WEDGE: recovery of gene expression values for sparse single-cell RNA-seq datasets using matrix decomposition

Yinlei Hu[1,2,4], Bin Li[1,4], Nianping Liu[1], Pengfei Cai[1], Falai Chen[2,*], Kun Qu[1,3,*]

Affiliation:

[1]Division of Molecular Medicine, Hefei National Laboratory for Physical Sciences at Microscale, The CAS Key Laboratory of Innate Immunity and Chronic Disease, Department of Oncology of the First Affiliated Hospital, Division of Life Sciences and Medicine, University of Science and Technology of China

[2]Department of Mathematics, University of Science and Technology of China

[3]CAS Center for Excellence in Molecular Cell Sciences, University of Science and Technology of China, Hefei, China

[4]Cofirst authors

*Correspondence: Kun Qu (qukun@ustc.edu.cn), Falai Chen (chenfl@ustc.edu.cn)

ABSTRACT

The low capture rate of expressed RNAs from single-cell sequencing technology is one of the major obstacles to downstream functional genomics analyses. Recently, a number of recovery methods have emerged to impute single-cell transcriptome profiles, however, restoring missing values in very sparse expression matrices remains a substantial challenge. Here, we propose a new algorithm, WEDGE (WEighted Decomposition of Gene Expression), which imputes expression matrix by using a low-rank matrix decomposition method. WEDGE successfully restored expression matrices, reproduced the cell-wise and gene-wise correlations, and improved the clustering of cells, performing impressively for applications with multiple cell type datasets with high dropout rates. Overall, this study demonstrates a potent approach for recovering sparse expression matrix data, and our WEDGE algorithm should help many researchers to more profitably explore the biological meanings embedded in their scRNA-seq datasets.

INTRODUCTION

Single-cell sequencing technology has been widely used in studies of many biological systems, including embryonic development [1-3], neuronal diversity [4-6] and a large variety of diseases [7-11]. Despite the rapid increase in sequencing throughput, the number of captured genes per cell is still limited by experiment noise [12-14]; hence, the gene expression matrices generated by single-cell sequencing techniques are sparse, and are difficult to use in subsequent analyses [13,15]. To overcome this, a variety of algorithms have be developed to impute the zero elements in the expression matrices [16-19].

For example, MAGIC [19] recovers gene expression by using data diffusion to construct an affinity matrix which attempts to represent the neighborhood of similar cells. Huang et al. combined Bayesian and Poisson LASSO regression methods into

SAVER [18] to estimate prior parameters and to restore missing elements of an expression data matrix, based on the assumption that gene expression follows a negative binomial distribution. Recently, they upgraded this approach to SAVER-X [20] by training a deep autoencoder model with gene expression patterns obtained from public single-cell data repositories. Eraslan et al. developed a deep neuron network model, DCA [17], which can denoise scRNA-seq data by learning gene-specific parameters. Many other tools have also emerged recently, such as SCRABBLE [21], VIPER [22], ENHANCE [23], ALRA [24], and netNMF-sc [25], each of which seeks to improve recovery of the expression matrix for single-cell data. However, for datasets with high dropout rates—which therefore have very sparse expression matrices—it is still a challenge to abundantly recover gene expression data while avoiding over-imputing [13,17,20]. Here, we introduce a new imputation algorithm, WEDGE, to recover gene expression values for sparse single-cell data based on low-rank matrix decomposition [26-28]. To assess this new approach, we applied WEDGE to multiple scRNA-seq datasets and compared its results with existing single-cell sequencing data imputation algorithms.

RESULTS

Algorithm, performance, and robustness of WEDGE

In WEDGE, we adopted a lower weight for the zero elements in the expression matrix during the low-rank decomposition, and generated a convergent recovered matrix by the alternating non-negative least-squares algorithm [29] (Fig. 1a and Methods). Most zero elements in a typical single-cell expression matrix are caused by the low RNA capture rates during experimental sampling and processing [17,20], thus WEDGE used a lower weight parameter ($0 \leq \lambda \leq 1$) for the zero elements during the optimization process compare to the nonzero elements (weight = 1). We chose not to set λ as zero, since the contribution of the zero elements is not completely negligible

as was shown in previous studies, and some of them still represent low expression levels of the corresponding genes [17]. Notably, as WEDGE is a completely unsupervised learning algorithm, it allows us to impute expression data matrices without any prior information about genes or cell types.

To test the performance of WEDGE in restoring gene expression, we first applied it to a simulated dataset generated by Splatter [30] and compared WEDGE against DCA [17], MAGIC [19], and SAVER-X [20] (Fig. 1b). The reference data includes distinct marker genes for 6 different cell types and a dense expression matrix (original sparsity=10%, where sparsity is the percentage of zero elements). We randomly set 44% of the elements in the original expression matrix to zero (*i.e.*, dropout rate=0.44) and obtained a down-sampled matrix, which we refer to as the “observed” data. The dropout events obscured the significance of the differentially expressed (DE) genes, but WEDGE successfully restored their expression patterns, obtaining a recovered matrix apparently similar to the reference matrix, especially for the DE genes in cell type 1.

Also, we adopted the tSNE algorithm to explore the intercellular relationships in two-dimensional space, and used the adjusted random index (ARI) [18,31] to assess the accuracy of the cell clustering results (Fig. 1c), wherein a higher ARI value indicates that the clustering result is relatively closer to the “true” cell types. Using the expression matrix recovered by WEDGE, we can clearly distinguish these cell types. The ARI value of the cell clusters from the WEDGE imputed matrix is 0.99, higher than those from the other three recovery methods.

We further evaluated the robustness of WEDGE by applying it to restore "observed" matrices with different dropout rates (Supplementary Fig. 1). Interestingly, for the observed matrix with a low dropout rate (0.15), all the four methods (WEDGE, DCA, MAGIC, and SAVER-X) successfully recovered the distinctions between the cell types. However, for the data with higher sparsity (*i.e.*, dropout rate>0.44) only WEDGE can still delineate the cell identities, demonstrating the advantage of WEDGE on imputing scRNA profiles with low capture rate.

In addition, to check whether the algorithm leads to over-imputing—for example, erroneously restoring non-DE genes so that they appear as DE genes—we applied WEDGE on another Splatter simulated dataset comprising two cell types, each with 1000 cells, 38 DE genes, and 162 non-DE genes. We then down-sampled the dataset to 42% sparsity ("observed" data), and imputed it with WEDGE, DCA, MAGIC and SAVER-X. We found that WEDGE recovered DE genes drove the principle components of the expression matrix (ARI=0.99, higher than the rest methods) (Supplementary Fig. 2a). On the contrary, the non-DE genes still cannot distinguish cell types after WEDGE imputation (Supplementary Fig. 2b), indicating that WEDGE did not introduce over-imputing.

