UCLA Statistics

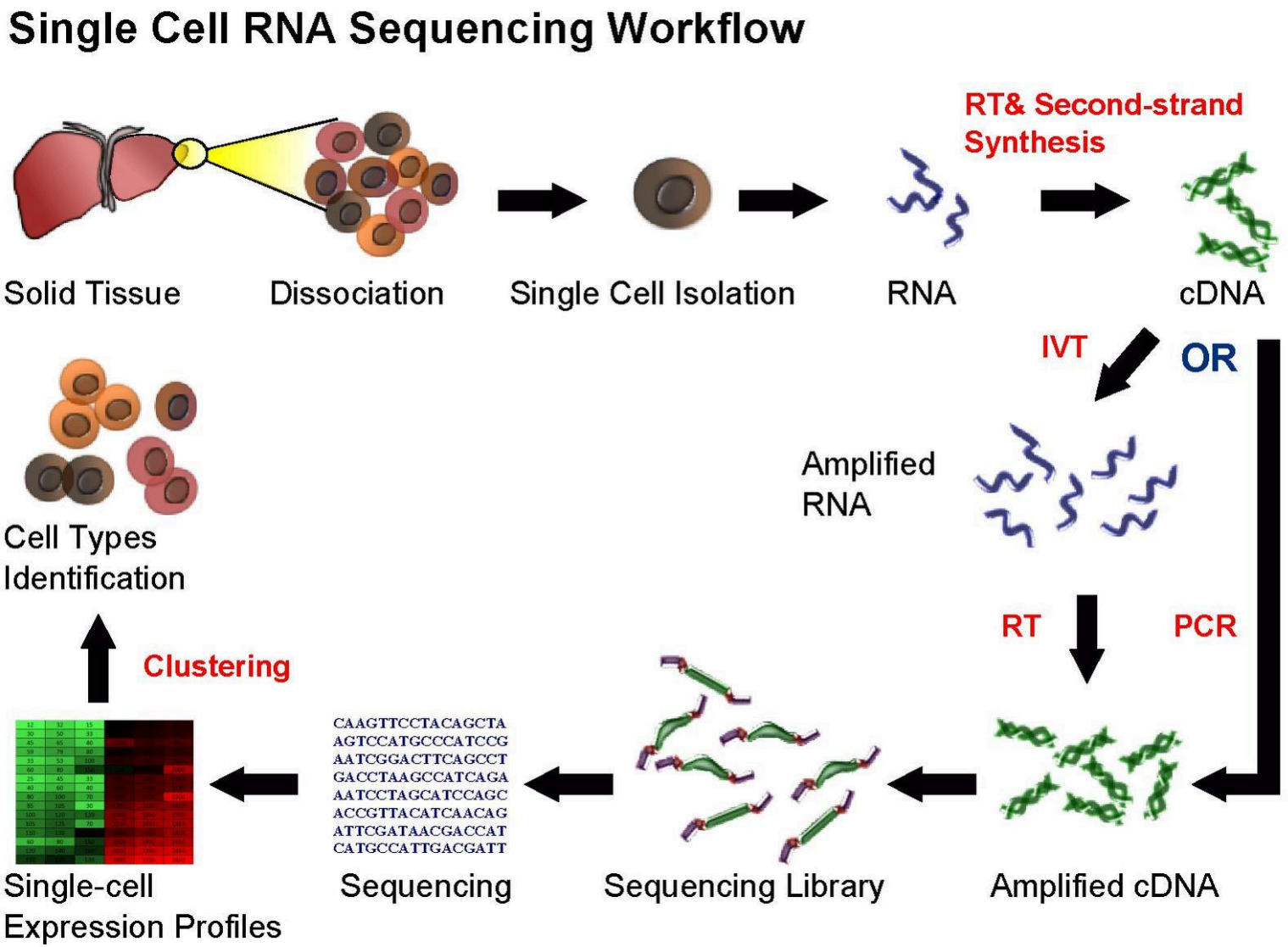
Imputation Methods for scRNA-seq Data

Wei (Vivian) Li

Department of Statistics
University of California, Los Angeles

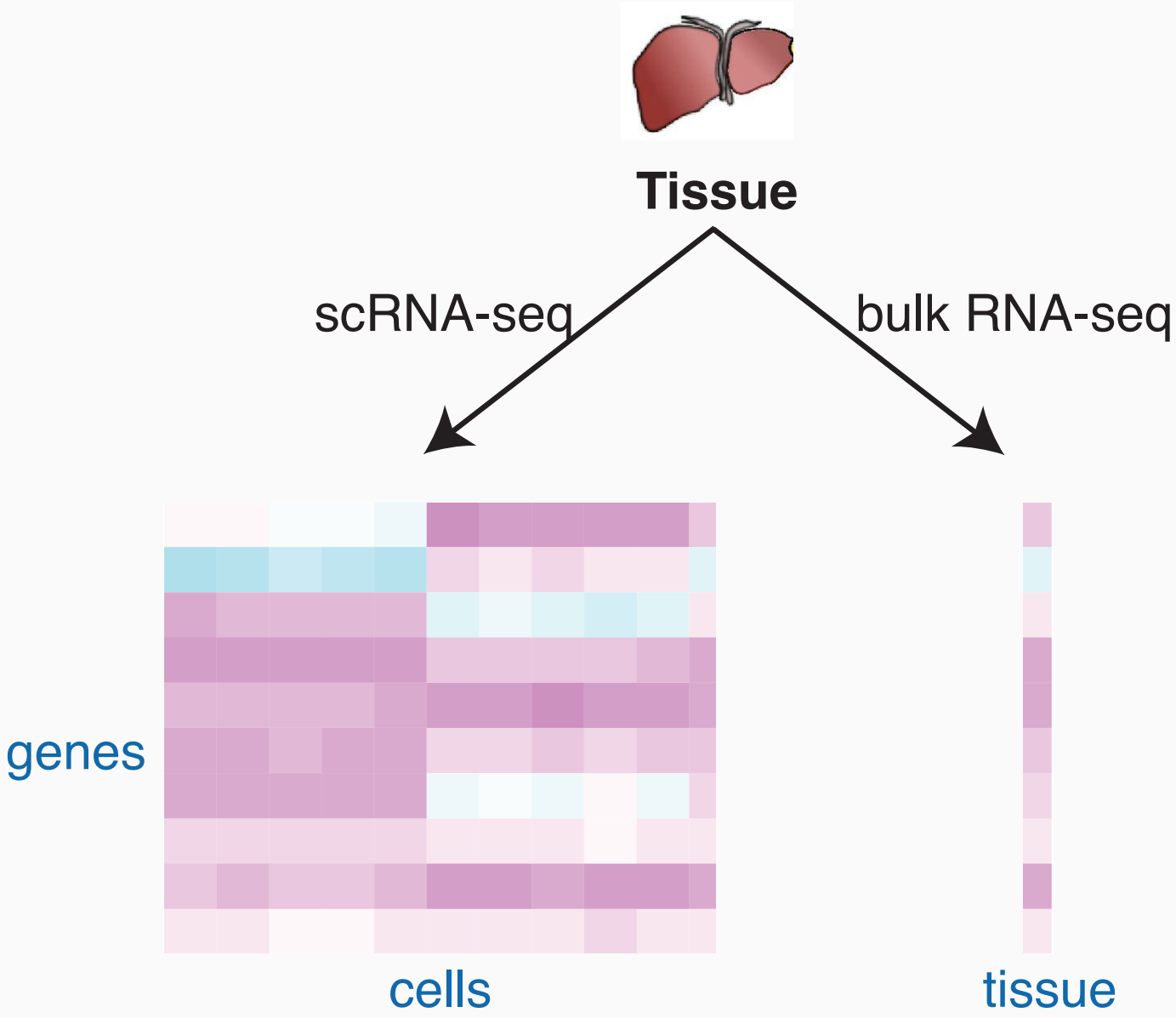
Introduction

Single Cell RNA Sequencing (scRNA-seq)

