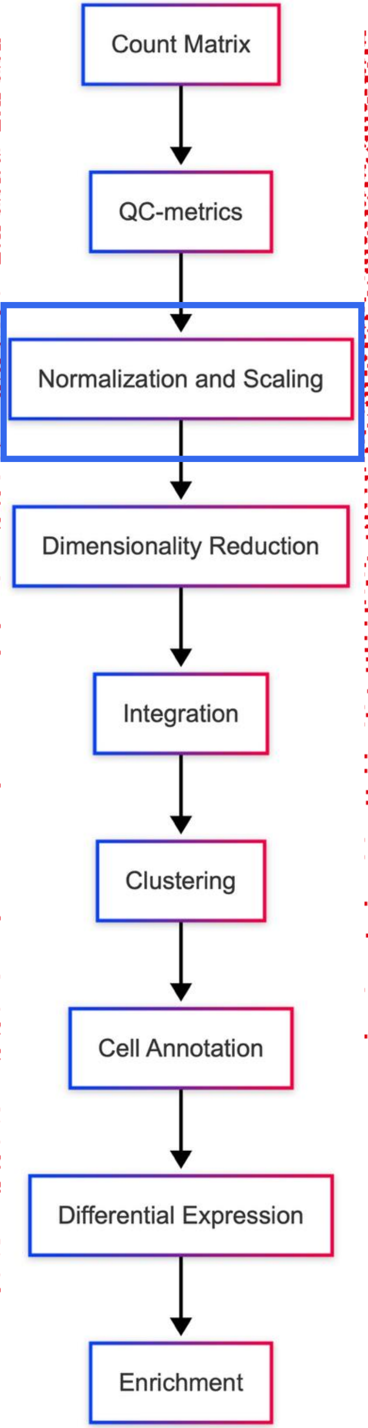
Swiss Institute of
Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Normalization and scaling



Deepak Tanwar

July 02-04, 2025

Adapted from previous year courses

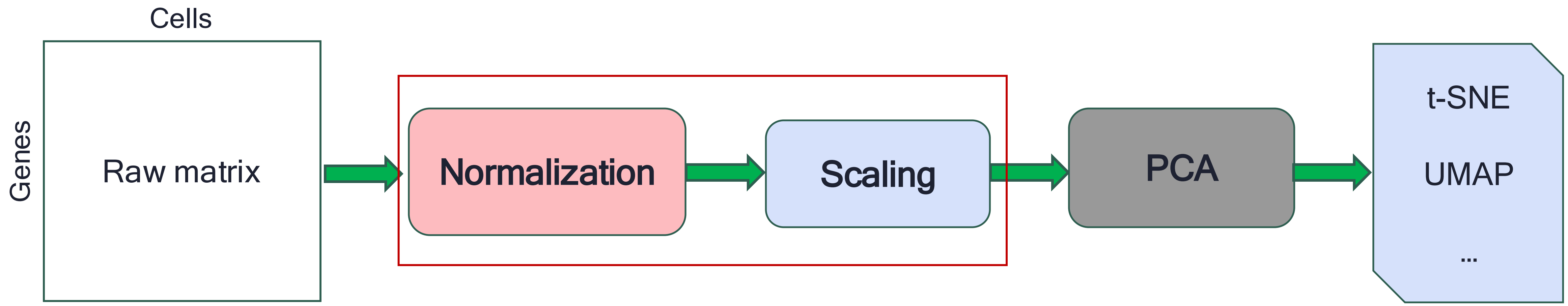
Learning objectives

Understand the importance of Normalization and Scaling

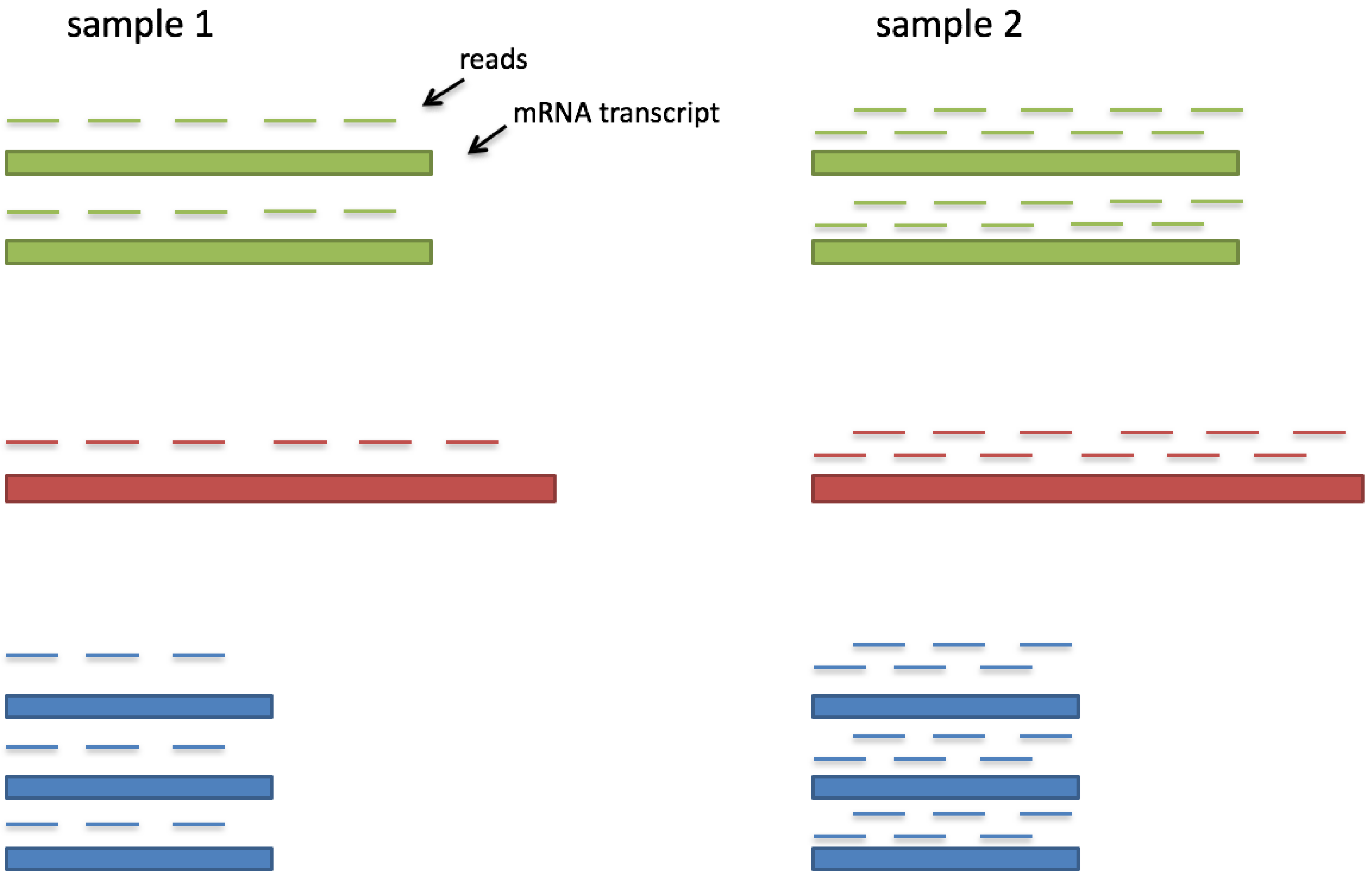
Identify and apply Normalization techniques

