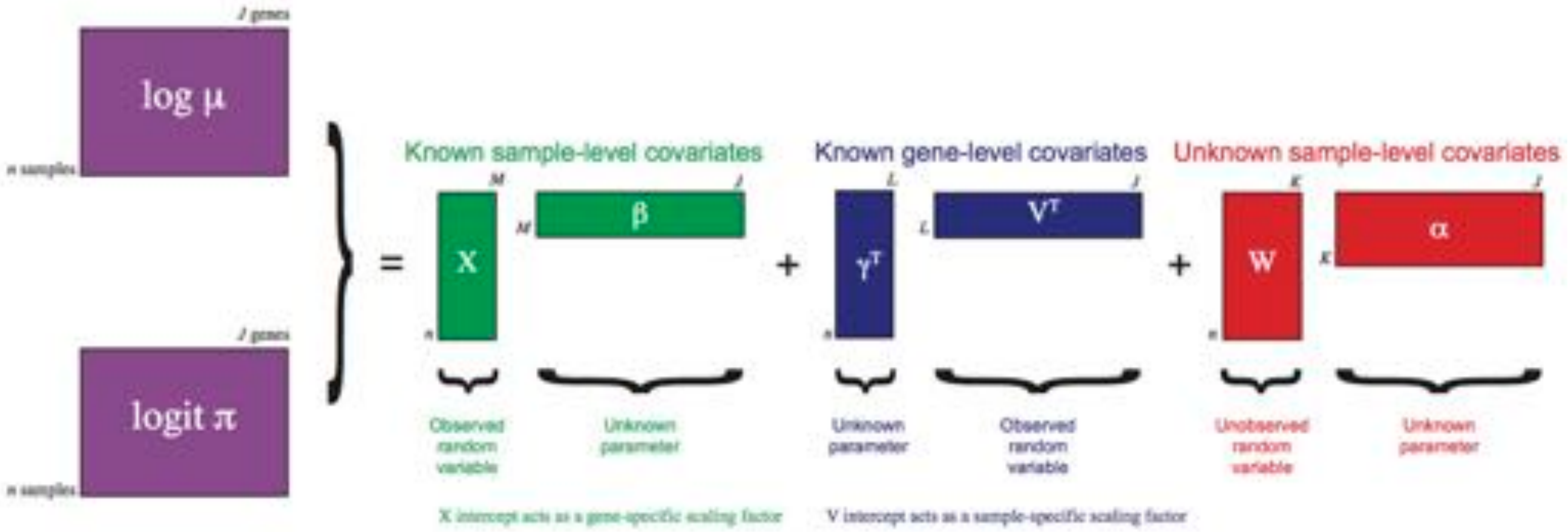


Sample quality affects PCA



The ZINB-WaVE model

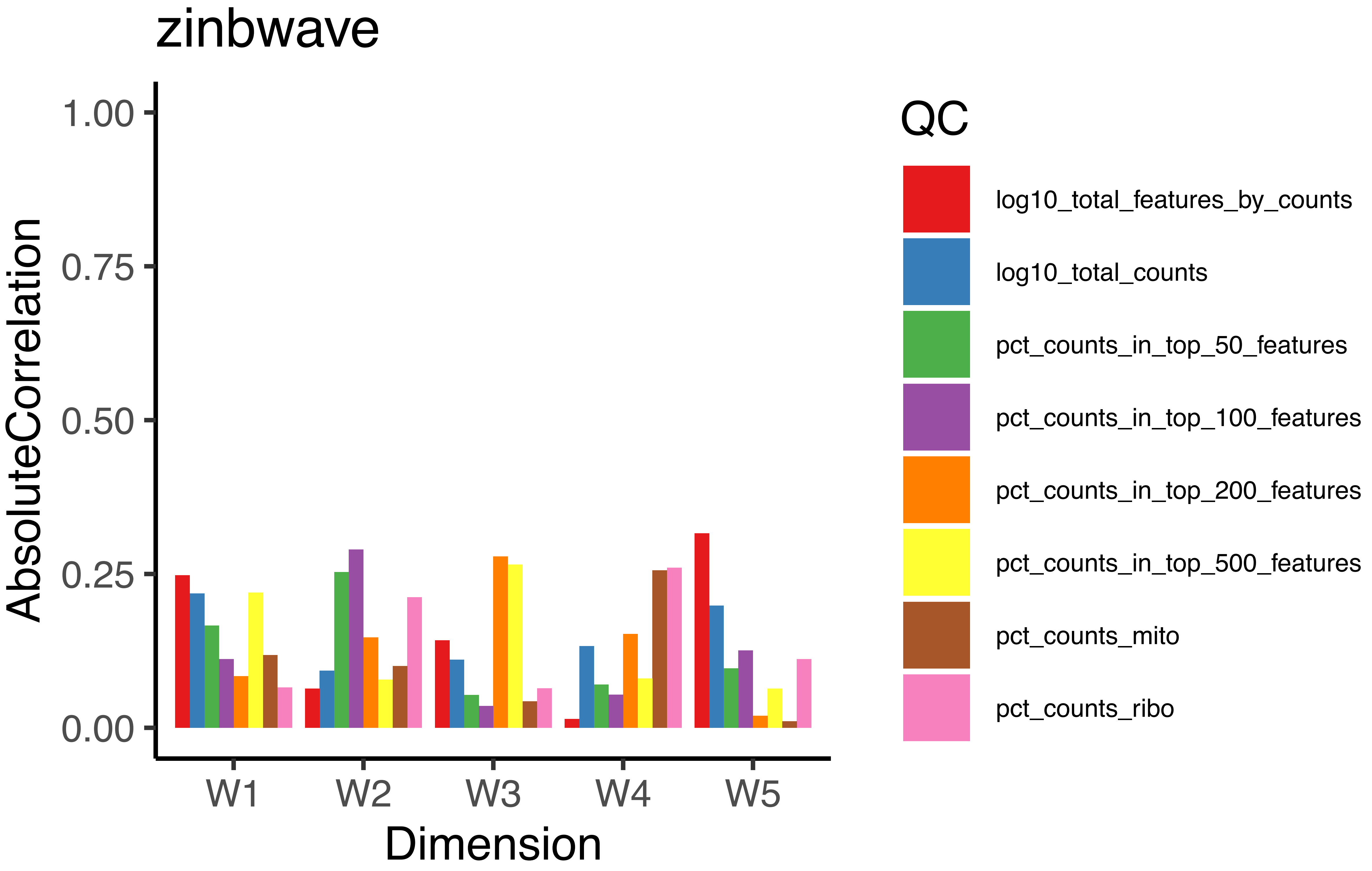
Given n samples and J genes, let Y_{ij} denote the count of gene j (for $j = 1, \dots, J$) for sample i (for $i = 1, \dots, n$).


J genes
log μ
n samples
J genes
logit π
n samples
=
Known sample-level covariates
M
J
X
M
β
n
Observed random variable
Unknown parameter
+
Known gene-level covariates
L
J
γ^T
L
V^T
n
Unknown parameter
Observed random variable
+
Unknown sample-level covariates
K
J
W
K
α
n
Unobserved random variable
Unknown parameter
X intercept acts as a gene-specific scaling factor
V intercept acts as a sample-specific scaling factor