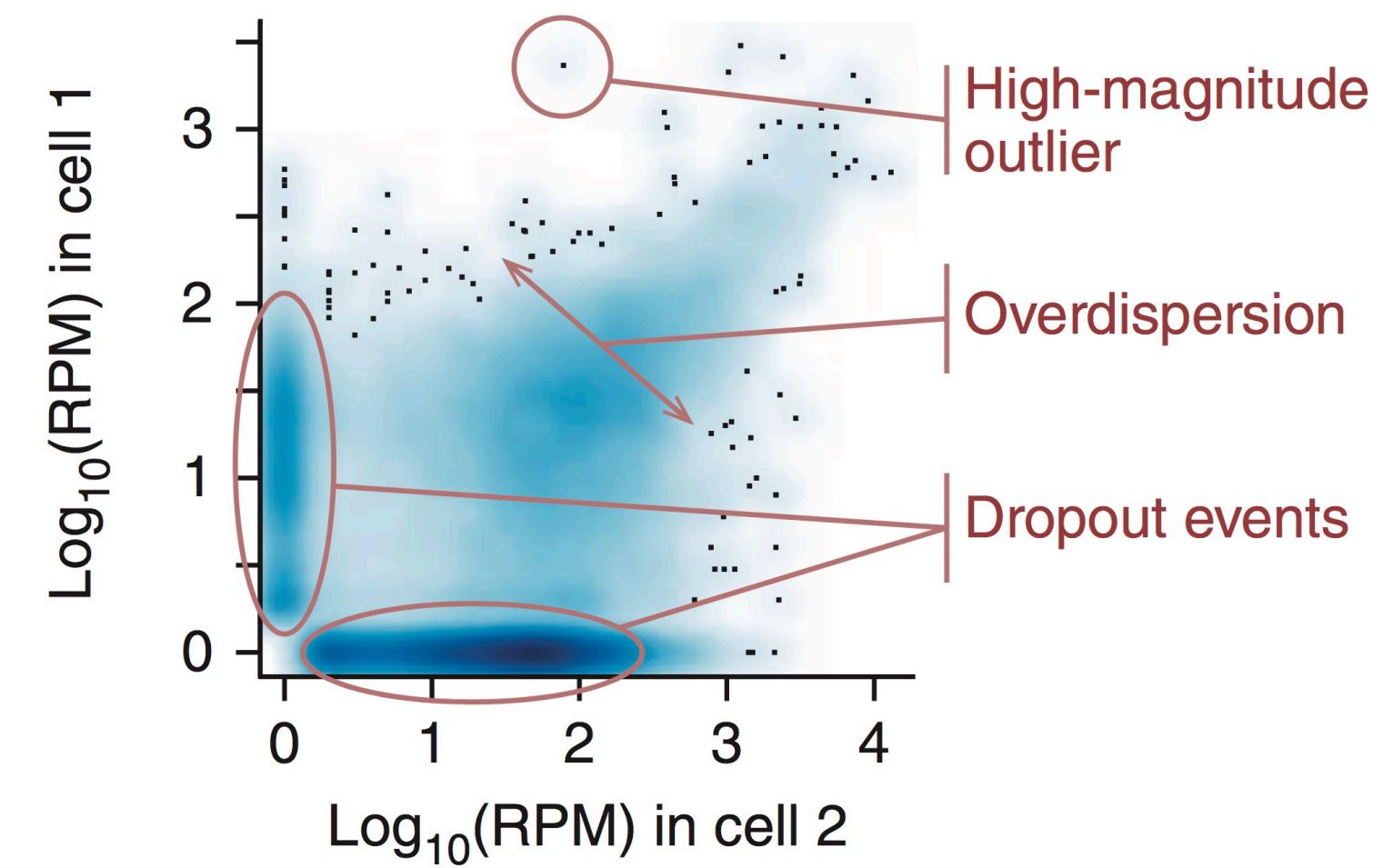


from *Wikipedia*, Single Cell Sequencing

scRNA-seq vs. Bulk RNA-seq



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

No Imputation or Implicit Imputation for Dropouts

- Clustering / cell type identification
  CIDR: incorporates implicit imputation of dropout values
  [Lin et al., 2017]

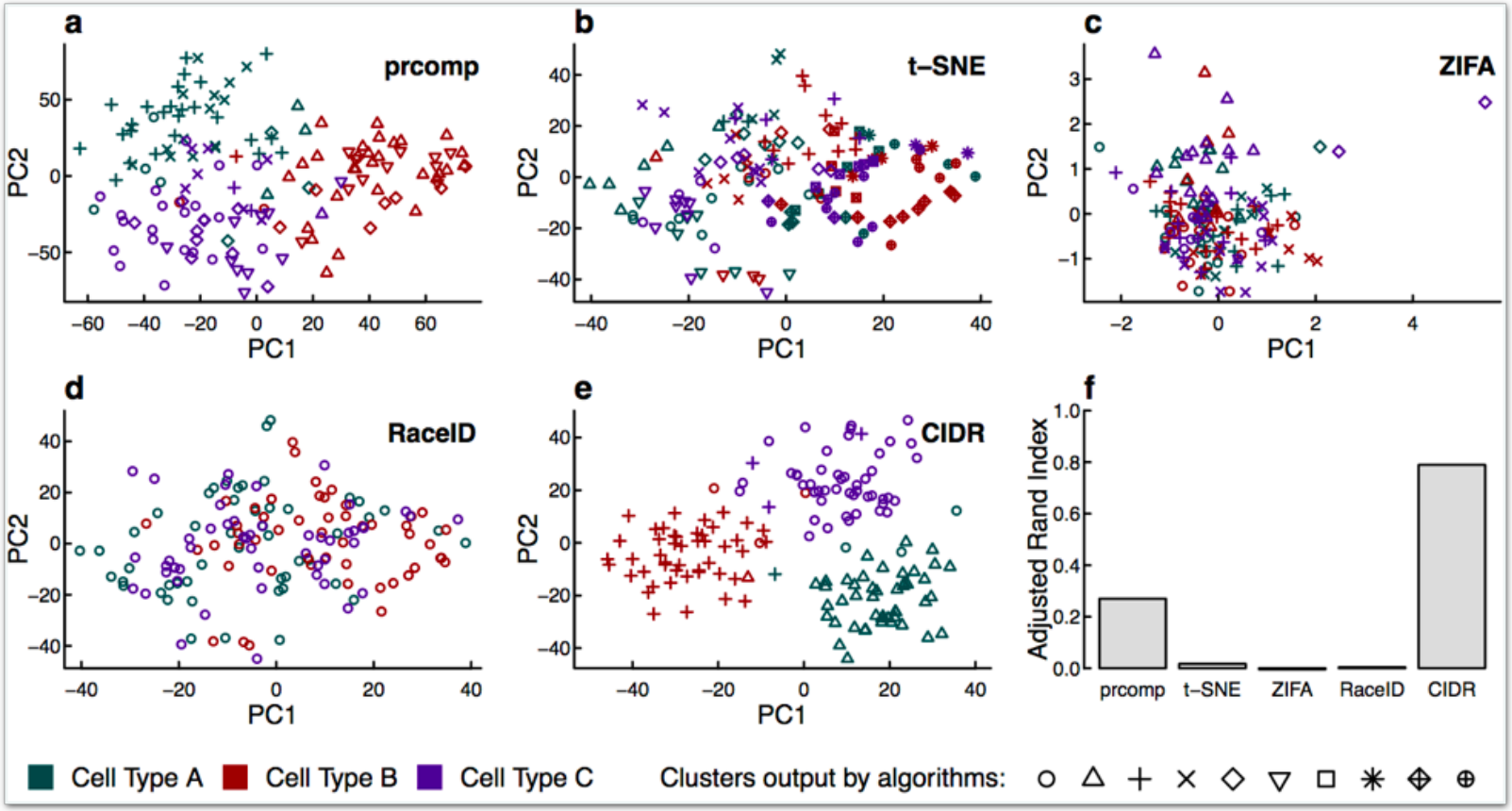
No Imputation or Implicit Imputation for Dropouts

- Clustering / cell type identification
  CIDR: incorporates implicit imputation of dropout values [Lin et al., 2017]
  - for each cell i, find a threshold T_i
    - entries larger than T_i are expressed entries
    - entries smaller than T_i are candidate dropouts
  - fit empirical dropout rate vs. average expressed entries: $\hat{P}(u)$

No Imputation or Implicit Imputation for Dropouts

- Clustering / cell type identification
  CIDR: incorporates implicit imputation of dropout values [Lin et al., 2017]
  - for each cell i, find a threshold T_i
    - entries larger than T_i are expressed entries
    - entries smaller than T_i are candidate dropouts
  - fit empirical dropout rate vs. average expressed entries: $\hat{P}(u)$
  - for gene k and each pair of cells i and j,
    $\hat{X}_{ki} = (1 - \hat{P}(X_{kj}))X_{kj} + \hat{P}(X_{kj})X_{ki}$
  - calculate dissimilarity measure between $\hat{\boldsymbol{X}}_{\cdot i}$ and $\hat{\boldsymbol{X}}_{\cdot j}$

No Imputation or Implicit Imputation for Dropouts



Comparison of clustering methods.

No Imputation or Implicit Imputation for Dropouts

- Cell relationship reconstruction
  - **Seurat**: 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 [Satija et al., 2015]

- Dimension reduction
  - **ZIFA**: accounts for dropout events based on an empirical observation: dropout rate of a gene depends on its mean expression level in the population [Pierson and Yau, 2015]
    - Dropout rate $p = \exp(-\lambda\mu^2)$.

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 differentiation trajectory

It is important to correct the false zero expression due to dropout events.