Recovery performance for real scRNA-seq datasets

To examine the performance of WEDGE on real scRNA-seq data, we applied it to Zeisel's [6] dataset on mouse brain scRNA-seq. We first constructed the reference matrix by extracting all the cells with more than 10000 UMI counts and all the genes detected in more than 40% of cells, and then generated an "observed" matrix with high sparsity by randomly setting a large proportion of the non-zero elements to zeros (dropout rate=0.85). From the heatmaps of gene expression matrices (Fig. 2a), we can see that WEDGE restored the expressions of the DE genes, especially those differentially expressed between interneurons and S1 pyramidal cells.

We also used other tools, including SCRABBLE[21], VIPER[22], ENHANCE [23], ALRA [24], and netNMF-sc [25] to assess the same "observed" matrix (Supplementary Fig. 3a). To quantify the similarity between the reference and recovered expression matrices, we calculated the cell-wise and gene-wise Pearson correlations between them [18], where higher correlation coefficients indicate better recovery performance. For cell-wise correlation coefficients, the WEDGE result (median value=0.81) is the highest among all the tested methods (Fig. 2b). The gene-wise correlation coefficients from WEDGE was also higher than that from the rest of the methods. Moreover, we computed the correlation matrix distances (CMDs) [18,32] between the reference and recovered data, where a lower CMD indicates that the recovered data is closer to the

131 reference data (Fig. 2c). For the matrix generated by WEDGE, the cell-to-cell CMD is
132 0.03 and the gene-to-gene CMD is 0.12, which are each tied for the lowest of all the
133 tested methods. These comparisons together highlight that our WEDGE approach can
134 restore both the cell-cell and gene-gene correlations from sparse single-cell RNA-seq
135 datasets.

136 In the tSNE map of cells, WEDGE can clearly distinguish interneurons, S1
137 pyramidal neurons, and CA1 pyramidal neurons, and the ARI value of 0.56 for the
138 clustering result calculated from its recovered matrix is the highest among all tested
139 recovery methods (Fig. 2d, Supplementary Fig. 3b). In particular, in visualizing the
140 expression of an interneuron marker gene *Gad1* [6] and a S1 pyramidal marker gene
141 *Tbr1*, WEDGE appropriately recovered their expression levels in the corresponding
142 cell types, without overestimating their expressions in other cell types (Fig. 2e,
143 Supplementary Fig. 3c).

144 As another example to confirm the utility of WEDGE, we applied the same
145 procedures described above to Baron's pancreas single-cell dataset [33], and compared
146 WEDGE with other methods for recovering gene expression data. We found that
147 WEDGE recovered the expressions of most of the DE genes, especially those of the
148 ductal and activated stellate cells (Supplementary Fig. 4a). Similarly, the cell-wise and
149 gene-wise Pearson correlation coefficients from WEDGE are both greater than those
150 from any other tested methods, emphasizing its strong recovery performance
151 (Supplementary Fig. 4b). Moreover, WEDGE's cell-to-cell and gene-to-gene CMDs of,
152 respectively, 0.02 and 0.11 were each the lowest for any of the tested methods
153 (Supplementary Fig. 4c). Finally, in terms of cell clustering, WEDGE clearly classified
154 alpha, beta, delta, ductal, acinar, and gamma cells, with an ARI value of 0.80, higher
155 than those from all the other methods (Supplementary Fig. 4d).

156 **Scalability and efficiency**

157 Lastly, to assess the computer resources that WEDGE spent on variously sized
158 datasets, we applied it to recover expression data matrices from datasets comprising

different numbers of cells (from 5000 to 1000000 cells) sampled from a very large dataset available from the mouse brain atlas project (see Methods). We found that the runtime of WEDGE increased linearly with the number of cells, and its speed is very close to DCA and MAGIC (Supplementary Fig. 5). For the dataset comprising 2000 genes and 1 million cells, WEDGE finished the imputation of missing values in an average of 24 minutes on a computer with 28 cores and 128 GB memory. In addition, for the convenience of other researchers, we have uploaded WEDGE and the datasets used in this study to GitHub (https://github.com/QuKunLab/WEDGE).

Discussion

In this study, we presented a new approach for recovering missing information in single-cell sequencing data, based on the combination of low-rank matrix decomposition and a low weight parameter for the zero elements in the expression matrix. For scRNA-seq datasets with high dropout rates and sparsity, WEDGE effectively recovered the expressions of undetected genes, and substantially promoted our understanding of sparse single-cell profiles, including but not limited to the accuracy of cell clustering. However, it is still challenging to restore information of low population cell subtypes for many existing imputation methods, and further improvements are still required.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2017YFA0102900 to K.Q.) and by National Natural Science Foundation of China grants (81788101, 91640113, 31771428 to K.Q.; 11571338 to F.C.). It was also supported by the Fundamental Research Funds for the Central Universities (to K.Q.) and Anhui Provincial Natural Science Foundation grant BJ2070000097 (to B.L.). We thank the USTC supercomputing center and the School of Life Science Bioinformatics Center for providing computing resources for this project.

Author contributions

K.Q. and F.C. conceived and supervised the project; Y.H. and B.L. designed implemented, and validated WEDGE with the help from N.L. and P.C.; B.L., Y.H. and K.Q. wrote the manuscript with inputs from all the authors.

Competing interests

The authors declare no competing interests.

