Swiss Institute of Bioinformatics

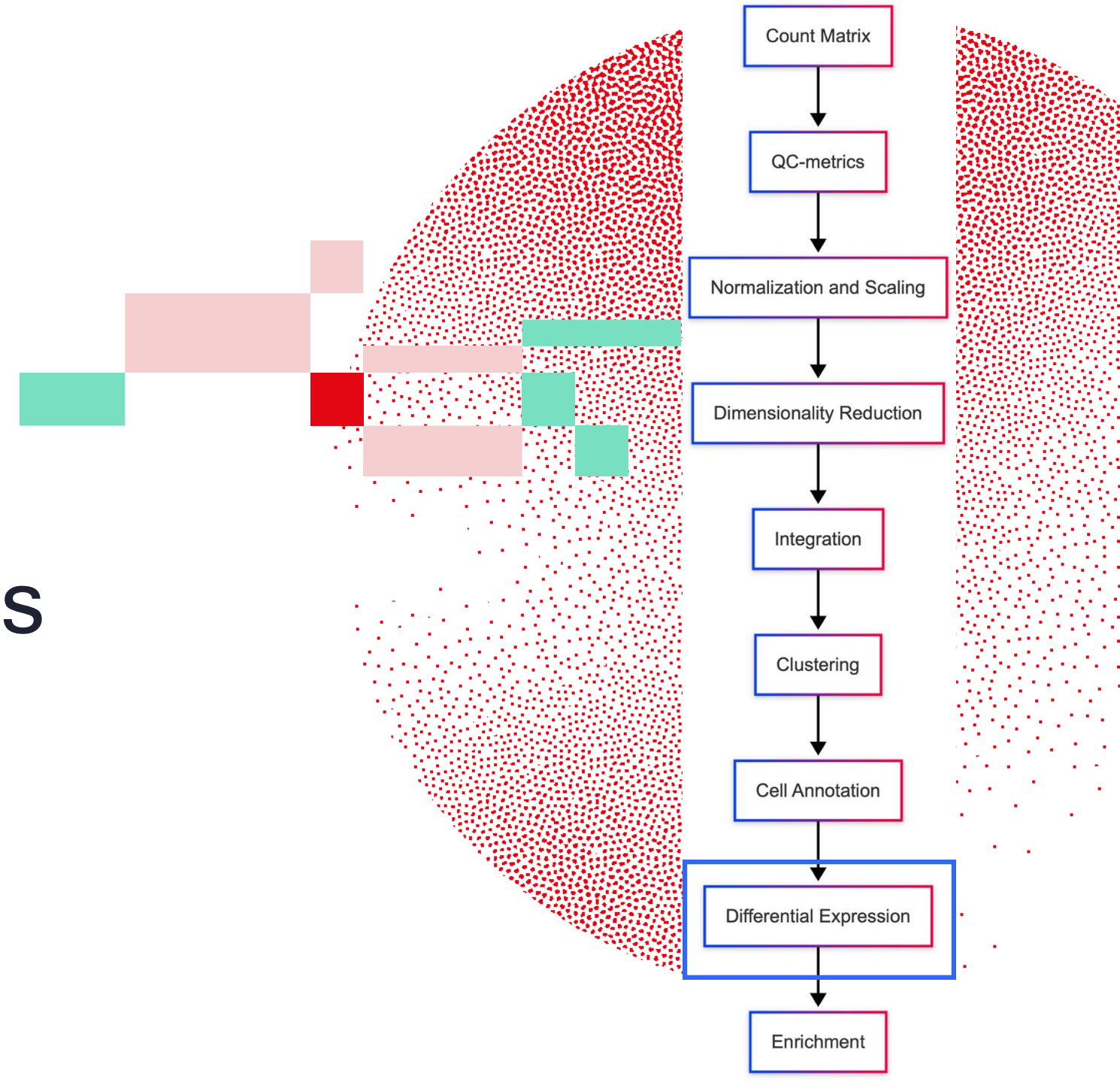
SINGLE-CELL TRANSCRIPTOMICS WITH R

Differential gene expression analysis

Deepak Tanwar

July 02-04, 2025

Adapted from previous year courses



Learning objectives

Differential gene expression analysis

- Finding markers
- Pseudo-bulk analysis

Differential gene expression analysis

DGE can be divided into 2 sub-groups:

a. Single-Cell-Level DGE Analysis (finding markers)

Goal: Identify genes that are differentially expressed between cell populations (e.g., cell types, clusters, or conditions) while accounting for the single-cell nature of the data (e.g., sparsity, dropout events).

