UCLA

scImpute

Accurate and Robust Imputation for Single-cell RNA-seq data

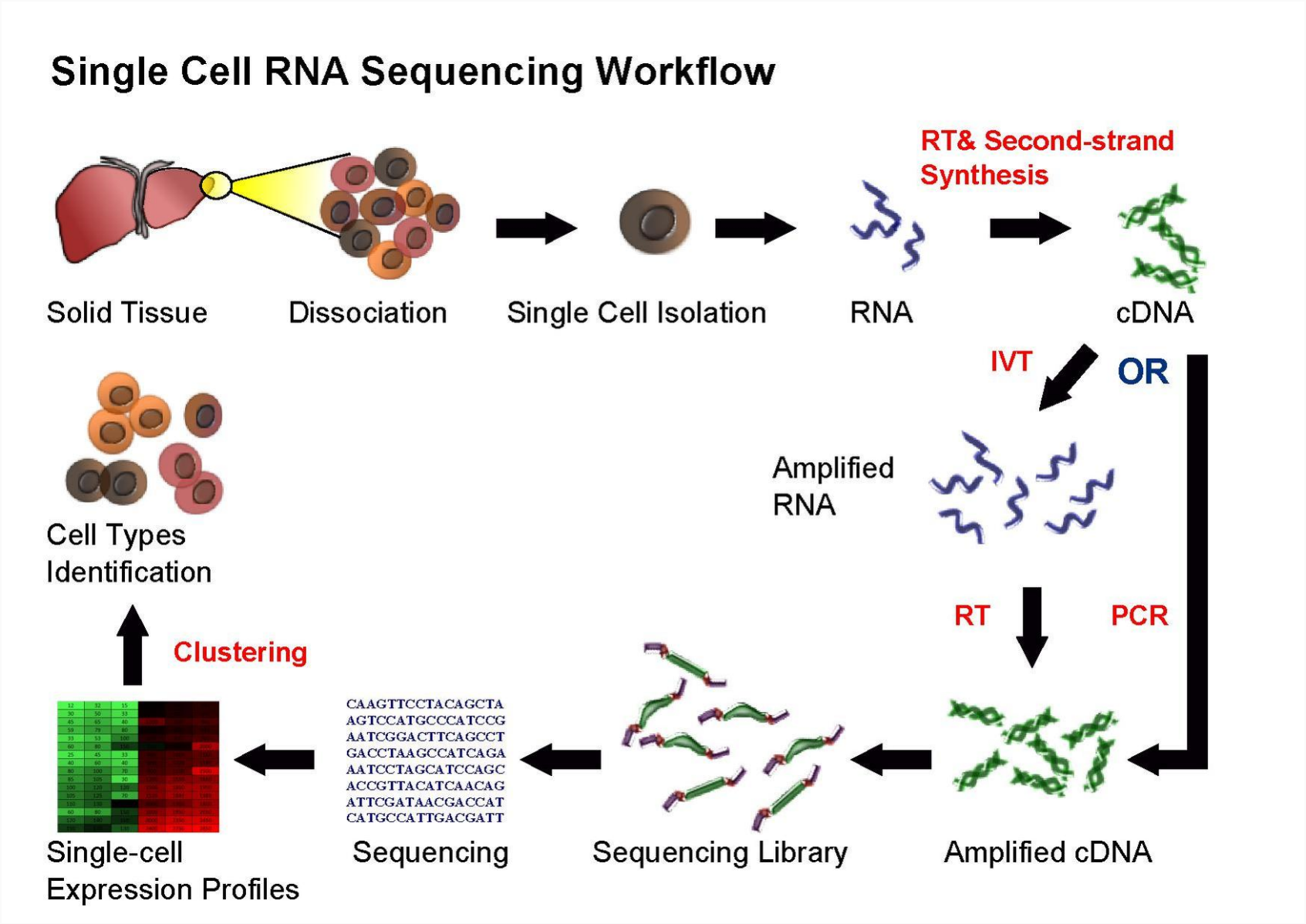
Wei Vivian Li and **Jingyi Jessica Li**

Department of Statistics
University of California, Los Angeles

http://jsb.ucla.edu

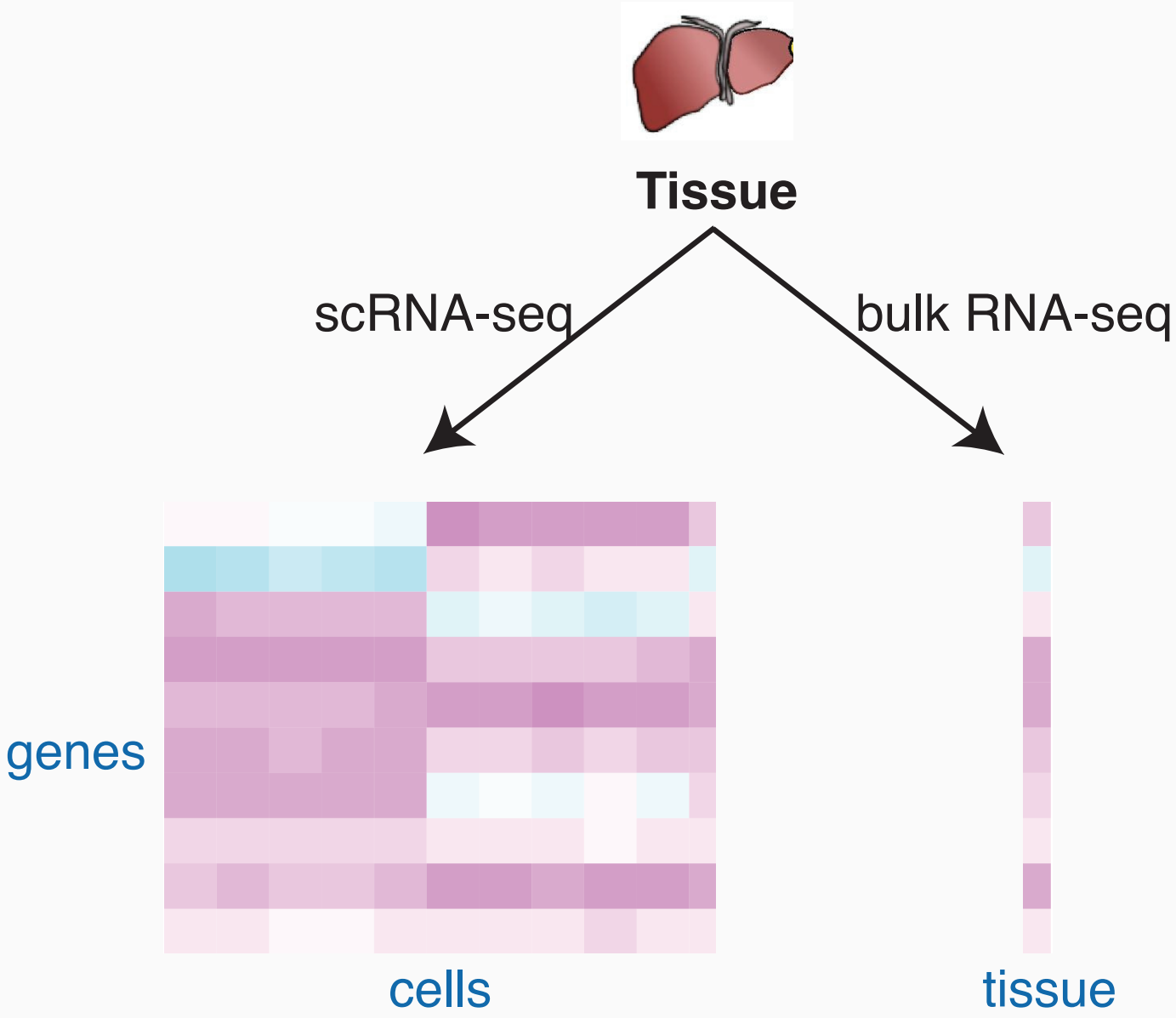
Background

Single-cell RNA Sequencing (scRNA-seq)