Genome-wide Imputation Methods for scRNA-seq

MAGIC: the first method for explicit and genome-wide imputation of scRNA-seq gene expression data [van Dijk et al., 2017]

- imputes missing expression values by sharing information across similar cells
- similarity between two cells $A_{ij} = e^{-(\frac{\text{Dist}_{ij}}{\sigma})^2}$
- transform the similarity matrix $\boldsymbol{A}$ into a Markov transition matrix $\boldsymbol{M}$
- raise the Markov matrix to the power of t: $\boldsymbol{M}^t$, which determines the weights of the cells

Genome-wide Imputation Methods for scRNA-seq

SAVER:

- borrows information across genes using a Bayesian approach [Huang et al., 2017]

DrImpute:

- borrows information across cells by averaging multiple imputation results [Kwak et al., 2017]

Genome-wide Imputation Methods for scRNA-seq

Limitations of aforementioned methods:

- It is not ideal to impute all gene expressions.
  - imputing expressions unaffected by dropout would 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 true biological non-expression
  - zero expressions can be resulted from gene expression stochasticity

Genome-wide Imputation Methods for scRNA-seq

Limitations of aforementioned methods:

- It is not ideal to impute all gene expressions.
  - imputing expressions unaffected by dropout would 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 true biological non-expression
  - zero expressions can be resulted from gene expression stochasticity

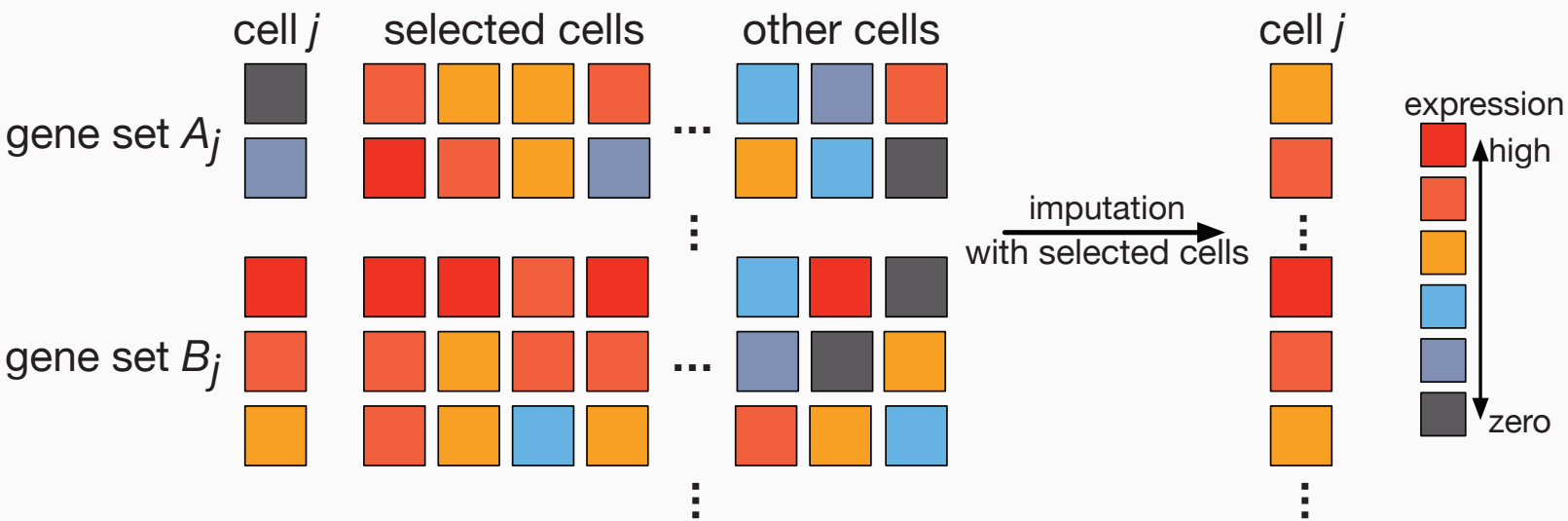
How to determine which values are affected by the dropout events?

scImpute

Main Ideas

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

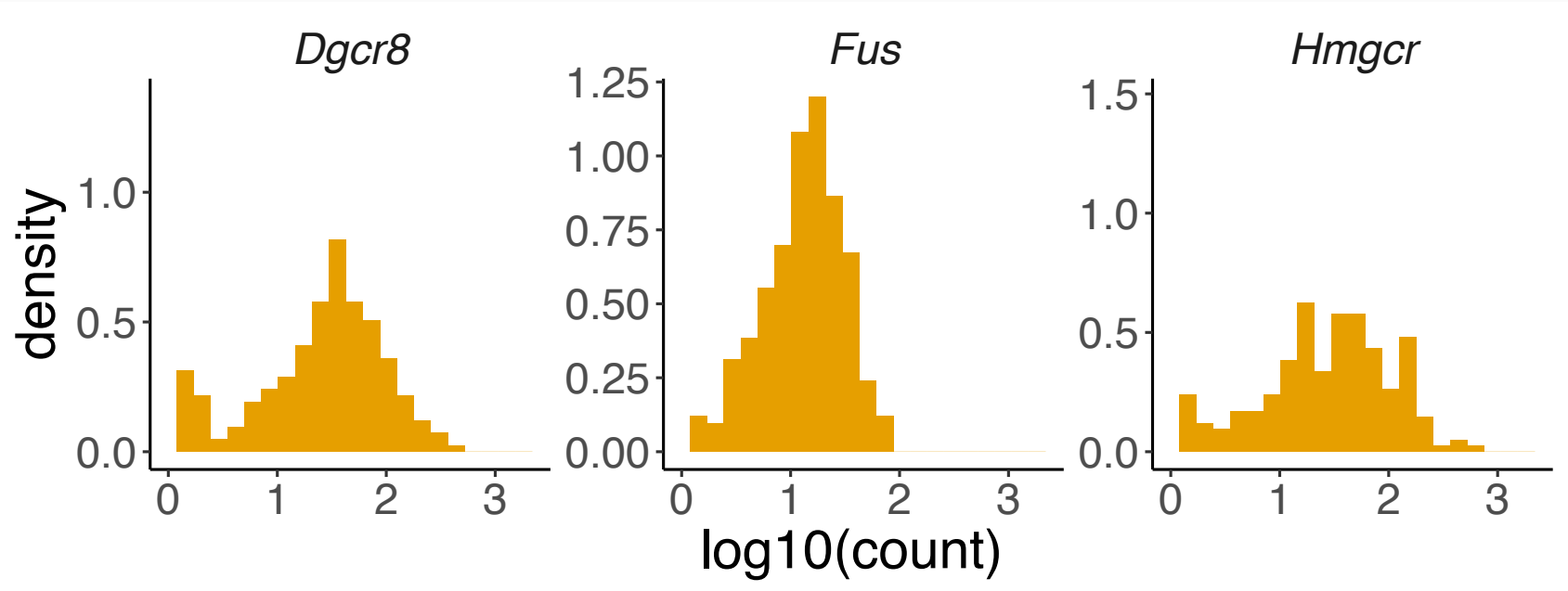
2. For each cell, to impute the highly likely dropout values by borrowing information from the same genes' expression in similar cells



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: $X_{ij} \geq 0$

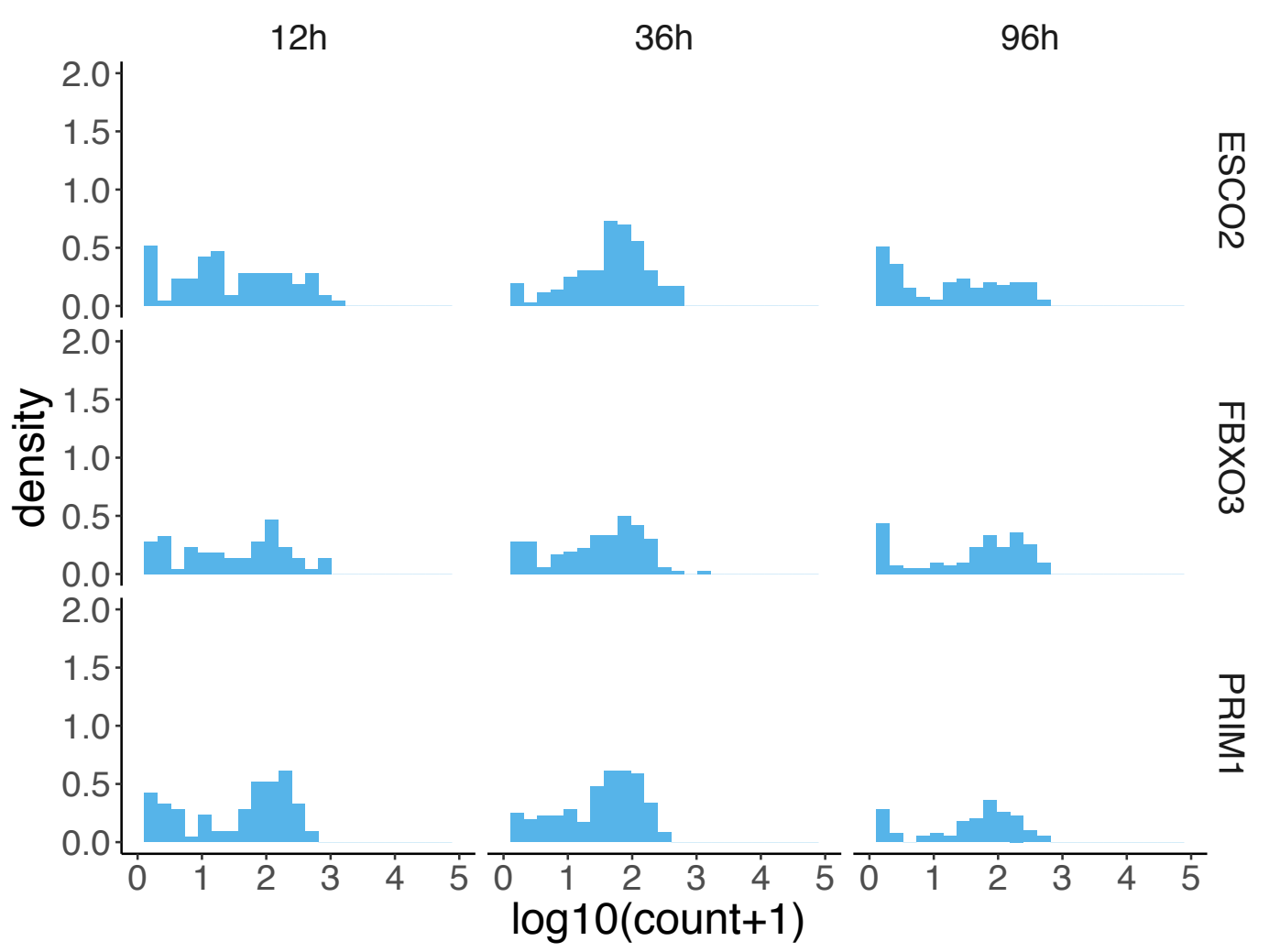


Three example mouse genes and the distributions of their expressions across 268 single cells [Deng et al., 2014]

scImpute Step I: Detection of Cell Subpopulations and Outliers

1. Perform PCA (principal component analysis) on matrix $\mathbf{X}$ for dimension reduction.
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.

scImpute Step II: Identification of Dropout Values



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

scImpute 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

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

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

scImpute 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

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

   - $\lambda_i^{(k)}$ is gene i's dropout rate in cell population k.
   - $z_{ij} = 1$ if gene i is a dropout in cell j; $z_{ij} = 0$ otherwise.
   - $P(z_{ij} = 1) = \lambda_i^{(k)}$

scImpute 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

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

   - $\lambda_i^{(k)}$ is gene i's dropout rate in cell population k.
   - $z_{ij} = 1$ if gene i is a dropout in cell j; $z_{ij} = 0$ otherwise.
   - $P(z_{ij} = 1) = \lambda_i^{(k)}$

$$\text{log-likelihood} = \sum_{j=1}^{J_k} \left\{ \mathbb{I}\{z_{ij=1}\} \log\left(\text{Gamma}\left(X_{ij}; \alpha_i^{(k)}, \beta_i^{(k)}\right)\right) + \mathbb{I}\{z_{ij=0}\} \log\left(\text{Normal}\left(X_{ij}; \mu_i^{(k)}, \sigma_i^{(k)}\right)\right) \right\}.$$

scImpute Step II: Identification of Dropout Values

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

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

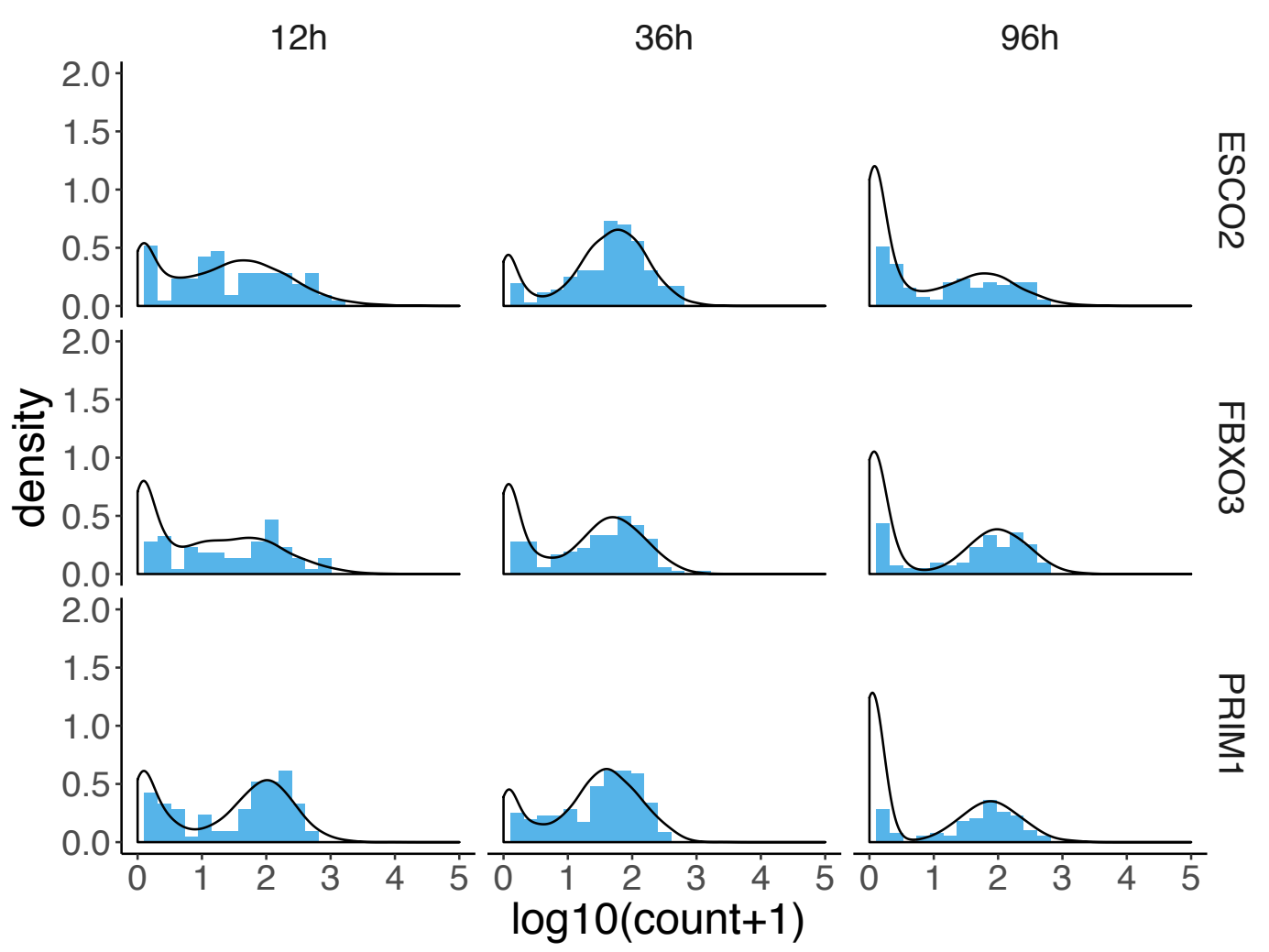
scImpute Step II: Identification of Dropout Values

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

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

- Each gene i has an estimated overall dropout rate $\hat{\lambda}_i$, which does not depend on individual cells.
- The estimated dropout probabilities d_{ij} $(j = 1, 2, \ldots, J_k)$ may vary among different cells.

scImpute Step II: Identification of Dropout Values



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

scImpute Step II: Identification of Dropout Values

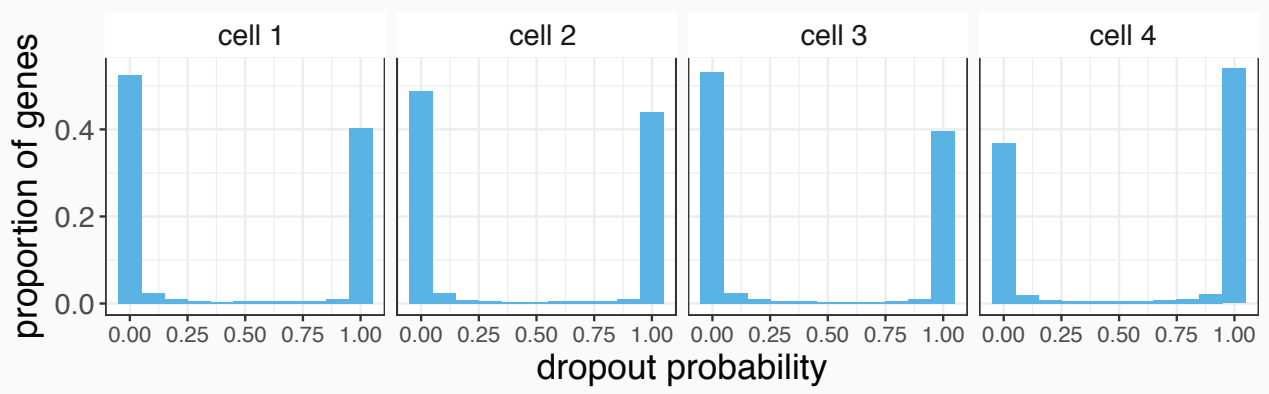
1. For each cell j, we select a gene set A_j in need of imputation:

