Briefings in Bioinformatics, 22(5), 2021, 1–13

https://doi.org/10.1093/bib/bbab085
Problem Solving Protocol

WEDGE: imputation of gene expression values from single-cell RNA-seq datasets using biased matrix decomposition

Yinlei Hu†, Bin Li†, Wen Zhang, Nianping Liu, Pengfei Cai, Falai Chen and Kun Qu

Corresponding authors: Kun Qu, Department of Oncology, The First Affiliated Hospital of USTC, Division of Molecular Medicine, Hefei National Laboratory for Physical Sciences at Microscale, School of Basic Medical Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230027, China, Tel.: +86-551-63606257; E-mail: qukun@ustc.edu.cn; Falai Chen, School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, China, Tel.: +86-551-63607537; E-mail: chenfl@ustc.edu.cn.
†These authors contributed equally to this work.

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 imputation methods have emerged for single-cell transcriptome data, however, recovering missing values in very sparse expression matrices remains a substantial challenge. Here, we propose a new algorithm, WEDGE (WEighted Decomposition of Gene Expression), to impute gene expression matrices by using a biased low-rank matrix decomposition method. WEDGE successfully recovered expression matrices, reproduced the cell-wise and gene-wise correlations and improved the clustering of cells, performing impressively for applications with sparse datasets. Overall, this study shows a potent approach for imputing sparse expression matrix data, and our WEDGE algorithm should help many researchers to more profitably explore the biological meanings embedded in their single-cell RNA sequencing datasets. The source code of WEDGE has been released at https://github.com/QuKunLab/WEDGE.

Key words: single-cell RNA-seq; imputation; denoising; matrix decomposition

Introduction

Single-cell RNA sequencing technology (scRNA-seq) can detect gene expression information at single-cell resolution. In recent years, many different scRNA-seq methods have emerged [1]. The differences among them lie in how the original transcripts are labeled and used to generate a sequencing library, which leads to their different detection efficiencies. For plate-based scRNA-seq techniques such as SMART-seq2 and CEL-seq2, the throughput (i.e. the number of cells captured in a single experiment) can be hundreds of cells, and for bead-based techniques such as 10X Chromium and Drop-seq, the throughput can reach thousands or even tens of thousands of cells ([1]; Supplementary Tables S1 and S2). scRNA-seq technology has been widely used in studies on many biological systems, including (but not limited to)

Yinlei Hu is a graduate student from Falai Chen's Lab.
Bin Li is an associate professor at Division of Life Sciences and Medicine, USTC. He was a post-doc research associate in the Chemistry Department of Columbia University from 2014 to 2016.
Wen Zhang is a graduate student from Kun Qu's Lab.
Nianping Liu is a graduate student from Kun Qu's Lab.
Pengfei Cai is a graduate student from Kun Qu's Lab.
Falai Chen is currently a professor at the School of Mathematical Sciences, University of Science and Technology of China. His research interests include computer aided geometric design, geometric modeling and computer graphics.
Kun Qu is currently a professor of Genomics and Bioinformatics at Division of Life Sciences and Medicine, USTC. From 2010 to 2016, he was a Bioinformatics Scientist and then the Director of Bioinformatics at Stanford University, USA.
Submitted: 31 October 2020; **Received (in revised form):** 3 February 2021

Algorithm 1. Optimization of WEDGE

Step 1: generate the initial $H \in \mathbb{R}^{+}_{r\times n}$ from singular value decomposition.
Step 2: from a given H, solve W in parallel with a non-negative least-square method.
Step 3: from the W obtained in step 2, calculate a new H.
Step 4: iteratively return back to steps 2 and 3 until the relative difference in the object function between two adjacent loops is $<1\times10^{-5}$ or the maximum specified number of iterations is reached.

Algorithm 2. Estimating the rank of the expression matrix

Input: the singular values of matrix A: $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_{\min(m,n)}$
Output: the final rank retained for A: r
Algorithm:
  $r = 1$;
  while $f(\sigma_r)$ and $r \leq \min(m,n) - 11$ do
    $r = r + 1$;
  end while

embryonic development [2–4], neuronal diversity [5–7], the immune system [8–10] and a large variety of diseases [11–15].

Despite the rapid increase in the applications of scRNA-seq technology, the number of detected genes per cell is still limited by technical challenges [16–19]. In single-cell sequencing experiments, a large amount of mRNA is lost during the cDNA generation process, and only a small amount of cDNA is amplified [20]. These technical deficiencies can result in dropout events, in which a gene is not detected in some cells even though it is expressed. Dropout events lead to a problem: many zero elements in the gene expression matrix are false zeros [21] that do not represent the true expression level of the genes in the cells [18, 22, 23]. To overcome this, a variety of algorithms have been developed to impute the zero elements in the expression matrices in order to restore the expression of the corresponding genes [21, 23–30].

For example, MAGIC [28] 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 Least Absolute Shrinkage and Selection Operator regression methods into SAVER [25] to estimate prior parameters and to restore missing elements of an expression data matrix, based on the assumption that gene expression follows the negative binomial distribution. Recently, they upgraded this approach to SAVER-X [30] 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 [23], which can denoise scRNA-seq data by learning gene-specific parameters. Many other tools have also emerged recently, such as SCRABBLE [27], VIPER [21], ENHANCE [29], ALRA [26], scImpute [31], scVI [32], DrImpute [33] and netNMF-sc [24], each of which seeks to improve recovery of the expression matrix for single-cell data (Supplementary Table S3). However, it is still a challenge to abundantly recover gene expression data while avoiding over-imputing [22, 24].

Here, we introduce a new algorithm, WEDGE, to impute gene expression values for sparse single-cell data based on low-rank matrix decomposition [34–36]. We applied WEDGE to multiple scRNA-seq datasets and compared its performance against several state-of-the-art methods. We also examined WEDGE's ability to distinguish cell subpopulations in the Tabula Muris dataset [37], and assessed its performance for accurately imputing marker gene expression in recently released datasets for peripheral blood mononuclear cells (PBMCs) from coronavirus disease 2019 (COVID-19) patients [38, 39]. Finally, we assessed the computer resources WEDGE consumes when analyzing large datasets.

Methods

Imputation model of WEDGE

WEDGE takes an expression matrix $A_{m\times n}$ as input, where the element 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. $a_{ij} = \log(\frac{a_{ij}\times 10\ 000}{\sum_i a_{ij}} + 1)$. After imputation, WEDGE outputs two non-negative matrices, W and H; we can then take their matrix product as the final imputed expression matrix, V. In the WEDGE algorithm, we imputed single-cell sequencing data through the following optimization framework:

$$\begin{aligned} Obj &= \min_{W,H} \textstyle\sum_{i,j\in\Omega} \left|a_{ij} - v_{ij}\right|^2 + \lambda \sum_{ij\in\overline{\Omega}} \left|v_{ij}\right|^2 \\ \text{subject to } v_{ij} &= \textstyle\sum_{k=1}^{r} w_{ik}h_{kj}, H \in \mathbb{R}^{+}_{r\times n}, W \in \mathbb{R}^{+}_{m\times r}, \end{aligned} \tag{1}$$

where $\Omega = \{(i,j)|a_{ij} > 0, i = 1 \sim m, j = 1 \sim n\}$, $\overline{\Omega}$ is the complementary set of Ω when the universe is $\{(i,j)|i = 1 \sim m, j = 1 \sim n\}$, v_{ij} is the element of the ith row and jth column of V, w_{ik} represents the element of matrix W and h_{kj} is the element of matrix H. In this model, the first term in the objective function guarantees minimization of approximation error between the nonzero elements of the original matrix A and the corresponding elements in the imputed matrix V, whereas the second term tends to minimize the elements in V which correspond to zero elements in A. The non-negative constraints in model (1) ensure that all the entries of V are non-negative.