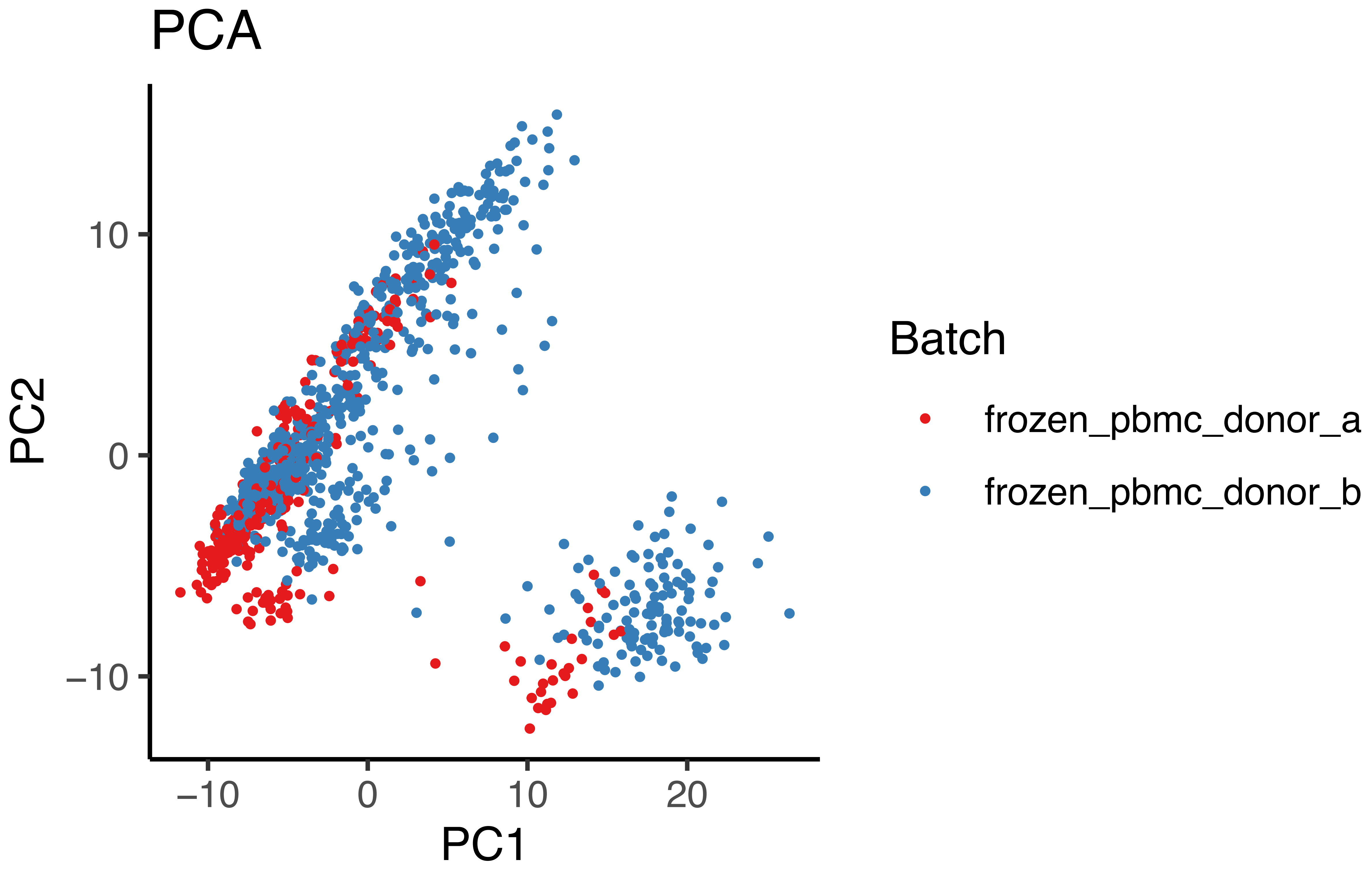
Risso et al. (2018). Nat Comm.