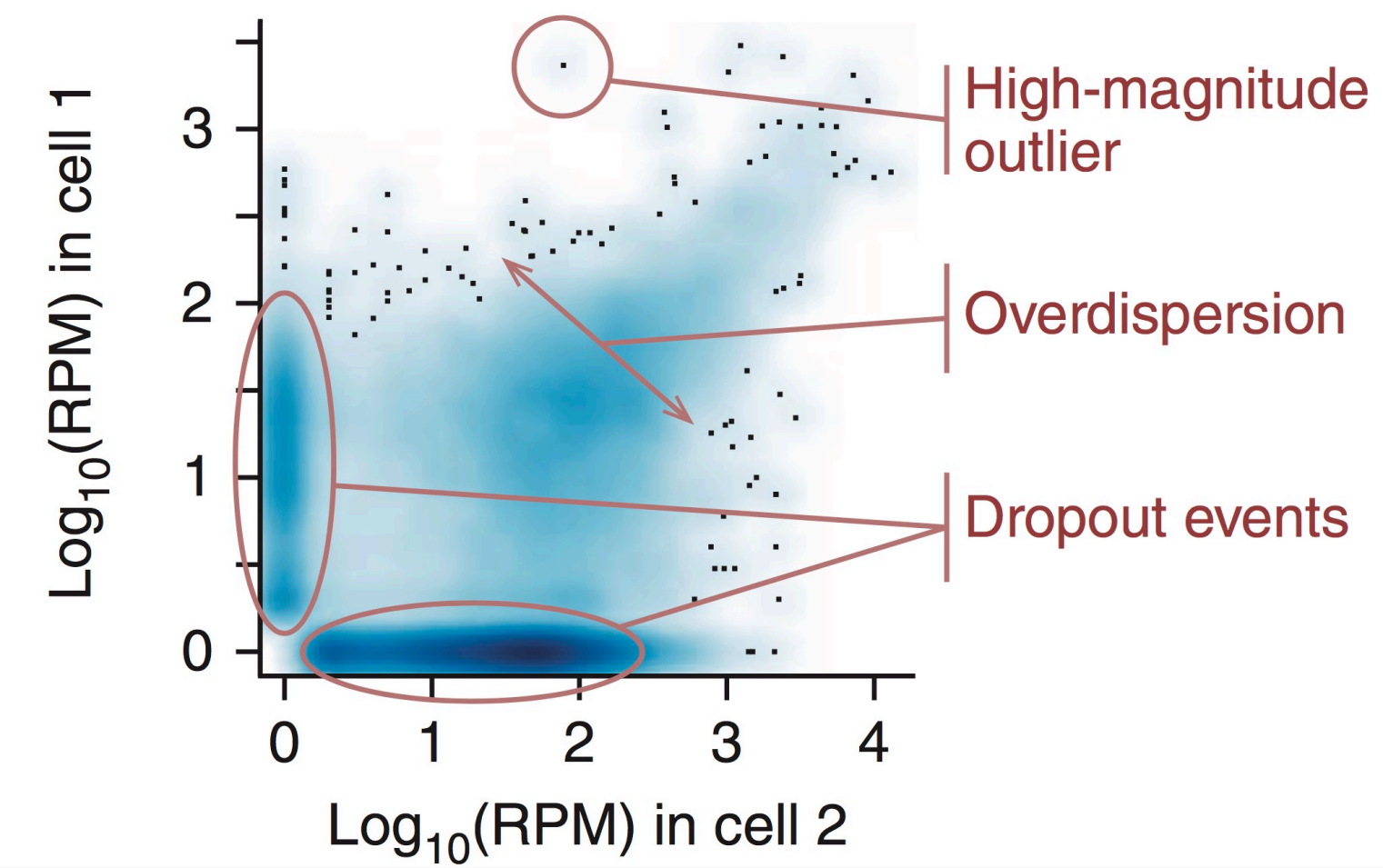


from *Wikipedia*, Single Cell Sequencing

scRNA-seq vs. Bulk RNA-seq for Gene Quantification



Dropout Events in scRNA-seq



from [Kharchenko et al., 2014] *Nature Methods*

Dropout Events in scRNA-seq

- A dropout event occurs when a transcript is expressed in a cell but is entirely undetected in its mRNA profile

- Dropout events occur due to low amounts of mRNA in individual cells

- The frequency of dropout events depends on scRNA-seq protocols
  - Fluidigm C1 platform: ~ 100 cells, ~ 1 million reads per cell
  - Droplet microfluidics: $\sim 10,000$ cells, ~ 100K reads per cell

  [Zilionis et al., 2017]

- **Trade-off**: given the same budget, more cells, more dropouts

Example Statistical Methods for scRNA-seq Data

- Clustering / cell type identification
  - **SNN-Cliq** [Xu and Su, 2015]: uses the ranking of genes to construct a graph and learn cell clusters
  - **CIDR** [Lin et al., 2017]: incorporates implicit imputation of dropout values
- Cell relationship reconstruction
  - **Seurat** [Satija et al., 2015]: infers the spatial origins of cells from their scRNA-seq data and a spatial reference map of landmark genes, whose expressions are imputed based on highly variable genes
- Dimension reduction
  - **ZIFA** [Pierson and Yau, 2015]: accounts for dropout events based on an empirical observation: dropout rate of a gene depends on its mean expression level in the population

Genome-wide Explicit Imputation for Dropouts

Why do we need genome-wide explicit imputation methods?

Downstream analyses relying on the accuracy of gene expression measurements:

- differential gene expression analysis
- identification of cell-type-specific genes
- reconstruction of cell differentiation trajectory
- and more

It is important to adjust the false zero expression values due to dropouts

Genome-wide Imputation Methods for scRNA-seq

MAGIC [van Dijk et al., 2017]:

- the first method for explicit and genome-wide imputation of scRNA-seq gene expression data
- imputes missing expression values by sharing information across similar cells
- creates a Markov transition matrix, which determines the weights of the cells

SAVER [Huang et al., 2017]:

- borrows information across genes using a Bayesian approach

DrImpute [Kwak et al., 2017]:

- borrows information across cells by averaging multiple imputation results

Our motivations

- It is not ideal to alter all gene expressions
  - altering values unlikely affected by dropouts might introduce new bias
  - could also eliminate meaningful biological variation

- It is inappropriate to treat all zero expressions as missing values
  - some zero expressions may reflect truly biological non-expression
  - zero expressions can be resulted from gene expression stochasticity

Our motivations

- It is not ideal to alter all gene expressions
  - altering values unlikely affected by dropouts might introduce new bias
  - could also eliminate meaningful biological variation

- It is inappropriate to treat all zero expressions as missing values
  - some zero expressions may reflect truly biological non-expression
  - zero expressions can be resulted from gene expression stochasticity

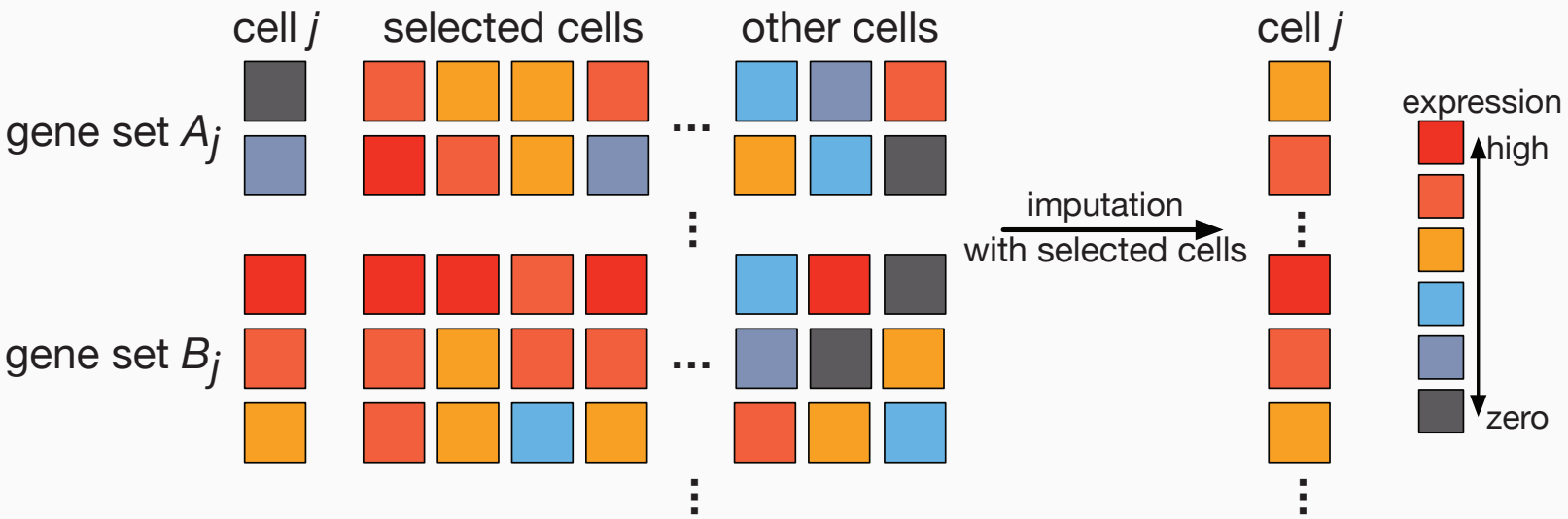
How to determine which values are affected by the dropout events?