Differential gene expression analysis

DGE can be divided into 2 sub-groups:

a. Single-Cell-Level DGE Analysis (finding markers)

Goal: Identify genes that are differentially expressed between cell populations (e.g., cell types, clusters, or conditions) while accounting for the single-cell nature of the data (e.g., sparsity, dropout events).

b. Pseudo-Bulk DGE Analysis (differential analysis)

Goal: Aggregate single-cell data into pseudo-bulk profiles to perform DGE analysis using bulk RNA-seq methods, reducing noise and leveraging biological replicates.

Marker gene identification

Methods:

- **Log-Fold Change (LFC) Analysis**: Calculate the log-fold change in expression between a target cluster and all other cells (e.g., in Seurat or Scanpy).

- **Wilcoxon Rank-Sum Test**: Test for genes with significantly higher expression in one group compared to others (e.g., in Seurat).

Marker gene identification

Methods:

- **Log-Fold Change (LFC) Analysis**: Calculate the log-fold change in expression between a target cluster and all other cells (e.g., in Seurat or Scanpy).

- **Wilcoxon Rank-Sum Test**: Test for genes with significantly higher expression in one group compared to others (e.g., in Seurat).

Use Cases:

- Annotating cell types (e.g., identifying CD3 as a marker for T cells)

- Discovering novel cell states or subpopulations

Marker gene identification

Methods:

- **Log-Fold Change (LFC) Analysis**: Calculate the log-fold change in expression between a target cluster and all other cells (e.g., in Seurat or Scanpy).

- **Wilcoxon Rank-Sum Test**: Test for genes with significantly higher expression in one group compared to others (e.g., in Seurat).

Use Cases:

- Annotating cell types (e.g., identifying CD3 as a marker for T cells)

- Discovering novel cell states or subpopulations

Challenges:

- Marker genes may not be unique to a single cell type, requiring careful validation