Users can tune the bias parameter $\lambda \in [0,1]$ 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. For more analysis on how to set λ value, users can refer to the second paragraph of the Discussion section.

Optimization of the model

In the WEDGE algorithm, the matrix W and H were separately considered, which means that we fixed H to optimize W, and then fixed W to generate the new H. First, we defined that g_i is the ith row of A, H_i^+ and H_i^0 are composed of the H columns that correspond to the nonzero and zero elements of g_i respectively, and g_i^+ is the vector after deleting all zero elements from g_i. Then

we rewrote the objective function of solving W, as,

$$\min_{w_i}\left\|w_i H_i^+ - g_i^+\right\|_2^2 + \lambda\left\|w_i H_i^0\right\|_2^2 = \min_{w_i}\left\|w_i \tilde{H}_i - g_i\right\|_2^2 \quad (2)$$

where $\tilde{H}_i$ is the combination of H_i^+ and λH_i^0 according to the original order of their elements in H, and w_i is the ith row of W. In this case, optimizing W is equivalent to solving m non-negative least-squares problems (2) in parallel [40]. After W was obtained, we fixed it and solved H using similar algorithm as described above.

Estimating the rank of the expression matrix

During the optimization process for WEDGE, we designed a heuristic algorithm to determine the rank of $A_{m\times n}$ 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

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

where ε is a small non-negative constant (0.085 by default). In Algorithm 2 we provided the details of the process for the evaluation of rank r:

Correlation matrix distance

Correlation matrix distance (CMD) is usually used to determine the difference between two correlation matrices. It denotes the error of imputation process [25, 30], as a larger CMD value indicates greater difference between the reference/raw matrix and the imputed matrix. The CMD of two correlation matrices R_1 and R_2 is expressed as $d(R_1, R_2) = 1 - \frac{\mathrm{tr}(R_1 R_2)}{\|R_1\|_f \|R_2\|_f}$, where $\mathrm{tr}(R_1 R_2)$ represents the trace of matrix $R_1 * R_2$ and $\|\cdot\|_f$ is the Frobenius norm of a matrix.

Jaccard index

The Jaccard index can be used to evaluate the similarity between two clusters [41]. The Jaccard index of two clusters A and B is defined as $r = \frac{|A \cap B|}{|A \cup B|}$, here $|\cdot|$ represented the number of elements in a set.

Expression bias

The expression bias of gene α in cell clusterC_i refers to the proportion of cells (in clusterC_i) whose expression level of gene α is higher than the average expression of other clusters:

$$EB(\alpha, C_i) = \frac{1}{N(C_i)} \sum_{k \in C_i} \delta\left[E(\alpha, k) - \overline{E}(\alpha, k' \notin C_i)\right]$$

where k is a cell belonging to cluster C_i, $E(\alpha, k)$ is the expression value of gene α in cell k, k' is a cell belonging to other clusters, $\overline{E}(\alpha, k' \notin C_i)$ is the average expression value of gene α in all other clusters, $N(C_i)$ is the number of cells in cluster C_i and δ function is defined as

$$\delta(x) = \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ if } x \leq 0 \end{cases}$$

Generation of the simulated scRNA-seq datasets

We used the *splatSimulate()* function in the Splatter R package [42] to generate simulated datasets. For the dataset containing 6 cell types, 500 genes and 2000 cells (shown in Figure 1B and C and Supplementary Figure S1), we set seed = 42 and dropout. Shape = −1, and the down-sampling rate was tuned by parameter dropout. Mid values ranging from 1 to 6. For the dataset D1 with 2 cell types, 200 genes and 2000 cells (shown in Supplementary Figure S2A), we set seed = 42 and dropout. Shape = −1 and dropout. Mid = 2. For the dataset D2 with 3 cell types, 200 genes and 2000 cells (shown in Supplementary Figure S2B), we set seed = 42 and dropout. Shape = −1 and dropout. Mid = 4.

Processing of different datasets

For Zeisel *et al.*'s dataset of the mouse cortex and hippocampus cells (GSE60361) [7], we generated a reference dataset that contains high quality cells and genes (performed identically with the previously described filtering step of SAVER [25]), and retained all of the marker genes described in the initial study (*Tbr1*, *Spink8*, *Aldoc*, *Gad1*, *Mbp* and *Thy1*). Then, we randomly set 85% of the nonzero elements of the reference data to zeros to generate observed data with dropouts. For Baron *et al.*'s dataset of human pancreatic islet cells (GSE84133) [43], 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 nonzero values. We then randomly set 65% of the nonzero elements to zeros to simulate dropout events. For Tabula Muris dataset (GSE109774; [37]), we applied the pipeline provided by Tabula Muris to filter the genes and cells, and obtained an expression matrix with 23 341 genes and 55 656 cells. For Guo *et al.*'s COVID-19 dataset, we filtered genes and cells using the same method as Guo *et al.*, to obtain an expression matrix containing 23 324 genes and 68 190 cells. For the COVID-19 dataset released by Schulte-Schrepping *et al.* [39], we used the same preprocessing method as the original paper to obtain a gene expression matrix containing 46 584 genes and 99 049 cells.

Scalability analysis

Scalability analysis was performed on a super computer with four Intel Xeon E7-8860 v4 2.20 GHz CPUs (72 cores in total) and 1TB memory. We down-sampled the mouse brain atlas dataset downloaded from 10X Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons) to construct the benchmark datasets with different number of cells (from 1000 to 1000 000). First, we filtered out the genes 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, 10 000, 100 000, 500 000 and 1000 000 cells from the raw dataset to simulate experiments of different scales.

Data availability

There are no new data associated with this article. Published datasets used in this study: Zeisel *et al.*'s dataset ([7]; GSE60361) of the mouse cortex and hippocampus cells is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60361; Baron