Figures

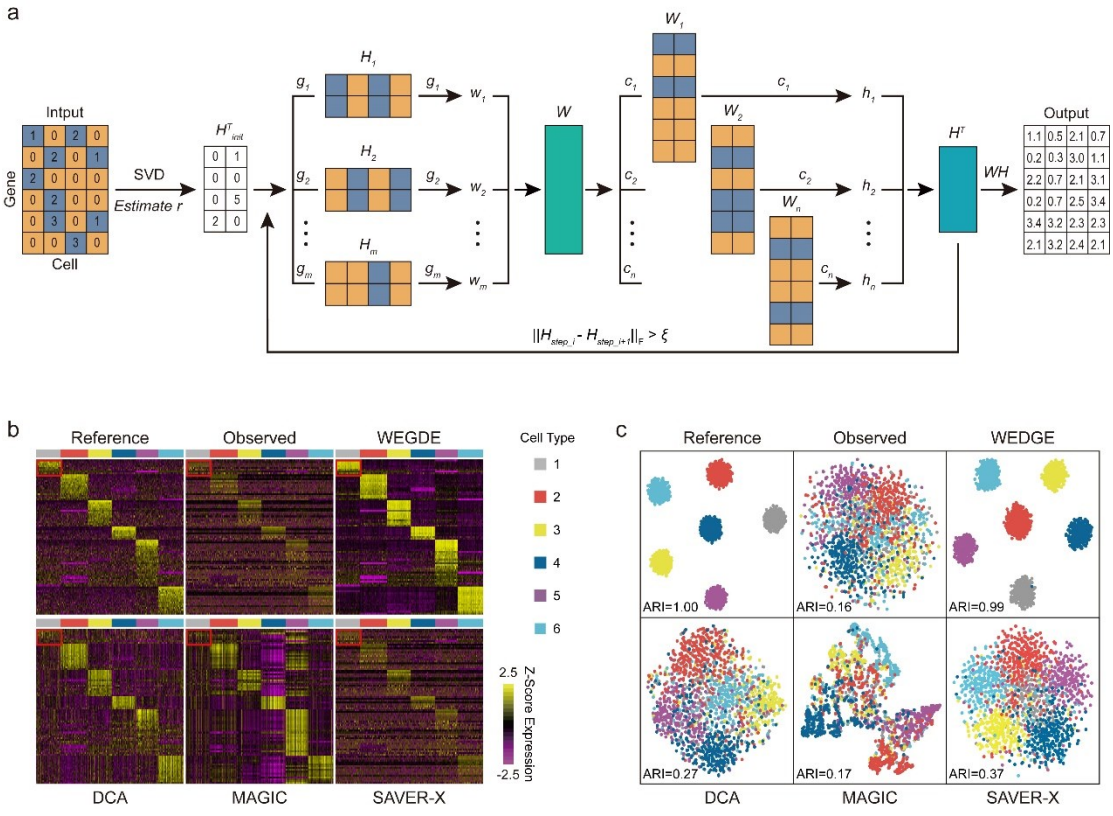


Figure 1. Design of the WEDGE algorithm and its performance for the simulated dataset. (**a**) Conceptual overview of the WEDGE workflow, where $\boldsymbol{H}$ and $\boldsymbol{W}$ are the low-dimensional representations of cells and genes respectively, $\boldsymbol{g}_i$ is the ith gene, $\boldsymbol{c}_j$ is the jth cell, $\boldsymbol{w}_i$ is the ith row of $\boldsymbol{W}$, and $\boldsymbol{h}_j$ is the jth column of $\boldsymbol{H}$. The yellow elements in $\boldsymbol{H}_i$ and $\boldsymbol{W}_j$ correspond to the zero elements in column i and row j of the original expression matrix, respectively, while the steel-blue elements represent the non-zero elements. ξ is the convergence criterion (=1×10^{-5} by default). (**b**) Expression matrices of the top DE genes of the simulated reference and observed data (with dropout rate=0.44), and the results generated by WEDGE, DCA, MAGIC, and SAVER-X. The color bar at the top indicates different cell types. (**c**) tSNE (t-distributed Stochastic Neighbor Embedding) maps of the cells from the expression matrices restored by different methods.

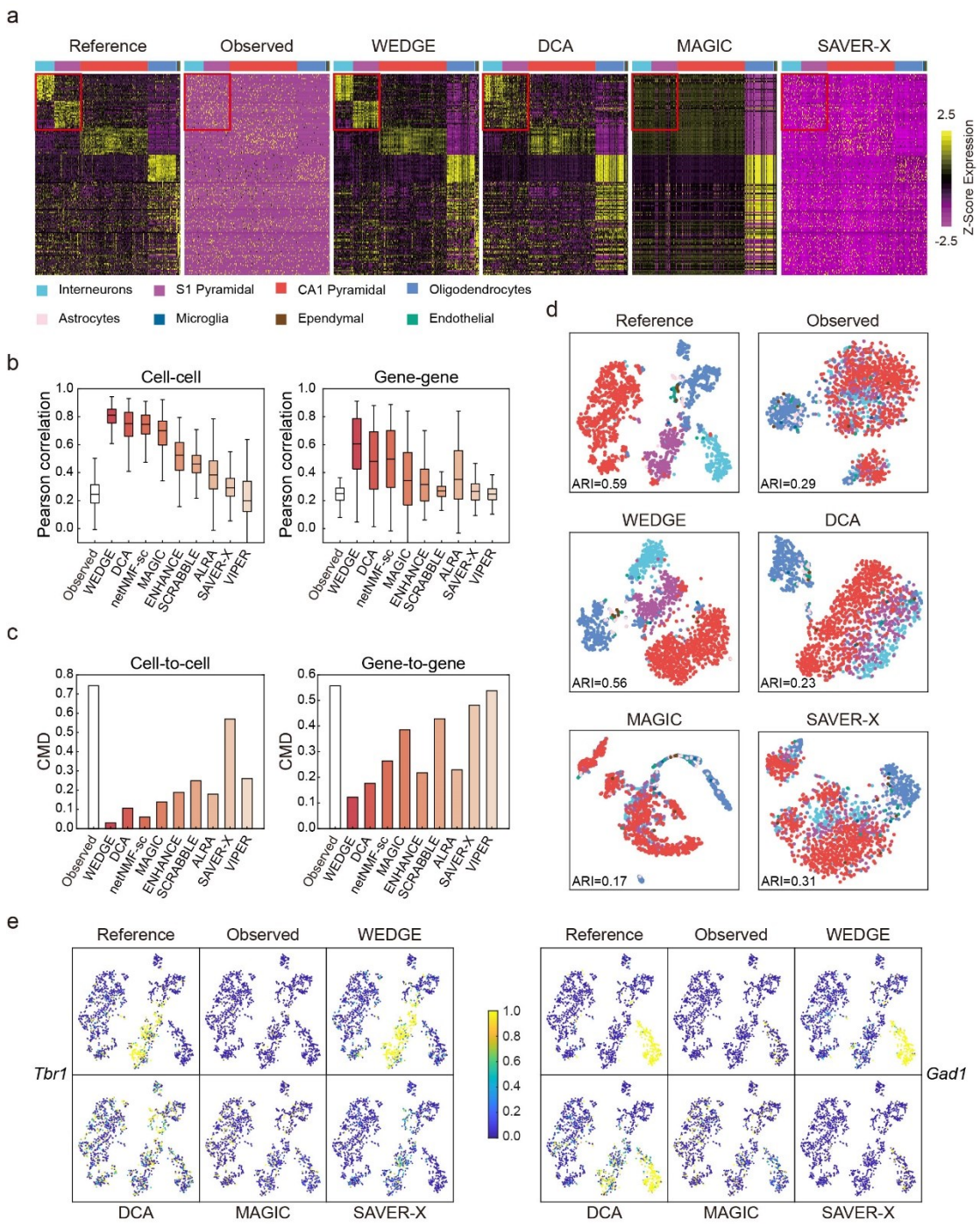


Figure 2. Application and performance assessment of WEDGE for Zeisel's single-cell sequencing dataset, compared to existing methods. (**a**) Visualizations of the expression matrices of the top DE genes of different cell types, including the reference data, the observed data (dropout rate=0.85), and the recovered data generated by four different methods. The color bar at the top indicates known cell types. (**b**) Pearson correlation coefficients between the reference and recovered matrices for cells (left panel) and genes (right panel). Center line, median; box limits, upper and lower

quartiles; whiskers, 1.5x interquartile range. (**c**) Distances of the cell-to-cell (left panel) and gene-to-gene (right panel) correlation matrices between the reference and recovered data. (**d**) 2-D tSNE maps of the cells from the reference, observed, and recovered datasets. The color scheme is the same as in (a). (**e**) Expression of the *Tba1* gene (left panel, a marker of S1 Pyramidal cells) and the *Gad1* gene (right panel, a marker of interneurons) as recovered by different methods, rendered in tSNE space.