- Dropout events can obscure marker gene detection

Research | Open access | Published: 26 February 2024

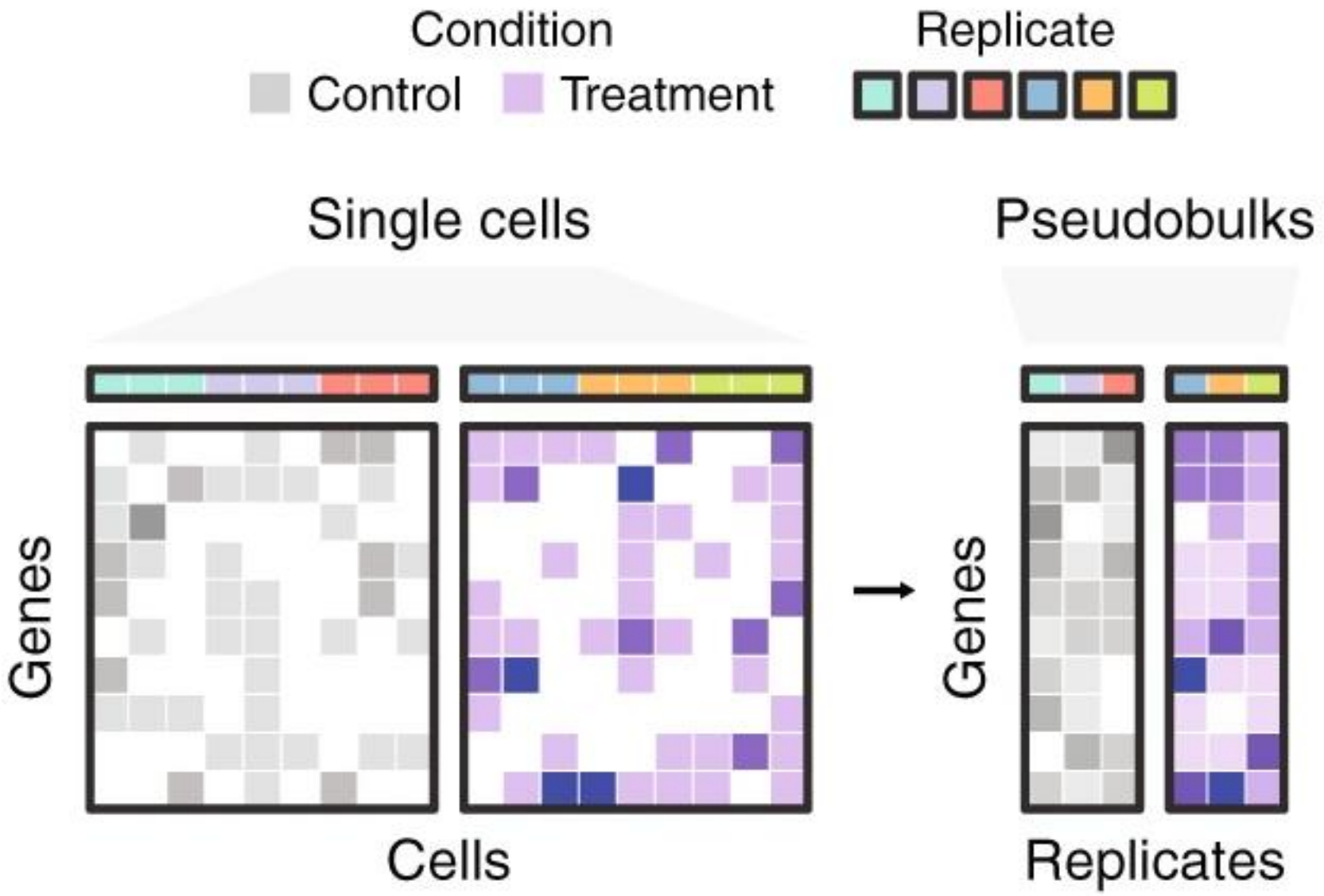
A comparison of marker gene selection methods for single-cell RNA sequencing data

Jeffrey M. Pullin & Davis J. McCarthy

Genome Biology **25**, Article number: 56 (2024) | Cite this article

Methods based on logistic regression, Student's *t*-test and the Wilcoxon rank-sum test all have strong performance

b. Pseudo-Bulk DGE Analysis



b. Pseudo-Bulk DGE Analysis

Methods:

- Aggregate gene expression counts within groups (e.g., by cell type and sample) and use bulk RNA-seq tools like DESeq2, edgeR, or limma

- Tools like muscat in Bioconductor are specifically designed for pseudo-bulk DGE analysis in scRNA-seq

b. Pseudo-Bulk DGE Analysis

Methods:

- Aggregate gene expression counts within groups (e.g., by cell type and sample) and use bulk RNA-seq tools like DESeq2, edgeR, or limma

- Tools like muscat in Bioconductor are specifically designed for pseudo-bulk DGE analysis in scRNA-seq

Advantages:

- Reduces noise and dropout effects

- Leverages well-validated bulk RNA-seq tools

b. Pseudo-Bulk DGE Analysis

Methods:

- Aggregate gene expression counts within groups (e.g., by cell type and sample) and use bulk RNA-seq tools like DESeq2, edgeR, or limma

- Tools like muscat in Bioconductor are specifically designed for pseudo-bulk DGE analysis in scRNA-seq

Advantages:

- Reduces noise and dropout effects

- Leverages well-validated bulk RNA-seq tools

Limitations: Loses single-cell resolution and cannot detect cell-to-cell variability

Quiz

What is a key limitation of pseudo-bulk analysis compared to single cell-level analysis?

