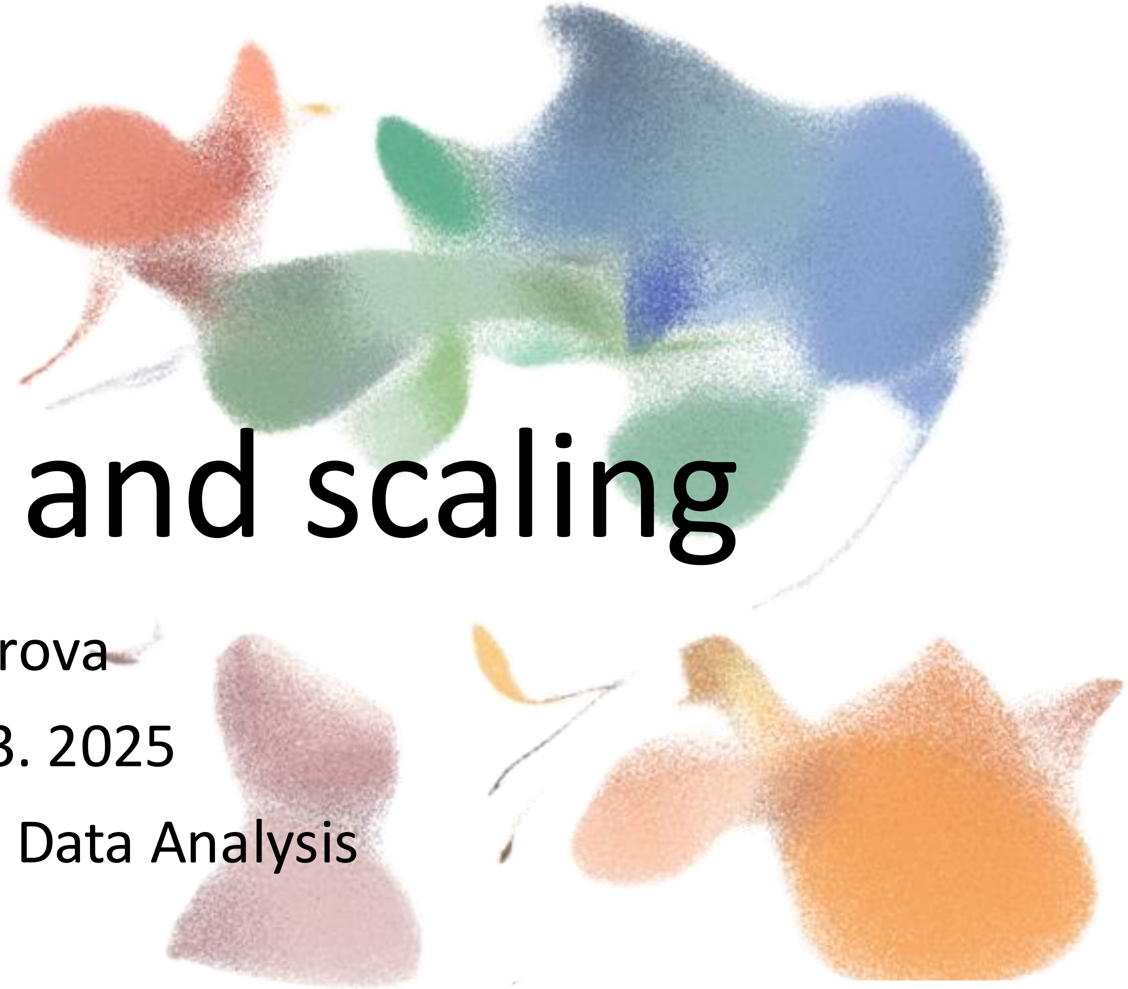


Normalization and scaling

Lucie Pfeiferova

December 1.-3. 2025

Course on scRNA-seq Data Analysis

Motivation and Learning objectives

- Normalization reduces technical differences so that differences between cells are not technical but biological, allowing meaningful comparison of expression profiles between cells.

