


Gene expression

EnImpute: imputing dropout events in single-cell RNA-sequencing data via ensemble learning

Xiao-Fei Zhang ¹, Le Ou-Yang^{2,*}, Shuo Yang³, Xing-Ming Zhao⁴, Xiaohua Hu⁵ and Hong Yan⁶

¹Department of Statistics, School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China, ²Guangdong Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China, ³Department of Respiratory Medicine, Wuhan Number 1 Hospital, Wuhan 430022, China, ⁴Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China, ⁵Department of Computer Science, College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA and ⁶Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on January 21, 2019; revised on April 10, 2019; editorial decision on May 18, 2019; accepted on May 21, 2019

Abstract

Summary: Imputation of dropout events that may mislead downstream analyses is a key step in analyzing single-cell RNA-sequencing (scRNA-seq) data. We develop EnImpute, an R package that introduces an ensemble learning method for imputing dropout events in scRNA-seq data. EnImpute combines the results obtained from multiple imputation methods to generate a more accurate result. A Shiny application is developed to provide easier implementation and visualization. Experiment results show that EnImpute outperforms the individual state-of-the-art methods in almost all situations. EnImpute is useful for correcting the noisy scRNA-seq data before performing downstream analysis.

Availability and implementation: The R package and Shiny application are available through Github at <https://github.com/Zhangxf-ccnu/EnImpute>.

Contact: leouyang@szu.edu.cn

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Single-cell RNA-sequencing (scRNA-seq), which measures global gene expression of individual cells, provides new opportunities to discover the mechanisms that cannot be seen from bulk RNA-sequencing data. Due to technical factors, dropout events, where transcripts are present in a cell but not detected, often occur in scRNA-seq experiments. This may hinder downstream analyses such as data visualization, cell clustering, cellular trajectory inference and differential expression analysis.

To address this issue, several computational methods have been developed recently to impute dropout events in scRNA-seq data (Chen *et al.*, 2018; Eraslan *et al.*, 2019; Huang *et al.*, 2018; Kwak

et al., 2018; Li and Li, 2018; Linderman *et al.*, 2018; Satija *et al.*, 2015; van Dijk *et al.*, 2018). These methods have shown diverse characteristics in terms of model assumptions and imputation strategies. For example, some methods impute dropout values for each cell by borrowing information from similar genes, while the remaining ones pool the data for each gene across similar cells. In addition, some methods use global strategies to impute the observed data, whereas the others adopt local strategies. Individual imputation methods may fail to recover the true gene expression levels when the model assumptions are not accurate. The performances of these methods depend heavily on the underlying data structures and evaluation approaches. None of these imputation methods is an

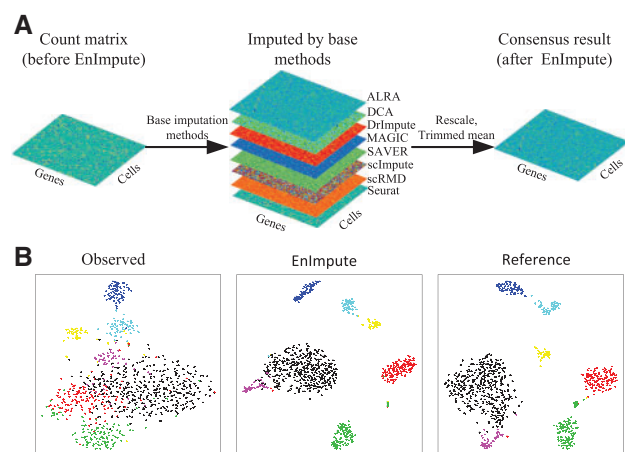


Fig. 1. Overview of EnImpute. **(A)** The workflow of EnImpute. **(B)** t-SNE visualization of the observed, imputed (by EnImpute) and reference datasets, color-coded by the clusters identified from the reference dataset (Color version of this figure is available at *Bioinformatics* online.)

apparent winner in all situations (Zhang and Zhang, 2018). Therefore, it is difficult to choose an optimal method for new data when there is no prior knowledge about the true expression levels.

By combining multiple results derived from different methods into a consensus result, ensemble learning is an effective strategy to deal with the above challenge. In this study, we introduce an ensemble learning method, named EnImpute, for imputing dropout values in scRNA-seq data. In the current implementation, EnImpute combines the results of eight individual methods using the trimmed mean. To provide a user-friendly tool, we develop an R package to implement the ensemble method. A Shiny application is also developed to facilitate easier implementation and visualization. Extensive experiments have shown the advantages of EnImpute over the individual imputation methods.

2 Materials and methods

EnImpute uses the trimmed mean to combine results from eight individual imputation methods: Adaptively-thresholded low rank approximation (ALRA) (Linderman et al., 2018), Deep count autoencoder network (DCA) (Eraslan et al., 2019), DrImpute (Kwak et al., 2018), Markov affinity-based graph imputation of cells (MAGIC) (van Dijk et al., 2018), Single-cell analysis via expression recovery (SAVER) (Huang et al., 2018), scImpute (Li and Li, 2018), scRMD (Chen et al., 2018) and Seurat (Satija et al., 2015). An overview of EnImpute is presented in Figure 1. The input of EnImpute is a raw read count matrix where the columns represent cells and the rows correspond to genes. EnImpute first runs the eight individual imputation methods independently to generate the base results. For ALRA, DrImpute, MAGIC, scRMD and Seurat, which do not implement library size normalization and log transformation in the imputation function, the input data is preprocessed using the normalization functions provided by the authors. After obtaining the imputation results from individual methods, EnImpute rescales the imputed data before combining them. First, the imputed expression levels in log-scale (ALRA, DrImpute, MAGIC, scRMD and Seurat) are exponentiated. Second, to eliminate the method-to-method variations in scale, EnImpute rescales the imputed results by

performing library normalization such that library size for each cell is 10 000. Third, the rescaled imputed data are log transformed with pseudocount 1. Finally, the consensus result is obtained by calculating the trimmed mean of the imputations generated by the base methods. The trimmed mean is used since it is more robust to outliers than the widely used mean that is computed as the sample average. Details are provided in Supplementary Section S1.

3 R package and Shiny web application

An R package, named EnImpute, is developed to implement the ensemble method. The main function of the package is EnImpute. The inputs include a raw read count matrix. Users can specify the library size for re-scaling the imputed results and the fraction of observations to be discarded before calculating the trimmed mean by setting the `scale.factor` and `trim` parameters. Users can also choose the individual imputation methods and set their parameters according to the arguments.

To provide a user-friendly tool, EnImpute is also implemented in a web application using the Shiny R package (Chang et al., 2017). The Shiny application interface is divided into three panels. The count matrix (a .csv file) can be uploaded, and the tuning parameters of each imputation method can be specified in the left panel. When the imputation is finished, t-SNE visualization of the raw data and the data imputed by EnImpute will be shown in the middle panel. The imputed data can be downloaded with the download button in the right panel. The detailed descriptions of the R package and Shiny application are presented in Supplementary Sections S2 and S3.

4 Results

We assess the performance through down-sampling experiments, differential expression analysis, and clustering and visualization analysis. The down-sampling experiments are conducted by following the method of Huang et al. (2018). Down-sampling on four scRNA-seq datasets are performed to generate the reference and observed data. We run EnImpute and the eight individual imputation methods on each of the observed data, and evaluate their performance by comparing the imputed data with the reference data. Experiment results show that EnImpute performs better than the individual methods on all datasets in terms of the correlation with reference data (both on the gene level and on the cell level), the recovery of cell-to-cell and gene-to-gene correlation matrices, and cell clustering and visualization (Supplementary Section S4). Differential expression analysis experiments show that EnImpute outperforms the other methods in increasing the agreement between bulk and single-cell differential expression analysis (Supplementary Section S5). Clustering and visualization analysis on three scRNA-seq datasets also reveal the advantage of EnImpute over individual methods (Supplementary Section S6). The effect of parameter `trim`, which specifies the fraction of observations to be trimmed, is analyzed in Supplementary Section S7.

5 Conclusions

We have developed an R package to introduce an ensemble method for imputing dropout events in scRNA-seq data. Besides EnImpute, DrImpute (Kwak et al., 2018) is also an ensemble learning-based imputation method. DrImpute integrates the results from the same

type of base imputation methods, whereas EnImpute combines the results from different types of base imputation methods that rely on different model assumptions. When compared with DrImpute, EnImpute can make full use of the advantages of different types of methods. In the future, more newly developed imputation methods will be integrated into EnImpute to improve the performance. Since imputing dropout events is becoming a routine step in scRNA-seq data analysis, EnImpute will serve a wide range of users for denoising the raw scRNA-seq data.

In this study, we integrate the base results from different imputation methods by taking the trimmed mean. Here, we use the trimmed mean since it is simple and can improve the performance of individual methods in most situations. In our opinion, more sophisticated ensemble methods that can fully take advantage of the strengths and weaknesses of the individual methods may produce better results. For example, a weighted ensemble approach, which learns a weight for each base imputation method and combines their results using a weighted mean, can be considered. In addition, it is well known that the zeros in scRNA-seq data can be divided into technical zeros that are caused by dropouts and biological zeros that reflect true biological non-expression. If we do not distinguish between technical and biological zeros and impute all zeros, the biologically non-expressed values will be altered incorrectly. However, determining which zeros are affected by dropouts is not easy. ALRA, SAVER, scImpute and scRMD have tried to distinguish the two types of zeros using different strategies and only impute the technical zeros. But experiment results show that they do not always outperform the methods that impute all the zeros (e.g. DCA, DrImpute, MAGIC and Seurat). This might be partially due to the problem that the two types of zeros are not distinguished accurately. In the future, we will develop new methods to decide which zeros are caused by dropouts and then impute these elements based on the imputations from the base methods.

Funding

This work was supported by the National Natural Science Foundation of China [11871026, 61532008, 61602309, 61772368, 61602347, 91530321 and 61572363], Natural Science Foundation of Hubei province [2018CFB521], self-determined research funds of CCNU from the colleges basic research and operation of MOE [CCNU18TS026], Shenzhen Research and Development program [JCYJ20170817095210760], Natural Science Foundation of SZU [2017077], Natural Science Foundation of Shanghai [17ZR1445600] and Hong Kong Research Grants Council [Projects C1007-15G and 11200818].

Conflict of Interest: none declared.

References

- Chang, W. *et al.* (2017) Shiny: web application framework for R. *R package*, page version 1.0.5.
- Chen, C. *et al.* (2018) scRMD: Imputation for single cell RNA-seq data via robust matrix decomposition. *bioRxiv*, 459404, doi: 10.1101/459404.
- Eraslan, G. *et al.* (2019) Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, **10**, 390.
- Huang, M. *et al.* (2018) Saver: gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, **15**, 539–542.
- Kwak, I.-Y. *et al.* (2018) Drimpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, **19**, 220.
- Li, W.-V. and Li, J.-J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, **9**, 997.
- Linderman, G. C. *et al.* (2018) Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv*, 397588; doi:10.1101/397588.
- Satija, R. *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
- van Dijk, D. *et al.* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, **174**, 1–14.
- Zhang, L. and Zhang, S. (2018) Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, doi: 10.1109/TCBB.2018.2848633.