Method: scImpute

Main Ideas

1. For each gene, to determine which expression values are most likely affected by dropout events

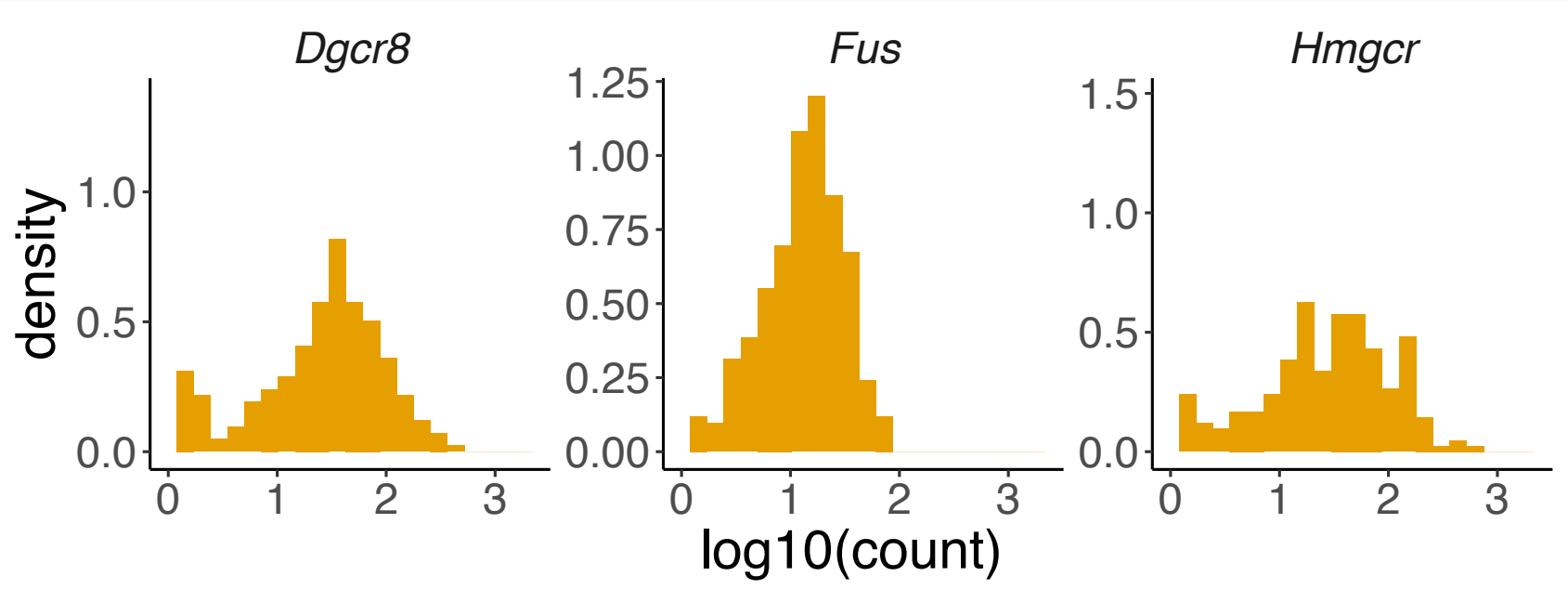
2. For each cell, to impute the highly likely dropout values by borrowing information from the same genes' expression in similar cells



Data Preprocessing

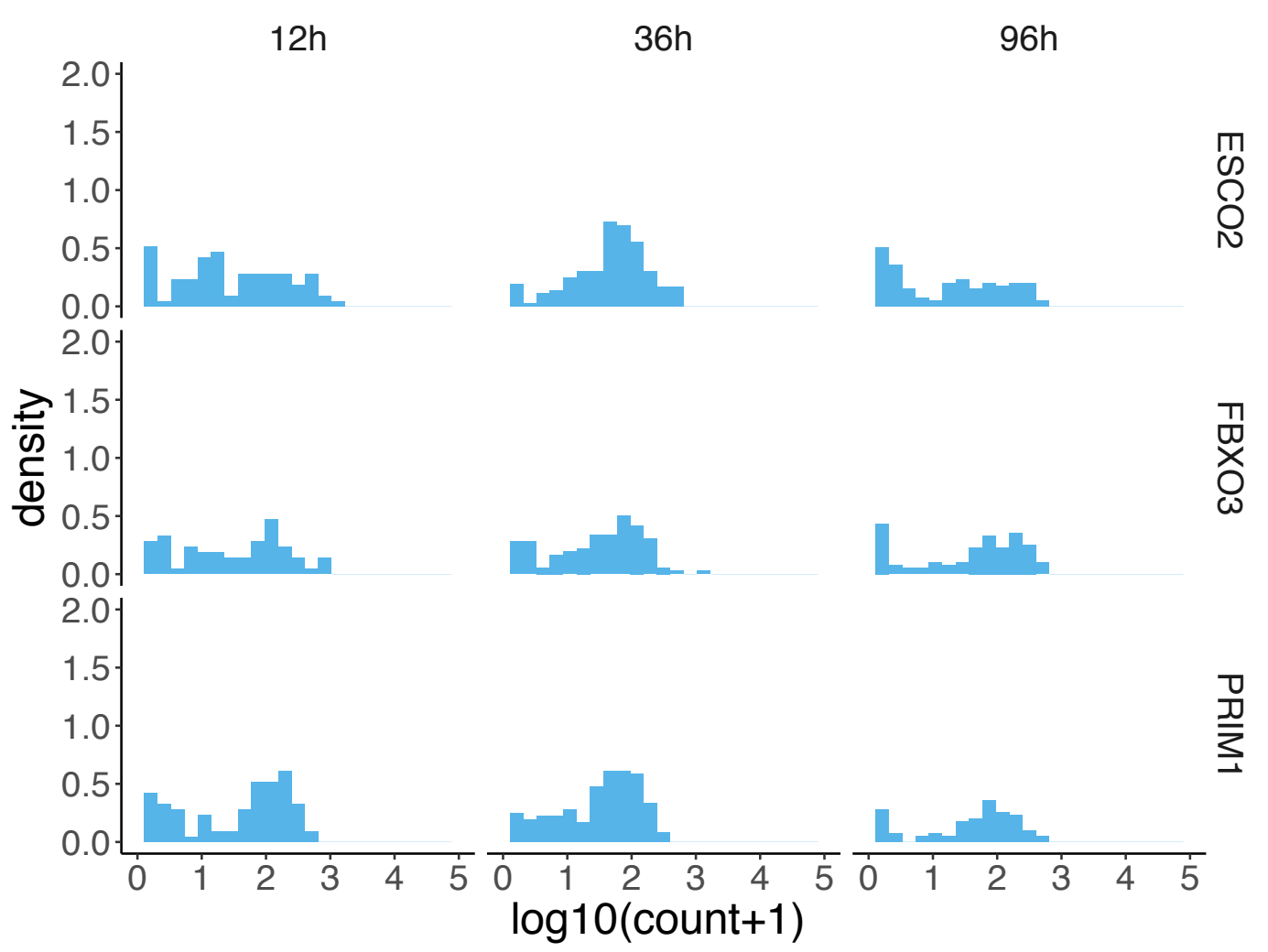
Input: A normalized and log transformed gene expression matrix $\mathbf{X}_{I \times J}$

- I genes
- J cells
- Expression of gene i in cell j: $X_{ij} \geq 0$



Three example mouse genes and the distributions of their expressions across 268 single cells [Deng et al., 2014]

Data Preprocessing



Observed expression distribution under three cell conditions in the human ESC data [Chu et al., 2016].

Step I: Detection of Cell Subpopulations and Outliers

1. Perform PCA (principal component analysis) on matrix $\mathbf{X}$ for dimension reduction (project every cell to a two-dimensional space)

2. Calculate the Euclidean distance matrix $\mathbf{D}_{J\times J}$ between the cells.

3. Detect outlier cells based on the distance matrix
   - The outlier cells could be a result of technical error or bias
   - The outlier cells may also represent real biological variation as rare cell types

4. Cluster the cells (excluding outliers) into K groups by spectral clustering
   - The candidate neighbor set of cell j is denoted as N_j

Step II: Identification of Dropout Values

1. For each gene i, we model its expression in cell population k as a random variable with density function

$$f_{X_i^{(k)}}(x) = \lambda_i^{(k)} \text{Gamma}\left(x; \alpha_i^{(k)}, \beta_i^{(k)}\right) + \left(1 - \lambda_i^{(k)}\right) \text{Normal}\left(x; \mu_i^{(k)}, \sigma_i^{(k)}\right),$$

where $\lambda_i^{(k)}$ is gene i's dropout rate in cell population k.