Understand Scaling and Transformation

Normalization and scaling

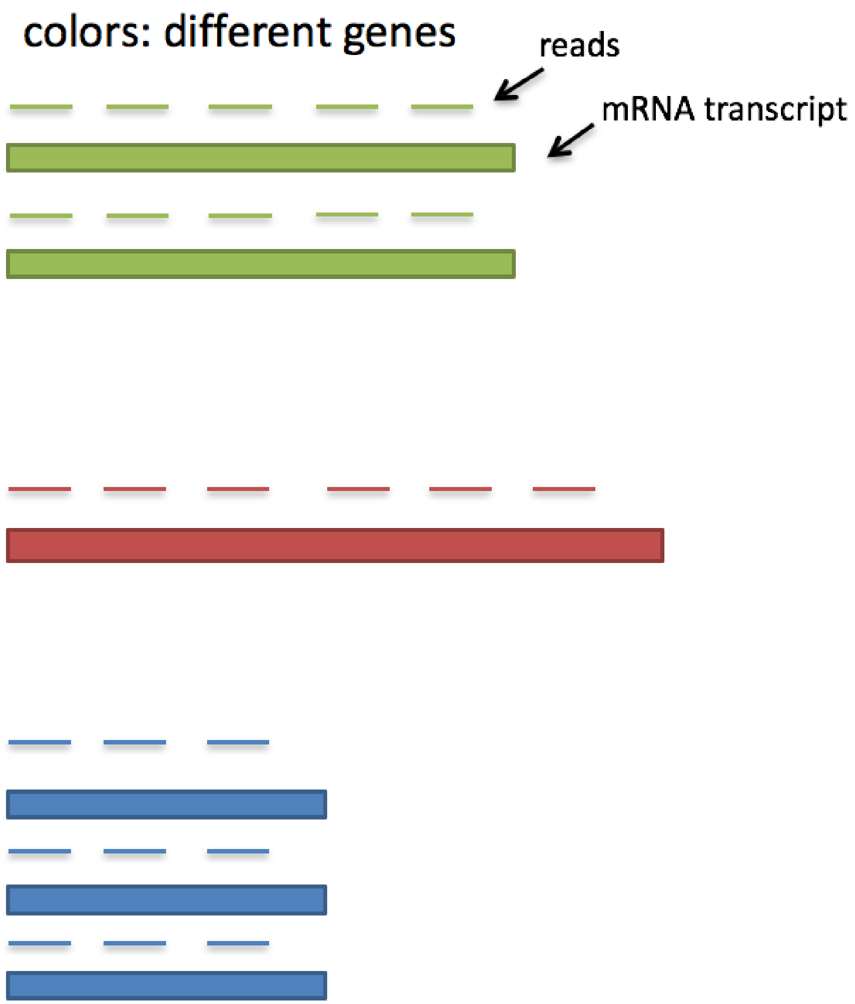


Understanding the differences



MI Love: RNA-seq statistical analysis

Understanding the differences



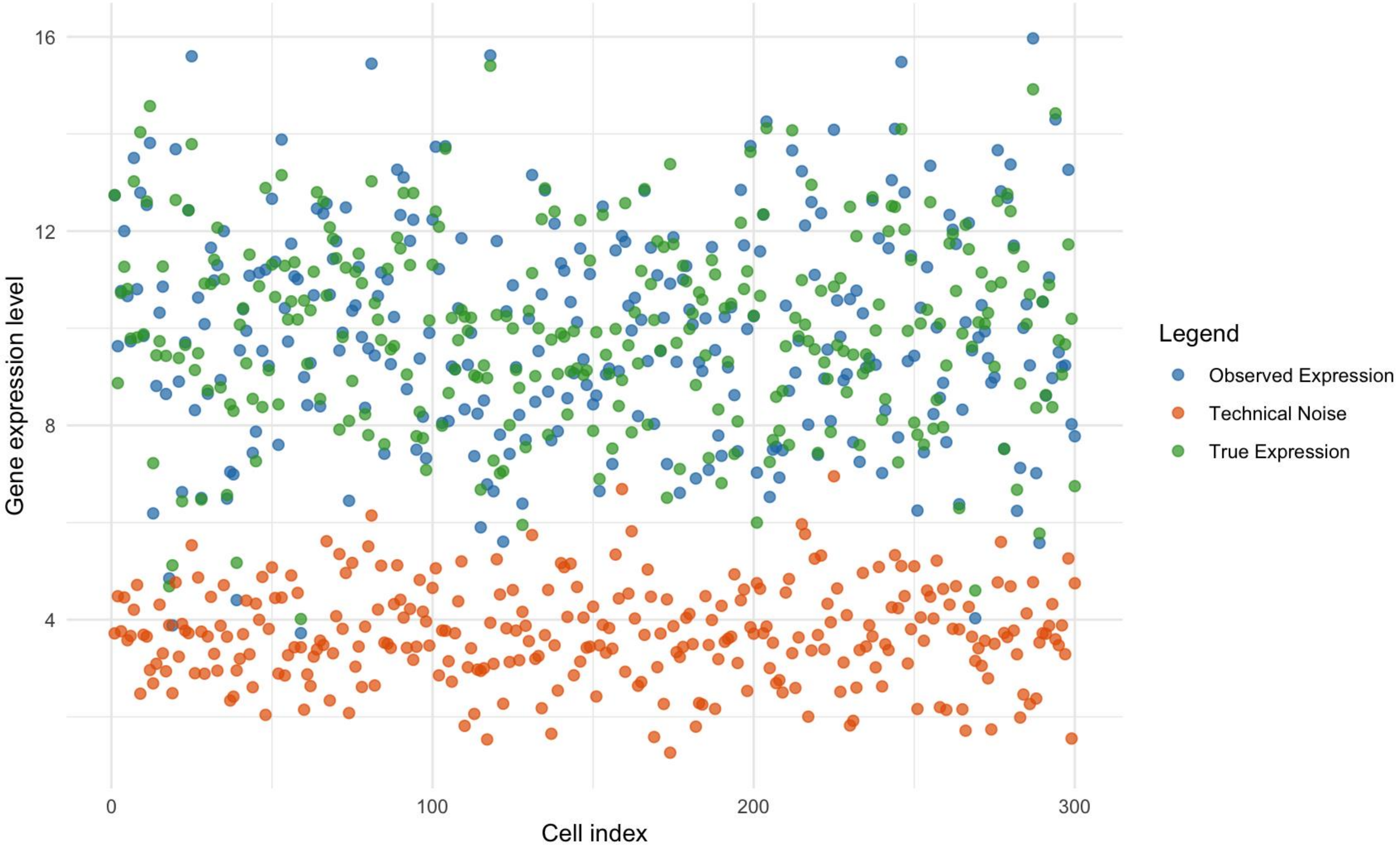
Slide adapted from MI Love: RNA-seq statistical analysis