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

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

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

that 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].

scImpute Step III: Imputation of Gene Expressions Cell by Cell

2. We learn which cells in the candidate neighbor set N_j are similar to cell j from the gene set B_j by the non-negative least squares (NNLS) regression:

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

where

- $\boldsymbol{X}_{B_j,j}$ 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 $|B_j| \times |N_j|$
- cell m in the neighbor set is selected to impute cell j only if $\hat{\beta}_m^{(j)} > 0$

scImpute Step III: Imputation of Gene Expressions Cell by Cell

2. We learn which cells in the candidate neighbor set N_j are similar to cell j from the gene set B_j by the non-negative least squares (NNLS) regression:

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

   where

   - $\boldsymbol{X}_{B_j,j}$ 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 $|B_j| \times |N_j|$
   - cell m in the neighbor set is selected to impute cell j only if $\hat{\beta}_m^{(j)} > 0$

3. 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:

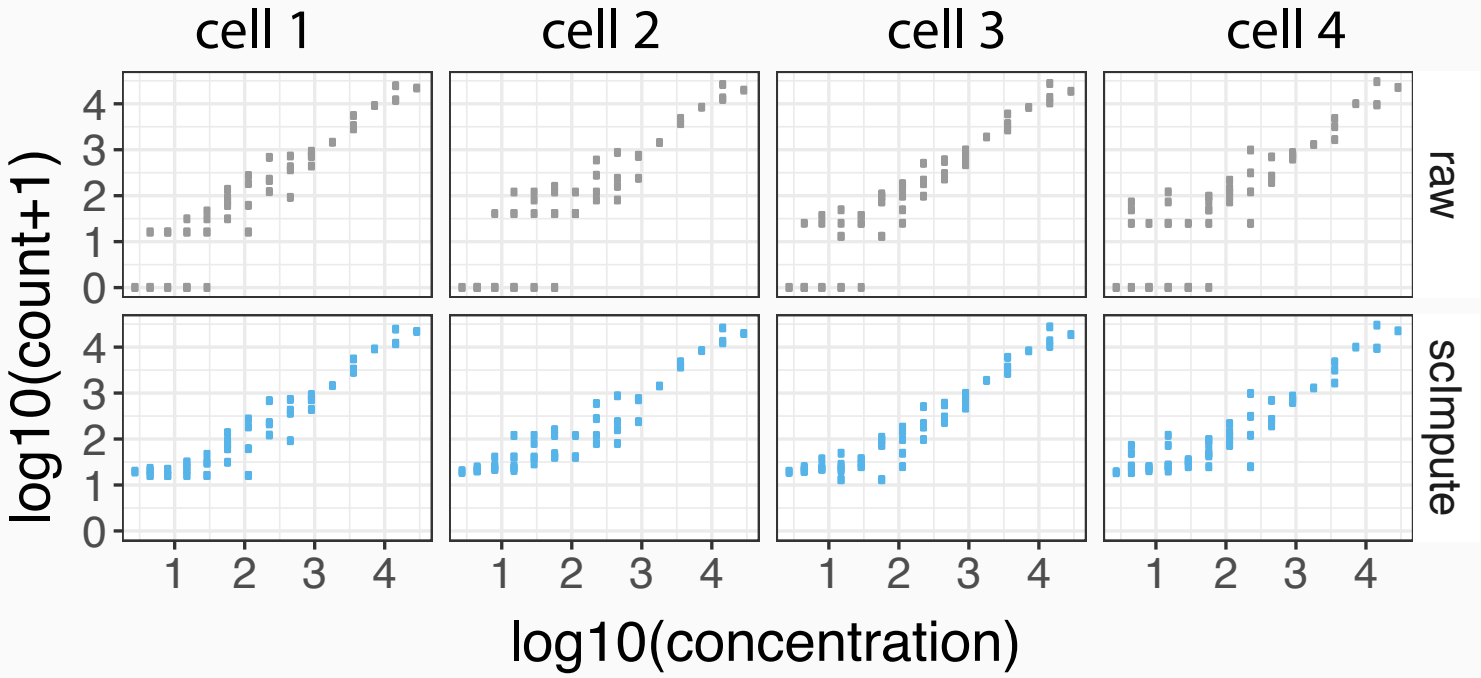
$$\hat{X}_{ij} = \begin{cases} X_{ij}, & i \in B_j, \\ X_{i,N_j}\hat{\boldsymbol{\beta}}^{(j)}, & i \in A_j. \end{cases}$$

Results

scImpute Recovers the Dropout Events

scImpute recovers the true expression of the ERCC spike-in transcripts, especially low abundance transcripts that are impacted by dropout events.

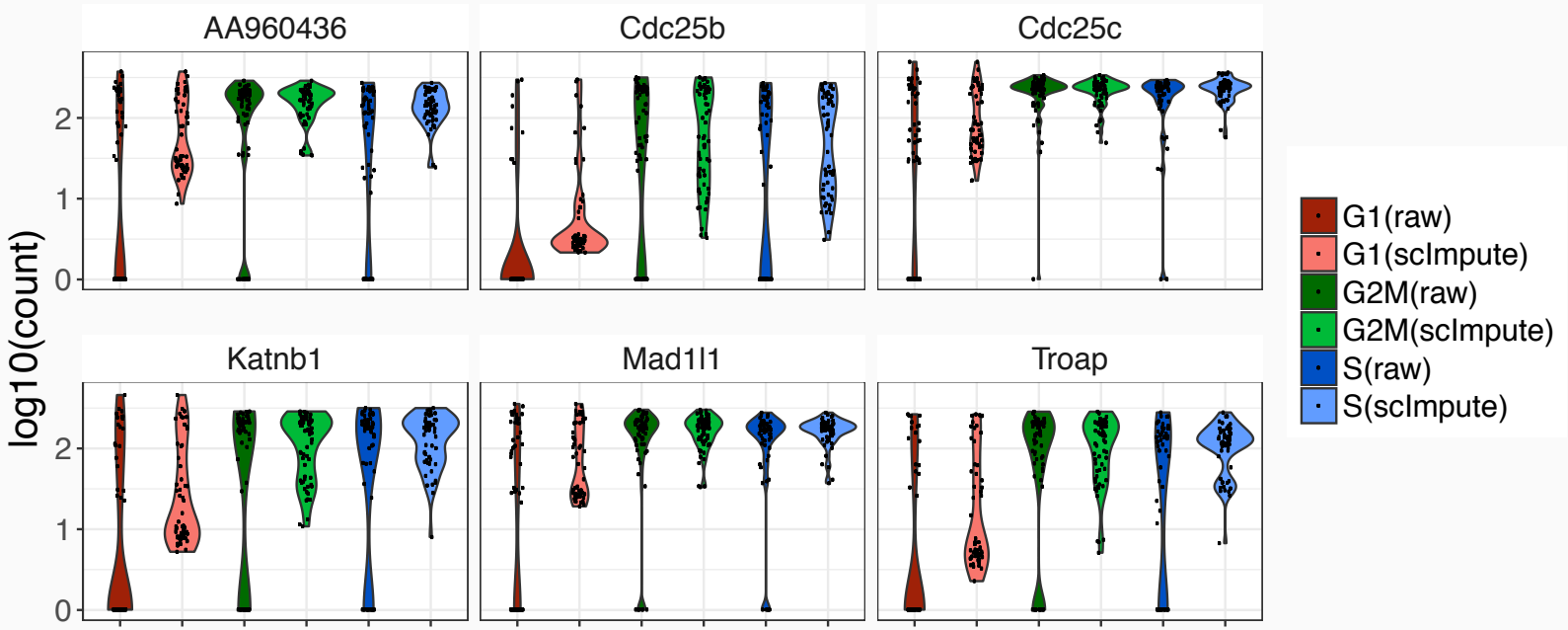
- 3,005 cells from the mouse somatosensory cortex region
- 57 ERCC transcripts



scImpute Recovers the Dropout Events

scImpute correctly imputes the dropout values 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)



scImpute Recovers the Dropout Events

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

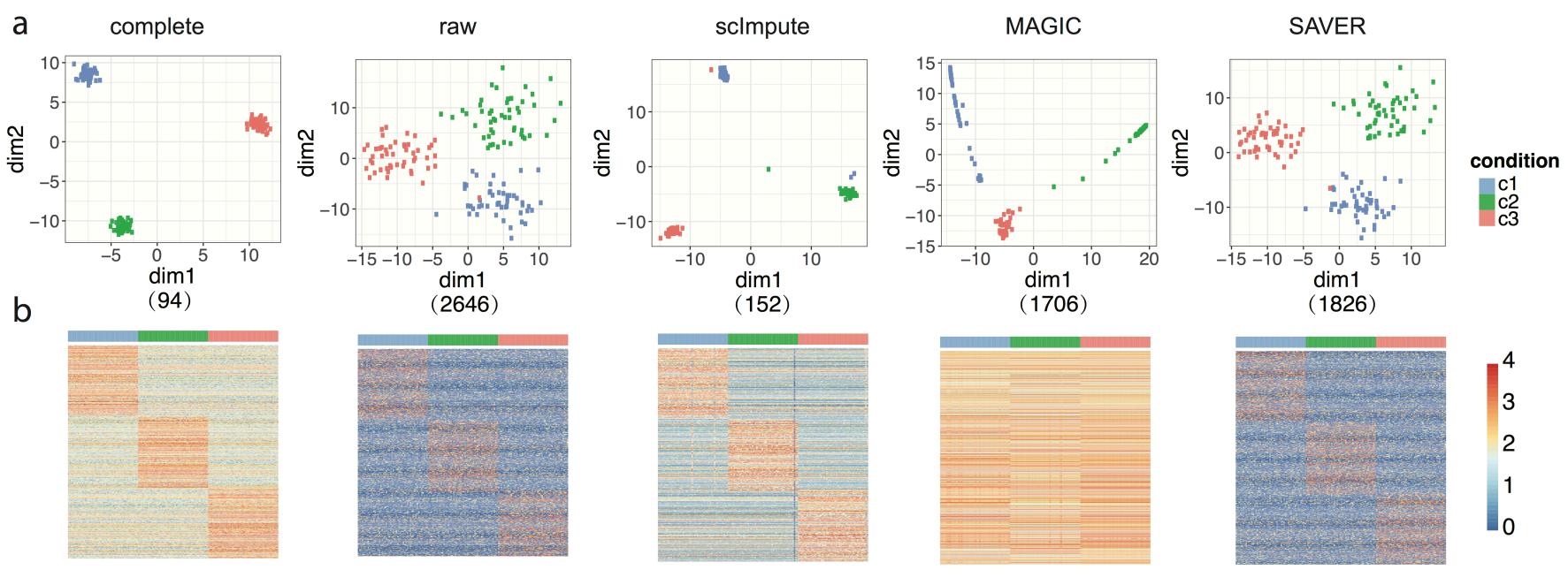
Procedures

- 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]

scImpute Recovers the Dropout Events



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

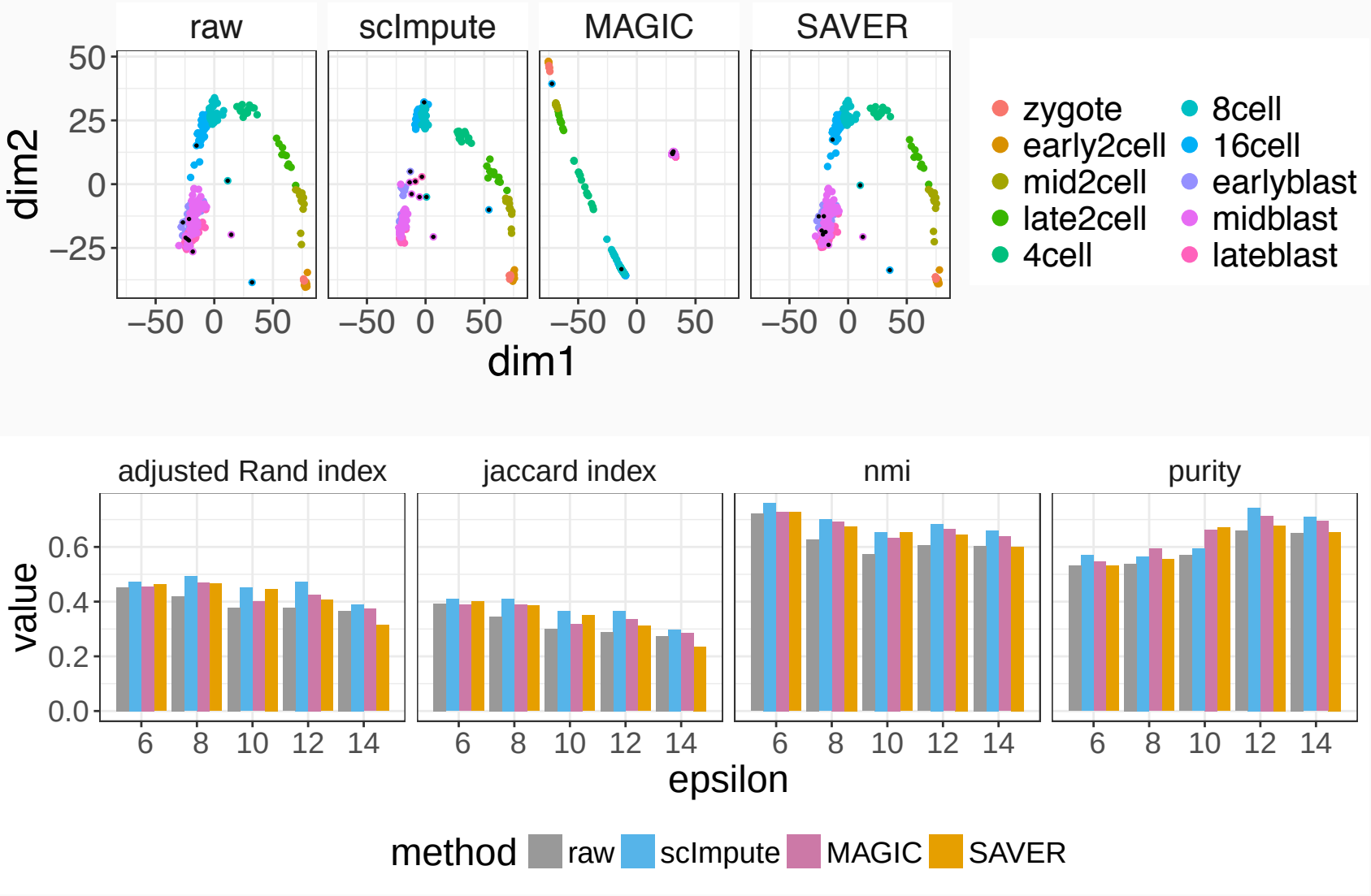
scImpute Helps Define Cell Types in Real Data

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 0

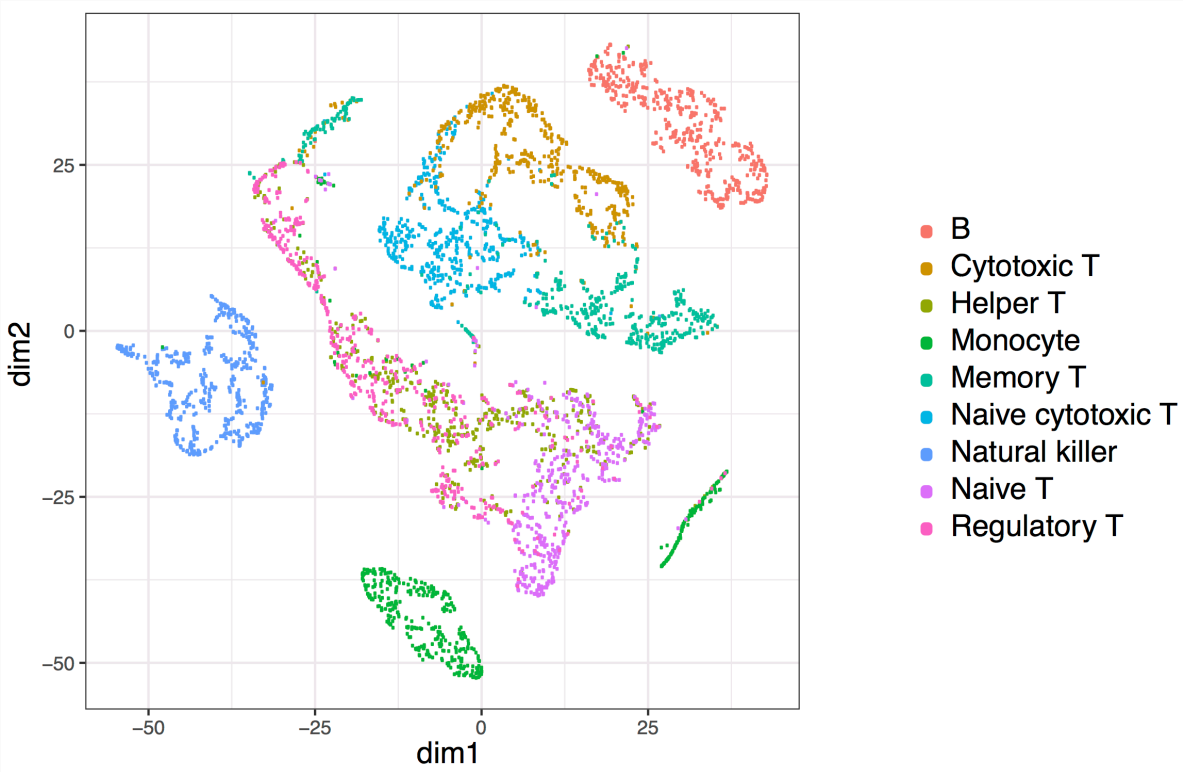
scImpute Helps Define Cell Types in Real Data



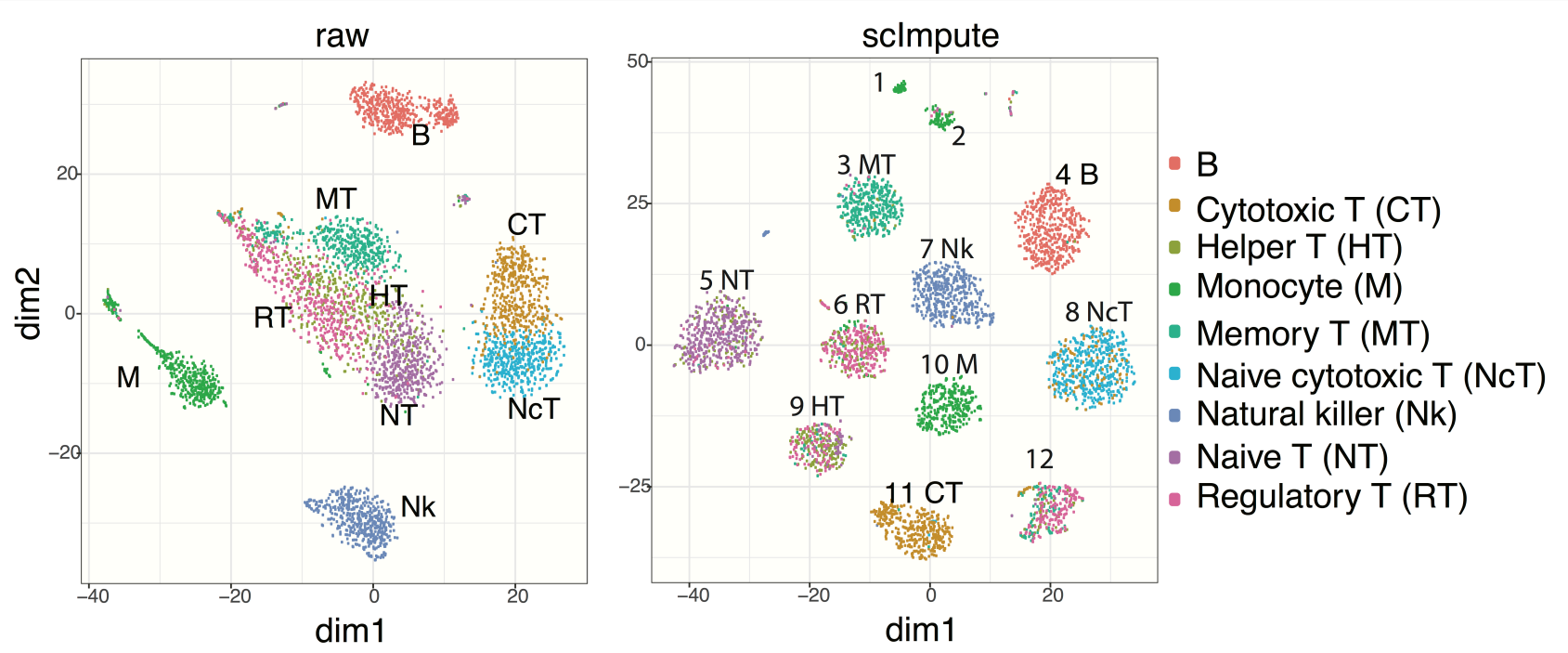
scImpute Helps Define Cell Types in Real Data

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%



scImpute Helps Define Cell Types in Real Data



The first two dimensions of the t-SNE results calculated from raw and imputed PBMC dataset.

scImpute Assists Differential Gene Expression Analysis

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)

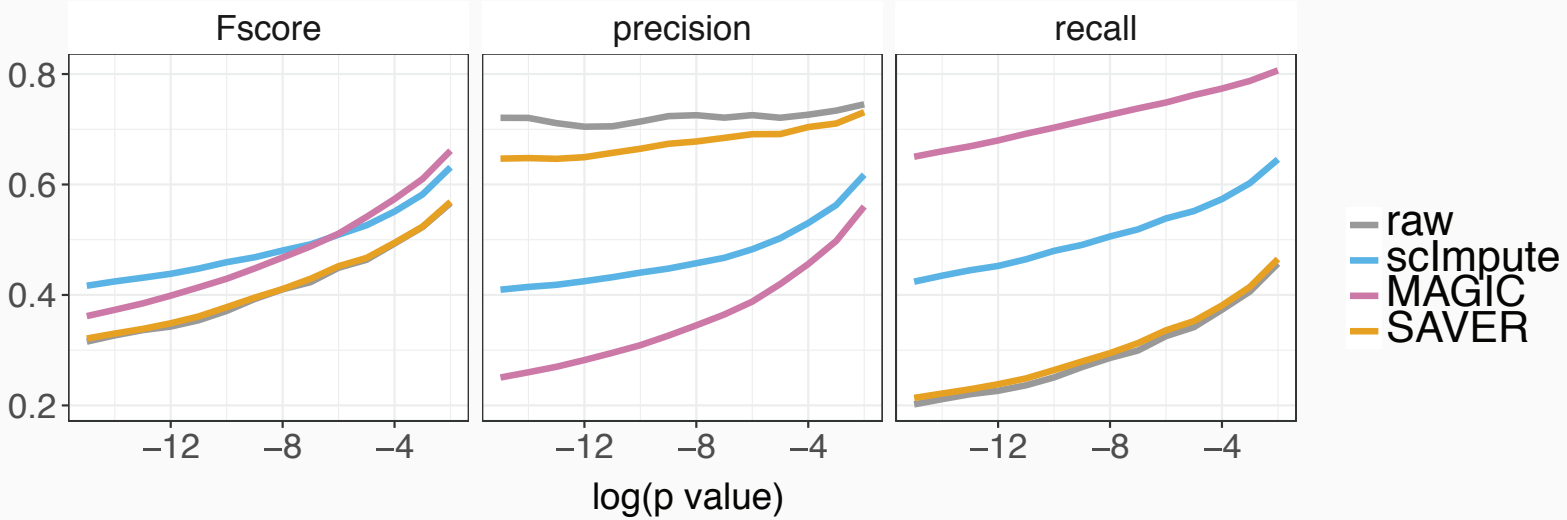
The percentage of zero gene expression

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

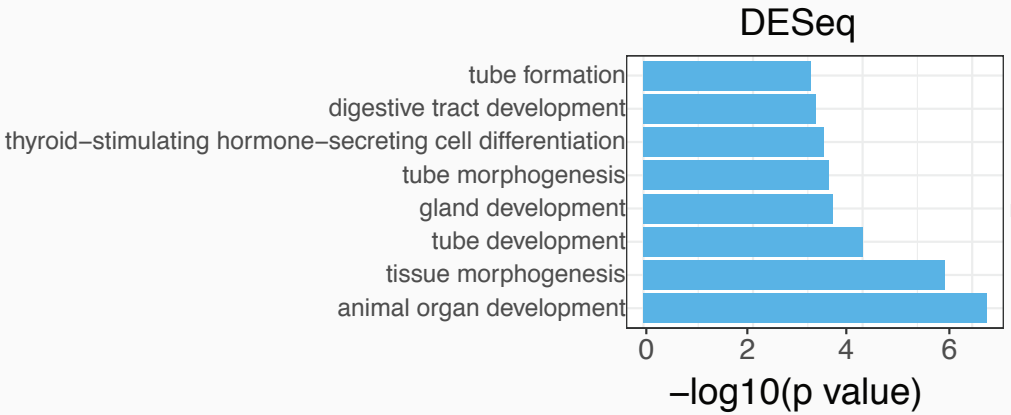
Differentially expressed (DE) genes are identified using `DESeq2` and `MAST`

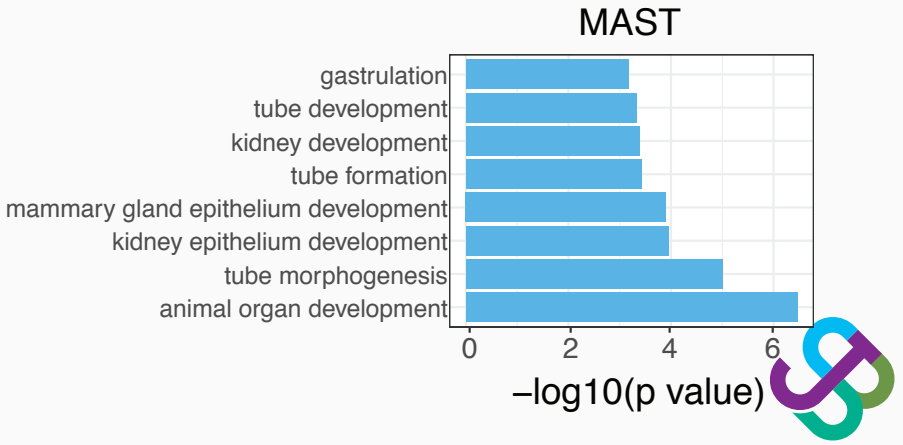
scImpute Assists Differential Gene Expression Analysis

Differential gene expression analysis



Gene ontology enrichment analysis





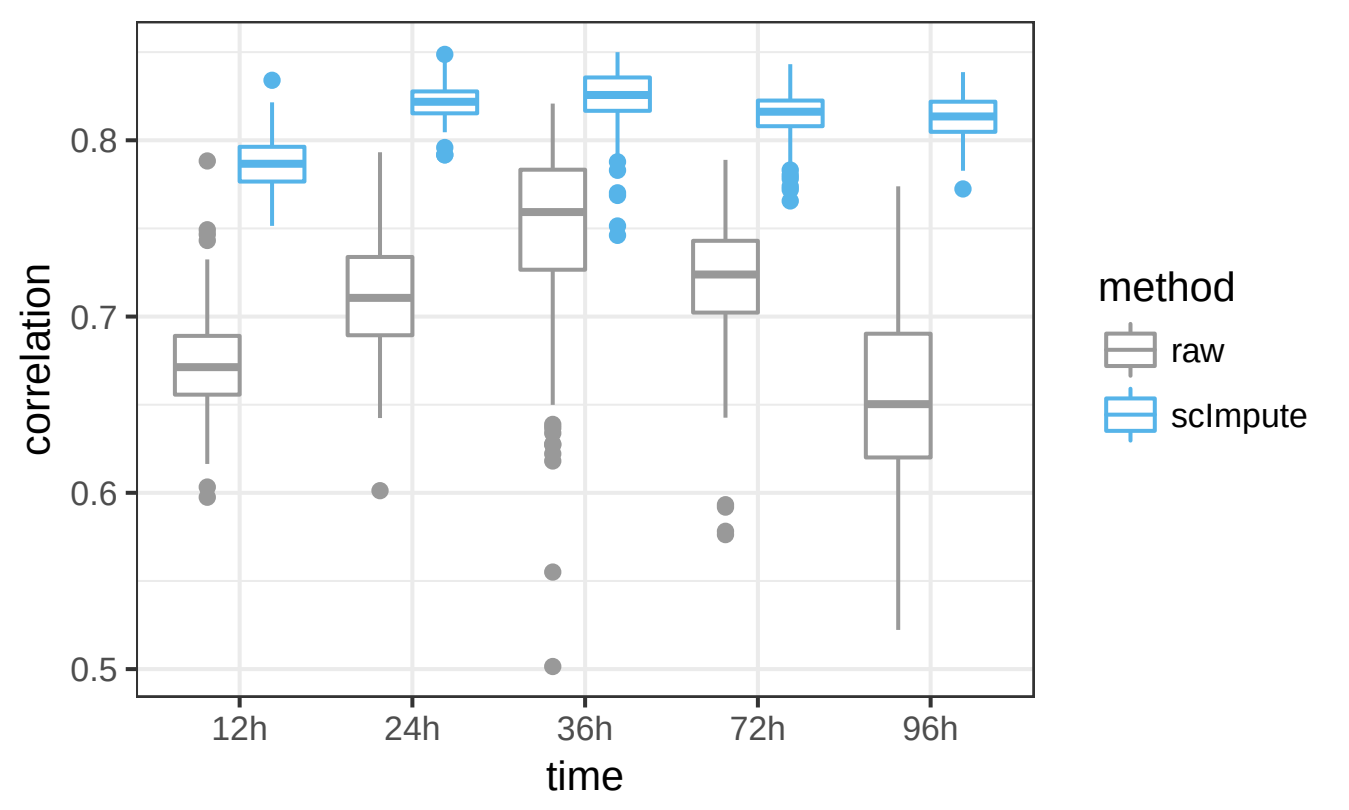
scImpute Assists Pattern Recognition in Timecourse scRNA-seq Data

Bulk and single-cell time-course RNA-seq data profiled at 0, 12, 24, 36, 72, and 96 h of differentiation during DEC emergence [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

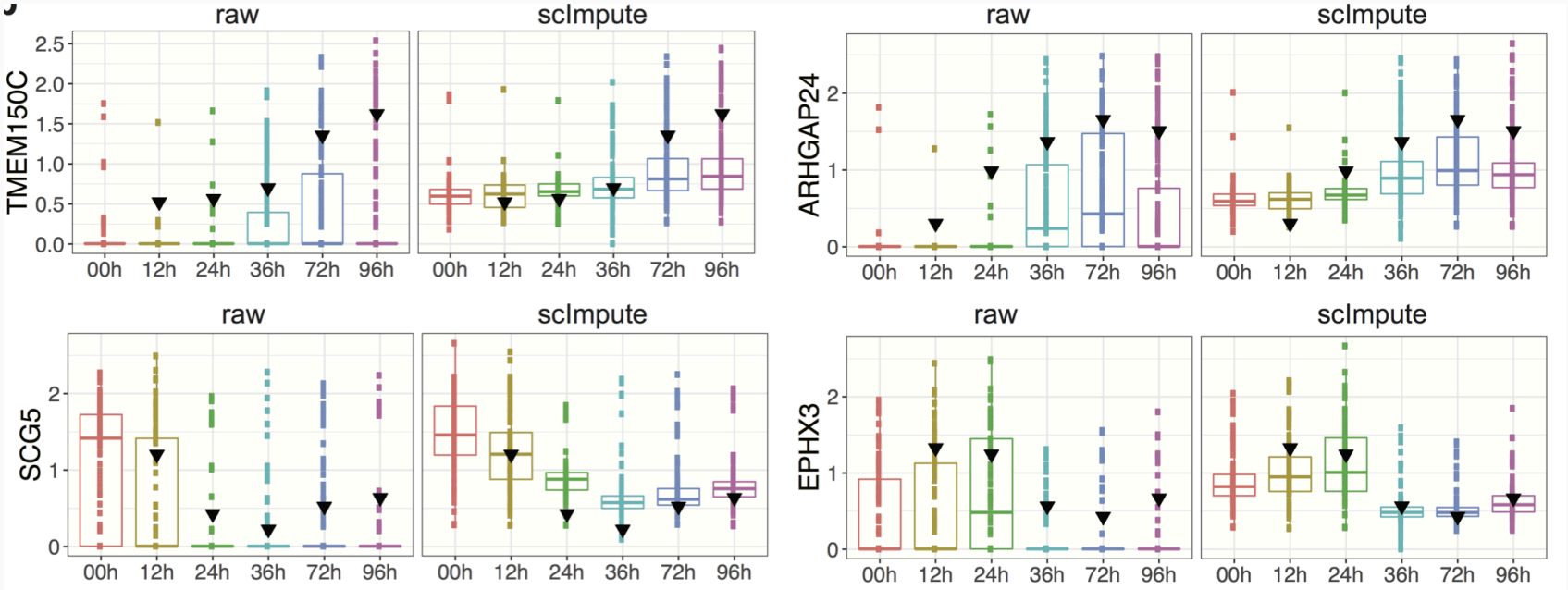
scImpute Assists Pattern Recognition in Timecourse scRNA-seq Data

Correlation between gene expression in single-cell and bulk data



scImpute Assists Pattern Recognition in Timecourse scRNA-seq Data

Imputed read counts reflect more accurate transcriptome dynamics along the time course.



Paper and Software

scImpute: Accurate And Robust Imputation For Single Cell RNA-Seq Data

by Wei Vivian Li and Jingyi Jessica Li

`https://doi.org/10.1101/141598`

(accepted by *Nature Communications*)

R package `scImpute`

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