Gene expression

2DImpute: imputation in single-cell RNA-seq data from correlations in two dimensions

Kaiyi Zhu *1,2 and Dimitris Anastassiou1,2,3,*

1Department of Systems Biology, Columbia University, New York, NY 10032, USA, 2Department of Electrical Engineering and 3Data Science Institute, Columbia University, New York, NY 10027, USA

*To whom correspondence should be addressed.

Received on October 31, 2019; revised on February 20, 2020; editorial decision on February 24, 2020; accepted on February 25, 2020

Abstract

Summary: We developed 2DImpute, an imputation method for correcting false zeros (known as dropouts) in single-cell RNA-sequencing (scRNA-seq) data. It features preventing excessive correction by predicting the false zeros and imputing their values by making use of the interrelationships between both genes and cells in the expression matrix. We showed that 2DImpute outperforms several leading imputation methods by applying it on datasets from various scRNA-seq protocols.

Availability and implementation: The R package of 2DImpute is freely available at GitHub (https://github.com/kzy0708/2DImpute).

Contact: d.anastassiou@columbia.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The explosive growth of single-cell RNA-sequencing (scRNA-seq) data in recent years has led to important biological discoveries (Rozenblatt-Rosen et al., 2017; Tiros and Suva, 2019; Zhong et al., 2018). Nevertheless, due to the low starting amounts of transcripts, a problem manifested in scRNA-seq data is the ‘dropout effect’ which results in many false zero values in the data (Neu et al., 2017) posing challenges for downstream analyses (Kiselev et al., 2019). To alleviate this problem, in addition to efforts to resolve technical issues, e.g. improving the transcript capture efficiency, imputation algorithms specific to scRNA-seq data have been developed for predicting those missing values due to dropouts. Some methods have been proposed integrating imputation with other processing tasks, such as normalization, batch effect correction and clustering (Prabhakaran et al., 2016; Tang et al., 2019). It is important, however, that imputation is also treated in isolation as simple dropout correction, so that users can have the option of building analysis pipelines processing dropout-corrected data specifically designed for their own particular tasks. The following five imputation methods: MAGIC (van Dijk et al., 2018); SAVER (Huang et al., 2018); scImpute (Li and Li, 2018); DrImpute (Gong et al., 2018); and VIPER (Cheng et al., 2013) are among the most widely used, and they rely on predicting values based on information from either similar cells or similar genes, using various approaches. All of these methods are constrained by several restrictions (Supplementary Material).

2 Materials and methods

Here we propose a novel scRNA-seq imputation approach, called 2DImpute (see detailed description in Supplementary Material). In brief, 2DImpute distinguishes dropout-suspected events from true biological zeros by using Jaccard distance-based cell-to-cell relationships and then imputes those predicted dropouts by leveraging the interrelationships across both genes and cells as follows. It first identifies co-expression signatures using the unsupervised ‘attractor metagene’ algorithm (Cheng et al., 2013) and imputes spurious zeros for the genes that are involved in such signatures. It then performs imputation on the remaining dropout values based on cell-to-cell relationships through k nearest neighbor regression across cells, making use of the previously imputed values of the co-expressed genes.

Because scRNA-seq datasets profiled from different protocols have distinct characteristics (Svensson et al., 2017), analyzing the performance of imputation on data profiled by various protocols could provide a more comprehensive assessment for each method. Therefore, to evaluate methods, we selected recently published, high quality, real scRNA-seq datasets (Azizi et al., 2018; Jerby-Arnon et al., 2018; Zheng et al., 2017) profiled on three prevalent scRNA-seq protocols (Smart-seq2, inDrop and 10× Genomics Chromium). Each dataset contains ~3000 cells and all data were pre-processed accordingly before imputation (Supplementary Material). To evaluate and compare different imputation techniques, we used three types of figure of merit based on several known facts serving as ground truth (Supplementary Material).
3 Results

We measured the accuracy of recovering artificially masked non-zeros, using two methods. First, we compared imputation methods in the particular case of fluorescence-activated cell sorting purified cells (Azizi et al., 2018), in which case we have the ground truth that all zeros in the fluorescent labeling gene (which is CD45 in this case) are false and in need of imputation. As shown in Figure 1b, 2DImpute is the only method that combines (i) the correct prediction that all zeros in CD45 are dropouts and has them imputed, while also (ii) retaining a separation between predicted true and false zeros for other genes (e.g. macrophage marker CD68). In addition, we validated the results in cases of other genes and datasets (Supplementary Figs S2 and S3). Second, we used Splatter-simulated (Zappia et al., 2017) datasets to provide ground truth for true-versus-false zeros (Supplementary Material). The results showed that 2DImpute is the only method that had balanced performance (Supplementary Table S1), by simultaneously exhibiting high sensitivity and high specificity (>0.5 for both).

Furthermore, we assessed imputation methods by examining their ability to recover cell-to-cell relationships by only strengthening ‘intra-cell’ correlations using a dataset consisting of a 50%:50% mixture of Jurkat cells and 293T cells (Zheng et al., 2017). 2DImpute was the only method that not only successfully reinforced the intra-cell correlation but also kept the inter-cell correlation to a low level (Fig. 1c). We obtained consistent results by using Pearson correlation (Supplementary Fig. S4). We also assessed the effects of imputation on gene-to-gene correlations by designing experiments based on permuted real data (Supplementary Table S2).

Imputation techniques are not required to be implemented in real-time; therefore, low computational complexity is not a fundamental requirement as long as it is reasonable. We compared the computational efficiencies of various methods, concluding that 2DImpute has reasonable complexity (Supplementary Table S3). We found that MAGIC was the fastest method, whereas VIPER had particularly slow performance in some cases.

4 Discussion

In conclusion, each of the prevalent methods has several drawbacks, while 2DImpute tends to balance the benefit of effective imputation and the risk of over imputation. It distinguishes dropouts from true zeros and leverages interrelationships both among genes and among cells. The algorithm does not require prior knowledge of the number of cell sub-populations; it does not make arbitrary assumptions of statistical models for expression distributions and is robust for use in various prevalent scRNA-seq platforms. Improvements in scRNA-seq imputation techniques will benefit subsequent analyses in many aspects, including network analysis, which is currently limited by false zero values.

Financial Support: none declared.

Conflict of Interest: none declared.

References

Azizi,E. et al. (2018) Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell, 174, 1293–1308.e36.
Chen,M. and Zhou,X. (2018) VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. Genome Biol., 19, 196.
Cheng,W.-Y. et al. (2013) Biomolecular events in cancer revealed by attractor metagenes. PLoS Comput. Biol., 9, e1002920.
Gong,W. et al. (2018) 2DImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics, 19, 220.
Huang,M. et al. (2018) SAVER: gene expression recovery for single-cell RNA sequencing. Nat. Methods, 15, 539–542.
Jerby-Aronn,L. et al. (2018) A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell, 175, 994–997.e24.
Kiselev,Y.V. et al. (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet., 20, 273–282.
Li,W.V. and Li,J.J. (2018) An accurate and robust imputation method scmpute for single-cell RNA-seq data. Nat. Commun., 9, 997.
Neu,K.E. et al. (2017) Single-cell genomics: approaches and utility in immunology. Trends Immunol., 38, 140–149.
Perou,C. et al. (2013) ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol., 16, 241.
Prabhakaran,S. et al. (2016) Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. JMLR Workshop Conf. Proc., 48, 1070–1079.
Rozenblatt-Rosen,O. et al. (2017) The Human Cell Atlas: from vision to reality. Nature, 550, 451–453.
Svensson,V. et al. (2017) Power analysis of single-cell RNA-sequencing experiments. Nat. Methods, 14, 381–387.
Tang,W. et al. (2019) bayNorm: Bayesian gene expression recovery, imputation and normalisation for single cell RNA-sequencing data. Bioinformatics, 35, 151–166.
Tirosh,I. and Suva,M.L. (2019) Deciphering human tumor biology by single-cell expression profiling. Annu. Rev. Cancer Biol., 3, 151–166.
van Dijk,D. et al. (2018) Recovering gene interactions from single-cell data using data diffusion. Cell, 174, 716–729.e27.
Zappia,L. et al. (2017) Splatter: simulation of single-cell RNA-sequencing data. Genome Biol., 18, 174.
Zheng,G.-Y. et al. (2017) Massively parallel digital transcriptional profiling of single cells. Nat. Commun., 8, 14049.
Zhong,S. et al. (2018) A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. Nature, 555, 524–528.