Methods

Imputation model of WEDGE. WEDGE takes an expression matrix $\boldsymbol{A}_{m\times n}$ as input, where the element $\boldsymbol{a}_{ij}$ represents the expression of the ith gene in the jth cell. By default, it normalizes the total expression of each cell to 10,000, and updates the expression value by performing a logarithm on it, *i.e.*, $\boldsymbol{a_{ij}} = \log\left(\frac{\boldsymbol{a_{ij}}\times 10000}{\sum_i \boldsymbol{a_{ij}}} + 1\right)$. After imputation, WEDGE outputs two non-negative matrices, $\boldsymbol{W}$ and $\boldsymbol{H}$; we can then take their matrix product as the final recovered expression matrix, $\boldsymbol{V}$. In the WEDGE algorithm, we recovered single-cell sequencing data through the following optimization framework:

$$\begin{aligned} &Obj = \min_{W,H} \sum_{i,j\in\Omega} \left|\boldsymbol{a}_{ij} - \boldsymbol{v}_{ij}\right|^2 + \lambda \sum_{ij\in\bar{\Omega}} \left|\boldsymbol{v}_{ij}\right|^2 \\ &\text{subject to } \boldsymbol{v}_{ij} = \sum_{k=1}^{r} \boldsymbol{w}_{ik}\boldsymbol{h}_{kj}, \boldsymbol{H} \in \mathbb{R}^{+}_{r\times n}, \boldsymbol{W} \in \mathbb{R}^{+}_{m\times r}, \end{aligned} \tag{1}$$

where $\boldsymbol{\Omega} = \{(i,j)|\boldsymbol{a}_{ij} > 0, i = 1\sim m, j = 1\sim n\}$, $\bar{\boldsymbol{\Omega}}$ is the complementary set of $\boldsymbol{\Omega}$ when the universe is $\{(i,j)|i = 1\sim m, j = 1\sim n\}$, $\boldsymbol{v}_{ij}$ is the element of the ith row and jth column of $\boldsymbol{V}$, $\boldsymbol{w}_{ik}$ represents the element of matrix $\boldsymbol{W}$, and $\boldsymbol{h}_{kj}$ is the element of matrix $\boldsymbol{H}$. In this model, the first term in the objective function guarantees minimization of approximation error between the non-zero elements of the original matrix $\boldsymbol{A}$ and the corresponding elements in the recovered matrix $\boldsymbol{V}$, while the second term tends to minimize the elements in $\boldsymbol{V}$ which correspond to zero elements in $\boldsymbol{A}$. We can tune $\lambda \in [0,1]$ to balance the contributions of the two terms of the objective functions. We set $\lambda = 0.15$ for all the datasets used in this study, which is also the default value for WEDGE.

Optimization of the model. In the WEDGE algorithm, the matrix $\boldsymbol{W}$ and $\boldsymbol{H}$ were separately considered, which means that we fixed $\boldsymbol{H}$ to optimize $\boldsymbol{W}$, and then fixed $\boldsymbol{W}$ to generate the new $\boldsymbol{H}$. First, we defined that $\boldsymbol{g}_i$ is the ith row of $\boldsymbol{A}$, $\boldsymbol{H}_i^{+}$ and $\boldsymbol{H}_i^{0}$ are composed of the $\boldsymbol{H}$ columns that correspond to the non-zero and zero elements of $\boldsymbol{g}_i$ respectively, and $\boldsymbol{g}_i^{+}$ is the vector after deleting all zero elements from $\boldsymbol{g}_{\mathrm{i}}$. Then we rewrote the objective function of solving $\boldsymbol{W}$, as,

$$\min_{\boldsymbol{w}_i}\|\boldsymbol{w}_i\boldsymbol{H}_i^+ - \boldsymbol{g}_i^+\|_2^2 + \lambda\left\|\boldsymbol{w}_i\boldsymbol{H}_i^0\right\|_2^2 = \min_{\boldsymbol{w}_i}\left\|\boldsymbol{w}_i\widetilde{\boldsymbol{H}}_i - \boldsymbol{g}_i\right\|_2^2 \tag{2}$$

where $\widetilde{\boldsymbol{H}}_i$ is the combination of $\boldsymbol{H}_i^+$ and $\lambda\boldsymbol{H}_i^0$ according to the original order of their elements in $\boldsymbol{H}$. In this case, optimizing $\boldsymbol{W}$ is equivalent to solving m non-negative least-squares problems in parallel [29]. After $\boldsymbol{W}$ was obtained, we fixed it and solved $\boldsymbol{H}$ using similar algorithm as described above.

Algorithm 1. Optimization of WEDGE

Step1: generate the initial $\boldsymbol{H} \in \mathbb{R}_{r\times n}^+$ from singular value decomposition.

Step2: from a given $\boldsymbol{H}$, solve $\boldsymbol{W}$ in parallel with a non-negative least-square method.

Step3: from the $\boldsymbol{W}$ obtained in step 2, calculate a new $\boldsymbol{H}$.

Step4: iteratively return back to step 2 and 3 until the relative difference in the object function between two adjacent loops is less than 1×10^{-5} or the maximum specified number of iterations is reached.