- Distinguish Normalization, Transformation, and Scaling
- Identify and apply Normalization techniques

Source of differences

- Biological
  - Cell subtype differences - size and transcriptional activity, variation in gene expression
- Technical: scRNA data is inherently noisy
  - Low mRNA content per cell
  - Cell-to-cell differences in mRNA capture efficiency
  - Variable sequencing depth
  - PCR amplification efficiency

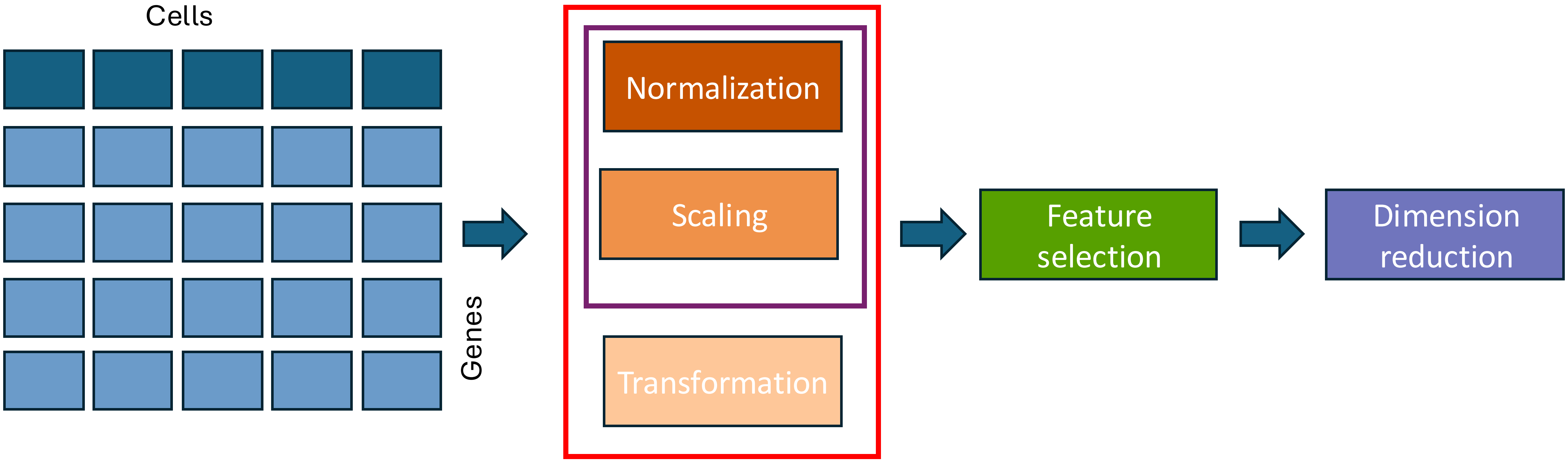
Raw counts problematic

- 1. Library size bias → deeper sequenced cells dominate
- 2. Zero-inflation → most genes = 0
- 3. High dynamic range

Normalization, transformation, scaling

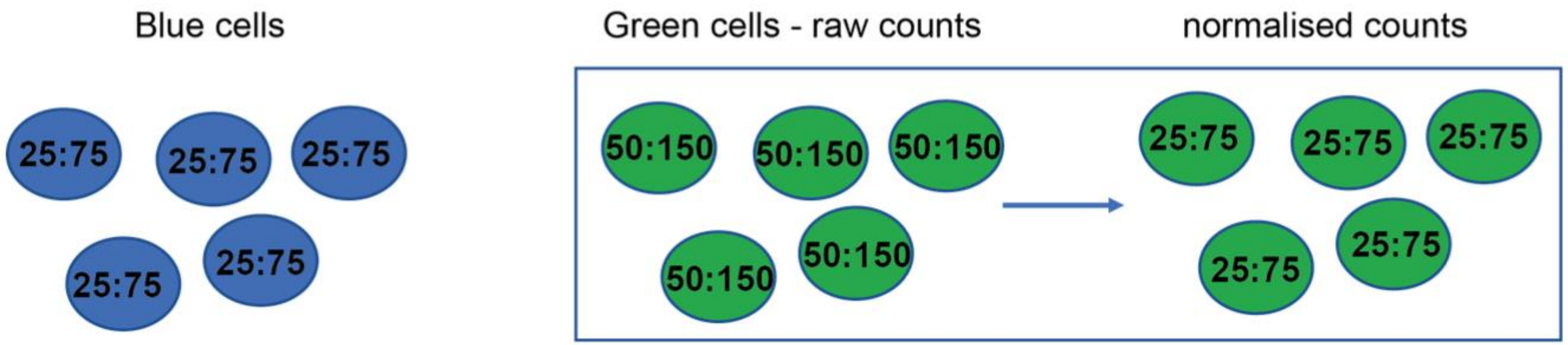
- **Normalization = correct for sequencing depth**
  → “Make cells comparable.”
- **Scaling = equalize gene variance**
  → “Make genes comparable.”
- **Transformation = stabilize distribution**
  → “Make the numbers behave statistically.”