A) It is computationally more intensive.

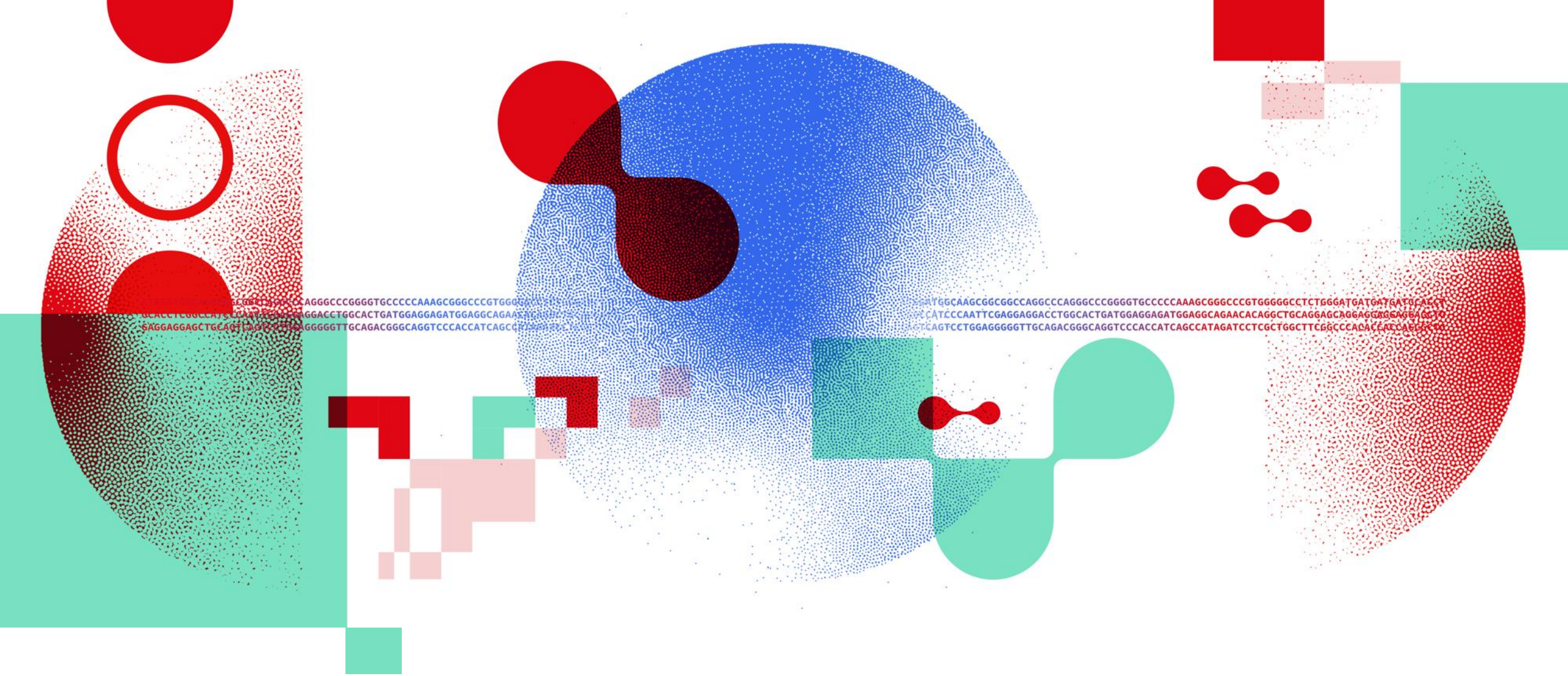
B) It loses single-cell resolution and cannot detect cell-to-cell variability.

C) It cannot be used for differential expression analysis.

D) It is less robust to dropout events.

Single Cell-Level: Finding markers

Pseudo-Bulk: Differential expression



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss