Supplementary to "scRMD: Imputation for single cell RNA-seq data via robust matrix decomposition"

Chong Chen, Changjing Wu, Linjie Wu, Yishu Wang,
Minghua Deng and Ruibin Xi

1 Online Methods

1.1 Data Pre-processing

The data pre-processing step is similar to the methods used in the literature(Wang *et al.*, 2017; Li and Li, 2018). Let $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ be the raw count matrix for single cell RNA-seq data with rows representing genes and columns representing cells. We normalize the count matrix by calculating the log10 reads per million plus 1. Specifically, denote by $\boldsymbol{Y} = (y_{ij})_{p\times n}$ the normalized matrix, we set

$$y_{ij} = \log_{10}\left(\frac{x_{ij}}{\sum_{k=1}^{p} x_{kj}} \times 1,000,000 + 1\right), \quad i = 1,\ldots,p, j = 1,\ldots,n. \tag{1}$$

We also remove genes that are expressed ($y_{ij} > 0$) in less than 5% of the cells.

1.2 The ADMM algorithm

First, we have the objective function

$$\begin{array}{rl} \min_{\boldsymbol{L},\boldsymbol{S}} & \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{L} + \boldsymbol{S}\|_F^2 + \lambda\|\boldsymbol{L}\|_* + \tau\|\boldsymbol{S}\|_1, \\ \text{subject to} & \mathcal{P}_{\Omega}(\boldsymbol{S}) \geq \boldsymbol{0}, \mathcal{P}_{\Omega^c}(\boldsymbol{S}) = \boldsymbol{0}, \boldsymbol{L} \geq \boldsymbol{0}, \end{array} \tag{2}$$

By introducing an auxiliary variable $\boldsymbol{Z}$ that satisfies $\boldsymbol{Z} = \boldsymbol{L}$, we write the Lagrange function as

$$L(\boldsymbol{L},\boldsymbol{S},\boldsymbol{Z};\boldsymbol{\Lambda}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Z} + \boldsymbol{S}\|_F^2 + \lambda\|\boldsymbol{L}\|_* + \tau\|\boldsymbol{S}\|_1 + \langle\boldsymbol{\Lambda}, \boldsymbol{Z} - \boldsymbol{L}\rangle + \frac{\rho}{2}\|\boldsymbol{Z} - \boldsymbol{L}\|_F^2,$$

where $\boldsymbol{\Lambda}$ is the Lagrange multiplier, $\langle\cdot,\cdot\rangle$ stands for inner product and ρ is a tuning parameter whose choice does not affect convergence of the algorithm. Now the ADMM iterates as follows

$$(\boldsymbol{Z}^{k+1}, \boldsymbol{S}^{k+1}) := \arg\min_{\boldsymbol{Z},\boldsymbol{S}} L(\boldsymbol{L}^k, \boldsymbol{S}, \boldsymbol{Z}; \boldsymbol{\Lambda}^k) \quad \text{s.t. } \mathcal{P}_\Omega(\boldsymbol{S}) \geq \boldsymbol{0}, \mathcal{P}_{\Omega^c}(\boldsymbol{S}) = \boldsymbol{0}, \boldsymbol{Z} \geq \boldsymbol{0}. \tag{3}$$

$$\boldsymbol{L}^{k+1} := \arg\min_{\boldsymbol{L}} L(\boldsymbol{L}, \boldsymbol{S}^{k+1}, \boldsymbol{Z}^{k+1}; \boldsymbol{\Lambda}^k). \tag{4}$$

$$\boldsymbol{\Lambda}^{k+1} := \boldsymbol{\Lambda}^k + \rho(\boldsymbol{Z}^{k+1} - \boldsymbol{L}^{k+1}).$$

As the update of $\boldsymbol{\Lambda}$ is trivial, we next elaborate on methods for solving (3) and (4). To update $\boldsymbol{Z}$ and $\boldsymbol{S}$, we rewrite (3) as

$$\min_{\boldsymbol{Z},\boldsymbol{S}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Z} + \boldsymbol{S}\|_F^2 + \tau\|\boldsymbol{S}\|_1 + \frac{\rho}{2}\|\boldsymbol{Z} - \boldsymbol{L}^k + \boldsymbol{\Lambda}^k/\rho\|_F^2 \tag{5}$$
$$\text{s.t. } \mathcal{P}_\Omega(\boldsymbol{S}) \geq \boldsymbol{0}, \mathcal{P}_{\Omega^c}(\boldsymbol{S}) = \boldsymbol{0}, \boldsymbol{Z} \geq \boldsymbol{0}.$$