Goal of normalization

Remove technical noise while preserving biological signal

- Library size

Understanding the differences



Normalization techniques applied in scRNA-seq

UMI (Unique Molecular Identifiers):

- Cells with **extremely high UMI counts** could be **doublets** (two or more cells captured in a single droplet).

- Cells with **very low UMI counts** might be **low-quality or empty droplets**.

Normalization techniques applied in scRNA-seq

UMI (Unique Molecular Identifiers):

- Cells with **extremely high UMI counts** could be **doublets** (two or more cells captured in a single droplet).

- Cells with **very low UMI counts** might be **low-quality or empty droplets**.

Detected genes:

- A healthy cell will express a moderate number of genes.

- Very **low gene count** could indicate a dead cell or an empty droplet.

- High gene count could indicate a doublet.

Normalization techniques applied in scRNA-seq

% Mitochondrial UMI:

1. Mitochondrial genes are usually expressed at **low levels**.
2. **High mitochondrial RNA percentage (>10-20%)** indicates **stressed or dying cells**.

Normalization techniques applied in scRNA-seq

% Mitochondrial UMI:

1. Mitochondrial genes are usually expressed at **low levels**.
2. **High mitochondrial RNA percentage (>10-20%)** indicates **stressed or dying cells**.

% Ribosomal UMI:

1. High ribosomal content may suggest technical artifacts or certain cell types (e.g., rapidly dividing cells).

Normalization techniques applied in scRNA-seq

% Mitochondrial UMI:

1. Mitochondrial genes are usually expressed at **low levels**.
2. **High mitochondrial RNA percentage (>10-20%)** indicates **stressed or dying cells**.

% Ribosomal UMI:

1. High ribosomal content may suggest technical artifacts or certain cell types (e.g., rapidly dividing cells).

% Globin UMI:

1. In blood samples, **high globin content** comes from red blood cells (RBCs).
2. Filtering out these cells is often necessary when focusing on immune or other cell types.