Step II: Identification of Dropout Values

1. For each gene i, we model its expression in cell population k as a random variable with density function

$$f_{X_i^{(k)}}(x) = \lambda_i^{(k)} \text{Gamma}\left(x; \alpha_i^{(k)}, \beta_i^{(k)}\right) + \left(1 - \lambda_i^{(k)}\right) \text{Normal}\left(x; \mu_i^{(k)}, \sigma_i^{(k)}\right),$$

where $\lambda_i^{(k)}$ is gene i's dropout rate in cell population k.

2. After estimating the parameters with the Expectation-Maximization (EM) algorithm, the dropout probability of gene i in cell j can be estimated as

$$d_{ij} = \frac{\hat{\lambda}_i^{(k)} \text{Gamma}\left(X_{ij}; \hat{\alpha}_i^{(k)}, \hat{\beta}_i^{(k)}\right)}{\hat{\lambda}_i^{(k)} \text{Gamma}\left(X_{ij}; \hat{\alpha}_i^{(k)}, \hat{\beta}_i^{(k)}\right) + \left(1 - \hat{\lambda}_i^{(k)}\right) \text{Normal}\left(X_{ij}; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{(k)}\right)}.$$

Step II: Identification of Dropout Values

1. For each gene i, we model its expression in cell population k as a random variable with density function

$$f_{X_i^{(k)}}(x) = \lambda_i^{(k)} \text{Gamma}\left(x; \alpha_i^{(k)}, \beta_i^{(k)}\right) + \left(1 - \lambda_i^{(k)}\right) \text{Normal}\left(x; \mu_i^{(k)}, \sigma_i^{(k)}\right),$$

where $\lambda_i^{(k)}$ is gene i's dropout rate in cell population k.

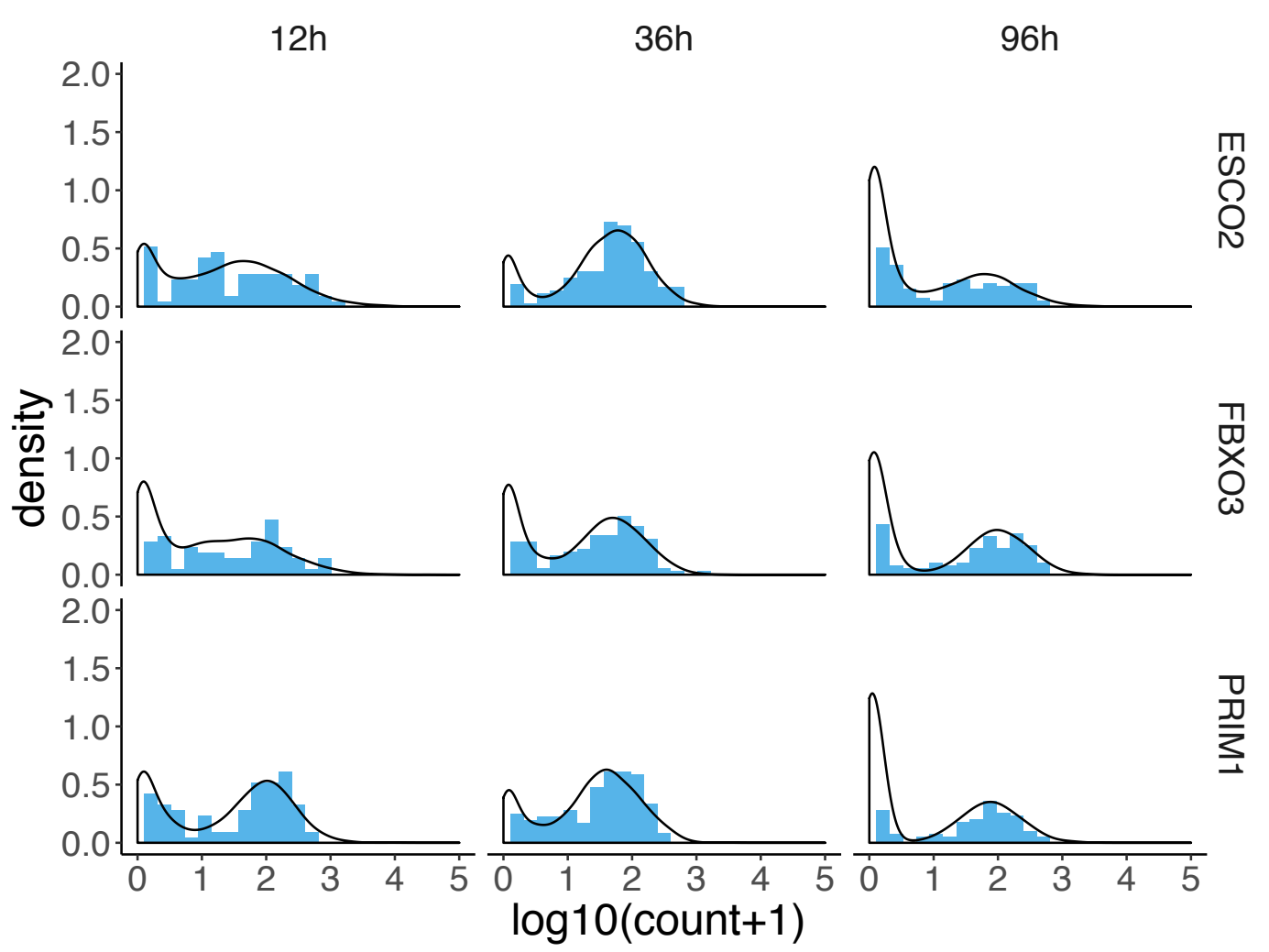
2. After estimating the parameters with the Expectation-Maximization (EM) algorithm, the dropout probability of gene i in cell j can be estimated as

$$d_{ij} = \frac{\hat{\lambda}_i^{(k)} \text{Gamma}\left(X_{ij}; \hat{\alpha}_i^{(k)}, \hat{\beta}_i^{(k)}\right)}{\hat{\lambda}_i^{(k)} \text{Gamma}\left(X_{ij}; \hat{\alpha}_i^{(k)}, \hat{\beta}_i^{(k)}\right) + \left(1 - \hat{\lambda}_i^{(k)}\right) \text{Normal}\left(X_{ij}; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{(k)}\right)}.$$

Remarks:

- The estimated dropout rates $\hat{\lambda}_i$ only depend on genes but not individual cells
- The estimated dropout probabilities d_{ij} depend on both genes and cells

Step II: Identification of Dropout Values



Observed and fitted expression distribution under three cell conditions in the human ESC data [Chu et al., 2016]

Step II: Identification of Dropout Values

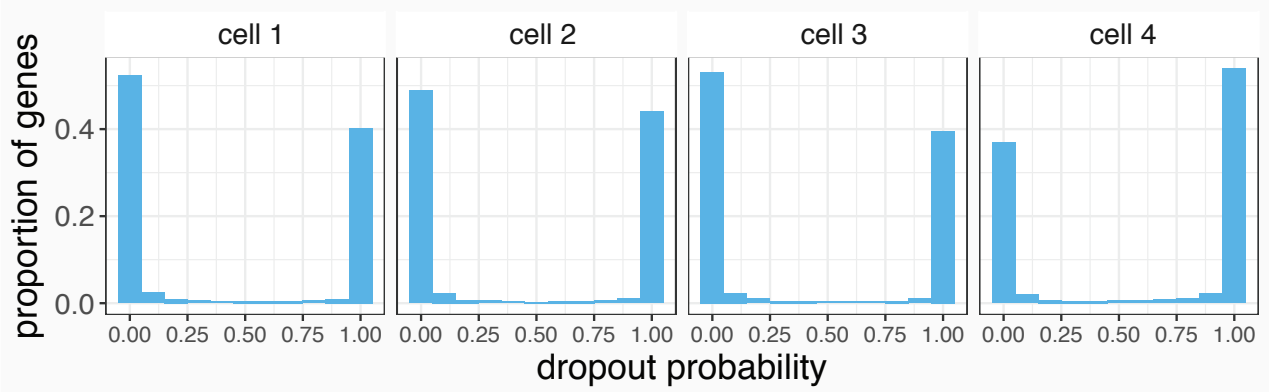
3. For each cell j, we select a gene set A_j in need of imputation:

$$A_j = \{i : d_{ij} \geq t\},$$

where t is a threshold on dropout probabilities. This also results in a gene set

$$B_j = \{i : d_{ij} < t\},$$

which have accurate gene expression with high confidence and do not need imputation.



The distribution of dropout probabilities in four randomly selected cells from the mouse embryo data [Deng et al., 2014]

Step III: Imputation of Gene Expressions Cell by Cell

1. For each cell j, we learn which cells in the candidate neighbor set N_j are similar to it based on the gene set B_j by the non-negative least squares (NNLS) regression:

$$\hat{\boldsymbol{\beta}}^{(j)} = \underset{\boldsymbol{\beta}^{(j)}}{\arg\min} ||\boldsymbol{X}_{B_j,j} - \boldsymbol{X}_{B_j,N_j}\boldsymbol{\beta}^{(j)}||_2^2, \text{ subject to } \boldsymbol{\beta}^{(j)} \geq \mathbf{0}\,.$$

   where

   - N_j represents the indices of cells that are candidate neighbors of cell j
   - $\boldsymbol{X}_{B_j,j}$ is a vector representing the B_j rows in the j-th column of $\boldsymbol{X}$
   - $\boldsymbol{X}_{B_j,N_j}$ is a sub-matrix of $\boldsymbol{X}$ with dimensions $|B_j| \times |N_j|$

Step III: Imputation of Gene Expressions Cell by Cell

1. For each cell j, we learn which cells in the candidate neighbor set N_j are similar to it based on the gene set B_j by the non-negative least squares (NNLS) regression:

$$\hat{\boldsymbol{\beta}}^{(j)} = \underset{\boldsymbol{\beta}^{(j)}}{\arg\min} ||\boldsymbol{X}_{B_j,j} - \boldsymbol{X}_{B_j,N_j}\boldsymbol{\beta}^{(j)}||_2^2, \text{ subject to } \boldsymbol{\beta}^{(j)} \geq \mathbf{0}\,.$$

   where

   - N_j represents the indices of cells that are candidate neighbors of cell j
   - $\boldsymbol{X}_{B_j,j}$ is a vector representing the B_j rows in the j-th column of $\boldsymbol{X}$
   - $\boldsymbol{X}_{B_j,N_j}$ is a sub-matrix of $\boldsymbol{X}$ with dimensions $|B_j| \times |N_j|$

2. The estimated coefficients $\hat{\boldsymbol{\beta}}^{(j)}$ from the set B_j are used to impute the expression of gene set A_j in cell j:

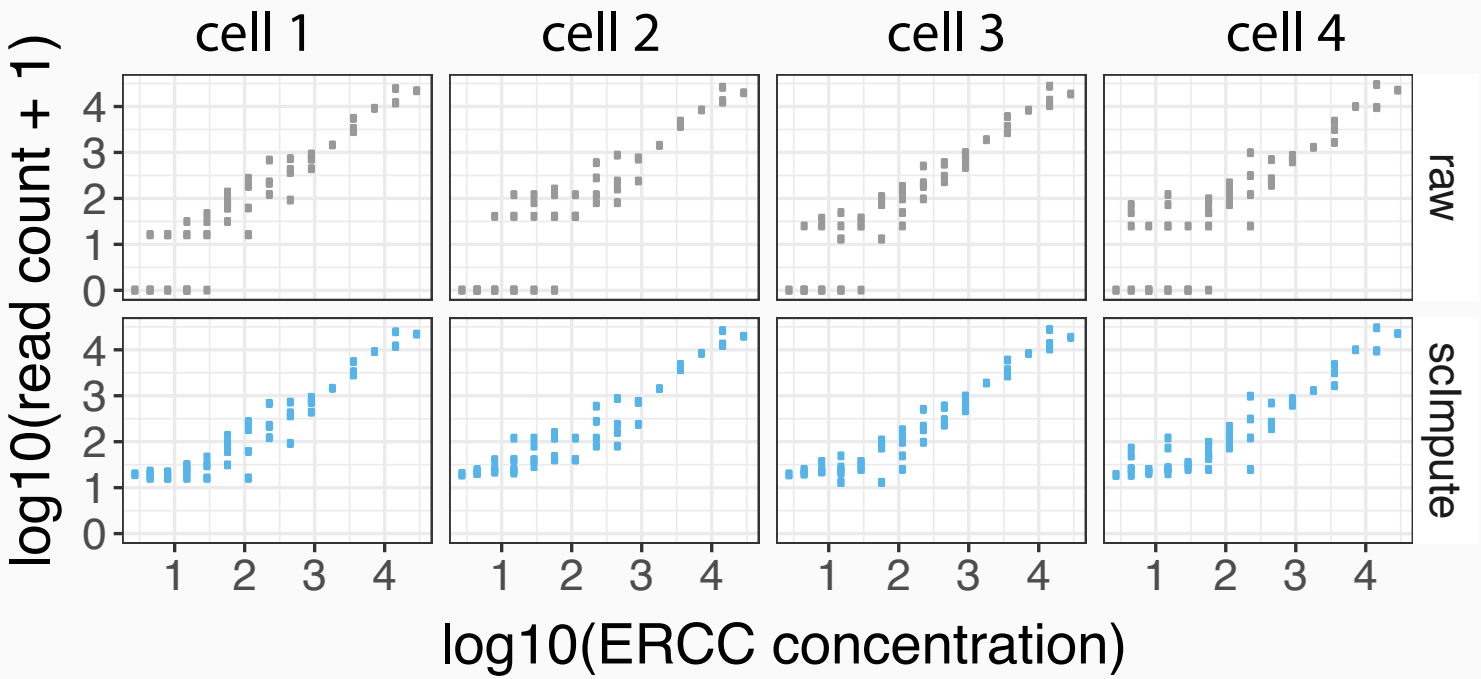
$$\hat{X}_{ij} = \begin{cases} X_{ij}, & i \in B_j, \\ X_{i,N_j}\hat{\boldsymbol{\beta}}^{(j)}, & i \in A_j. \end{cases}$$

Results

Case Study 1: ERCC Spike-ins

scImpute recovers the true expression of the ERCC spike-in transcripts [Jiang et al., 2011], especially low abundance transcripts impacted by dropout events

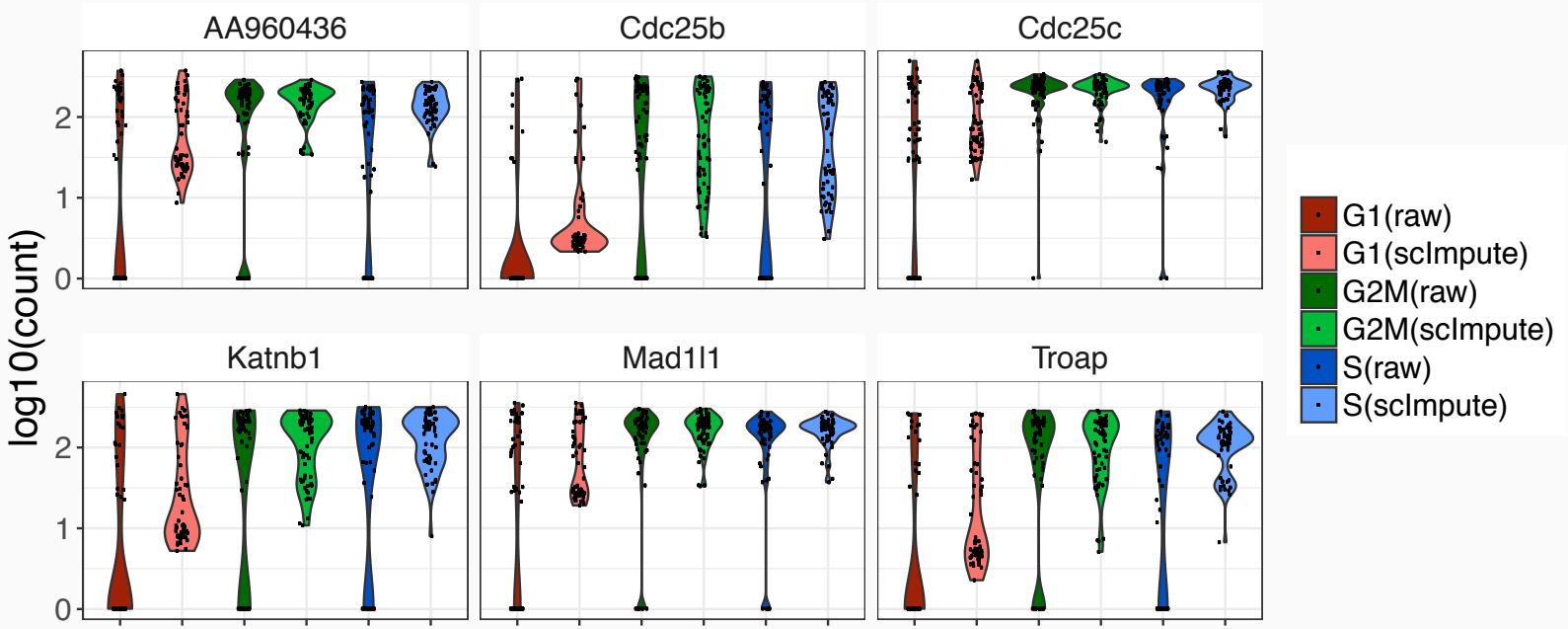
- 3,005 cells from the mouse somatosensory cortex region
- 57 ERCC transcripts



Case Study 2: Cell-cycle Gene Expression

scImpute correctly imputes the missing expressions of cell-cycle genes

- 892 annotated cell-cycle genes
- 182 embryonic stem cells (ESCs) that had been staged for cell-cycle phases (G1, S and G2M) [Buettner et al., 2015]



Case Study 3: Differential Gene Expression (Simulation)

Settings

- Three cell types c_1, c_2, and c_3, each with 50 cells
- Among a total of $20,000$ genes, 810 genes are truly differentially expressed, with 270 having higher expression in each cell type

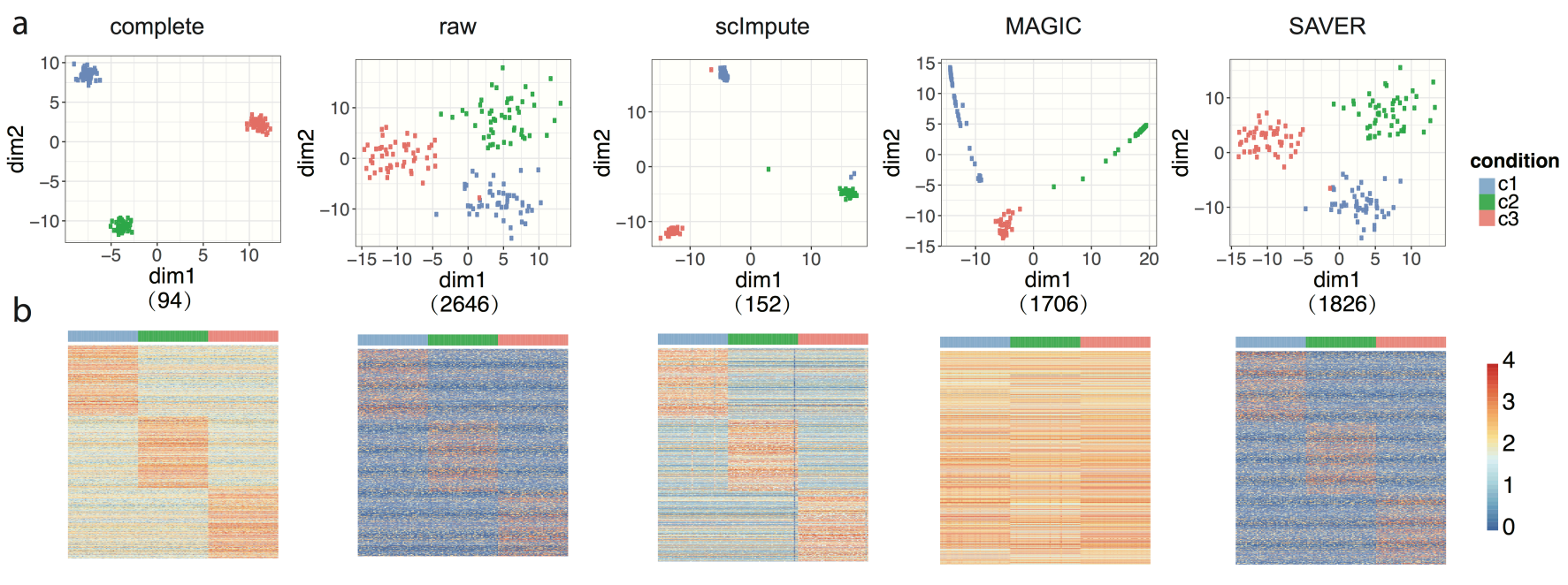
Procedures

- complete data: simulate gene expression values from normal distributions and shift the mean expression of DE genes.
- raw data: zero values are randomly introduced into the count matrix. The dropout rate of gene i is

$$\lambda_i = \exp\left(-0.1 \times (\bar{X}_{i\cdot})^2\right) ,$$

as assumed in [Pierson and Yau, 2015]

Case Study 3: Differential Gene Expression (Simulation)



- The relationships among the 150 cells are clarified after we apply scImpute
- The imputed data by scImpute lead to a clearer contrast between the up-regulated genes in different cell types

Case Study 4: Differential Gene Expression (Real Data)

Both single-cell and bulk RNA-seq data from human embryonic stem cells (ESC) and definitive endorderm cells (DEC) [Chu et al., 2016]

- 6 samples of bulk RNA-seq (4 in H1 ESC and 2 in DEC)
- 350 samples (cells) of scRNA-seq (212 in H1 ESC and 138 in DEC)

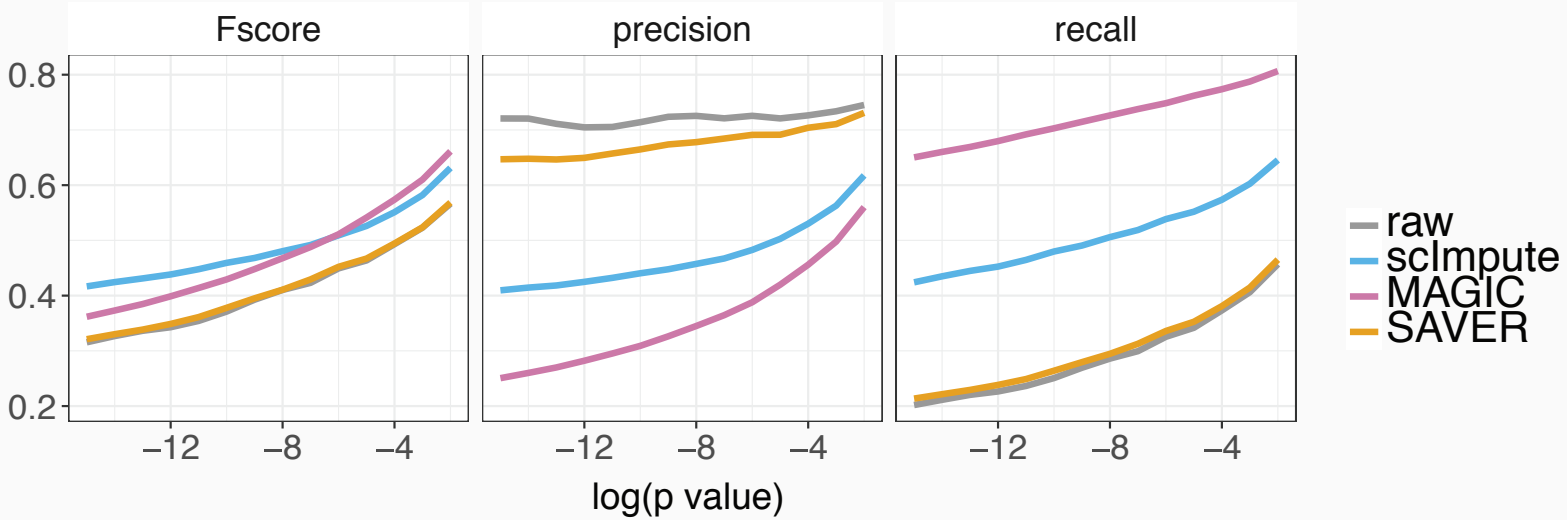
The percentage of zero gene expression

- 14.8% in bulk data
- 49.1% in single-cell data

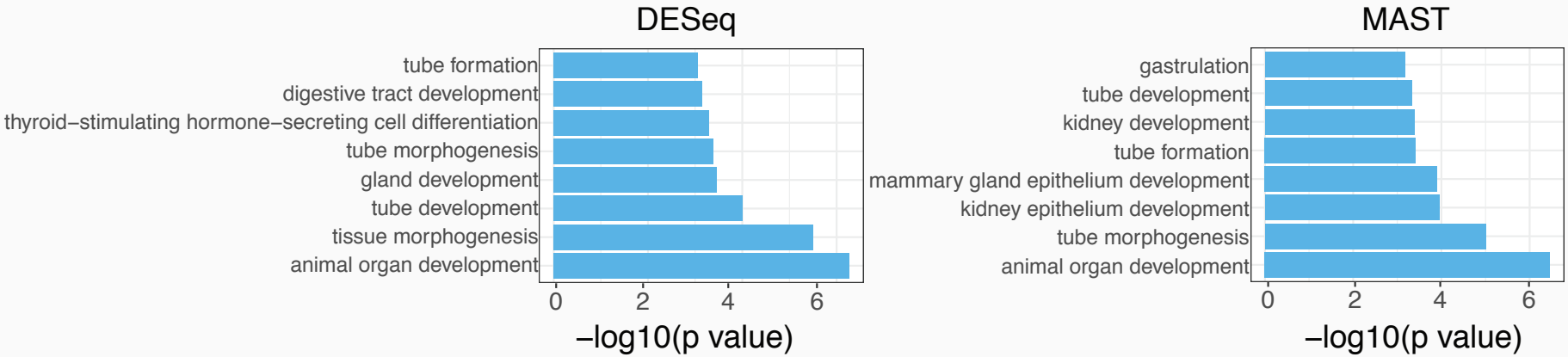
Differentially expressed (DE) genes are identified using `DESeq2` [Love et al., 2014] and `MAST` [Finak et al., 2015]