For notational simplicity, denote by $\boldsymbol{S}_\Omega = (s_{ij} : (i,j) \in \Omega)^T$, then we have $\boldsymbol{S}_{\Omega^c}^{k+1} = \boldsymbol{0}$, hence

$$\boldsymbol{Z}_{\Omega^c}^{k+1} = \max\left\{\frac{1}{1+\rho}(\boldsymbol{Y}_{\Omega^c} + \rho\boldsymbol{L}_{\Omega^c}^k - \boldsymbol{\Lambda}_{\Omega^c}^{k+1}), 0\right\}, \tag{6}$$

where the max operator should be understood as taking maximum entrywisely. To update elements on Ω, for a fixed $\boldsymbol{Z}_\Omega^{k+1}$, $\boldsymbol{S}_\Omega^{k+1}$ can be obtained by taking the sub-differential of the ℓ_1 norm, which implies

$$\boldsymbol{S}_\Omega^{k+1} = \max\{\boldsymbol{Z}_\Omega^{k+1} - \boldsymbol{Y}_\Omega - \tau, 0\}. \tag{7}$$

Therefore for $(i,j) \in \Omega$, we have

$$\frac{1}{2}(y_{ij} - z_{ij}^{k+1} - s_{ij}^{k+1})^2 + \tau|s_{ij}^{k+1}| = \begin{cases} \frac{1}{2}(z_{ij}^{k+1} - y_{ij})^2, & \text{if } z_{ij}^{k+1} - y_{ij} \leq \tau, \\ -\frac{\tau^2}{2} + \tau(z_{ij}^{k+1} - y_{ij}), & \text{if } z_{ij}^{k+1} - y_{ij} > \tau. \end{cases}$$

Combining the last quadratic term of $\boldsymbol{Z}$ in (5), we obtain

$$z_{ij}^{k+1} = \begin{cases} \max\{l_{ij}^k - \lambda_{ij}^k/\rho - \tau/\rho, 0\}, & \text{if } l_{ij}^k - \lambda_{ij}^k/\rho - \tau/\rho - y_{ij} \leq \tau, \\ \max\{(y_{ij} + \rho l_{ij}^k - \lambda_{ij}^k)/(1+\rho), 0\}, & \text{otherwise.} \end{cases} \tag{8}$$

This display, together with (6) and (7), complete the update of $\boldsymbol{Z}$ and $\boldsymbol{S}$. As for the nuclear norm minimization problem (4), its solution is well-known and given by the soft singular thresholding (SVT) algorithm. Specifically, we aim to solve

$$\min_{\boldsymbol{L}} \lambda\|\boldsymbol{L}\|_* + \frac{\rho}{2}\|\boldsymbol{L} - \boldsymbol{Z}^{k+1} - \boldsymbol{\Lambda}^k/\rho\|_F^2.$$

Let $\boldsymbol{Z}^{k+1} + \boldsymbol{\Lambda}^k/\rho = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ be the singular value decomposition (SVD) of $\boldsymbol{Z}^{k+1} + \boldsymbol{\Lambda}^k/\rho$, where $\boldsymbol{U} \in \mathbb{R}^{p\times r}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$ are orthogonal matrices with r being the rank of $\boldsymbol{Z}^{k+1} + \boldsymbol{\Lambda}^k/\rho$, and $\boldsymbol{D} = \text{diag}\{d_1, \ldots, d_r\}$ is a diagonal matrix. Then the solution to (4) is given by

$$\boldsymbol{L}^{k+1} = \text{svt}(\boldsymbol{Z}^{k+1} + \boldsymbol{\Lambda}^k/\rho, \lambda/\rho) := \boldsymbol{U}\tilde{\boldsymbol{D}}\boldsymbol{V}^T, \tag{9}$$

where $\tilde{\boldsymbol{D}}$ is defined as $\tilde{\boldsymbol{D}} = \text{diag}\{\tilde{d}_1, \ldots, \tilde{d}_n\}$ with $\tilde{d}_i = \max\{d_i - \lambda/\rho, 0\}, i = 1, \ldots, r$.

Overall, the ADMM for solving (2) is summarized in Algorithm 1 and its global convergence is guaranteed (Boyd *et al.*, 2011).

Algorithm 1 ADMM for solving (2).

Require: The normalized matrix $\boldsymbol{Y}$ and parameters λ, τ.

Ensure: $\hat{\boldsymbol{L}}$ and $\hat{\boldsymbol{S}}$.

1. Initialize $\boldsymbol{Z}^0, \boldsymbol{L}^0, \boldsymbol{S}^0$ and $\boldsymbol{\Lambda}^0$ with some possible values, and set $k = 0$.
2. **While** not converged **do**
3. Update $\boldsymbol{Z}^{k+1}$ and $\boldsymbol{S}^{k+1}$ as in (6), (7), and (8).
4. Update $\boldsymbol{L}^{k+1}$ as in (9).
5. $\boldsymbol{\Lambda}^{k+1} = \boldsymbol{\Lambda}^k + \rho(\boldsymbol{Z}^{k+1} - \boldsymbol{L}^{k+1})$.
6. **End while**.

1.3 Tuning parameter selection

We propose a heuristic method to choose the tuning parameters λ and τ According to the Gordon's Theorem for Gaussian matrices (Vershynin, 2010), a $p \times n$ matrix $\boldsymbol{A}$ with entries being independent standard normal variables satisfies

$$|\sqrt{p} - \sqrt{n}| \leq \text{E}\phi_{\min}(\boldsymbol{A}) \leq \text{E}\phi_{\max}(\boldsymbol{A}) \leq \sqrt{p} + \sqrt{n},$$

where $\phi_{\min}(\boldsymbol{A})$ and $\phi_{\max}(\boldsymbol{A})$ are the smallest and largest singular values of $\boldsymbol{A}$. A proper choice of λ and τ should be able to suppress the noises while identifying the signals. Therefore, from the Gordons Theorem, we should set λ roughly at the order of the maximum singular value of the noise matrix. For the lasso penalty parameter τ, available statistical theory show that it should be at the order of the standard deviation of the noises(Zhao and Yu, 2006). In the meanwhile, it follows from the above inequalities that the ratio between the singular value and magnitude for a tall noise matrix ($p \gg n$) is about $\sqrt{p}$. Therefore, we first obtain a simple estimate of $\boldsymbol{L}$ by $\hat{\boldsymbol{L}}^0 = svt(\boldsymbol{Y}, \sqrt{p})$. Afterwards, $\boldsymbol{Y} - \hat{\boldsymbol{L}}^0$ can be viewed as an approximate residual noise matrix. We then set $\tau = sd(\boldsymbol{Y} - \hat{\boldsymbol{L}}^0)$ and $\lambda = \sqrt{p} \cdot sd(\boldsymbol{Y} - \hat{\boldsymbol{L}}^0)$, where $sd(\boldsymbol{Y} - \hat{\boldsymbol{L}}^0)$ is an estimate of the standard deviation of the noise, with the aim of eliminating noises in the residual matrix $\boldsymbol{Y} - \hat{\boldsymbol{L}}^0$. Such choices of tuning parameters avoid computational intensive procedures such as the cross validation or information criterion, and enjoys superior performance in practice.

2 Supplementary Methods

2.1 MAGIC

We download the R version of MAGIC from https://github.com/pkathail/magic. We directly use the source code with the parameters as the default parameters.

2.2 scImpute

We download the R version of scImpute form https://github.com/Vivianstats/scImpute. We also directly use the source code with the parameters as the default parameters.

2.3 CIDR

We installed the R package from https://github.com/VCCRI/CIDR. We use the functions with the parameters as the default parameters.

2.4 SIMLR

We installed the R package from https://github.com/BatzoglouLabSU/SIMLR. We use the functions with the parameters as the default parameters.

2.5 PCA

For a given data, PCA reductions are calculated by the function svds in R package "rARPACK" after each row was normalized to zero mean.

2.6 tSNE

tSNE reductions are calculated by the function Rtsne in R package "Rtsne".

2.7 Adjusted Rand Index

Given two labels c and $\hat{c}$, denote $n_{ij} = |\{k|c_k == i, \hat{c}_k == j\}|$, $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$. Then we have

$$ARI(c, \hat{c}) = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}},$$

where $n = \sum_{i,j} n_{ij}$.

3 Data Availability

scRMD is an open source R package available at the GitHub repository https://github.com/ChongC1990/scRMD. The eight single cell data used in the paper are available at Gene Expression Omnibus: Ziegenhain (GSE75790), Chu (GSE75748), Kowalczyk (GSE59114), Usoskin (GSE59739), Ting (GSE51372), Deng (GSE45719), Alles(GSE89164), Karaiskos(GSE95025). The Pollen data is available in the NCBI Sequence Read Archive under accession number SRP041736.

4 Supplementary Table and Figure

Table S1: The Pearson correlations between the imputed ERCC genes expression and their relative concentration levels are first calculated. The p-values comparing correlations from every pairs of imputation algorithms by the two sample Mann-Whitney-Wilcoxon test are shown in the table.

	CELseq	MARSseq	SCRBseq	SmartSeq2	SmartSeq
RAW vs scRMD	3.4e-08	1.5e-11	2.8e-16	1.0e-32	1.2e-19
RAW vs scImpute	2.9e-14	5.7e-08	1.5e-06	2.1e-16	1.2e-07
RAW vs MAGIC	2.9e-22	8.0e-23	1.2e-23	3.4e-52	3.8e-44
scRMD vs scImpute	2.1e-3	1.1e-4	1.4e-11	3.6e-14	4.4e-05
scRMD vs MAGIC	1.7e-23	8.0e-23	0.033	3.8e-49	4.0e-38
scImpute vs MAGIC	1.3e-17	8.0e-23	4.5e-24	1.1e-51	2.1e-42

Table S2: The basic summary statistics of the four data sets used in clustering analysis. Mean zero rates under different perturbation are also listed.

Data Set	N	K	full-data	down-sampling	model-based-dropout	low-expression-gene
Usoskin	622	4	78%	83%	92%	97%
Ting	149	7	58%	80%	89%	80%
Pollen	249	11	51%	60%	76%	75%
Deng	268	10	46%	62%	79%	81%

Table S3: The ratio between the mean within-group and between-group distance of two dimension tSNE for the Usoskin and Ting data.

	raw	raw+scRMD	dropout model	dropout model+scRMD
Usoskin	0.41	0.37	0.39	0.36
Ting	0.23	0.22	0.29	0.17

Table S4: The ratio between the mean of within-group and between-group distance of two dimension tSNE for real data

Data	Raw	scRMD	MAGIC	scImpute
Alles	0.3162	0.3075	0.3774	0.4194
Karaiskos	0.6047	0.5532	0.7144	0.7960
Hrvatin	0.4666	0.3824	0.5992	0.5654

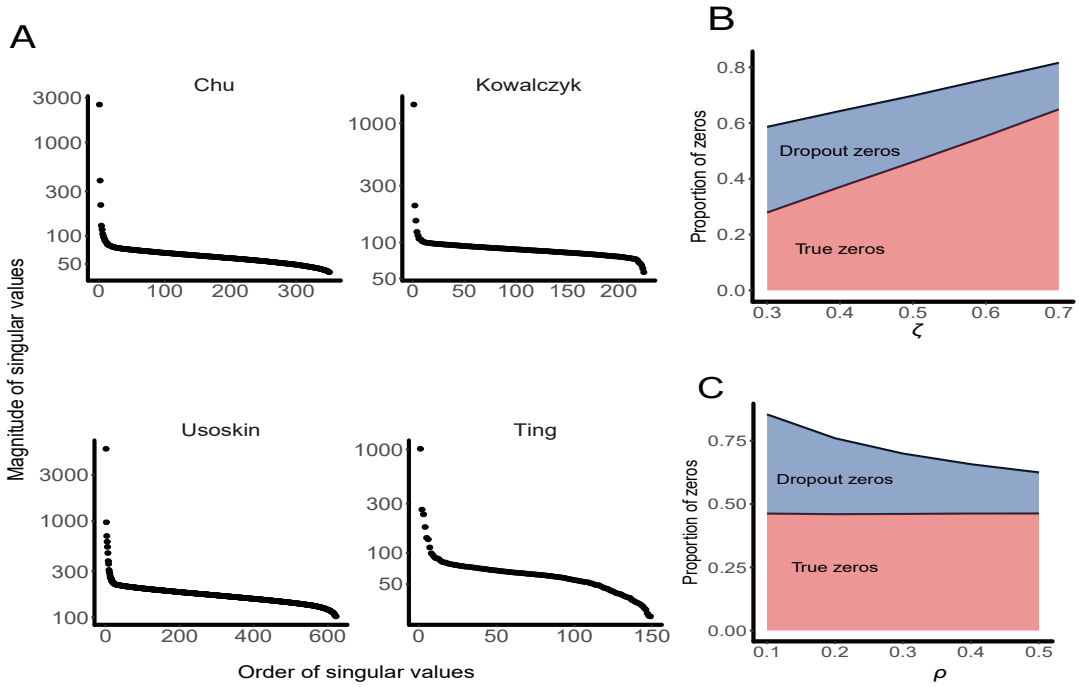

Figure S1: (A) Decay of singular values for the four data sets. Note that the y-axis is plotted in the log scale. (B-C) The proportions of true zeros and dropout in simulation when varying ζ and ρ.

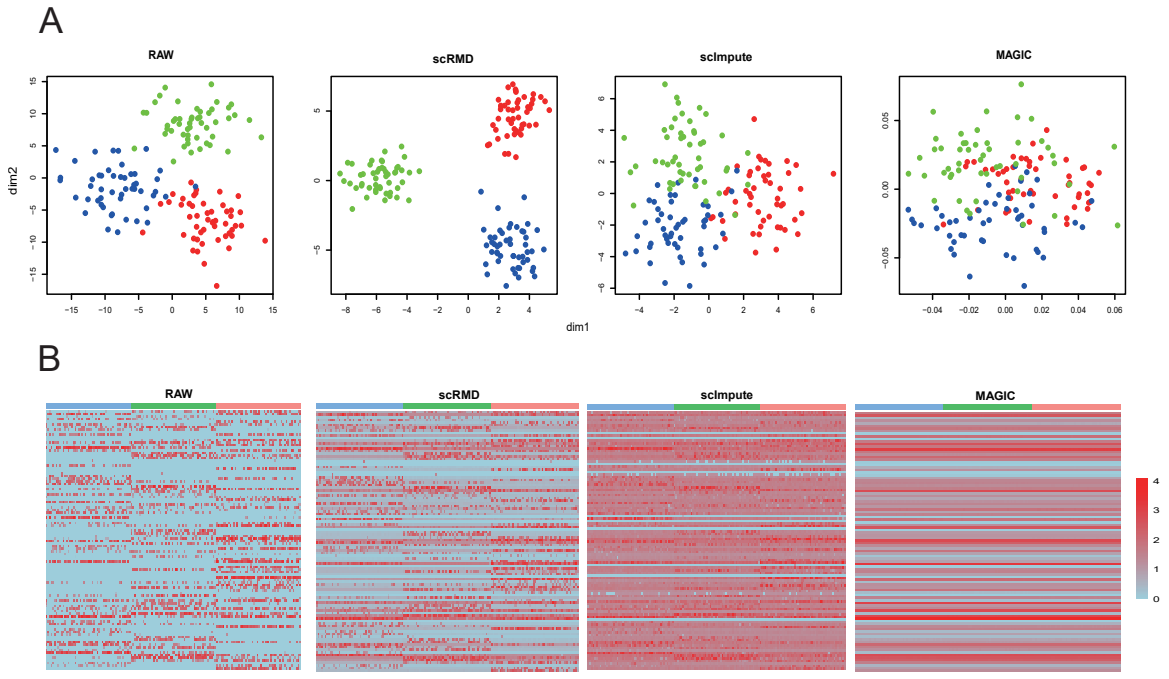

Figure S2: An example of simulation for clustering analysis (A) The visualization of two dimensional PCA of both raw data and imputed data. (B) The heatmap of the differential genes.

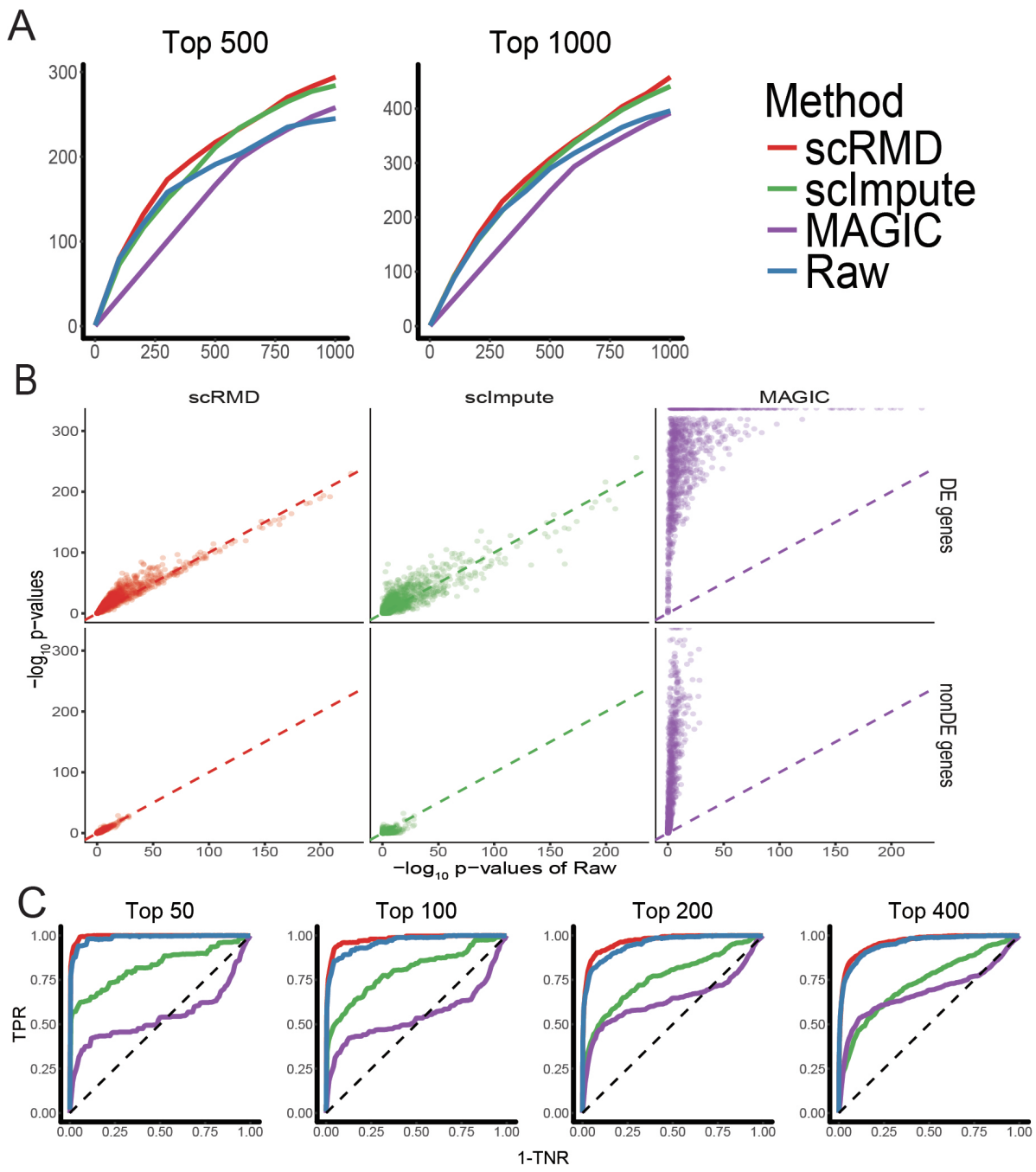

Figure S3: (A) The overlap of single cell DE genes with the golden standard gene sets. The golden standard gene sets are chosen as top 500 and 1,000 genes in the bulk data, respectively. (B) Similar to Figure 4 C, but for p-values given by limma. (C) The ROC curve of each algorithm for DE analysis on 10% subsampled data. The golden standards are chosen as top 50, 100, 200 and 400 genes detected by the corresponding algorithms.

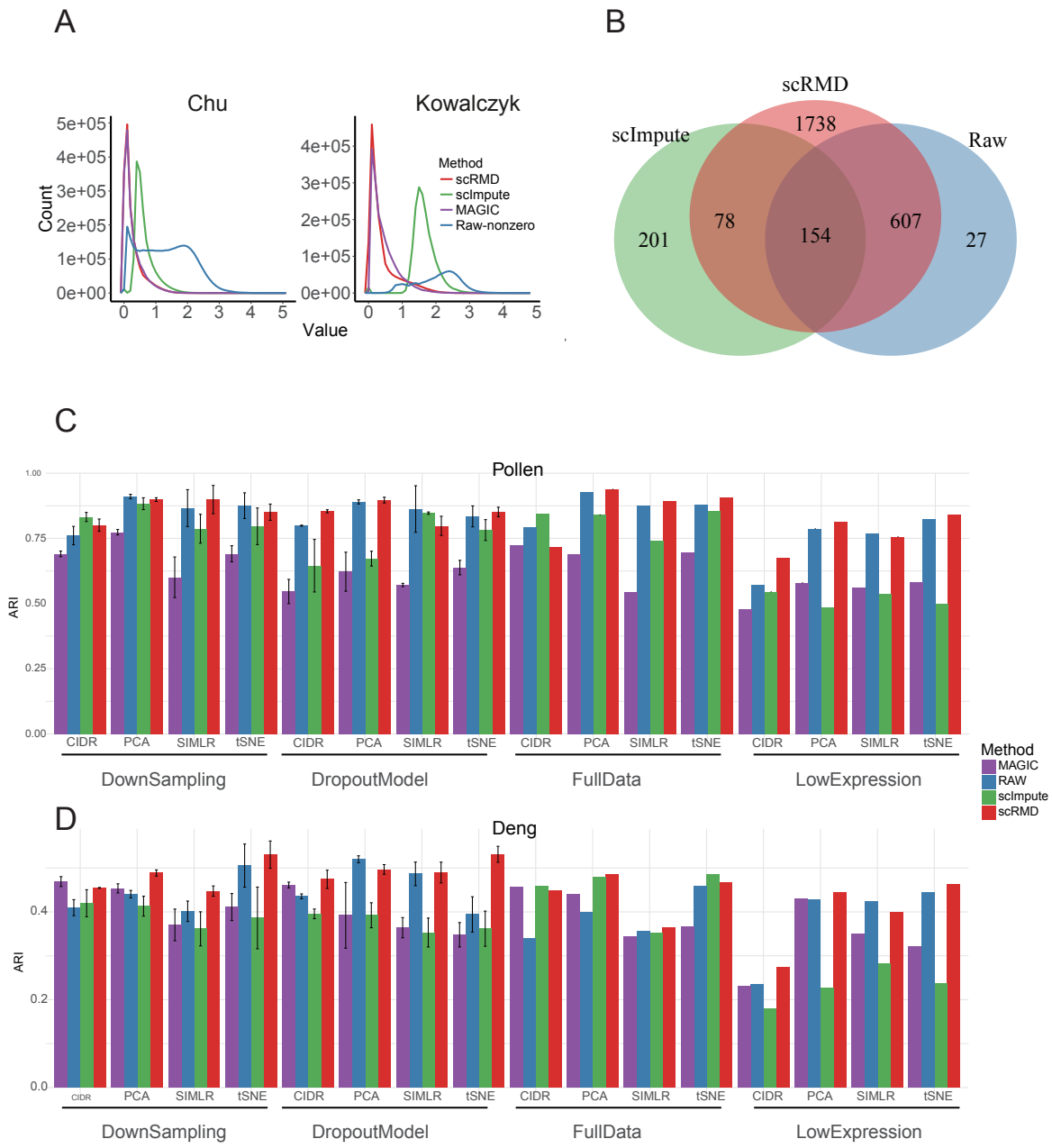

Figure S4: (A) Frequencies of non-zero values in raw single cell RNA-seq data and imputed values of the candidate dropouts by the three methods. (B) Venn diagram for DE genes detected by Raw, scRMD and scImpute. (C,D) Similar to Figure 5 C,D, but for the Pollen and the Deng data.

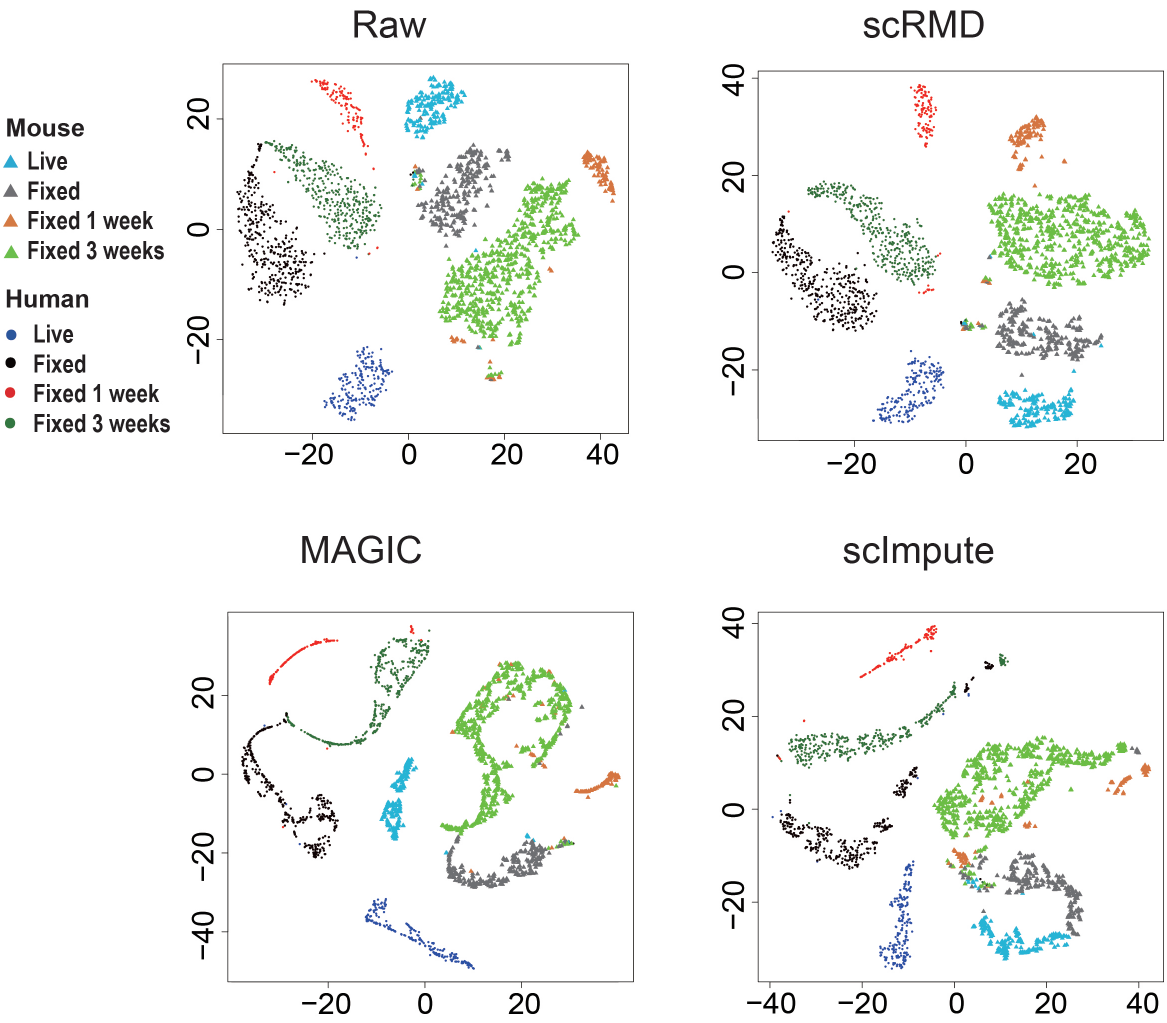

Figure S5: The tSNE visualization of the Alles data.