Normalization process in scRNA-seq



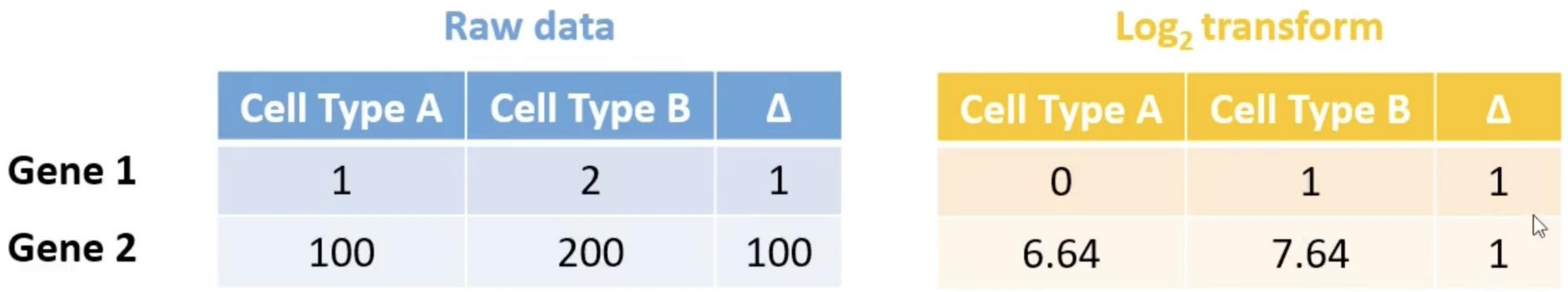
Normalization process in scRNA-seq

- 1. Remove sequencing-depth (library-size) differences so that cells are comparable.
- Scale to common factor



Normalization process in scRNA-seq

- 1. Remove sequencing-depth (library-size) differences so that cells are comparable.
- 2. Scale to common factor
- 3. Transform values (with log)



Raw data

	Cell Type A	Cell Type B	Δ
Gene 1	1	2	1
Gene 2	100	200	100

Log$_2$ transform

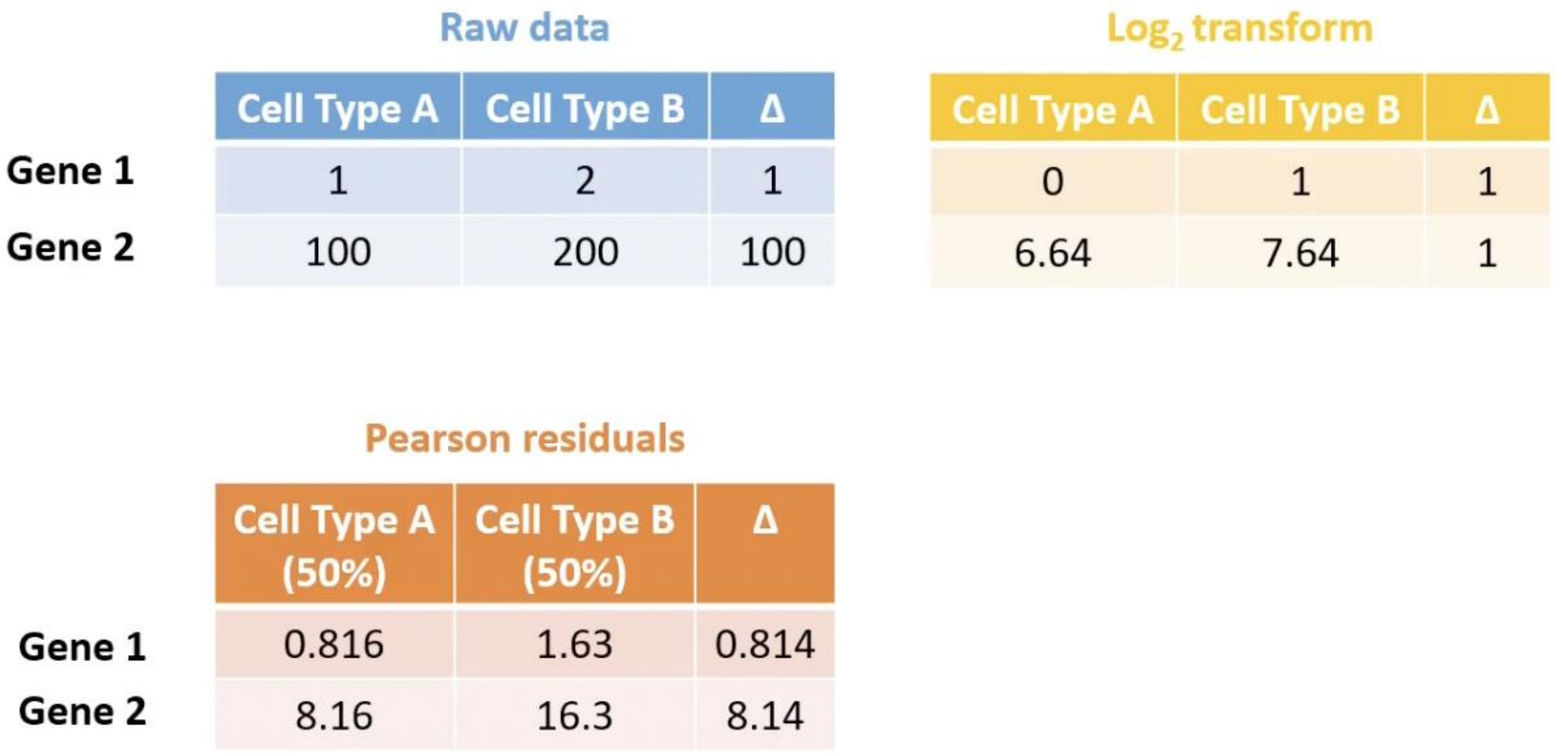
Cell Type A	Cell Type B	Δ
0	1	1
6.64	7.64	1

SCTransform

- **1. Model fitting:**
- Fit a **negative binomial generalized linear model (NB-GLM)** for each gene.
- Total UMI count per cell is used as a **covariate** (proxy for sequencing depth).
- **2. Parameter regularization:**
- Intercept, slope, and dispersion parameters are **regularized** based on their relationship with gene mean.
- Prevents **overfitting** and stabilizes parameter estimates, especially for lowly expressed genes.
- **3. Variance-stabilizing transformation:**
- Use the regularized model to compute **Pearson residuals**.
- These residuals represent normalized, variance-stabilized expression values.

SCTransform

- **1. Model fitting:**
- Fit a **negative binomia**
- Total UMI count per ce
- **2. Parameter regulariz**
- Intercept, slope, and d relationship with gene
- Prevents **overfitting** ar expressed genes.



Raw data

	Cell Type A	Cell Type B	Δ
Gene 1	1	2	1
Gene 2	100	200	100

Log$_2$ transform

Cell Type A	Cell Type B	Δ
0	1	1
6.64	7.64	1

Pearson residuals

	Cell Type A (50%)	Cell Type B (50%)	Δ
Gene 1	0.816	1.63	0.814
Gene 2	8.16	16.3	8.14

- **3. Variance-stabilizing transformation:**
- Use the regularized model to compute **Pearson residuals**.
- These residuals represent normalized, variance-stabilized expression values.

Summary

Step	What it does	Why it's needed	Typical methods
Normalization	Adjusts **library size / sequencing depth** differences between cells	Cells have different total RNA counts; normalization puts them on the same scale	CPM/TPM, Size-factor methods (e.g., scran), SCTransform’s model-based approach
Scaling	Adjusts **per-gene variance** across genes or cells	Prevents highly expressed genes from dominating analyses such as PCA	Z-score scaling, regression of covariates
Transformation	Changes the **distribution**of gene expression values	RNA counts are skewed; transformations stabilize variance and make data more normally distributed	log1p, variance stabilizing transformation (VST), SCTransform

Bonus – normalization by deconvolution

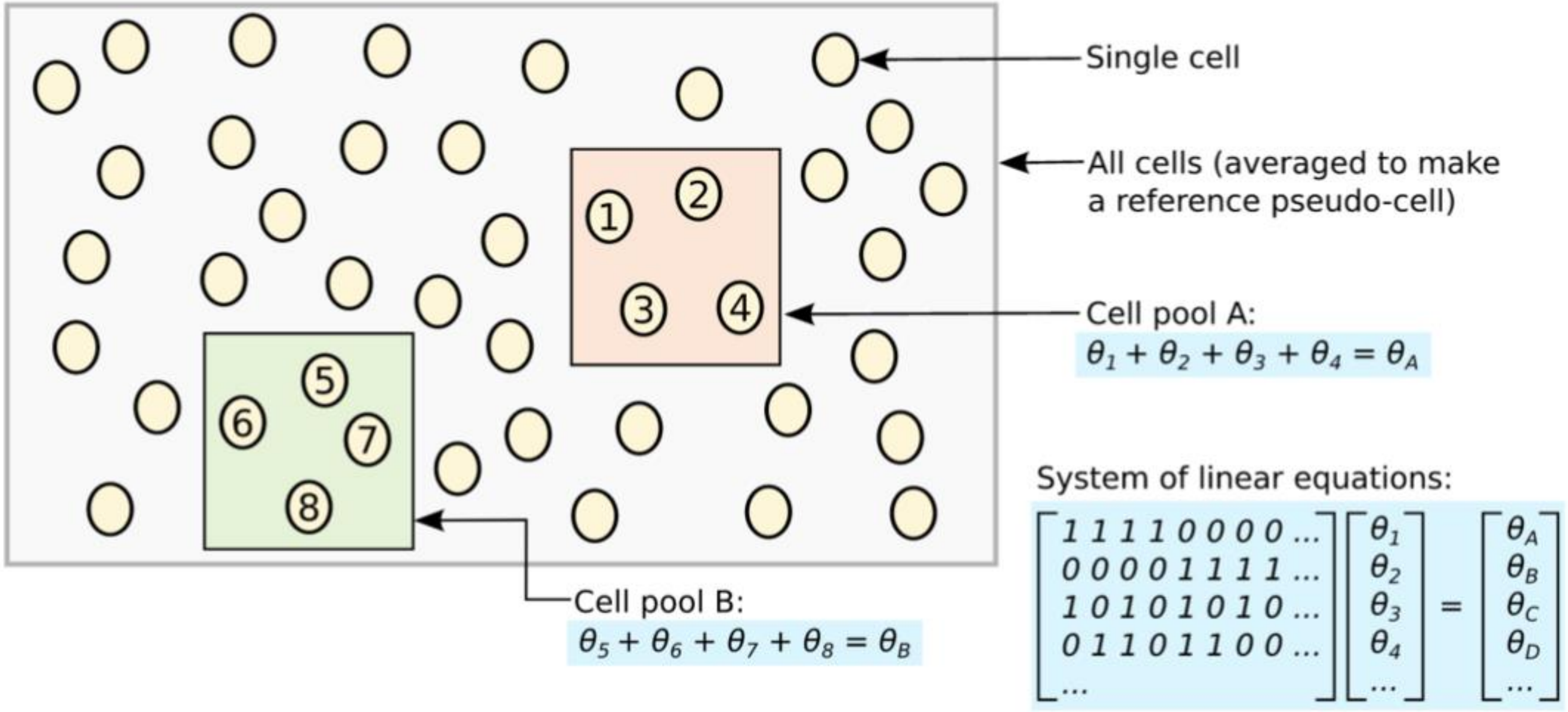
- **Why deconvolution?**
- Single cells have **low UMI counts**, making size factor estimation unstable.
- Many genes show **zero expression**, causing unreliable per-cell normalization.
- Deconvolution improves accuracy by **borrowing information across cells**

Bonus – normalization by deconvolution

- **1. Pooling:**
- Group cells into pseudo-bulk pools to create higher-coverage “pseudo-samples.”
- This reduces sparsity and increases signal for size factor estimation.
- **2. Compute pool-based size factors:**
- Estimate size factors on pooled data, where zeros are less problematic.
- **3. Deconvolution:**
- Mathematically solve for **individual cell size factors** from overlapping pools.
- Ensures cell-specific normalization factors consistent across all pools.
- **4. Normalize counts:**
- Divide each cell’s gene counts by its deconvolved size factor.

Bonus – normalization by deconvolution

- **1. Pooling:**
- Group cells into p
- This reduces spar
- **2. Compute pool**
- Estimate size fact
- **3. Deconvolution**
- Mathematically s
- Ensures cell-spec
- **4. Normalize cou**
- Divide each cell’s gene counts by its deconvolved size factor.



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:
$\theta_5 + \theta_6 + \theta_7 + \theta_8 = \theta_B$
1
2
3
4
5
6
7
8
System of linear equations:
$\begin{bmatrix} 1\,1\,1\,1\,0\,0\,0\,0 \dots \\ 0\,0\,0\,0\,1\,1\,1\,1 \dots \\ 1\,0\,1\,0\,1\,0\,1\,0 \dots \\ 0\,1\,1\,0\,1\,1\,0\,0 \dots \\ \dots \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \dots \end{bmatrix} = \begin{bmatrix} \theta_A \\ \theta_B \\ \theta_C \\ \theta_D \\ \dots \end{bmatrix}$

- https://sib-swiss.github.io/single-cell-r-training/
- Amezquita, R. A., Lun, A. T. L., Hicks, S. C., Marini, F., et al. (2020). Orchestrating single-cell analysis with Bioconductor. Nature Methods, 17(2), 137–145.
- https://bioconductor.org/books/3.14/OSCA.basic/
- https://bioconductor.org/books/release/OSCA/
- Hafemeister C, Satija R (2019). “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression.” Genome Biology, 20, 296. doi:10.1186/s13059-019-1874-1, https://doi.org/10.1186/s13059-019-1874-1.