zinbwave Bioconductor Package

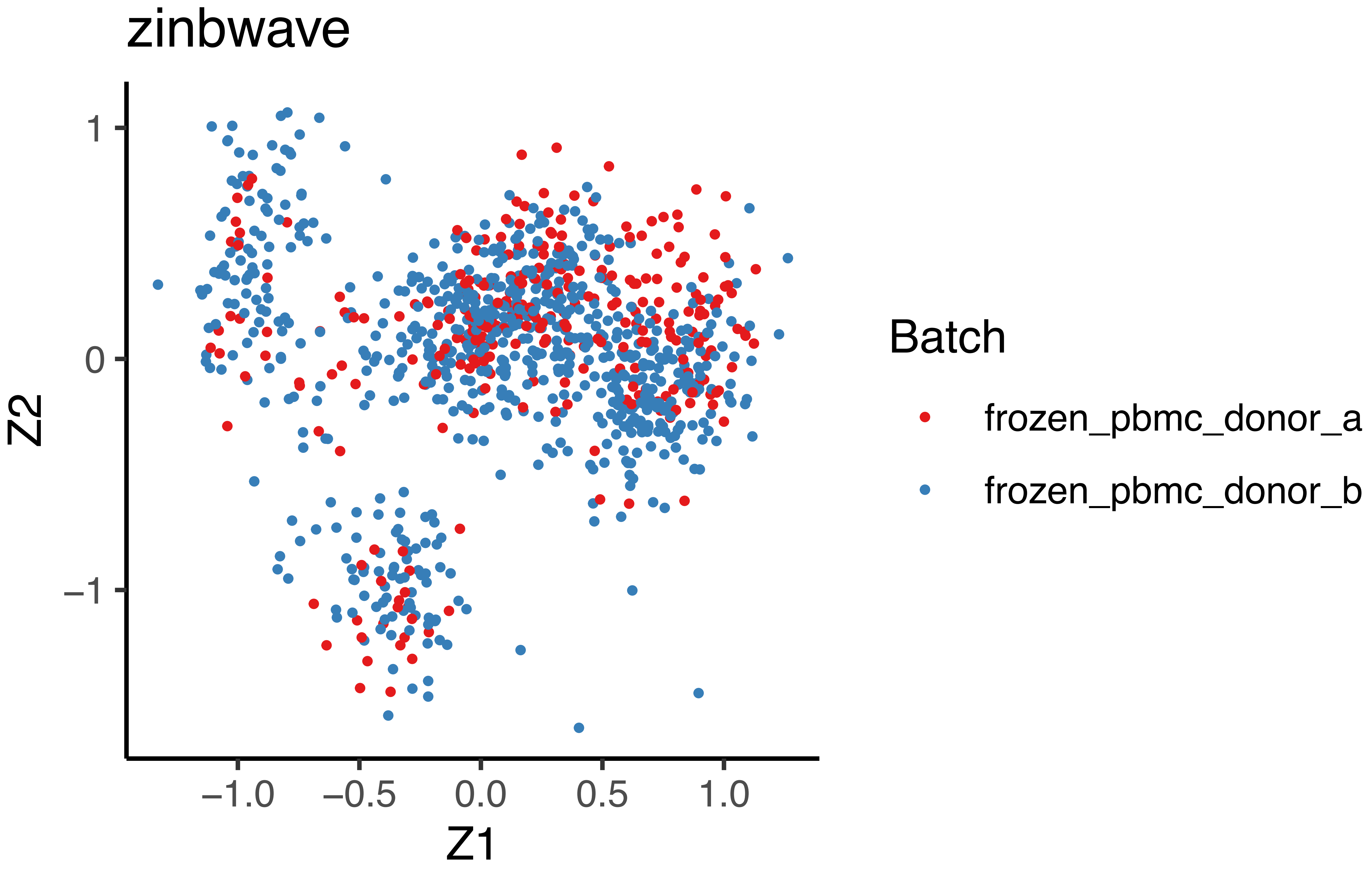
ZINB-WaVE adjusts for quality



Evident batch effects



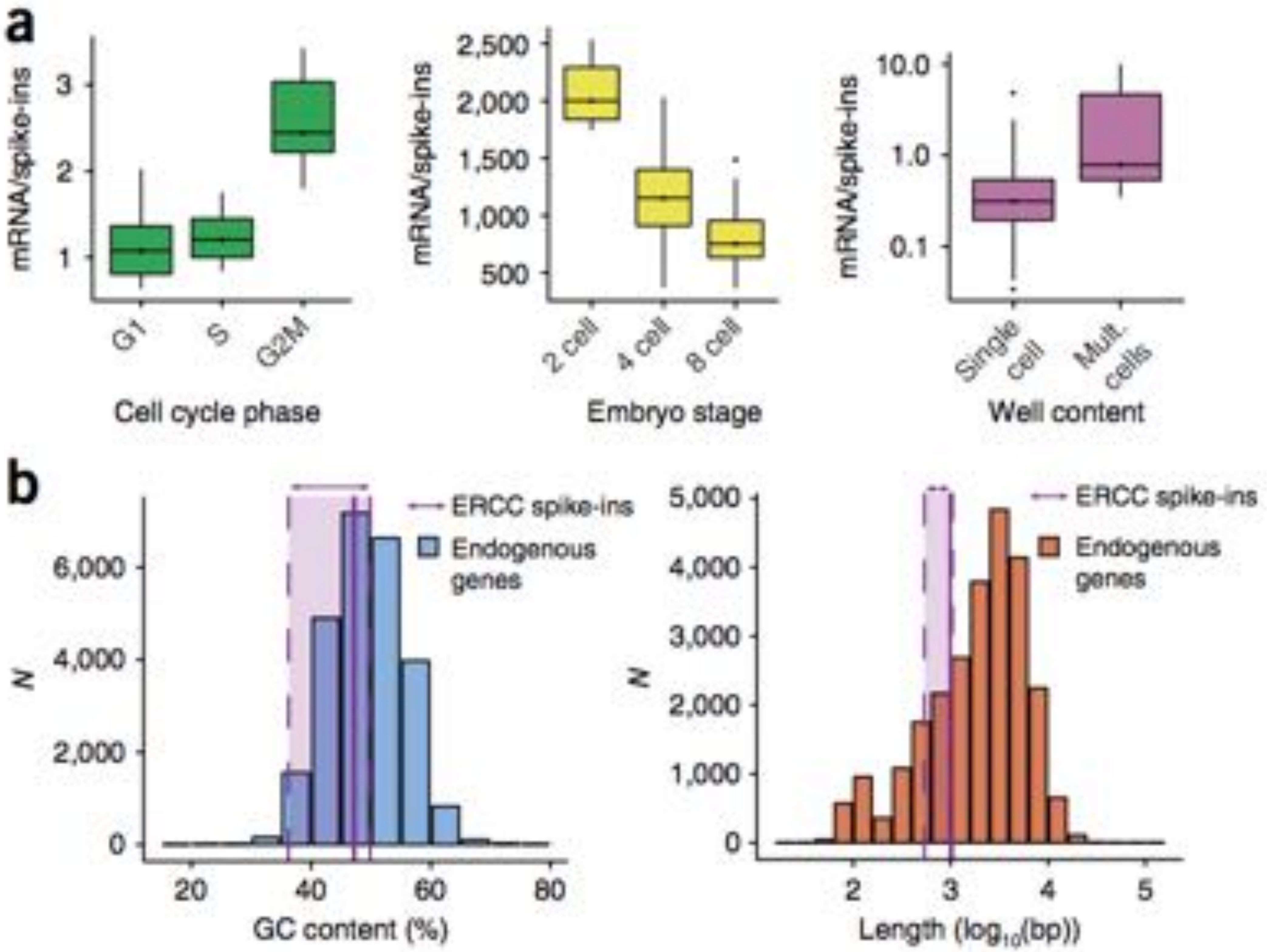
ZINB-WaVE adjusts for batch effects