Estimating the rank of the expression matrix. During the optimization process for WEDGE, we designed a heuristic algorithm to determine the rank of $\boldsymbol{A}_{m\times n}$ based on the relative variation of its singular values $\{\sigma_i: i = 1\sim\min(m,n)\}$. For a descending list of λ_i, we defined a function $f(\sigma_i)$ as

$$f(\sigma_i) = \begin{cases} 0, & if\ \ \frac{\sigma_i-\sigma_{i+1}}{\sigma_{i+1}} \geq \varepsilon\ \ and\ \ \frac{\sigma_{i+p}-\sigma_{i+p+1}}{\sigma_{i+p+1}} < \varepsilon\ \ for\ \ p \in [1,10], \\ 1, & otherwise, \end{cases} \tag{3}$$

where ε is a small non-negative constant (0.085 by default). The following provides the details of the algorithm for the evaluation of rank r:

Algorithm 2: Estimating the rank of the expression matrix

Input: the singular values of matrix $\boldsymbol{A}$: $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_{\min(m,n)}$

Output: the final rank retained for $\boldsymbol{A}$: r

Algorithm:

$r = 1$;

while $f(\sigma_r)$ and $r \leq \min(m,n) - 11$ do

$r = r + 1$;

---

end while

Generation of the simulated scRNA-seq datasets. We used the *splatSimulate()* function in the Splatter R package [30] to generate simulated datasets. For the dataset containing 6 cell types, 500 genes, and 2000 cells (shown in Fig. 1b-c and Supplementary Fig. 1), we set seed=42 and dropout.shape=-1, and its dropout rate was tuned by parameter dropout.mid values ranging from 1 to 6. For the dataset with two cell types, 200 genes, and 2000 cells (shown in Supplementary Fig. 2), we set seed=42, dropout.shape=-1, and dropout.mid=2.

Baron and Zeisel datasets. For Zeisel's dataset of the mouse cortex and hippocampus cells (GSE60361) [6], we generated a reference dataset that contains high quality cells and genes (performed identically with the previously described filtering step of SAVER [18]), and retained all of the marker genes described in the initial Zeisel study (*Tbr1*, *Spink8*, *Aldoc*, *Gad1*, *Mbp*, and *Thy1*). Then, we randomly set 85% of the non-zero elements of the reference data to zeros to generate observed data with dropouts. For Baron's dataset of human pancreatic islet cells (GSE84133) [33], we also used the same process to filter the high quality cells and genes from the original data to build the reference dataset, in which 54% of the elements had non-zero values. We then randomly set 65% of the non-zero elements to zeros to simulate dropout events.

Parameters for other tools. (1) For the application to all datasets in the paper, DCA (version 0.2.2) was performed on the expression matrices with default parameters (type = 'zinb-conddisp', hiddensize = '64,32,64' and learningrate = 0.001). (2) We ran SAVER-X (version1.0.0) on the expression matrices with default parameters. Specifically, we set "data.species = Others, use.pretrain = F" for the simulated datasets, "data.species = Mouse, use.pretrain = T, pretrained.weights.file = mouse_AdultBrain.hdf5, model.species = Mouse" for Zeisel's dataset, and "data.species = Human, use.pretrain = T, pretrained.weights.file = human_Devbrain_nomanno.hdf5, model.species = Human" for Baron's dataset. (3) MAGIC (Python version 1.5.5) was performed with normalized expression matrices (same as WEDGE) and default parameters (decay = 15, t = 'auto', and knn_dist =

291 'euclidean'). Specifically, we set the numbers of principal components to 20 and the
292 number of nearest neighbors to 15. (4) SCABBLE (MATLAB version) was performed
293 on the library size normalized expression matrices with default parameters (parameter
294 = [1,0,1e-4], nIter = 100). (5) VIPER (version 0.1.1) was performed on the expression
295 matrices with default parameters (num = 0.8 times the number of genes,
296 percentage.cutoff = 0.1, minbool = FALSE, and alpha = 0.5). (6) ALRA (RunALRA
297 function in Seurat 3.0 [34,35]) was performed on the expression matrices with default
298 parameters (k = NULL, q = 10). (7) ENHANCE was performed on the expression
299 matrices with default parameters (max-components = 50). (8) netNMF-sc (version 2.0)
300 was run on the expression matrices with default parameters and the outputs were used
301 for downstream analysis. Specifically, we set dimensions = 20 and normalize = False.
302 For the matrix preprocessing step of all methods, we adopted the algorithms
303 recommended by their respective tutorials. For tools that did not describe the
304 preprocessing algorithm, we used the same procedure as WEDGE.

305 **Parameters for cell clustering.** We used Scanpy [36] (version 1.4.0) and the default
306 parameters (n_neighbors = 15 and n_pcs=50 in neighbors() function) to cluster cells
307 for each dataset. Particularly, we set the "resolution" parameter (in louvain() function)
308 to 1 for the first simulated dataset (shown in Fig. 1b-c and Supplementary Fig. 1), to
309 0.5 for the second simulated dataset (shown in Supplementary Fig. 2), and to 0.3 for
310 both of the real datasets (Zeisel's and Baron's datasets), to get the highest ARI value
311 for the cell clustering of each reference matrix. For the clustering of the observed and
312 recovered matrices of a given dataset, we adopted the same parameters as we used
313 for the corresponding reference matrix.

314 **Scalability analysis.** Scalability analysis was performed on a server with two Intel
315 Xeon E5-2680 v4 2.40 GHz CPUs, which contains 28 processers in total. We used the
316 mouse brain atlas dataset of 10X Genomics (https://support.10xgenomics.com/single-
317 cell-gene-expression/datasets/1.3.0/1M_neurons) to construct the benchmark datasets
318 with different number of cells (from 1000 to 1000000). First, we filtered out the genes
319 that were only expressed in three or fewer cells, and normalized the library size of the

dataset. Then, we used the gene filtering function of Scanpy, *i.e.*, scanpy.pp.highly_variable_genes(), with min_mean=0.0125, max_mean=3, min_disp=0.5 and n_top_genes=2000 to obtain the top 2000 most variable genes. With the fixed number of genes, we sampled 1000, 5000, 10000, 100000, 500000, and 1000000 cells from the raw dataset to simulate experiments of different scales.

TSNE and heatmap visualization. (1) Settings for tSNE: for the reference data, the observed data, and the recovered data generated by the different recovery methods, we used the first 20 principal components to perform 2 dimension tSNE analysis. (2) Settings for heatmap: we calculated the Z-score of the expression of each gene, and used Seurat 3.0 [34,35] to extract the top DE genes from the reference matrix. The top 20 DE genes (only.pos = TRUE, min.pct = 0.1, and logfc.threshold = 0.25 in FindAllMarkers() function) of each cell type are shown in the heatmaps for Baron's and Zeisel's datasets (*i.e.*, Fig. 2a, Supplementary Fig. 3a & 4a), and the top 30 DE genes (only.pos = TRUE, min.pct = 0.1, and logfc.threshold = 0.15 in FindAllMarkers() function) for each cell type are shown in the heatmaps of a simulated dataset (*i.e.*, Fig. 1b). For the observed and recovered matrices of each dataset, we present the same DE genes in the same order as in the heatmap of the reference matrix.

References

1 Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).

2 Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593-597, doi:10.1038/nature12364 (2013).

3 Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* **20**, 1131-1139, doi:10.1038/nsmb.2660 (2013).