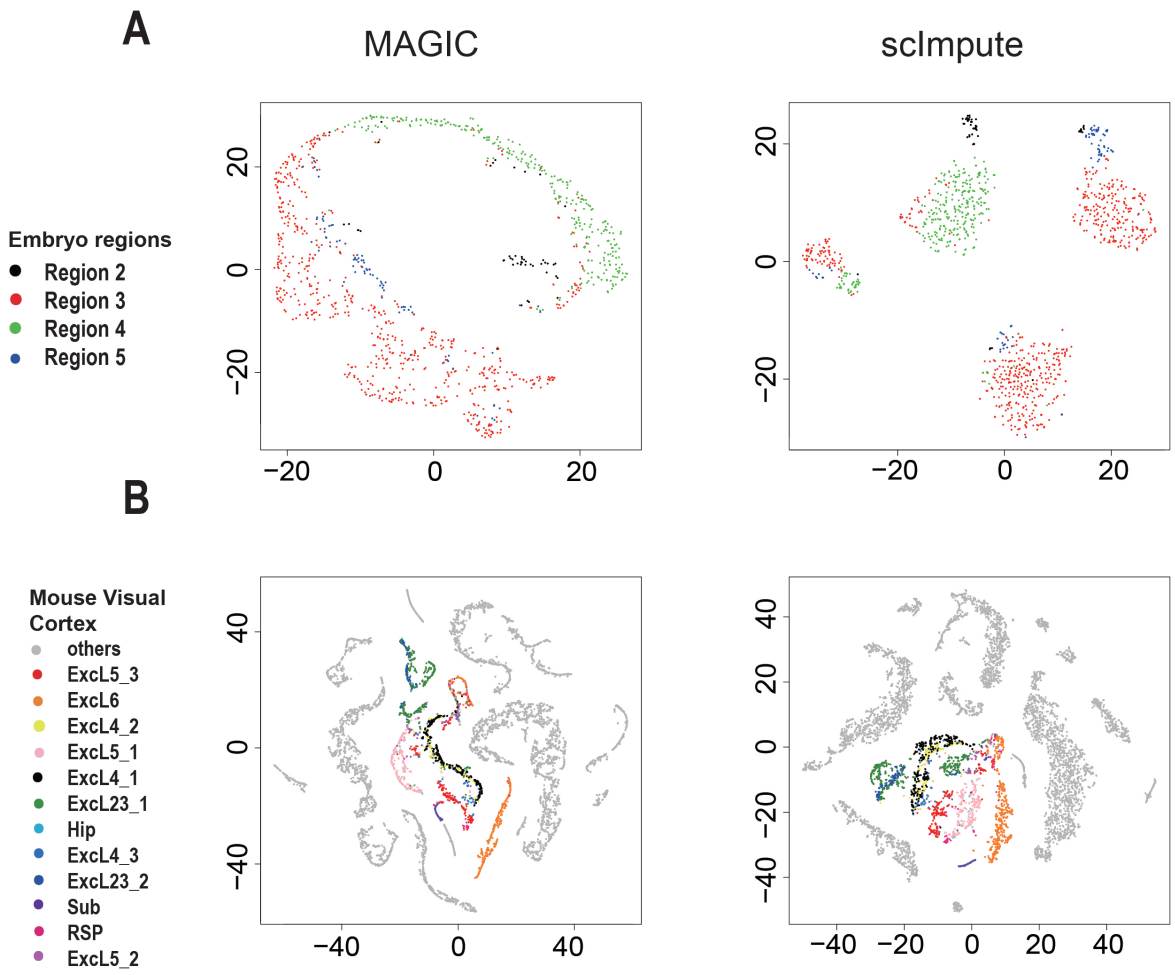

Figure S6: TThe tSNE visualization of the MAGIC and scImpute imputed data. (A) The Karaiskos data set. (B) The Hrvatin data.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, **3**(1), 1–122.

Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, **9**(1), 997.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and

analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods*, **14**(4), 414.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**(Nov), 2541–2563.