Case Study 4: Differential Gene Expression (Real Data)

Differential gene expression analysis



Gene ontology enrichment analysis



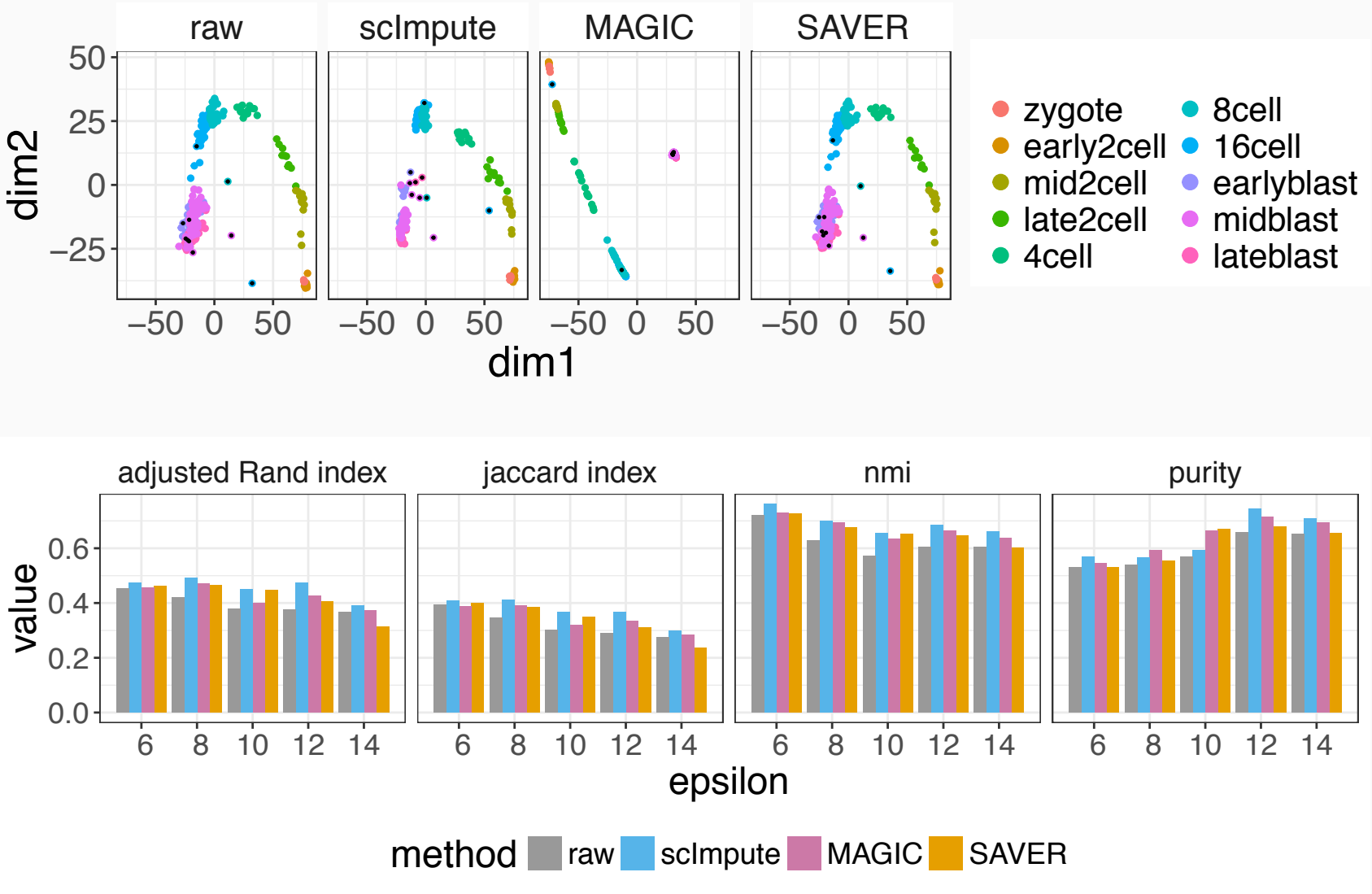
Case Study 5: Cell Clustering Example 1

268 single cells from mouse preimplantation embryos [Deng et al., 2014]

1. zygote (4 cells)
2. early 2-cell stage (8 cells)
3. middle 2-cell stage (12 cells)
4. late 2-cell stage (10 cells)
5. 4-cell stage (14 cells)
6. 8-cell stage (37 cells)
7. 16-cell stage (50 cells)
8. early blastocyst (43 cells)
9. middle blastocyst (60 cells)
10. late blastocyst (30 cells)

70.0% entries in the gene expression matrix are zeros

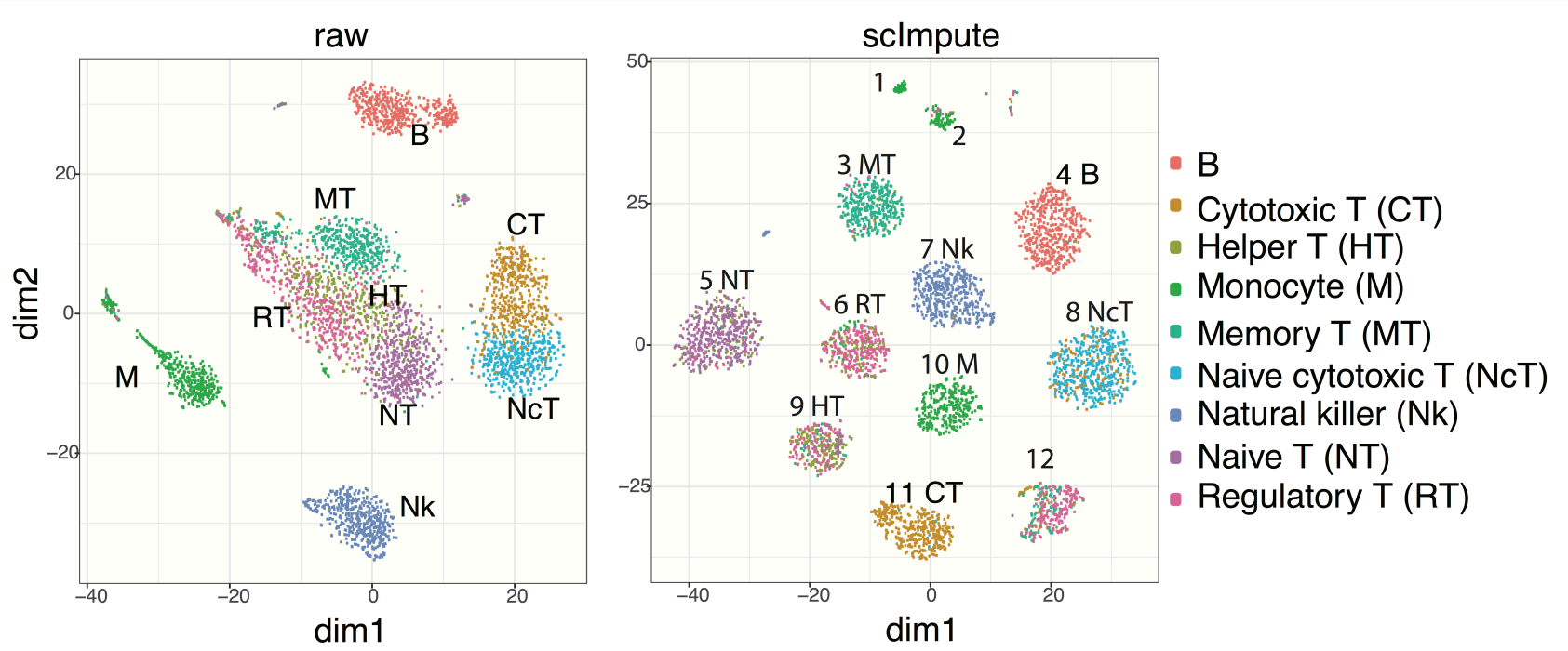
Case Study 5: Cell Clustering Example 1



Case Study 6: Cell Clustering Example 2

4,500 peripheral blood mononuclear cells (PBMCs) from high-throughput droplet-based system 10x genomics [Zheng et al., 2017]

Proportion of zero expression is 92.6%



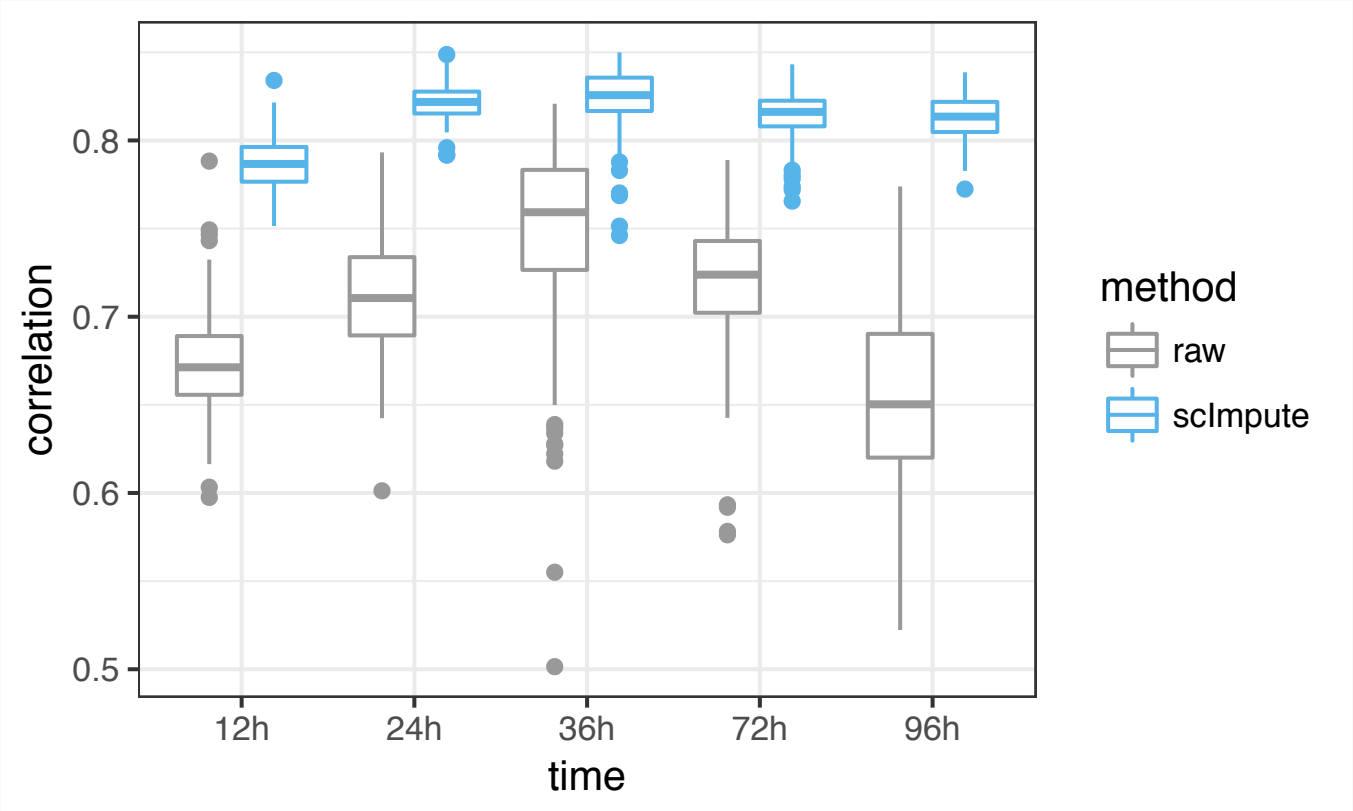
Case Study 7: Gene Expression Dynamics

Bulk and single-cell time-course RNA-seq data profiled at 0, 12, 24, 36, 72, and 96 h of the differentiation of embryonic stem cells into definitive endorderm cells [Chu et al., 2016]

time point	00h	12h	24h	36h	72h	96h	total
scRNA-seq (cells)	92	102	66	172	138	188	758
bulk RNA-seq (replicates)	0	3	3	3	3	3	15

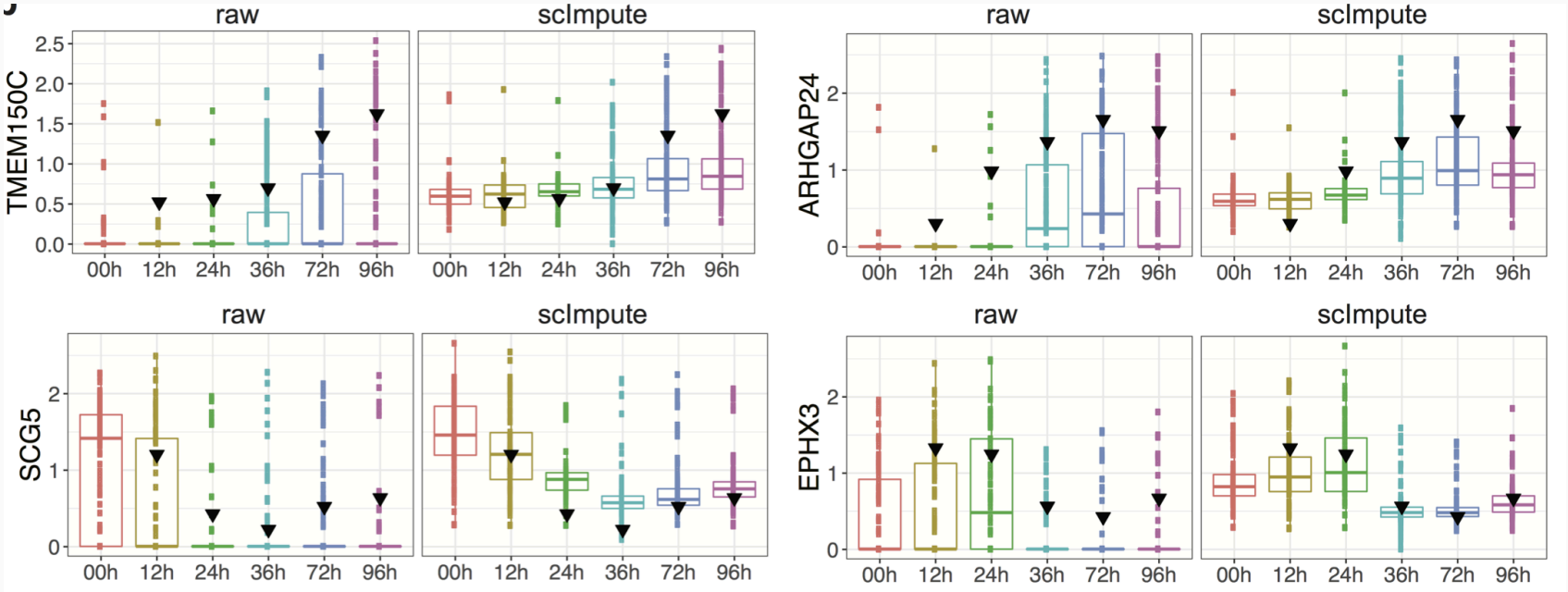
Case Study 7: Gene Expression Dynamics

Correlation between gene expression in single-cell and bulk data



Case Study 7: Gene Expression Dynamics

Imputed read counts reflect more accurate gene expression dynamics along the time course



Conclusions

- We propose a statistical method scImpute to address the dropout issue prevalent in scRNA-seq data
- scImpute focuses on imputing the missing expression values of dropout genes, while retaining the expression levels of genes that are largely unaffected by dropout events
- scImpute is compatible with existing pipelines or downstream analysis of scRNA-seq data, such as normalization, differential expression analysis, clustering and classification
- scImpute scales up well when the number of cells increases

Paper and Software

An accurate and robust imputation method scImpute for single-cell RNA-seq data

by Wei Vivian Li and Jingyi Jessica Li

Nature Communications 9:997

R package `scImpute`

`https://github.com/Vivianstats/scImpute`

Acknowledgements

- Dr. Mark Biggin (Lawrence Berkeley National Laboratory)
- Dr. Robert Modlin (UCLA)
- Dr. Matteo Pellegrini and Feiyang Ma (UCLA)
- Dr. Xia Yang and Douglas Arneson (UCLA)