4 Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590, doi:10.1126/science.aaf1204 (2016).

5 Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80, doi:10.1038/nbt.4038 (2018).

6 Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142 (2015).

7 Zhang, P. *et al.* Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. *Cell reports* **27**, 1934-1947. e1935 (2019).

8 Zhang, L. *et al.* Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* **564**, 268 (2018).

9 Guo, X. *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nature medicine* **24**, 978 (2018).

10 Zheng, C. *et al.* Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342-1356. e1316 (2017).

11 Peng, J. *et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research* **29**, 725-738 (2019).

12 Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature methods* **11**, 637 (2014).

13 Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* **20**, 273-282, doi:10.1038/s41576-018-0088-9 (2019).

14 Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16**, 133 (2015).

15 Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat Methods* **16**, 479-487, doi:10.1038/s41592-019-0425-8 (2019).

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

17 Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* **10**, 390, doi:10.1038/s41467-018-07931-2 (2019).

18 Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* **15**, 539-542, doi:10.1038/s41592-018-0033-z (2018).

19 van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729 e727, doi:10.1016/j.cell.2018.05.061 (2018).

20 Wang, J. *et al.* Data denoising with transfer learning in single-cell transcriptomics. *Nature*

Methods **16**, 875-878, doi:10.1038/s41592-019-0537-1 (2019).

21 Peng, T., Zhu, Q., Yin, P. & Tan, K. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biol* **20**, 88, doi:10.1186/s13059-019-1681-8 (2019).

22 Chen, M. & Zhou, X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol* **19**, 196, doi:10.1186/s13059-018-1575-1 (2018).

23 Wagner, F., Barkley, D. & Yanai, I. Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis. *BioRxiv*, 655365 (2019).

24 Linderman, G. C., Zhao, J. & Kluger, Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv*, 397588 (2018).

25 Elyanow, R., Dumitrascu, B., Engelhardt, B. E. & Raphael, B. J. netNMF-sc: Leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *bioRxiv*, 544346 (2019).

26 Wang, Z. *et al.* Orthogonal rank-one matrix pursuit for low rank matrix completion. *SIAM Journal on Scientific Computing* **37**, A488-A514 (2015).

27 Kim, H. & Park, H. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM journal on matrix analysis and applications* **30**, 713-730 (2008).

28 Kim, Y.-D. & Choi, S. in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing.* 1541-1544 (IEEE).

29 Lawson, C. L. & Hanson, R. J. *Solving least squares problems*. Vol. 15 (Siam, 1995).

30 Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* **18**, 174, doi:10.1186/s13059-017-1305-0 (2017).

31 Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* **14**, 483 (2017).

32 Herdin, M., Czink, N., Ozcelik, H. & Bonek, E. in *2005 IEEE 61st Vehicular Technology Conference.* 136-140 (IEEE).

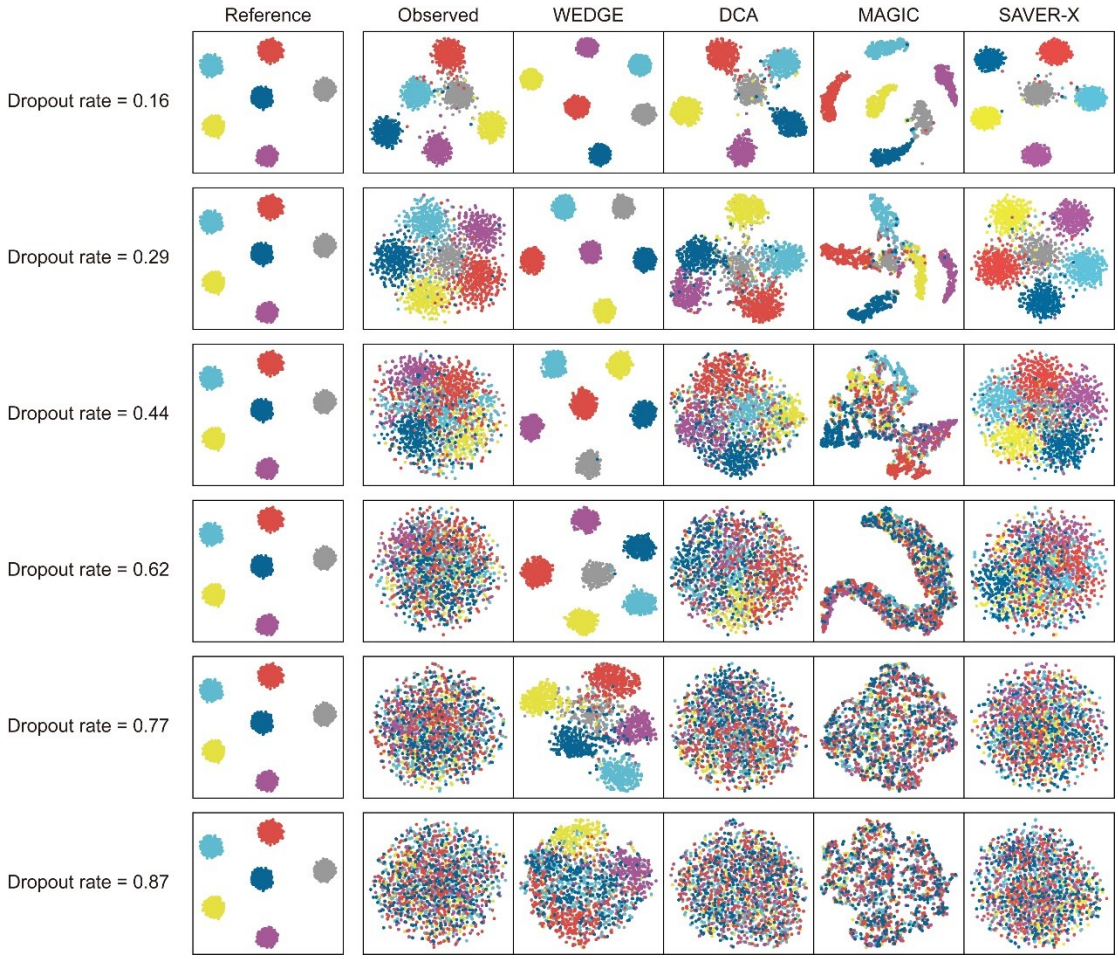
33 Baron, M. *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems* **3**, 346-360. e344 (2016).

34 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411 (2018).

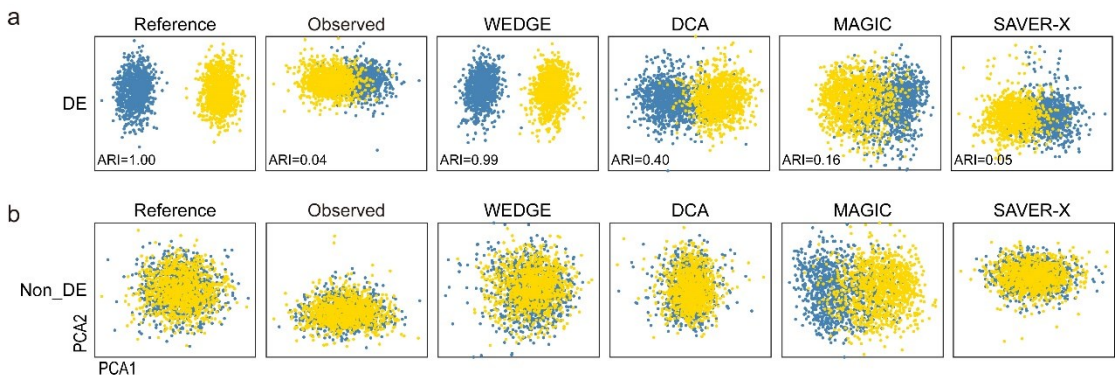
35 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* (2019).

36 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15, doi:10.1186/s13059-017-1382-0 (2018).

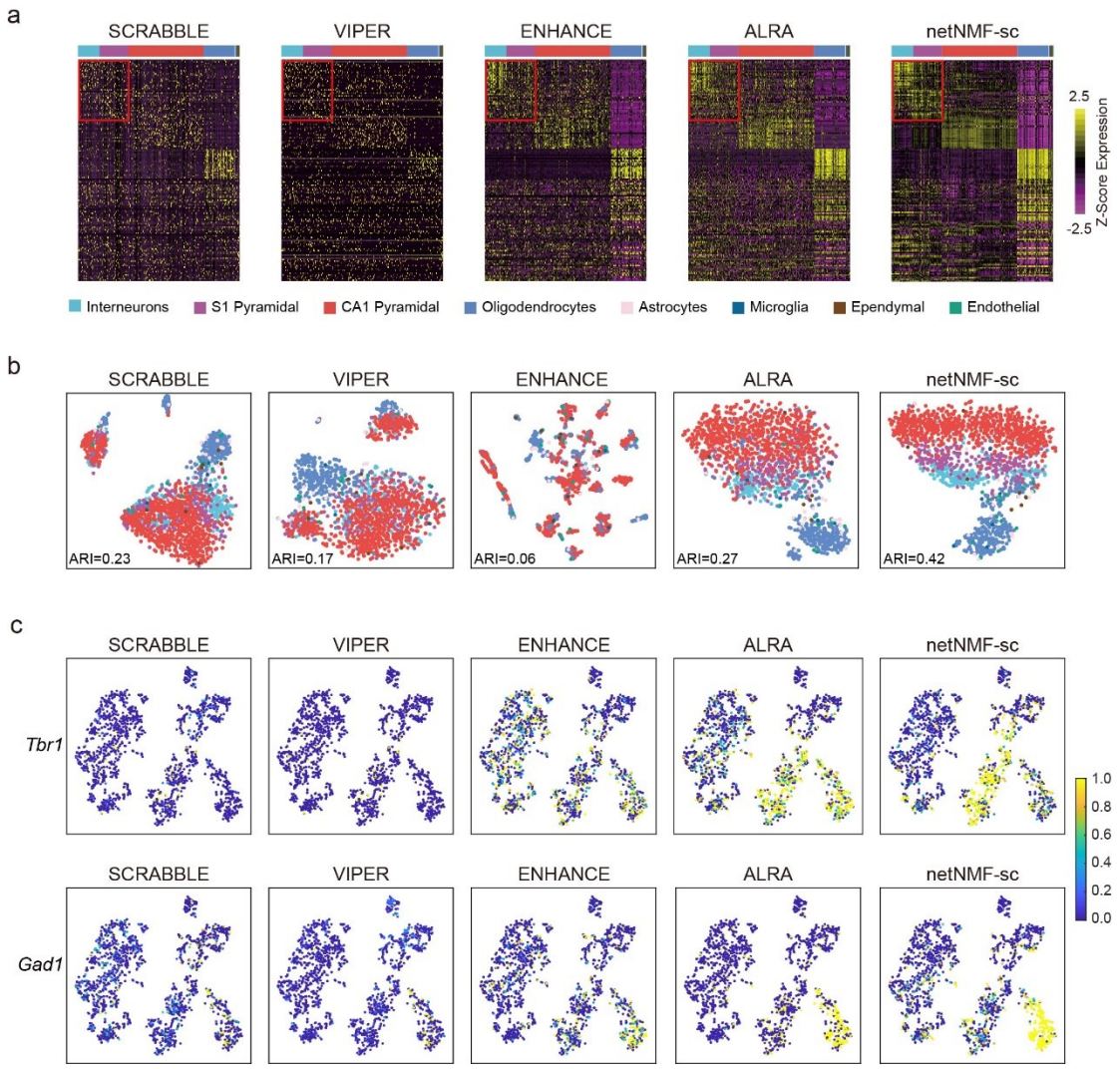
Supplementary Figures



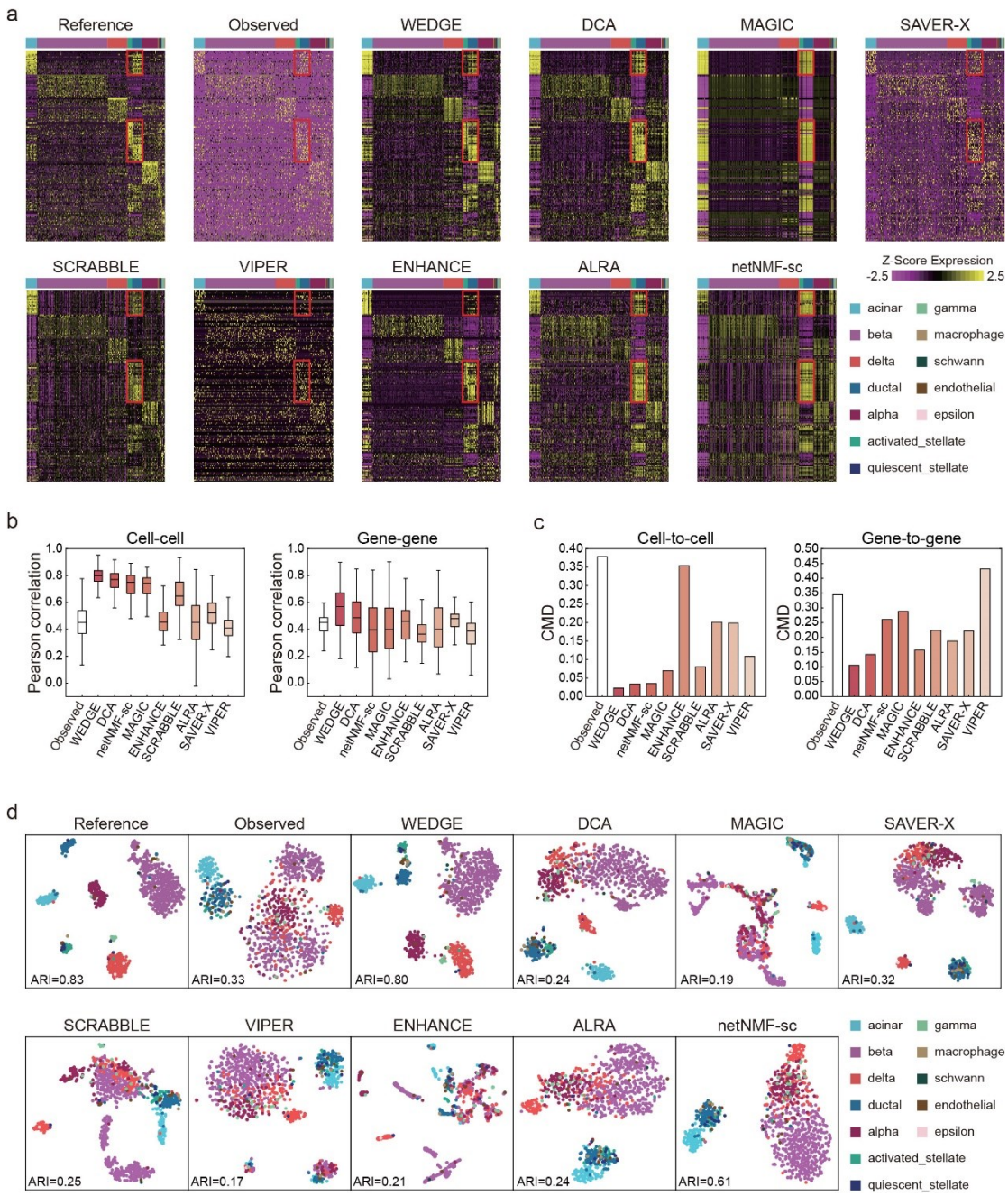
Supplementary Figure 1. Performance of different imputation algorithms for the recovery of the observed data with different dropout rates, visualized in tSNE space.



Supplementary Figure 2. Two dimension PCA visualization of the cells from the reference data, observed data, and data recovered with 4 different methods. (**a**) Scores plots of PCA results for cells, calculated using expression data for the cell-type-specific DE genes (as defined from the reference dataset), based on the data recovered using four different imputation methods. (**b**) PCA results calculated using the recovered expression data for the non-DE genes.

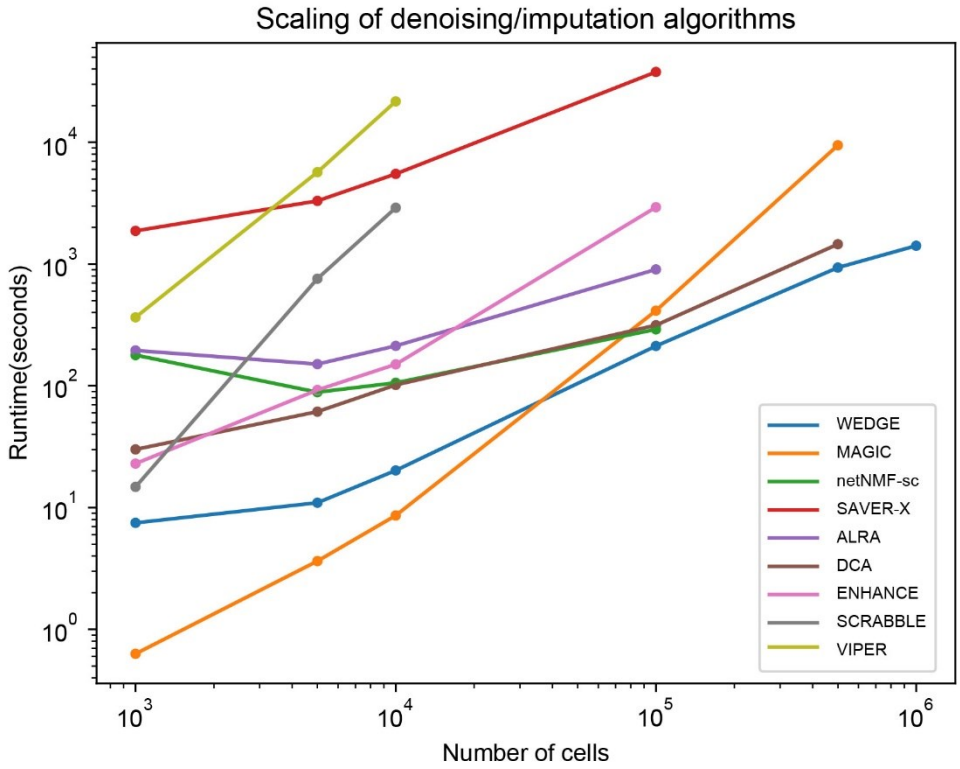


Supplementary Figure 3. The performance of SCRABBLE, VIPER, ENHANCE, ALRA, and netNMF-sc on Zeisel's dataset. (**a**) Visualization of the expression matrices of the top DE genes of different cell types, for the reference data, the observed data (dropout rate=0.85), and the recovered data generated by different methods. The color bar at the top indicates known cell types. (**b**) 2-D tSNE maps of the cells from the reference, observed, and variously recovered data. The color scheme is the same as in (a). (**c**) Expression of *Tba1* (a marker of S1 Pyramidal cells) and *Gad1* (a marker of interneurons), rendered in tSNE space.



Supplementary Figure 4. Application and performance of WEDGE to Baron's single-cell sequencing dataset, compared to existing methods. (**a**) Visualization of the expression matrices of the top DE genes of different cell types, for the reference data, the observed data (dropout rate=0.65), and the recovered data generated by different methods. The color bar at the top indicates known cell types. (**b**) Pearson correlation coefficients between the reference and recovered matrices (as shown in (a)) for cells (left panel) and genes (right panel). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. (**c**) Distances of the cell-to-cell (left panel)

and gene-to-gene (right panel) correlation matrices between the reference and recovered datasets. (**d**) 2-D tSNE maps of the cells from the reference, observed, and variously recovered data. The color scheme is the same as in (a).



Supplementary Figure 5. The runtime of WEDGE and other methods on single-cell datasets with the 2000 most highly variable genes and the indicated numbers of cells. All the calculations were performed on a computer with 28 cores and 128 GB memory. For the datasets with over 500,000 cells, some methods reported memory error, so their results are not shown here.