High UMI count + high gene count → **Doublet suspicion**

High UMI count + high gene count → **Doublet suspicion**

High mitochondrial percentage → **Apoptotic or stressed cell**

High UMI count + high gene count → **Doublet suspicion**

High mitochondrial percentage → **Apoptotic or stressed cell**

Low gene count + low UMI → **Low-quality or empty droplet**

High UMI count + high gene count → **Doublet suspicion**

High mitochondrial percentage → **Apoptotic or stressed cell**

Low gene count + low UMI → **Low-quality or empty droplet**

High ribosomal percentage → **Potential technical artifact**

High UMI count + high gene count → **Doublet suspicion**

High mitochondrial percentage → **Apoptotic or stressed cell**

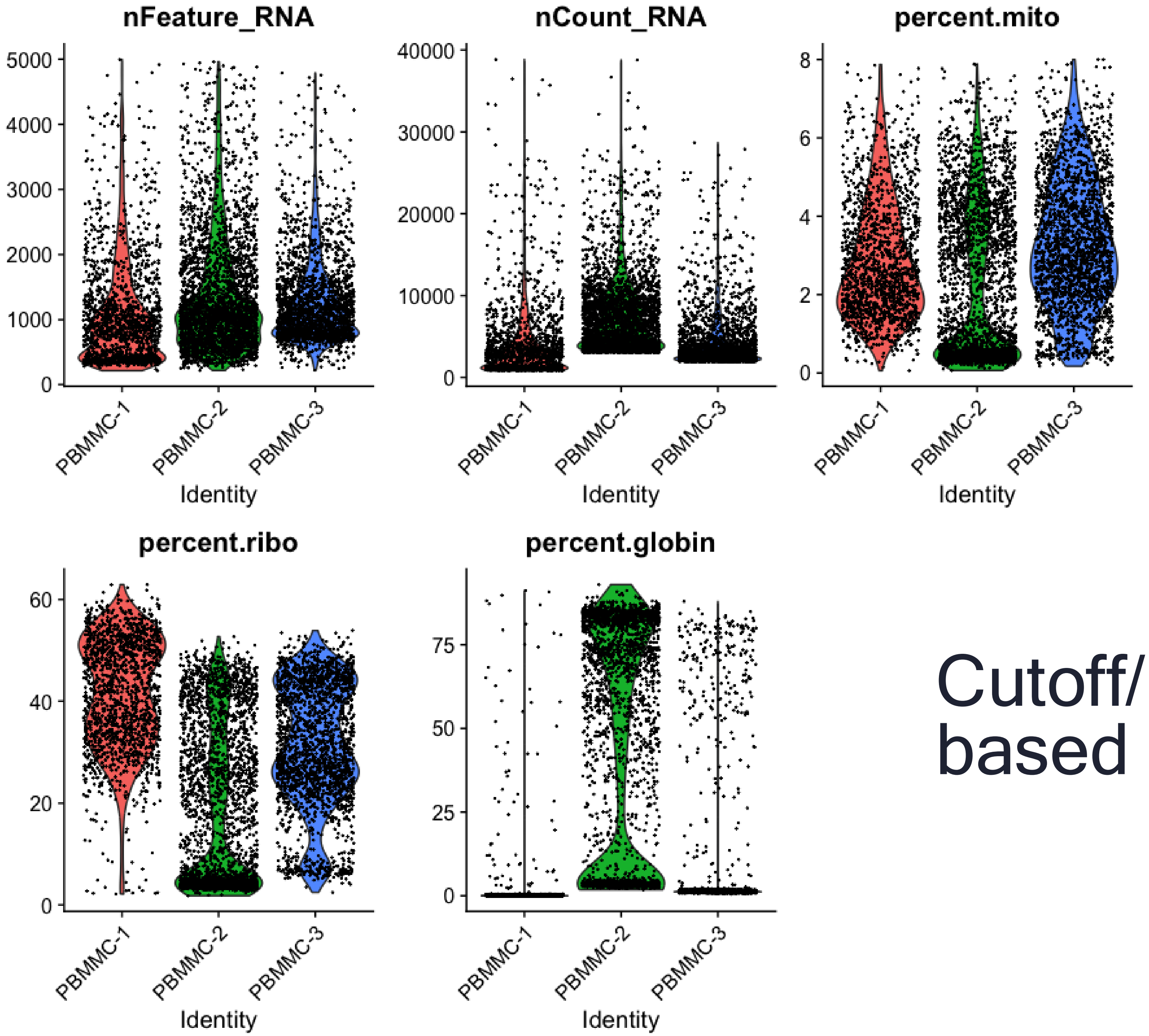
Low gene count + low UMI → **Low-quality or empty droplet**

High ribosomal percentage → **Potential technical artifact**

```
seurat_obj <- subset(seurat_obj,
                     subset = nFeature_RNA > 200 &
                              nFeature_RNA < 5000 &
                              nCount_RNA < 20000 &
                              percent.mt < 10 &
                              percent.ribo < 40 &
                              percent.globin < 5)
```

Summary with reasons/ references

Metric	Common Range	Reason/Reference
nFeature_RNA	200-5000	Seurat tutorials, debris filtering, doublet removal
nCount_RNA	<20,000	Heuristic; high UMI counts may be doublets
percent.mt	<10%	Damaged/apoptotic cells, Ilicic et al., 2016
percent.ribo	<40% (opt.)	Low-complexity transcripts, low-quality filtering
percent.globin	<5%-10%	Blood contamination, Bhattacherjee et al., 2019



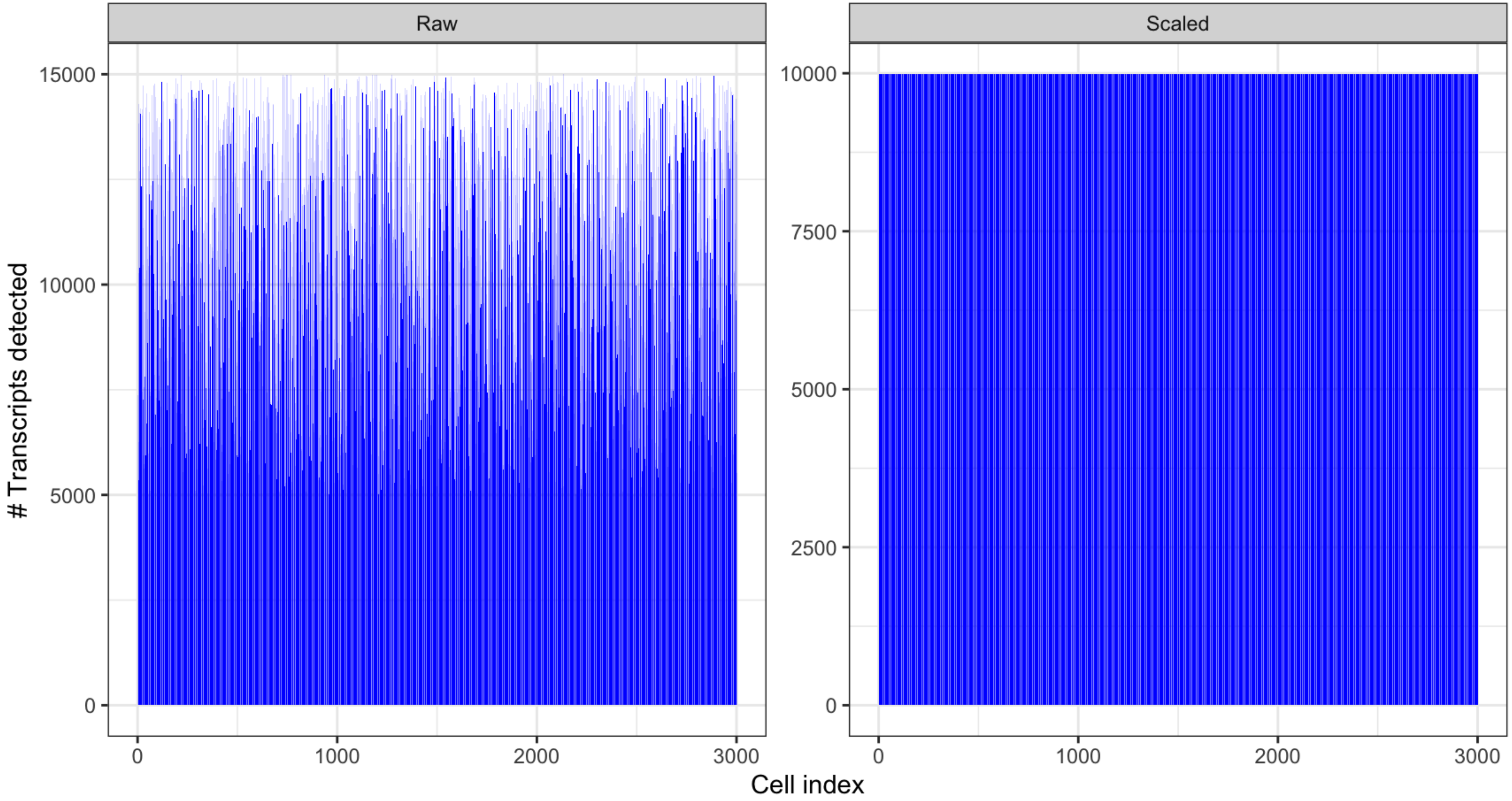
Cutoff/ range could vary based on the dataset

Scaling

Multiply each UMI count by a cell specific factor to get all cells to have the same UMI counts

Different cells have different amounts of mRNA; this could be due to differences between cell types or variation within the same cell type depending on how well the chemistry worked in one drop versus another.

Scaling: standardize range, mean and variance



Transformation

- Simple transformations
- Pearson residuals

Transformation : Simple transformations

Raw data

	Cell Type A	Cell Type B	Δ
Gene 1	1	2	1
Gene 2	100	200	100
Gene 3	5	25	20
Gene 4	400	800	400
Gene 5	10	60	50

Transformation : Simple transformations

	Raw data			Log$_2$ transform		
	Cell Type A	**Cell Type B**	**Δ**	**Cell Type A**	**Cell Type B**	**Δ**
Gene 1	1	2	1	0.00	1.00	1.00
Gene 2	100	200	100	6.64	7.64	1.00
Gene 3	5	25	20	2.32	4.64	2.32
Gene 4	400	800	400	8.64	9.64	1.00
Gene 5	10	60	50	3.32	5.91	2.58

Transformation : Simple transformations

	Raw data			Log$_2$ transform			Square root transform		
	Cell Type A	**Cell Type B**	**Δ**	**Cell Type A**	**Cell Type B**	**Δ**	**Cell Type A**	**Cell Type B**	**Δ**
Gene 1	1	2	1	0.00	1.00	1.00	1.00	1.41	0.41
Gene 2	100	200	100	6.64	7.64	1.00	10.00	14.14	4.14
Gene 3	5	25	20	2.32	4.64	2.32	2.24	5.00	2.76
Gene 4	400	800	400	8.64	9.64	1.00	20.00	28.28	8.28
Gene 5	10	60	50	3.32	5.91	2.58	3.16	7.75	4.58

Transformation : Simple transformations

-Log transformation

-Square root transformation

$$y_{ij} = f(x_{ij})$$

i: cell

j: gene

Quiz

Which of the simple transformation methods transform each measurements individually?

1. Log
2. Square root
3. None
4. Both

Transformation : Pearson residuals

$$y_{ij} = w_j \cdot x_{ij}$$

y_{ij} is the transformed expression value for gene j in cell i.

x_{ij} is the original expression value (e.g., UMI count).

w_j is a weight that adjusts for gene-specific variance.

$$w_j = \frac{1}{\sqrt{\mathrm{mean}(x_j)}}$$

$\mathrm{mean}(x_j)$ is the average expression of gene j across all cells.

Taking the **inverse square root** of the mean adjusts for differences in gene expression levels.

Transformation : Pearson residuals

Instead of transforming each measurements individually, Pearson residuals apply a weight to all measurements of a gene

Transformation : Pearson residuals

Raw data

Cell A: 75%, Cell B: 25%

	Cell Type A	Cell Type B	Δ
Gene 1	1.00	2.00	1.00
Gene 2	100.00	200.00	100.00
Gene 3	5.00	25.00	20.00
Gene 4	400.00	800.00	400.00
Gene 5	10.00	60.00	50.00

Log transform

Cell A: 75%, Cell B: 25%

	Cell Type A	Cell Type B	Δ
Gene 1	0.00	1.00	1.00
Gene 2	6.64	7.64	1.00
Gene 3	2.32	4.64	2.32
Gene 4	8.64	9.64	1.00
Gene 5	3.32	5.91	2.58

Square root transform

Cell A: 75%, Cell B: 25%

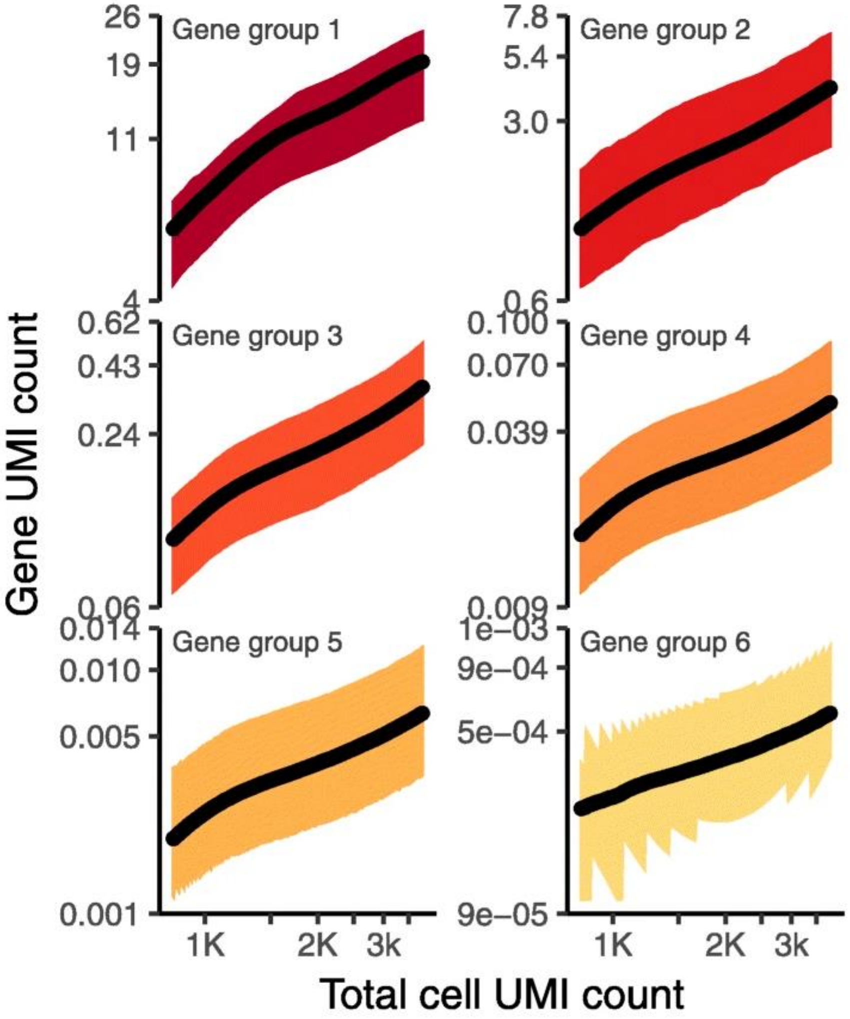
	Cell Type A	Cell Type B	Δ
Gene 1	1.00	1.41	0.41
Gene 2	10.00	14.14	4.14
Gene 3	2.24	5.00	2.76
Gene 4	20.00	28.28	8.28
Gene 5	3.16	7.75	4.58

Pearson Residuals

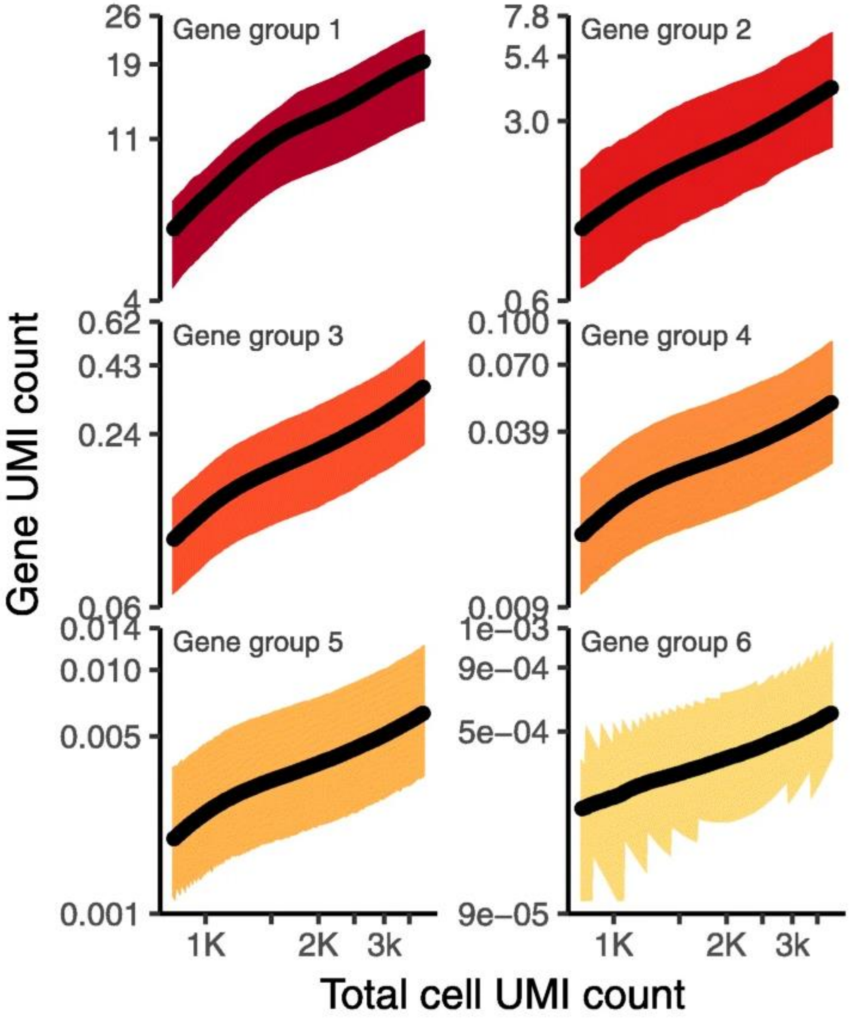
Cell A: 75%, Cell B: 25%

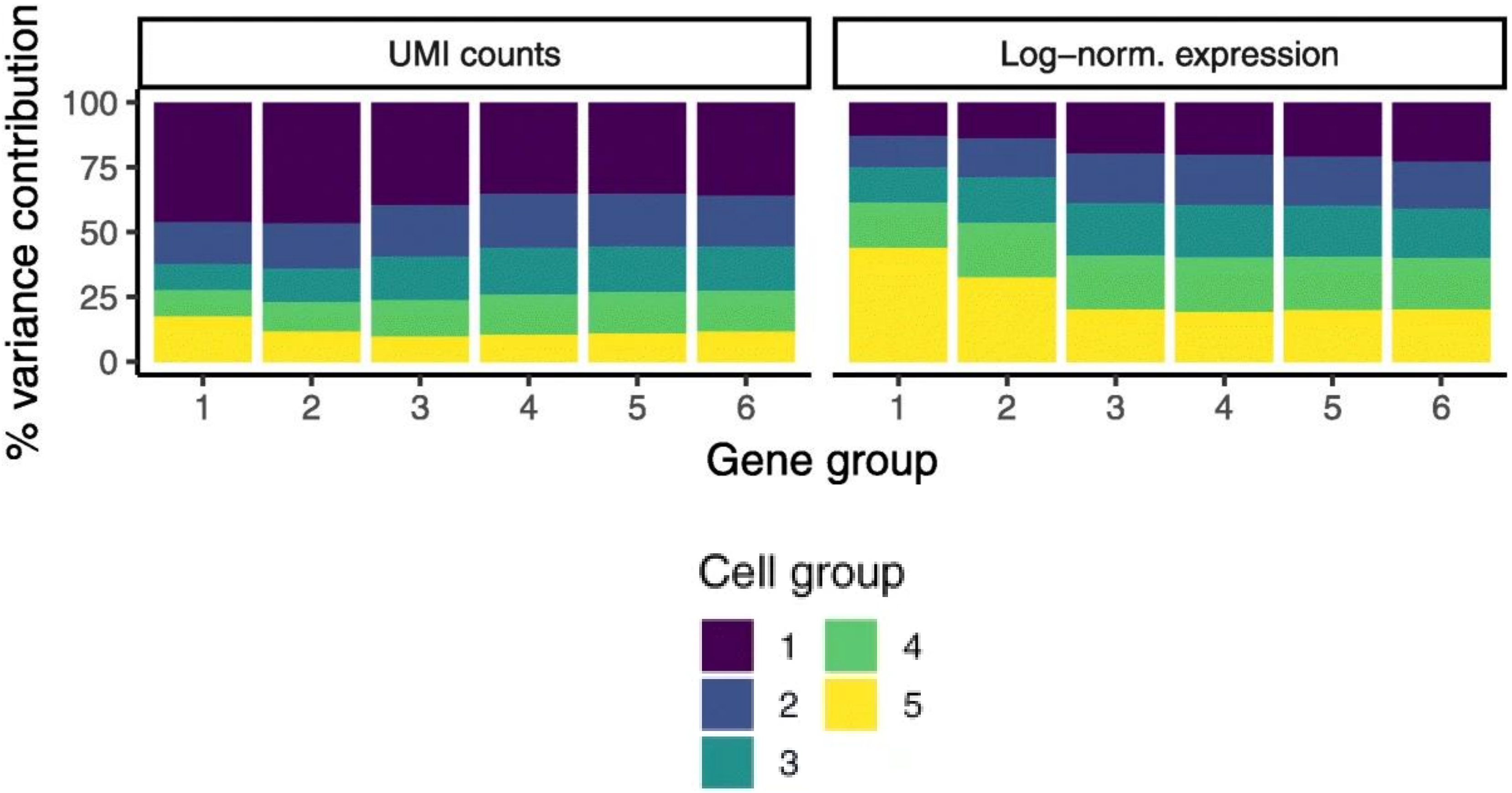
	Cell Type A	Cell Type B	Δ
Gene 1	-0.83	1.44	2.28
Gene 2	-8.33	14.43	22.77
Gene 3	-3.69	6.39	10.08
Gene 4	-16.67	28.87	45.53
Gene 5	-5.87	10.16	16.03