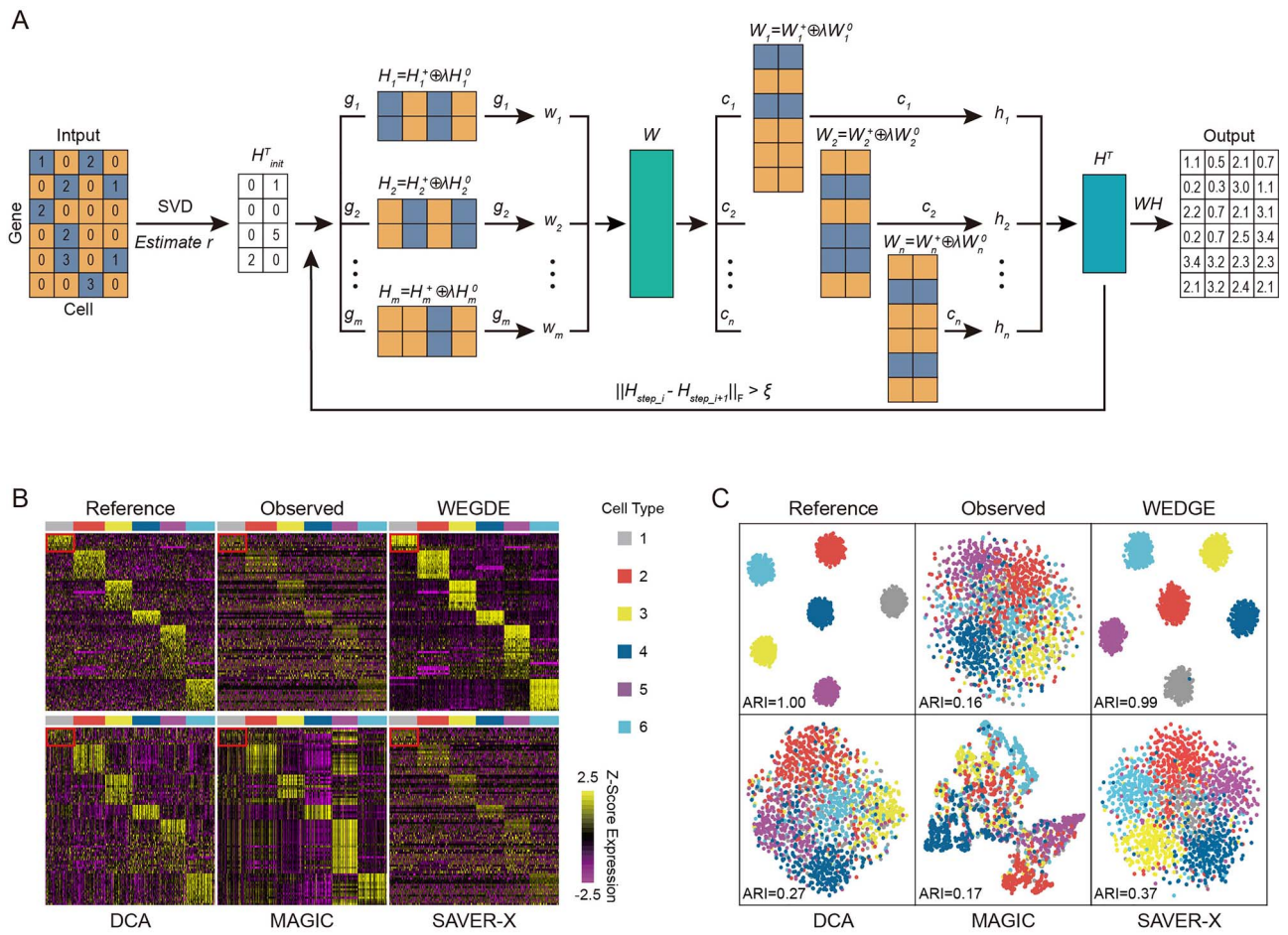


Figure 1. Design of the WEDGE algorithm and its performance for the simulated dataset. (**A**) Conceptual overview of the WEDGE workflow. The blue and yellow grids in the input expression matrix represent nonzero and zero elements, respectively. H and W are the coefficient matrix and feature matrix, respectively, calculated by using biased matrix decomposition. g_i is the ith gene, c_j is the jth cell, w_i is the ith row of W, and h_j is the jth column of H. H_i^+ and H_i^0 are composed of the columns of H that correspond to the nonzero and zero elements of g_i, respectively. W_j^+ and W_j^0 are composed of the rows of W that correspond to the nonzero and zero elements of c_j, respectively. H_i is the combination of H_i^+ and λH_i^0, where the blue and yellow grids are the elements of H_i^+ and λH_i^0, respectively. W_j is the combination of W_j^+ and λW_j^0, where the blue and yellow grids are the elements of W_j^+ and λW_j^0, respectively. ξ is the convergence criterion ($=1 \times 10^{-5}$ by default). (**B**) Expression matrices of the top DE genes of the simulated reference and observed data (sparsity = 0.50), 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 imputed by different methods.

et al.'s dataset ([43]; GSE84133) of human pancreatic islet cells is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133; Tabula Muris dataset ([37]; GSE109774) including 20 mouse organs is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109774; Guo *et al.*'s dataset ([38]; GSE150861) containing PBMCs from two COVID-19 patients is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150861, and the 10X dataset of two healthy donors is available at https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.1.0/5k_pbmc_NGSC3_aggr); Schulte-Schrepping's dataset ([39]; EGAS00001004571) containing PBMCs from 18 COVID-19 patients and 22 controls (obtained by 10X platform) is available at https://beta.fastgenomics.org/consent. The source code of WEDGE was released at https://github.com/QuKunLab/WEDGE.

Computational environment

All the experiments in this article were performed on a computer cluster with 4 Intel Xeon E78860v4 CPUs (2.2 GHz, 45 MB L3 cache and 72 CPU cores in total) and 1TB of memory (DDR4 2400 MHz).

Results

Algorithm, performance and robustness of WEDGE

The expression matrices obtained from single-cell sequencing experiments are sparse, caused by the low RNA capture rates during experimental sampling and processing [21, 23]. Weighted non-negative matrix factorization (WNMF) has demonstrated its potential for recovering missing elements from a sparse matrix [24, 35]. However, it should be noted that the contribution of the zero elements in the raw matrix is completely ignored in the WNMF optimization process. In WEDGE, we adopted a low weight ($0 \leq \lambda \leq 1$) for the zero elements in the raw expression matrix during the biased low-rank matrix decomposition method (bLRMD), and generated a convergent imputed matrix using an alternating non-negative least-squares algorithm (Figure 1A and section 'Methods'). We chose not to set λ as zero, as the contribution of the zero elements is not completely negligible [23]. Notably, as WEDGE is a completely unsupervised 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 assess a simulated dataset generated using Splatter [42] and compared WEDGE against DCA [23], MAGIC [28] and SAVER-X [30] (Figure 1B). The reference data include distinct marker genes for six different cell types and a dense expression matrix (original sparsity = 10%, where sparsity is the percentage of zero elements). Based on the assumption that gene expression values follow a negative binomial distribution [42], the Splatter simulation set 45% of the nonzero elements in the reference data to zero (i.e. down-sampling rate = 0.45 and sparsity = 0.50) to simulate dropout events and to obtain 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 recovered their expression patterns, obtaining an imputed 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; [25, 44]) to assess the accuracy of the cell clustering results (Figure 1C), wherein a higher ARI value indicates that the clustering result is relatively closer to the 'true' cell types. Using the expression matrix imputed 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 imputation methods.

We further evaluated the robustness of WEDGE by applying it to impute 'observed' matrices with different sparsities (Supplementary Figure S1). Interestingly, for the observed matrix with a low sparsity (=0.25), all the 12 methods successfully recovered the distinctions between the cell types. However, for very sparse data (i.e. sparsities >0.50) only WEDGE can still delineate the cell identities, suggesting the advantage of WEDGE on imputing scRNA profiles with low capture rate.

In addition, to determine whether the algorithm leads to over-imputing—for example, erroneously restoring non-DE genes so that they appear as DE genes—we applied WEDGE as well as all the other imputation methods on two simulated datasets: dataset D1, which comprised two cell types, and dataset D2, which comprised three cell types. After imputation, we found that the two cell types in dataset D1 could be clearly distinguished based on the expression of the DE genes after WEDGE imputation, with an ARI value of 0.99, higher than the ARI values generated from all the other methods (Supplementary Figure S2A). However, these two cell types could not be distinguished based on the imputed expression of the non-DE genes (ARI <0.01 for WEDGE). Similarly, when we applied WEDGE to dataset D2, we found that the three cell types could be classified by the DE genes (ARI = 0.99 after WEDGE imputation, the highest among all methods) but not the non-DE genes (ARI <0.01 after WEDGE imputation) (Supplementary Figure S2B). These results indicate that WEDGE did not erroneously impute the non-DE genes to DE genes, implying its robustness to 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 *et al.*'s dataset [7] on mouse brain scRNA-seq. We first constructed the reference matrix by extracting all the cells with $>10\,000$ UMI counts and all the genes detected in $>40\%$ of cells, and then generated an 'observed' matrix with high sparsity by randomly setting a large proportion of the nonzero elements to zeros (sparsity = 0.90). From the heatmaps of gene expression matrices (Figure 2A), we can see that WEDGE recovered the expression of the DE genes, especially those DE between interneurons and S1 pyramidal cells.

We also used other tools, including SCRABBLE [27], VIPER [21], ENHANCE [29], ALRA [26], scImpute [31], scVI [32], DrImpute [33] and netNMF-sc [24] to assess the same 'observed' matrix (Supplementary Figure S3A). To quantify the similarity between the reference and imputed expression matrices, we calculated the cell-wise and gene-wise Pearson correlations between them [25], 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 (Figure 2B). The gene-wise correlation coefficients from WEDGE were also higher than that from the rest of the methods. Moreover, we computed the correlation matrix distances (CMDs) [25, 45] between the reference and imputed data, where a lower CMD indicates that the imputed data are closer to the reference data (Figure 2C). For the matrix generated by WEDGE, the cell-to-cell CMD is 0.03 and the gene-to-gene CMD is 0.12, which are each tied for the lowest of all the tested methods. These comparisons together highlight that our WEDGE approach can recover both the cell–cell and gene–gene correlations from sparse single-cell RNA-seq datasets.

In the tSNE map of cells, WEDGE can clearly distinguish interneurons, S1 pyramidal neurons and CA1 pyramidal neurons, and the ARI value of 0.56 for the clustering result calculated from its imputed matrix is the highest among all tested methods (Figure 2D; Supplementary Figure S3B). In particular, in visualizing the expression of an interneuron marker gene *Gad1* [7] and an S1 pyramidal marker gene *Tbr1*, WEDGE appropriately recovered their expression levels in the corresponding cell types, without overestimating their expression in other cell types (Figure 2E; Supplementary Figure S3C). Furthermore, we also applied WEDGE to Zeisel *et al.*'s raw dataset [7], which includes 3005 cells and 19 973 genes. After processing with WEDGE, the clustering performance index ARI increased from 0.42 (raw data) to 0.56, the highest among all tested methods (Supplementary Figure S4).

As another example to confirm the utility of WEDGE, we applied the same procedures described above to Baron *et al.*'s pancreas single-cell dataset [43], and compared WEDGE with other imputation methods. WEDGE recovered the expression of most of the DE genes, especially those of the ductal and activated stellate cells (Supplementary Figure S5A). Similarly, the cell-wise and gene-wise Pearson correlation coefficients from WEDGE are both greater than those from any other tested methods, emphasizing its strong recovery performance (Supplementary Figure S5B). Also, the cell-to-cell and gene-to-gene CMDs of WEDGE are 0.02 and 0.11, respectively, which are the lowest and the second lowest (scVI's gene-to-gene CMDs is 0.10) among all methods (Supplementary Figure S5C). Finally, in terms of cell clustering, WEDGE clearly classified alpha, beta, delta, ductal, acinar, and gamma cells, with an ARI value of 0.80, higher than those from all the other methods (Supplementary Figure S5D).

To verify whether WEDGE aids in discovery of the biological functions of cells, we calculated the average expression of gene sets related to some GO terms in Zeisel *et al.*'s and Baron *et al.*'s datasets. For example, previous studies have shown that pyramidal neurons are associated with cognition [46]; however, no significant differential gene expression was discovered between the S1 pyramidal neurons and interneurons in the original Zeisel *et al.*'s dataset (P-value ≥ 0.04). After imputation, WEDGE and netNMF-sc enhanced the differential expression patterns of the cognition-related gene set between the two types of neurons

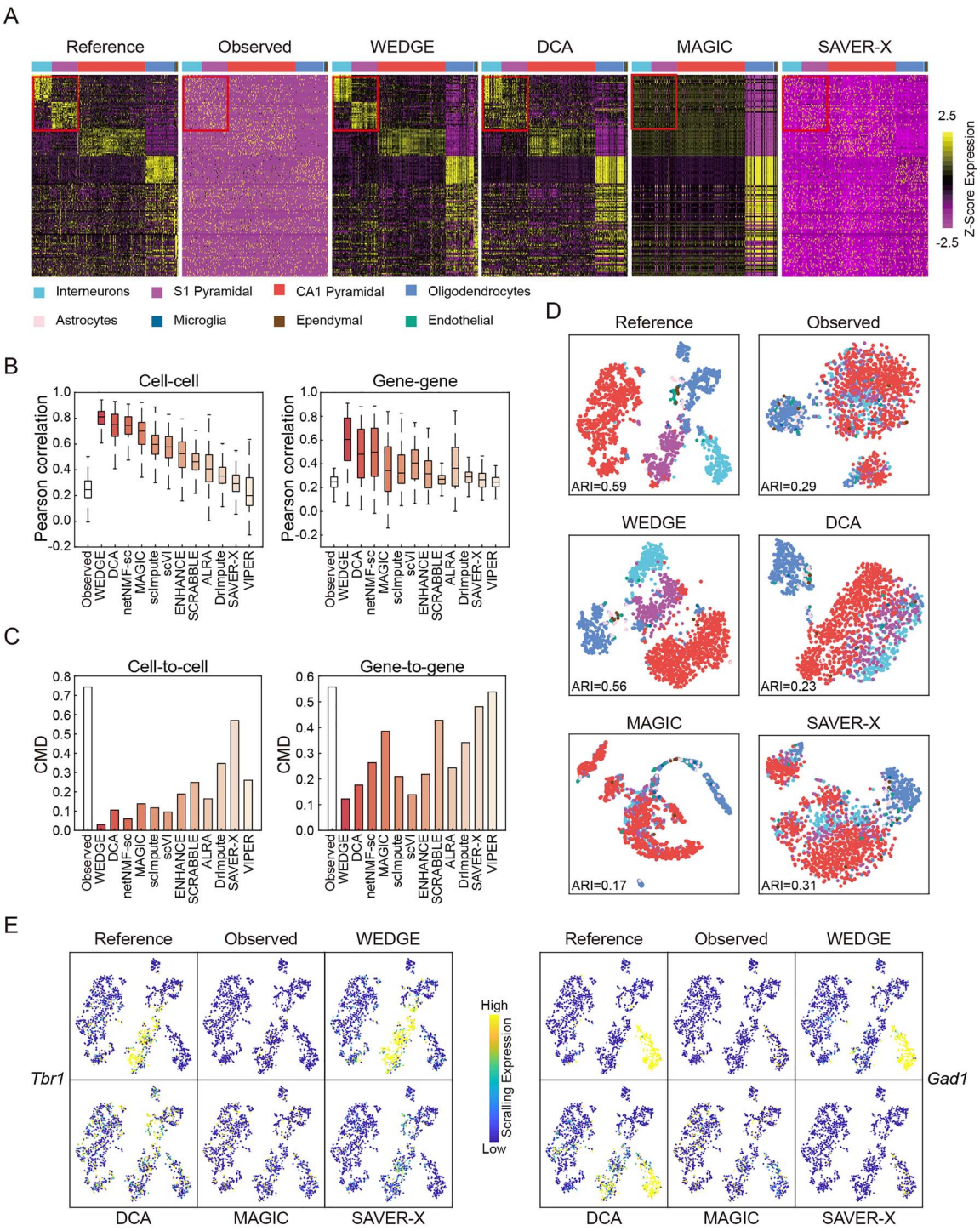


Figure 2. Application and performance assessment of WEDGE for dataset GSE60361 (Zeisel *et al.*), compared with existing methods. (**A**) Visualizations of the expression matrices of the top DE genes of different cell types, including the reference data, the observed data (sparsity = 0.90), and the imputed data generated by four different methods. The color bar at the top indicates known cell types. (**B**) Pearson correlation coefficients between the reference and imputed matrices for cells (left panel) and genes (right panel). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. (**C**) Distances of the cell-to-cell (left panel) and gene-to-gene (right panel) correlation matrices between the reference and imputed data. (**D**) 2-D tSNE maps of the cells from the reference, observed and imputed 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 imputed by different methods, rendered in tSNE space.

(P-value $< 10^{-6}$) (Supplementary Figure S6A). Similarly, the insulin secretion function of pancreatic beta cells has been reported in a previous study [47]; however, in Baron *et al.*'s dataset, no significant differential gene expression of insulin secretion-related genes was observed between beta and alpha cells in the original dataset (P-value ≥ 0.17). In the data imputed by WEDGE, the expression levels of insulin secretion-related genes in beta cells were significantly higher than those in

alpha cells (*P*-value $< 10^{-21}$) (Supplementary Figure S6B). Some benchmarked methods also slightly improved the differential expression patterns of this gene set in beta versus alpha cells (e.g. the *P*-values of VIPER and DrImpute imputed dataset were 0.0045 and 0.0046, respectively), whereas others did not enhance or even reverse this trend (such as DCA, SAVER-X and SCRABBLE).

Classification of cell subpopulations

To test that if WEDGE can be applied for large datasets with multiple tissues and organs, we used it to process the recently released Tabula Muris dataset of mouse (10X Chromium sequencing data; [37]). We used the k-nearest-neighbor graph-based method in Seurat to cluster cells for both the raw and imputed data, and presented the results in tSNE space (Supplementary Figure S7A and B). Among the 54 cell clusters generated from the raw data, 45 clusters have Jaccard index values > 0.5, indicating that they were also identified in the clustering of the WEDGE imputed data (Supplementary Figure S7C; see section 'Methods'). Notably, the WEDGE imputation improved the clustering resolution for some cell types: a main cluster of B cells (Supplementary Figure S8) was classified into four subclusters (clusters 1, 3, 50 and 53; Figure 3A), whereas these subclusters were not separated based on the raw data (Figure 3B). Most cells in clusters 1, 3 and 53 are from the spleen, whereas cluster 50 is dominated by B cells from lungs (Figure 3C).

Cluster 3 splenic B cells strongly expressed the marginal zone B cell (MZ) marker *Cr2*, whereas cluster 1 splenic B cells were characterized by strong expression of the follicular B cell (FO) marker *Fcer2a* [48] (Figure 3D and E). We developed a factor called 'expression bias' (see section 'Methods' for the formula) to assess how a given imputation method affects the differential enrichment trend for inter-cluster expression comparisons. WEDGE imputation increased the expression bias of *Cr2* in cluster 3 from 0.74 to 0.88, and increased the expression bias of *Fcer2a* in cluster 1 from 0.56 to 0.96. The marker gene sets characteristic of MZ and FO splenic B cell subsets as reported by Kleiman [49] were also DE in cluster 3 and 1, respectively, and WEDGE respectively amplified the expression bias of these marker gene sets from 0.59/0.60 to 0.97/0.89 (Figure 3F and G). Moreover, we tested the expression of the marker gene sets of the two cell types reported by Newman *et al.* [50], which also supported our classification of these two respective cell subpopulations as MZ and FO splenic B cells (Supplementary Figure S9A and B). We noted that neither the raw data nor the WEDGE imputed data showed any obvious expression of the transitional B cell marker *Cd93* [48] (Supplementary Figure S9C), and it was also notable that cluster 53 apparently represented an aggregation of cells with detected transcripts for < 600 genes (Supplementary Figure S9D).

We also used other state-of-the-art methods to impute the same dataset and checked their clustering results on the splenic B cells (Supplementary Figure S10). DCA, MAGIC, ALRA, ENHANCE and DrImpute enhanced the expression of *Cr2* and *Fcer2a* in some cells, but these methods also amplified batch effects, and clustering based on the imputation data from these methods did not clearly distinguish splenic B cells into FO and MZ subpopulations. SAVER-X did not classify the FO and MZ subpopulations based on differential expression trends for *Cr2* or *Fcer2a*. In addition, VIPER, SCRABBLE and scVI were unable to obtain imputation results from this dataset within 48 h on the computer with 72 CPU-cores (2.2 GHz) and 1TB memory, and netNMF-sc and scImpute did not complete because of memory errors.

Imputation of gene expression for the COVID-19 dataset

In a previous study of COVID-19 patients, Guo *et al.* [38] reported that one monocyte subset was found to strongly impact cytokine storms in patients classified as severe-stage. Cells of this severe-stage-specific monocyte subset strongly express many cytokines and related transcription factors such as *IL6*, *IL10*, *CXCL2*, *CXCL3*, *CCL4*, *ATF3*, *TNF* and *HIVEP2*. However, dropout events in single-cell sequencing experiments obscured this trend. Here, we applied WEDGE to this recently released COVID-19 dataset containing 13 239 PBMCs from patients and 54 951 PBMCs from healthy donors [38]. The clusters obtained from the WEDGE imputed data correspond to cell clusters reported in the original paper, with an ARI value of 0.70, the highest among all tested methods (Figure 4A and B). Guo *et al.* divided the monocyte cells into clusters 2, 9, 13 and 16, with the cells in cluster 9 apparently representing the severe-stage-specific monocyte subset [38]. Following WEDGE imputation, the severe-stage-specific cytokines and upstream transcription factors reported by Guo *et al.* had increased expression bias in cluster 9 cells (from 0.03–0.93 to 1.00; Figure 4D). Notably, the DE genes of cluster 9 generated from the WEDGE imputed data cover 99% of the DE genes in the raw data (Figure 4C). The WEDGE imputed data also increased the expression bias of the *APOE* and *CXCR3* genes in cluster 9 cells (Supplementary Figure S11A). Similar increases in expression bias were detected in the WEDGE imputed data for severe-stage-specific DE genes reported by Wilk *et al.* [51] (Supplementary Figure S11B).

In addition, the COVID-19 studies recently published by Schulte-Schrepping *et al.* [39] and Wilk *et al.* [51] have shown that the expression level of *CD11C* (i.e. *ITGAX*) in monocytes of patients with severe COVID-19 is lower than that in normal monocytes, while *CD163* and *ISG15* expression is upregulated in patients with severe COVID-19. However, among the 4 monocyte clusters in Guo *et al.*'s dataset, the expression level of *ITGAX* is not the lowest in severe stage-specific monocytes (i.e. cluster 9), and the high expression levels of *CD163* and *ISG15* in cluster 9 are not obvious (EB ≤ 0.65). After imputing Guo *et al.*'s dataset with WEDGE, we recovered the low *ITGAX* expression level and high *CD163*/*ISG15* expression level in cluster 9 (EB ≥ 0.88), implying that we can obtain the same conclusion from different datasets (Supplementary Figure S12A). Other methods also recovered the high expression level of *CD163* in cluster 9. However, in the imputed results of SAVER-X, MAGIC, ALRA and ENHANCE, the expression level of *ITGAX* in cluster 9 is not the lowest among the 4 monocyte clusters. Moreover, in the results of DCA, SAVER-X, MAGIC, ALRA and DrImpute, the expression level of *ISG15* in cluster 9 is not the highest. On the other hand, Guo *et al.* showed that severe stage-specific monocytes highly express *IL6*, *IL10*, *CXCL2*, *CXCL3*, *ATF3*, *TNF* and *HIVEP2*, but these trends are not obvious in Schulte-Schrepping *et al.*'s dataset [39] (EB ≤ 0.33). After imputing Schulte-Schrepping *et al.*'s dataset with WEDGE, we improved the differential expression patterns of these genes (EB ≥ 0.67) (Supplementary Figure S12B).

Scalability and efficiency

To quantify the scalability and efficiency of different imputation algorithms, we counted the time and memory consumption of WEDGE and other state-of-the-art methods when imputing Tabula Muris dataset that contains 55 656 cells and 23 341 genes (after filtering). WEDGE consumed 1 h and up to 38GB of memory on a 72-core computer to complete the imputation

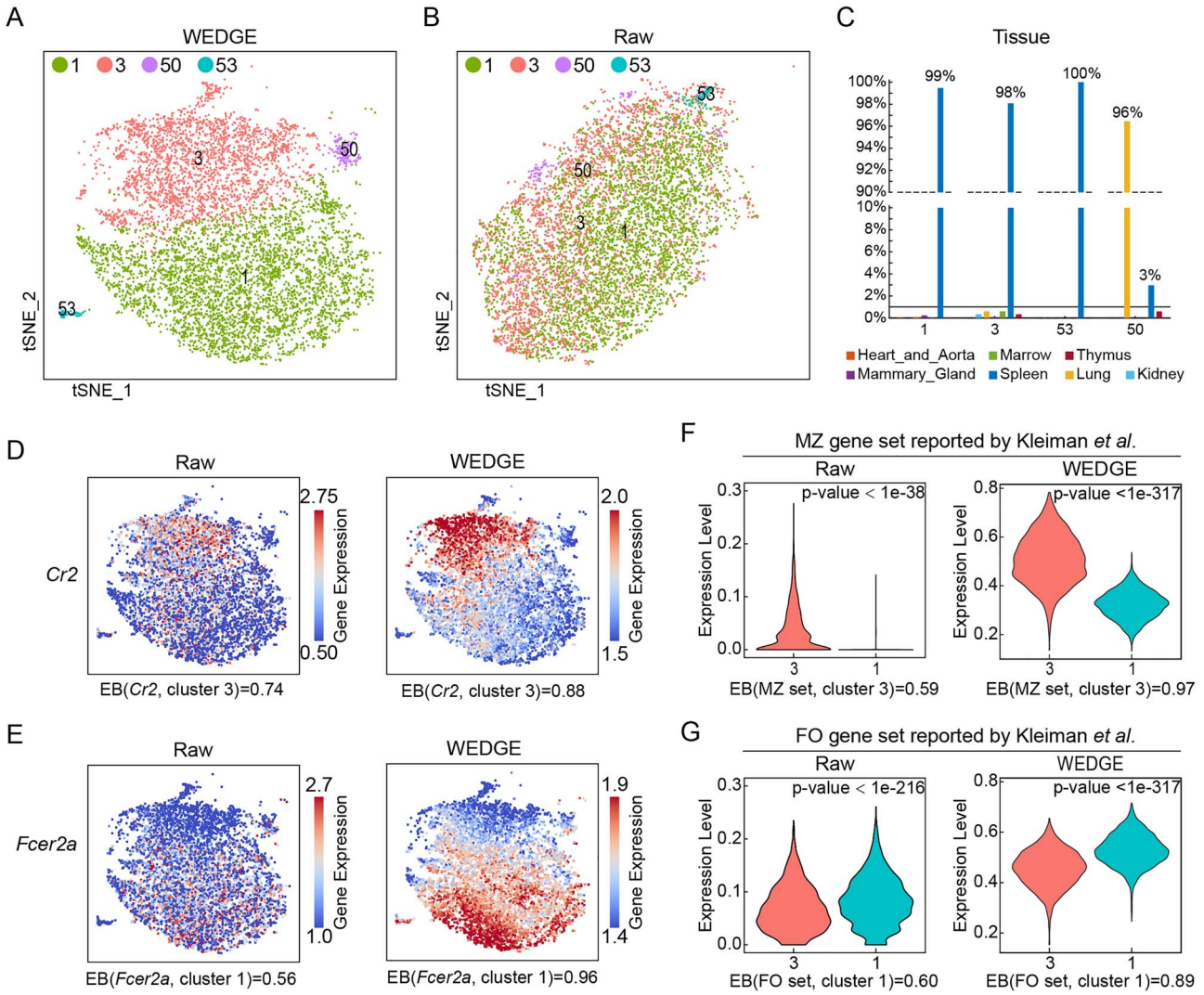


Figure 3. WEDGE imputation of dataset GSE109774 (Tabula Muris) facilitated the classification of splenic B cell subpopulations. (**A**, **B**) 2-D tSNE maps of the splenic B cells generated from the WEDGE imputed data (**A**) and the raw data (**B**). The colors indicate cell clusters from the WEDGE imputed data. (**C**) The ratio of cells from different organs in splenic B cell clusters. (**D**, **E**) The expression of *Cr2* (**D**) or *Fcer2a* (**E**) in the raw and WEDGE imputed data, rendered in the 2-D tSNE space. EB: expression bias (see section 'Methods'). (**F**, **G**) The average expression of the marker gene sets of MZ (**F**) and FO (**G**) cells (reported by Kleiman *et al.* [49]) from the raw and WEDGE imputed data.

process, which was close to the MAGIC method. (Figure 5A and B). To further assess the computer resources that WEDGE spends on datasets of various sizes, we applied it to impute datasets comprising different numbers of cells (5000–1000 000) but a fixed number of genes (2000), which were sampled from the mouse brain atlas project (see section 'Methods'). The runtime of WEDGE increased linearly with the number of cells (Supplementary Note S1), and its speed was close to DCA and MAGIC (Figure 5C). For the dataset containing 1 million cells and 2000 genes, WEDGE finished the imputation of missing values in 12 min. Notably, WEDGE offers a visual interactive interface, making it convenient for researchers to use. We have uploaded WEDGE and the datasets used in this study to GitHub (https://github.com/QuKunLab/WEDGE).

Discussion

WEDGE effectively recovered the expression of undetected genes in scRNA-seq data, performing especially well for datasets with high sparsities, thereby substantially promoting our ability to understand sparse single-cell profiles. Accordingly, WEDGE imputation increased the ARI values of the clustering results of scRNA-seq datasets, and facilitated high-resolution classification of cell subpopulations. Nevertheless, we found that other tools also showed different advantages. For example, MAGIC, ENHANCE, ALRA and WEDGE consumed less computer time than other methods during analysis of large-scale data, and scVI and MAGIC required the least computer memory among all these methods.

We tested the impact of the bias parameter λ on imputation and found that the performance of WEDGE is not sensitive to λ for datasets with sparsities < 0.6 (Supplementary Figure S13). For datasets with sparsities > 0.6, a λ value of 0.1–0.2 can produce the best recovery results. Therefore, we set $\lambda = 0.15$ as the default value in WEDGE and used it for all datasets in this study. The imputation contribution of the zero elements decreases with the increase of matrix sparsity, but it cannot be ignored, which implies that some zero elements may be related to the

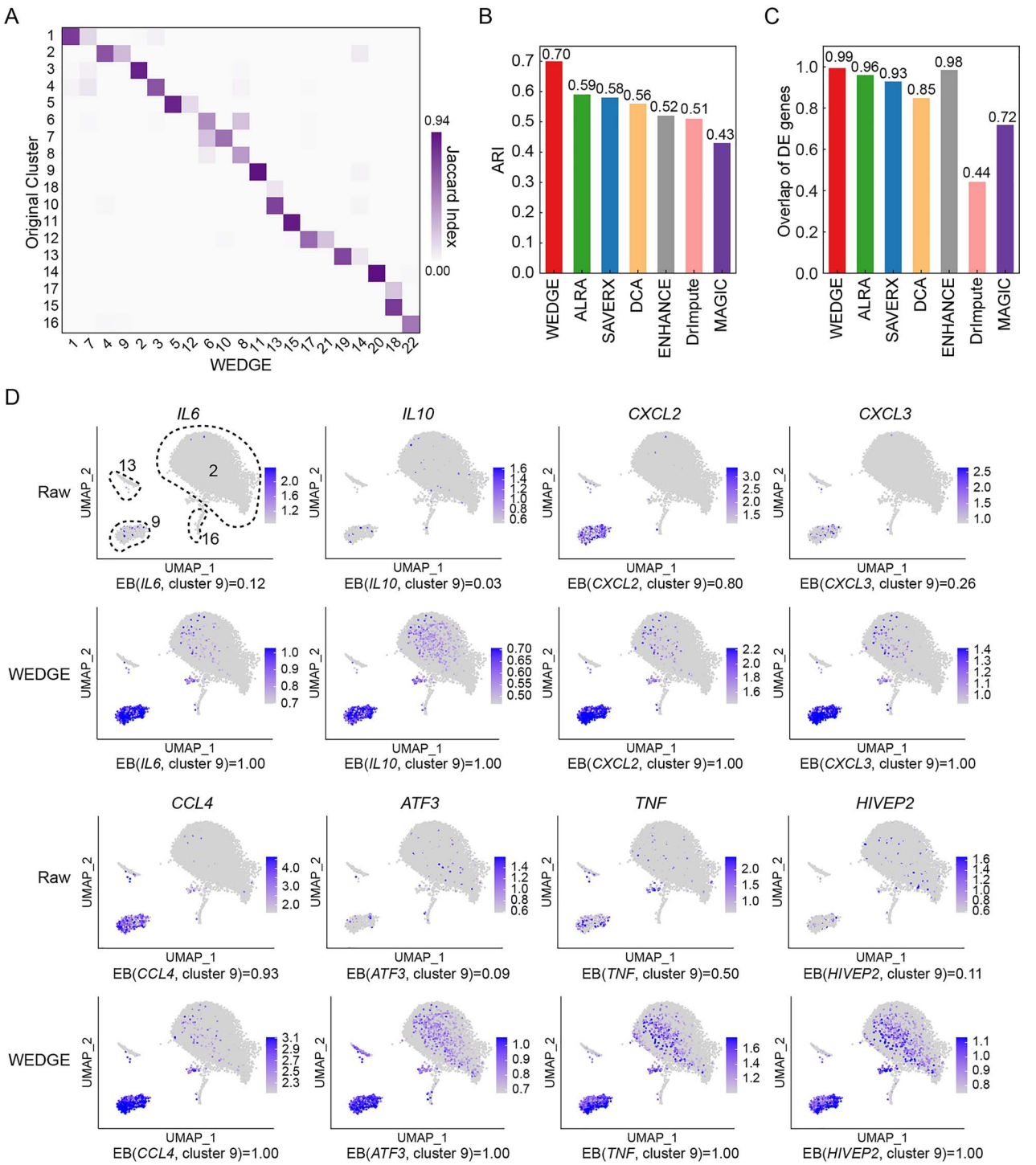
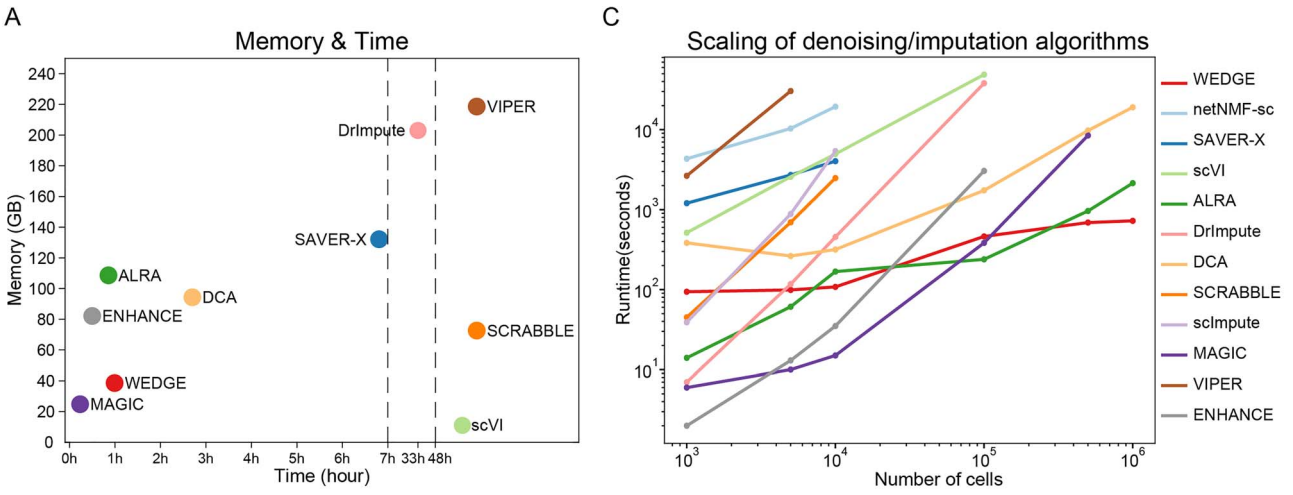


Figure 4. WEDGE enhanced marker gene expression for the COVID-19 dataset GSE150861 (Guo *et al.*). (**A**) Jaccard index between the cell clusters reported by Guo *et al.* [38] and the clusters generated from WEDGE imputed data. (**B**) ARI values calculated from the clustering results of the imputed data generated by different methods. The results of other methods are not shown, as SCRABBLE and scVI did not finish the imputation within 48 h on a computer with 72 CPU-cores (2.2 GHz) and 1TB memory, while VIPER, scImpute and netNMF-sc reported memory errors. (**C**) Proportion of the reported DE genes that can be regenerated from the imputed data, i.e. $\frac{|D_{raw} \cap D_{imputed}|}{|D_{raw}|}$, here, D_{raw} and $D_{imputed}$ are DE gene sets generated from the raw data and the imputed data, respectively. (**D**) The expression of known marker genes of severe-stage specific monocytes. Clusters 2, 9, 13 and 16 are cell clusters reported by the original paper. EB: expression bias (see section 'Methods').

low expression of certain genes, rather than simply reflecting experimental noise.

Another question is whether WEDGE will introduce false positive signals in the DE gene detection process. To address this question, we used the set of DE genes detected from the reference data of different datasets as the gold standard (fold-change > 2 and P-value < 0.001) and plotted the ROC curves of the DE genes detected from the observed and imputed data.



B

Runtime & Memory requirement of imputation methods on GSE109774 (Tabula Muris)

Methods	WEDGE	DCA	MAGIC	SAVER-X	ALRA	ENHANCE	VIPER	SCRABBLE	DrImpute	scVI	netNMF-sc	scImpute
Memory (GB)	38.6G	94.3G	24.8G	132.1G	108.6G	82.1G	218.6G	72.6G	202.9G	10.9G	NA	NA
Time(s)	3567s	9710s	836s	24508s	3081s	1793s	>48h	>48h	118618s	>48h	NA	NA

Figure 5. Computer resources consumed by different imputation methods, on a computer with 72 CPU-cores and 1TB memory. (**A**, **B**) The runtime and memory spent by different methods for imputing dataset GSE109774 (Tabula Muris), where netNMF-sc and scImpute reported errors and stopped running. (**C**) The computational cost of different methods for imputing single-cell datasets of various sizes. Results that took > 9 h or reported memory errors are not shown here.

For the simulated data with high sparsity (=0.65), when the true positive rate was 0.8, the false positive rate of the imputed data was 0.2, lower than that of the observed data (=0.3) (Supplementary Figure S14A). The ROC curves of Zeisel *et al.*'s and Baron *et al.*'s datasets also showed that WEDGE did not introduce additional false positive values in DE gene detection (Supplementary Figure S14B and C). Therefore, we recommend that researchers impute the expression matrix before searching for DE genes.

There are still challenges for the informative imputation of scRNA-seq datasets, such as how to recover the heterogeneity between cell types instead of experimental batches, how to discover cell subtypes with very few cells from the imputed data, and how to use limited computer resources to process large datasets containing millions of cells. Moreover, it necessary to assess whether current imputation methods are applicable to datasets obtained using diverse bioanalytical methods beyond standard RNA-seq (e.g. single-cell ATAC-seq and profiling methods for various epigenomic modifications).

Conclusion

Here, we present an approach, WEDGE, to impute missing gene expression information in single-cell sequencing datasets that is based on the combination of low-rank matrix decomposition and biased weight parameters for the zero and nonzero elements in the expression matrix. We show that the usage of WEDGE significantly improves the clustering accuracy of many scRNA-seq datasets, amplifies the contribution of differential genes to identifying cell types, and helps distinguish more cell subpopulations from low-quality data.

Key Points

- WEDGE is an effective tool for imputing sparse single-cell data using biased low-rank matrix decomposition.
- WEDGE enhances DE gene expression and cell–cell/gene–gene correlation from the raw scRNA-seq data.
- In addition, WEDGE facilitates the cell subtype classification of sparse single-cell data.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

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 W.Z., N.L. and P.C.; B.L., Y.H. and K.Q. wrote the manuscript with inputs from all the authors.

Acknowledgments

We thank the USTC supercomputing center and the School of Life Science Bioinformatics Center for providing computing resources for this project. We thank CAS Interdisciplinary Innovation Team.

Funding

This work was supported by the National Key R&D Program of China (grant number 2020YFA0112200, 2017YFA0102900 to K.Q.), the National Natural Science Foundation of China (grant number 91940306, 81788101, 31970858, 31771428, 91640113 to K.Q.; grant number 61972368, 11571338 to F.C.) and the Fundamental Research Funds for the Central Universities (grant number YD2070002019, WK2070000158, WK9110000141 to K.Q.).

References

1. Ding J, Adiconis X, Simmons SK, *et al*. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat Biotechnol* 2020;**38**:737–46.
2. Yan L, Yang M, Guo H, *et al*. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* 2013;**20**:1131–9.
3. Klein AM, Mazutis L, Akartuna I, *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**:1187–201.
4. Xue Z, Huang K, Cai C, *et al*. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 2013;**500**:593–7.
5. Lake BB, Ai R, Kaeser GE, *et al*. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* 2016;**352**:1586–90.
6. Lake BB, Chen S, Sos BC, *et al*. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* 2018;**36**:70–80.
7. Zeisel A, Muñoz-Manchado AB, Codeluppi S, *et al*. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015;**347**:1138–42.
8. Shalek AK, Satija R, Adiconis X, *et al*. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 2013;**498**:236–40.
9. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* 2018;**18**:35.
10. Björklund ÅK, Forkel M, Picelli S, *et al*. The heterogeneity of human CD127+ innate lymphoid cells revealed by single-cell RNA sequencing. *Nat Immunol* 2016;**17**:451–60.
11. Guo X, Zhang Y, Zheng L, *et al*. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat Med* 2018;**24**:978.
12. Peng J, Sun B-F, Chen C-Y, *et al*. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res* 2019;**29**:725–38.
13. Zhang L, Yu X, Zheng L, *et al*. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* 2018;**564**:268.
14. Zhang P, Yang M, Zhang Y, *et al*. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. *Cell Rep* 2019;**27**:1934, e1935–47.
15. Zheng C, Zheng L, Yoo J-K, *et al*. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* 2017;**169**:1342, e1316–56.
16. Grün D, Kester L, Van Oudenaarden A. Validation of noise models for single-cell transcriptomics. *Nat Methods* 2014;**11**:637.
17. Hou W, Ji Z, Ji H, *et al*. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol* 2020;**21**:218.
18. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;**20**:273–82.
19. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015;**16**:133.
20. Ziegenhain C, Vieth B, Parekh S, *et al*. Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* 2017;**65**:631, e634–43.
21. Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol* 2018;**19**:196.
22. Tian L, Dong X, Freytag S, *et al*. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat Methods* 2019;**16**:479–87.
23. Eraslan G, Simon LM, Mircea M, *et al*. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**:390.
24. Elyanow R, Dumitrascu B, Engelhardt BE, *et al*. netNMF-sc: leveraging gene–gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Res* 2020;**30**:195–204.
25. Huang M, Wang J, Torre E, *et al*. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 2018;**15**:539–42.
26. Linderman GC, Zhao J, Kluger Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv* 2018, doi: https://doi.org/10.1101/397588.
27. Peng T, Zhu Q, Yin P, *et al*. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biol* 2019;**20**:88.
28. van Dijk D, Sharma R, Nainys J, *et al*. Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;**174**:716, e727–29.
29. Wagner F, Barkley D, Yanai I. Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis. *BioRxiv* 2019, doi: https://doi.org/10.1101/655365.
30. Wang J, Agarwal D, Huang M, *et al*. Data denoising with transfer learning in single-cell transcriptomics. *Nat Methods* 2019;**16**:875–8.
31. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* 2018;**9**:997.
32. Lopez R, Regier J, Cole MB, *et al*. Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8.
33. Gong W, Kwak IY, Pota P, *et al*. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* 2018;**19**:220.
34. Kim H, Park H. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J Matrix Anal Appl* 2008;**30**:713–30.
35. Kim Y-D, Choi S. Weighted nonnegative matrix factorization. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, p. 1541–4. IEEE Xplore, New York, NY, US.
36. Wang Z, Lai M-J, Lu Z, *et al*. Orthogonal rank-one matrix pursuit for low rank matrix completion. *SIAM J Sci Comput* 2015;**37**:A488–514.
37. Tabula Muris C, Overall C, Logistical C, *et al*. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 2018;**562**:367–72.

38. Guo C, Li B, Ma H, *et al*. Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. *Nat Commun* 2020;**11**:1–11.
39. Schulte-Schrepping J, Reusch N, Paclik D, *et al*. Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell* 2020;**182**:1419, e1423–40.
40. Lawson CL, Hanson RJ. *Solving least squares problems*. SIAM, 1995;**15**:158–65.
41. Levandowsky M, Winter D. Distance between sets. *Nature* 1971;**234**:34–5.
42. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 2017;**18**:174.
43. Baron M, Veres A, Wolock SL, *et al*. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst* 2016;**3**:346, e344–60.
44. Kiselev VY, Kirschner K, Schaub MT, *et al*. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;**14**:483.
45. Herdin M, Czink N, Ozcelik H *et al*. Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels. In: *2005 IEEE 61st Vehicular Technology Conference*. 2005, p. 136–40. IEEE Xplore, New York, NY, US.
46. Spruston N. Pyramidal neurons: dendritic structure and synaptic integration. *Nat Rev Neurosci* 2008;**9**: 206–21.
47. Fu Z, Elizabeth RG, Liu D. Regulation of insulin synthesis and secretion and pancreatic Beta-cell dysfunction in diabetes. *Curr Diabetes Rev* 2013;**9**:25–53.
48. Martin F, Kearney JF. Marginal-zone B cells. *Nat Rev Immunol* 2002;**2**:323–35.
49. Kleiman E, Salyakina D, De Heusch M, *et al*. Distinct transcriptomic features are associated with transitional and mature B-cell populations in the mouse spleen. *Front Immunol* 2015;**6**:30.
50. Newman R, Ahlfors H, Saveliev A, *et al*. Maintenance of the marginal-zone B cell compartment specifically requires the RNA-binding protein ZFP36L1. *Nat Immunol* 2017; **18**:683–93.
51. Wilk AJ, Rustagi A, Zhao NQ, *et al*. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat Med* 2020;**26**:1070–6.

Appendix A. Supplementary materials

The review of scRNA-seq techniques (Supplementary Table S1), The review of processing pipelines of scRNA-seq raw data (Supplementary Table S2), Introduction of imputation methods for scRNA-seq data (Supplementary Table S3), The time complexity analysis of WEDGE (Supplementary Note S1), Parameters for other imputation tools (Supplementary Note S2), Parameters for cell clustering (Supplementary Note S3), and settings for TSNE, UMAP, and heatmap visualization (Supplementary Note S4) can be found in Supplementary materials.