NORMALIZATION VS. BATCH CORRECTION

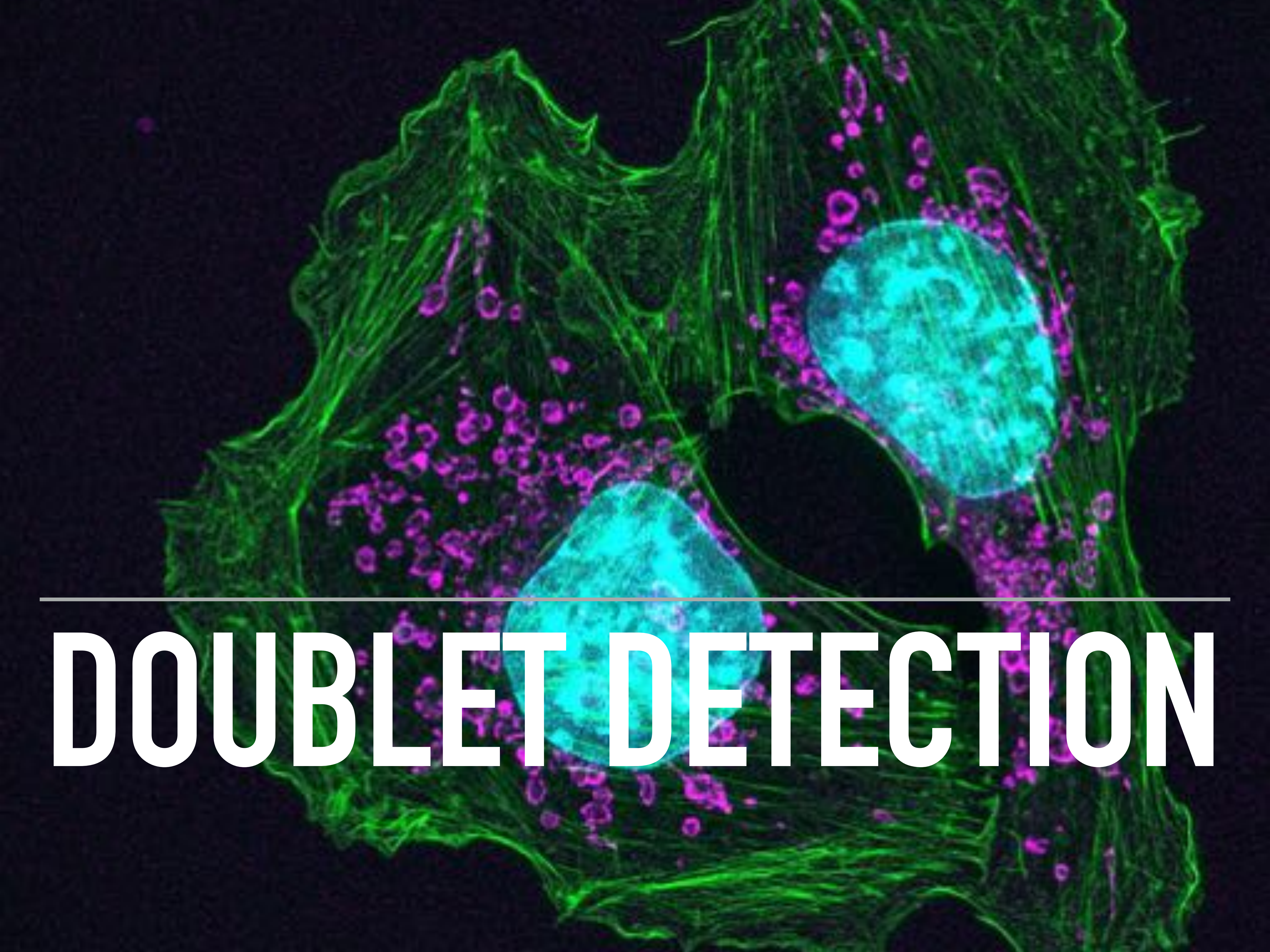
- Most people consider normalization and batch correction as two separate steps.
- However, some methods (e.g., ZINB-WaVE) aim at performing both steps simultaneously.
- For more on batch correction, see tomorrow's lecture!
- When we expect a lot of difference in gene expression among cell types scaling, normalization using spike-ins is attractive. However…

BEHAVIOR OF ERCC SPIKE-INS



Vallejos et al. (2017). Nat Methods.

DOUBLET DETECTION



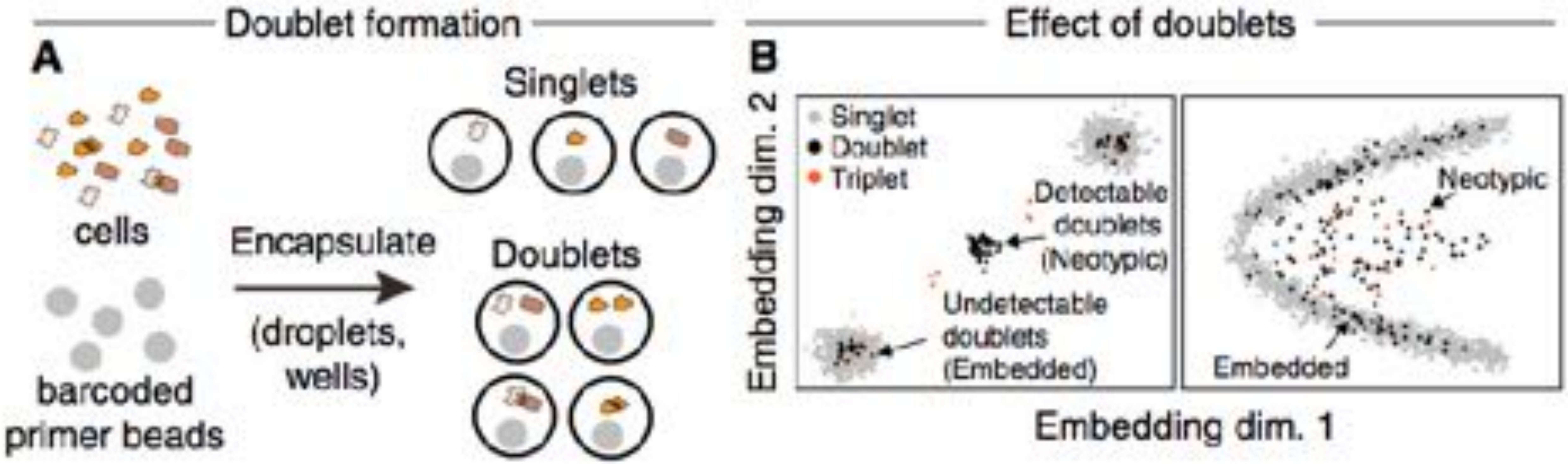
DOUBLET DETECTION

- Doublets occur when a library is made by two cells.
- This can happen if two cells occupy the same microwell (Fluidigm, plates) or if two cells are encapsulated in the same droplet.
- Doublets are problematic for two reasons:
    - Having twice as much RNA they appear as extremely high quality samples
    - They can appear as artifactual transition states between two cell types.

DOUBLET DETECTION

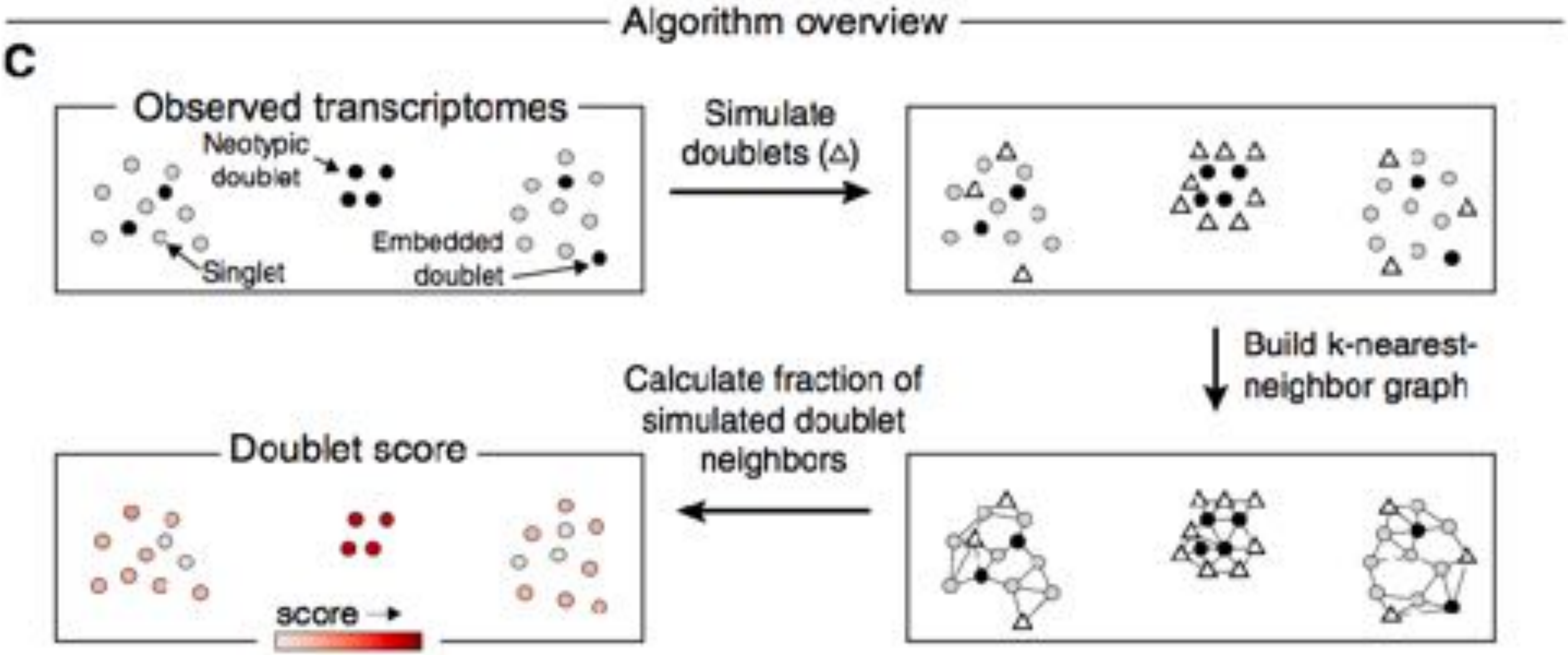
- There are several computational approaches that aim at detecting doublets.
- However, there is no consensus yet on the best approach.
- Published software include **scrublet** and **DoubletFinder**.
- They both employ a similar approach based on simulating *synthetic doublets*.
- As usual, careful experimental design can help, e.g., by mixing male and female individuals we can detect doublets by using sex-specific genes.

DETECTABLE DOUBLETS



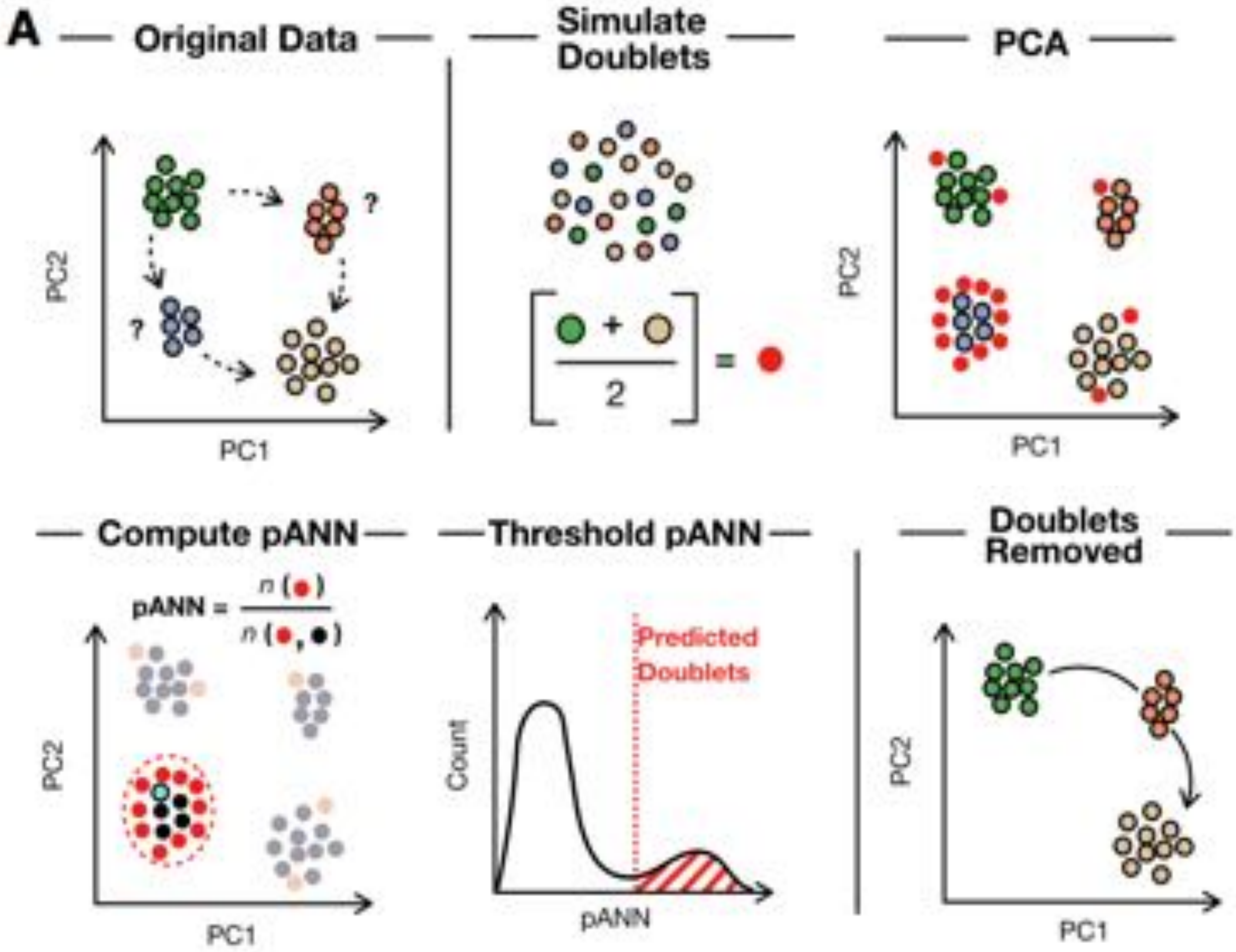
Wolock et al. (2019). Cell Systems.

SCRUBLET



Wolock et al. (2019). Cell Systems.

DOUBLETFINDER

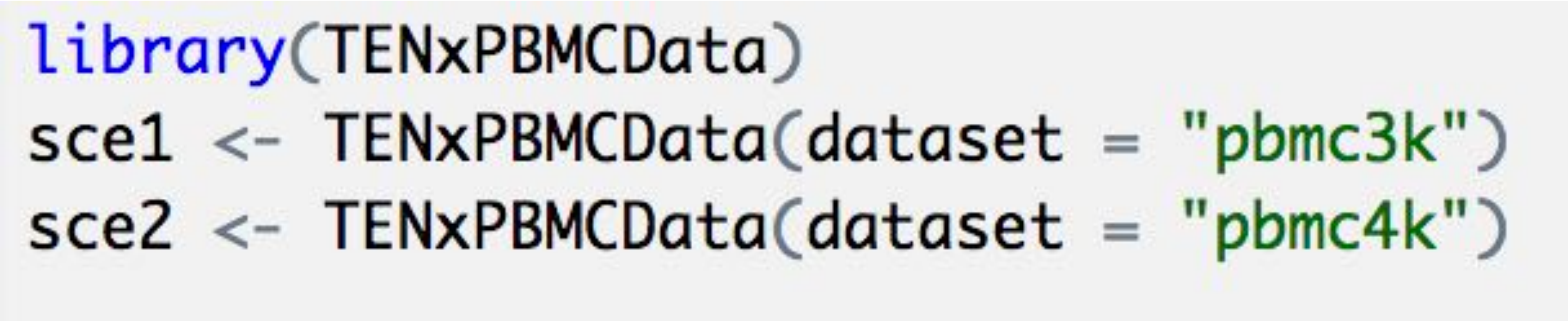


McGinnis et al. (2019). Cell Systems.

DOUBLET DETECTION IN BIOCONDUCTOR

- There are two strategies implemented in the *scran* package.
- One aims at giving a score to each cell similarly to the previous approaches.
- Another strategy is to mark *clusters* as being made of doublets.
- This is more efficiently computationally, but cannot identify doublets that look like transitional states.

FOR THE AFTERNOON LAB



```r
library(TENxPBMCData)
sce1 <- TENxPBMCData(dataset = "pbmc3k")
sce2 <- TENxPBMCData(dataset = "pbmc4k")
```

THANK YOU FOR YOUR ATTENTION!

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE STATISTICHE

Davide Risso

risso.davide@gmail.com

@drisso1893

@drisso