Gene counts are heavily influenced by sequencing depth

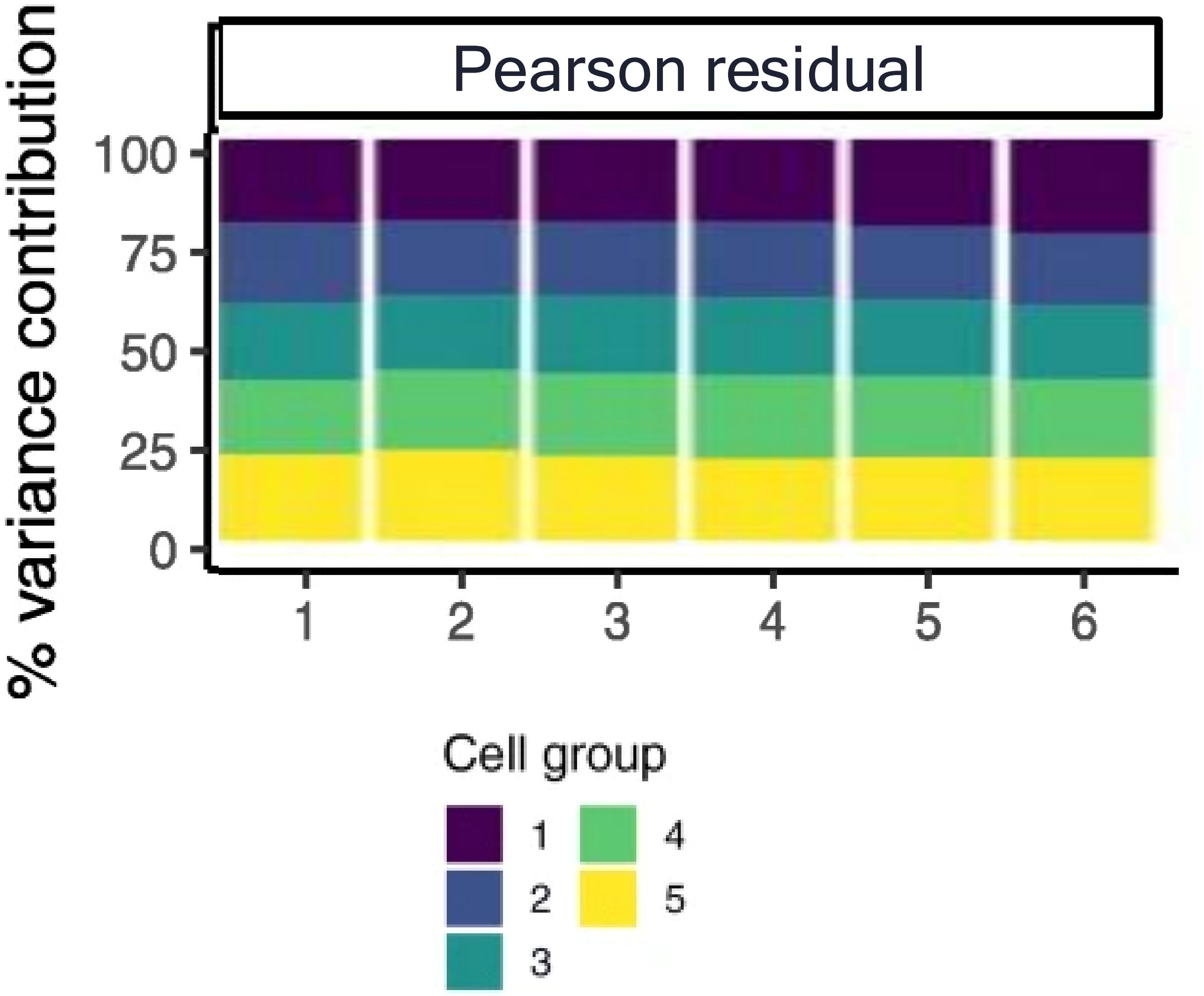


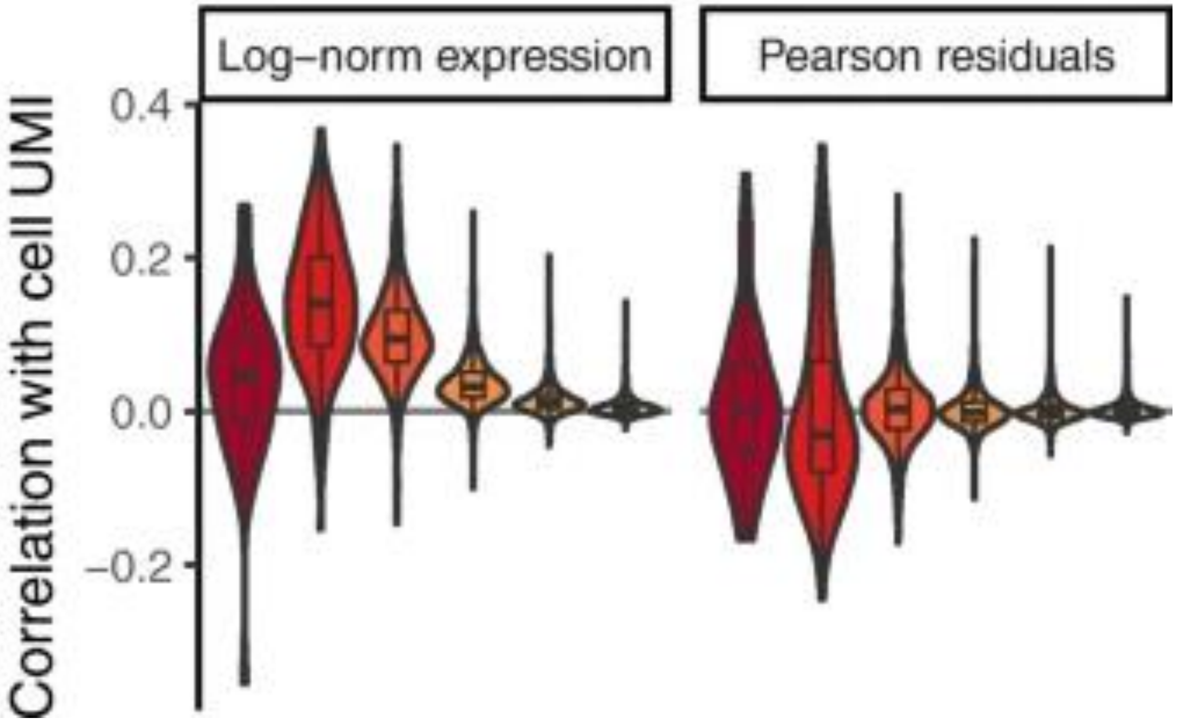
Gene counts are heavily influenced by sequencing depth





sctransform





Summary

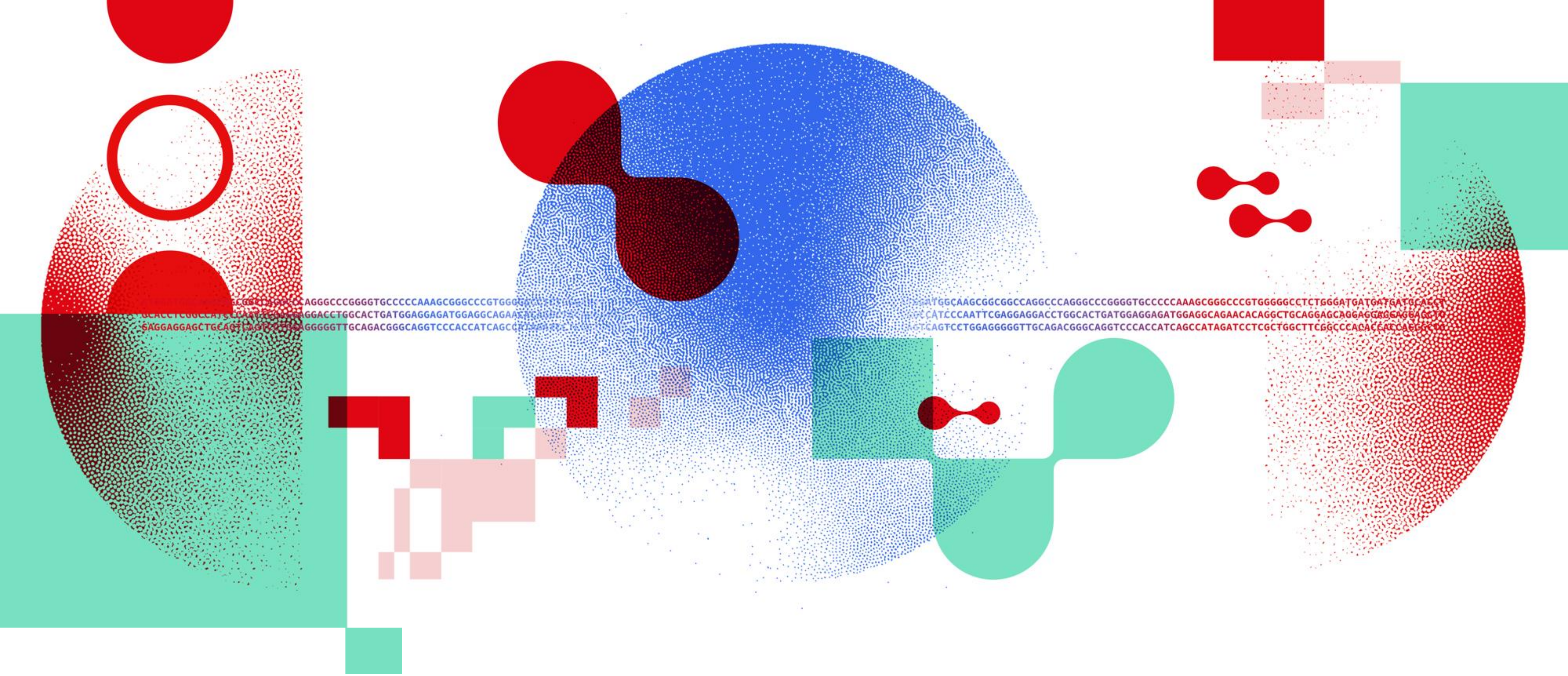
Normalization: Adjust UMI counts, mitochondrial, ribosomal, globin RNA percentages

Goal: Remove technical noise, preserve biological signals

Scaling: Standardize range, mean, variance

Transformations: Log, square root, Pearson residuals

Outcome: Reliable, meaningful scRNA-seq data analysis



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss