Single-cell RNA sequencing data imputation using deep neural network

Duc Tran
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
duct@nevada.unr.edu

Bang Tran
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
bang.t.s@nevada.unr.edu

Hung Nguyen
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
hungnp@nevada.unr.edu

Frederick C. Harris, Jr.
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
fred.harris@cse.unr.edu

Nam Sy Vo
Computational Biomedicine
Vingroup Big Data Institute
Hanoi, Vietnam
v.namvs@vintech.net.vn

Tin Nguyen*
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
tinn@unr.edu

***Abstract*—Recent research in biology has shifted the focus toward single-cell data analysis. The new single-cell technologies have allowed us to monitor and characterize cells in early embryonic stage and in heterogeneous tumor tissue. However, current single-cell RNA sequencing (scRNA-seq) technologies still need to overcome significant challenges to ensure accurate measurement of gene expression. One critical challenge is to address the dropout event. Due to the low amount of starting material, a large portion of expression values in scRNA-seq data is missing and reported as zeros. These missing values can greatly affect the accuracy of downstream analysis. Here we introduce scIRN, a neural network-based approach, that can reliably recover the missing values in single-cell data and thus can effectively improve the performance of downstream analyses. To impute the dropouts in single-cell data, we build a neural network that consists of two sub-networks: imputation sub-network and quality assessment sub-network. We compare scIRN with state-of-the-art imputation methods using 10 scRNA-seq datasets. In our extensive analysis, scIRN outperforms existing imputation methods in improving the identification of cell sub-populations and the quality of visualizing transcriptome landscape.**

***Index Terms*—single cell, scRNA-seq, imputation, sequencing, neural network, gene expression, residual network, dimension reduction, clustering, visualization**

I. Introduction

The ability to monitor and characterize biological samples at single-cell resolution has opened up many novel research fields, such as studying cells in early embryonic stage or decomposition heterogeneous environment of cancer tumor [1, 2]. These promising applications have led to the generation of a massive amount of single-cell data, where each dataset consists of hundreds of thousands of cells [3–5].

Current single-cell RNA sequencing (scRNA-seq) technologies still need to overcome significant challenges to ensure the accurate measurement of gene expression [6, 7]. One notable challenge of scRNA-seq is the dropout events, which happen when a highly expressed gene has no expression value in the sequencing data [8]. The sources of these errors can be attributed to the limitation of sequencing technologies. Due to the low amount of starting mRNA collected from individual cells, failed amplification can happen and causes the expression values to be inaccurately reported [9–11]. This leads to an excessive amount of zeros in the expression values of scRNA-seq data. On the other hand, the zero expression values can can also due to biological variability. Since downstream analyses of scRNA-seq are performed on gene expression data, it is essential to have a precise expression measurement. Therefore, imputing scRNA-seq data to recover the information loss caused by dropout would greatly improve the quality of downstream analyses.

To address the dropout challenge, a number of imputation methods have been developed to infer the missing data [12–19]. Those methods can be classified into two categories: (i) statistical-based methods, and (ii) diffusion smooth-based methods. Methods in the first category include bayNorm [12], SAVER [13], scImpute [14], scRecover [15], and RIA [17]. These methods typically model the data as a mixture of distributions. For example, scImpute models the gene expression as a mixture of two different distributions: the Gaussian distribution represents the actual gene expression while the Gamma distribution accounts for the dropout events. Similarly, SAVER [13] models read counts as a mixture of Poisson-Gamma and then uses a Bayesian approach to estimate true expression values of genes by borrowing information across genes. Another method, scRecover [15], uses uses the zero-inflated negative binomial model (ZINB) [20] to identify genes with zero-inflated expression. After identifying genes with true dropout, it uses the existing imputation methods such as scImpute, SAVER or MAGIC to impute the data. A more recent method, RIA [17], assumes that highly expressed genes follow a normal distribution and apply hypothesis testing method to identify true dropouts. Next, it imputes their values

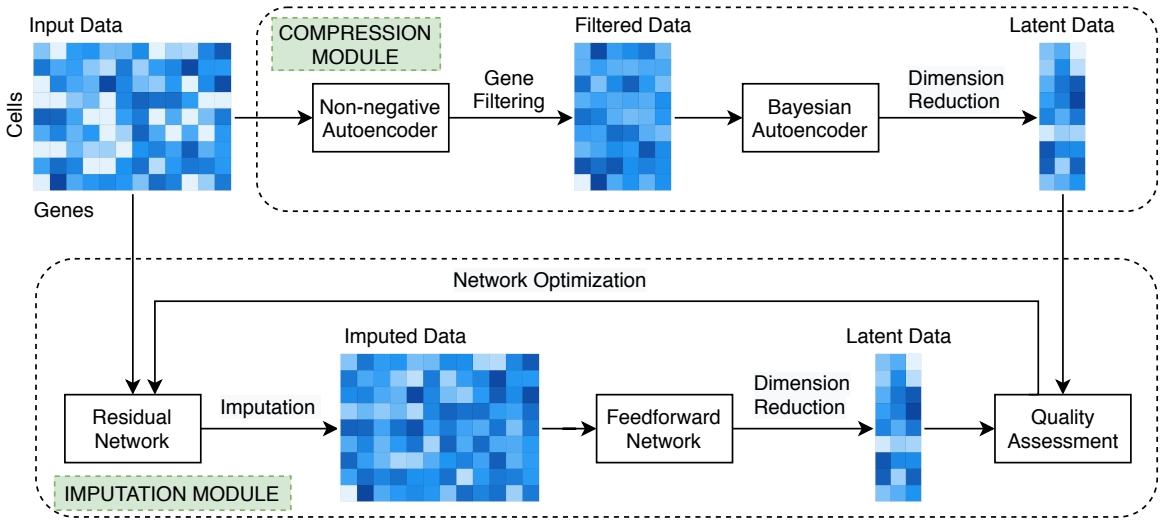

Fig. 1. The overall workflow of single-cell Imputation using Residual Network (scIRN). The first module (compression module) generates a compressed, low-dimensional representation of original data. The input data is first filtered (using an one-layer, non-negative kernel autoencoder) to remove genes that have insignificant contribution to the global structure of the data. After that, we project the data into a low-dimensional space to obtain a compressed data matrix (latent data). This latent data is used as the training target for the imputation process. In the second module (imputation module), zero values in input matrix are imputed using a neural network-based imputation model. These imputed values are added to original data without modifying the non-zeros values to produce the imputed data matrix. The imputed data is compressed to a low-dimensional space (latent data). The parameters of the imputation module is repeatedly optimized by minimizing the difference between the two latent matrices.

by using linear regression model. All of these methods assume the gene expression data follows a specific distribution, which does not always hold true in the reality. In addition, exiting methods involve the estimation of many parameters for all genes across the whole genome. This can potentially lead to overfitting and high time complexity.

Methods in the second category include DrImpute [16], MAGIC [18], and kNN-smoothing [19]. MAGIC imputes zero expression values using a heat diffusion algorithm [21]. It constructs the affinity matrix between cells using Gaussian kernel and then constructs a Markov transition matrix by normalizing the sc-RNA similarity matrix. Next, MAGIC estimates the weights of other cells using the transition matrix. Another method is DrImpute [16] that is based on the cluster ensemble [22] and consensus clustering [23, 24]. It performs clustering for a predefined number of times and imputes the data by averaging expression values of similar cells. If the number of clusters is not provided by users, DrImpute uses some default values that might not be optimal for the data. kNN-smoothing is designed to reduce noise by aggregating information from similar cells (neighbors). The method assumes that the zero counts of scRNA-seq data follows a Poisson distribution. For cells that contain zero counts, kNN-smoothing performs a smoothing step using each cell's k nearest neighbors either through the application of diffusion models or weighted sums respectively. The major drawback of these methods is that they rely on many parameters to fine-tune their model, which often leads to over-smoothing the data.

Here we propose a new approach, single-cell Imputation using Residual Network (scIRN), that can reliably impute missing values from single-cell data. Our method consists of two steps. The first step is to generate a compressed and accurate low-dimensional representation of the original data. The second step is to estimate the missing values using a neural network and information from the low-dimensional representation. The approach is tested using 10 single-cell datasets in comparison with four other methods. We demonstrate that scIRN outperforms existing imputation methods (MAGIC [18], scImpute [14], SAVER [13], and DrImpute [16]) in improving the identification of cell sub-populations and the quality of biological landscape.

II. Methods

The input of scIRN is an expression matrix, in which rows represent cells and columns represent genes or transcripts. The overall workflow of scIRN is described in Figure 1, which consists of two modules: (i) generating a low-dimensional, non-redundant representation of the original data, and (ii) imputing the dropout values. The purpose of the first module is to remove redundant signals and noise from the data. The output of the first module is a low-dimensional, non-redundant representation of the original data. This presentation is used as the target for the second module. In the second module, we impute the original data using a residual network. The parameters of the residual network are repeatedly adjusted so that the compressed representation of the imputed data is as similar to the non-redundant representation as possible. The details of each step are described in the following sections.

A. Generating low-dimensional, non-redundant representation

To generate a compressed, low-dimensional representation of original data, we apply our previously developed method, called scDHA [34]. scDHA consists of two core modules.

TABLE I
DESCRIPTION OF THE 10 SINGLE-CELL DATASETS USED TO ASSESS THE PERFORMANCE OF IMPUTATION METHODS.

Dataset	Tissue	Size	Class	Protocol	Accession ID	Reference
1. Deng	Mouse Embryo	268	6	Smart-Seq2	GSE45719	Deng *et al.*, 2014 [2]
2. Pollen	Human Tissues	301	11	SMARTer	SRP041736	Pollen *et al.*, 2014 [25]
3. Usoskin	Mouse Brain	622	4	STRT-Seq	GSE59739	Usoskin *et al.*, 2015 [26]
4. Kolodziejczyk	Mouse Embryo Stem Cells	704	3	SMARTer	E-MTAB-2600	Kolodziejczyk *et al.*, 2015 [27]
5. Xin	Human Pancreas	1,600	8	SMARTer	GSE81608	Xin *et al.*, 2016 [28]
6. Muraro	Human Pancreas	2,126	10	CEL-Seq2	GSE85241	Muraro *et al.*, 2016 [29]
7. Klein	Mouse Embryo Stem Cells	2,717	4	inDrop	GSE65525	Klein *et al.*, 2015 [30]
8. Romanov	Mouse Brain	2,881	7	SMARTer	GSE74672	Romanov *et al.*, 2017 [31]
9. Zeisel	Mouse Brain	3,005	9	STRT-Seq	GSE60361	Zeisel *et al.*, 2015 [32]
10. Baron	Human Pancreas	8,569	14	inDrop	GSE84133	Baron *et al.*, 2016 [33]

The first module is a non-negative kernel autoencoder that can filter out genes or components that have insignificant contributions to data representation. The second module is a Stacked Bayesian Self-learning Network that is built upon the Variational Autoencoder [35] to project the filtered data onto a much lower-dimensional space. The output of scDHA is a low-dimensional matrix that preserves the global structure of the original data. This representation is used as the training target for the imputation module.

B. Imputing dropout data using residual network

To impute the dropouts in single-cell data, we build a neural network that consists of two sub-networks. The first network aims to infer the true value of zeros in the data. The output is a matrix with the same size as the input, in which the values at zero positions are modified. The non-zero values remain the same as of the original data. The second network aims to compress the imputed data to a lower dimension. This compressed data has the same size as the representation generated in the first step. By minimizing the difference between the representation generated from imputed data and the representation from the first step, the imputed values are ensured to have high accuracy.

The formulation of the neural network can be written as:

$$\begin{aligned} X_I &= f_I(X) \\ Z' &= f_C(X_I) \end{aligned}$$

where $X \in R_+^n$ is the input of the model (X is simply the original data), f_I and f_C represent the transformation by the two sub-networks, f_I imputes the zero values in the data, f_C compresses the imputed data onto a lower-dimensional space, and $Z' \in R^m$ ($m << n$) is the compressed data. For the f_I transformation, we use residual network [36] for a more stable and accurate imputation process. The network is optimized by minimizing $\|Z' - Z\|_2^2$, where Z is the low-dimensional representation generated by scDHA.

III. RESULTS

We compares our method with four state-of-the-art imputation methods: MAGIC [18], scImpute [14], SAVER [13], and DrImpute [16]. Each of these methods represents a distinct strategy to single-cell data imputation: MAGIC is a Markov-based technique, DrImpute integrates clustering result from other software, while scImpute and SAVER use statistical models. Table I shows the 10 datasets used in our data analysis. The processed datasets were downloaded from Hemberg lab's website (https://hemberg-lab.github.io/scRNA.seq.datasets). In each dataset, the cell sub-populations are known. We used this information *a posteriori* to assess how the imputation methods improve the identification of cell populations, and how they enhance the visualization of transcriptome landscapes.

For each dataset, we used the above methods to impute the data. The quality of the imputed data is assessed using two downstream analyses, clustering and visualization. For clustering, we partitioned the data using k-means and compared the obtained partitioning against the true cell types using Adjusted Rand index (ARI) [37]. For visualization, we used UMAP [38] to generate the 2D representation and then calculated the silhouette index (SI) [39] of the 2D representation. SI measures the cohesion among cells of the same type, as well as the separation between different cell types.

A. scIRN improves the identification of sub-populations

Given a dataset, we used the five methods to impute the data. After imputation, we have 6 matrices: the raw data and five imputed matrices (from MAGIC, scImpute, SAVER, DrImpute, and scIRN). To assess how separable the cell types in each matrix is, we reduced the number of dimensions using PCA and then clustered the data using k-means. The accuracy of cluster assignments is measured by ARI.

Figure 2 shows the ARI values for the raw and imputed data. Existing methods improve cluster analysis in some datasets but decreases the ARI values in some others. For example, MAGIC has higher ARIs than the raw data for the Deng, Usoskin, Muraro, Klein, Romanov, and Baron but has lower ARIs in the remaining 4 datasets. scIRN is the only method able to improve the clustering performance compared to raw data in every dataset. Moreover, scIRN has the highest ARIs in all but Usoskin datasets. The average ARI of scIRN-imputed data is 0.77, which is higher than those obtained from raw data and data imputed by MAGIC, scImpute, SAVER, DrImpute (0.44, 0.41, 0.46, 0.43, 0.58, respectively).

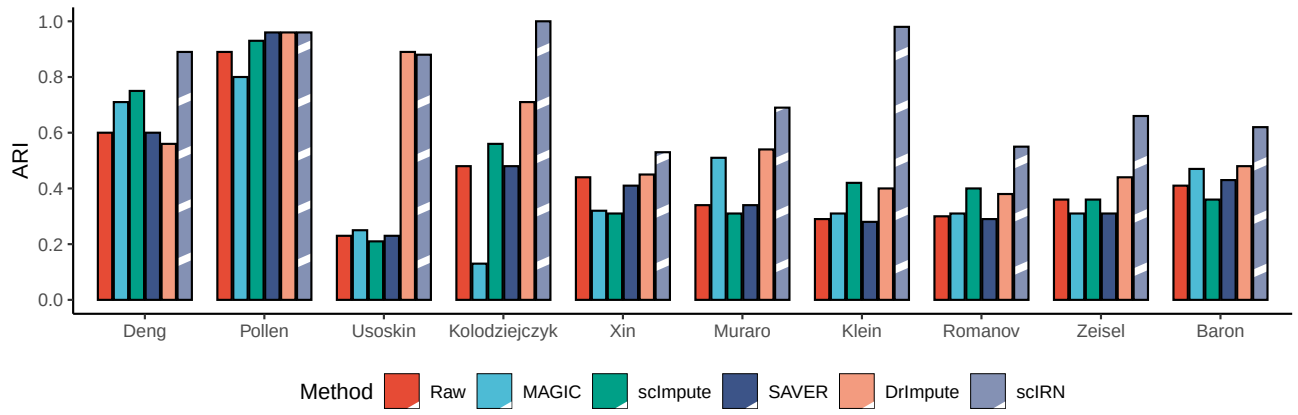

Fig. 2. Adjusted Rand index (ARI) obtained from clustering on raw data and data imputed by MAGIC, SAVER, scImpute, DrImpute, and scIRN. The x-axis shows the names of the datasets while the y-axis shows ARI value of each method. scIRN outperforms other methods in all datasets except Usoskin.

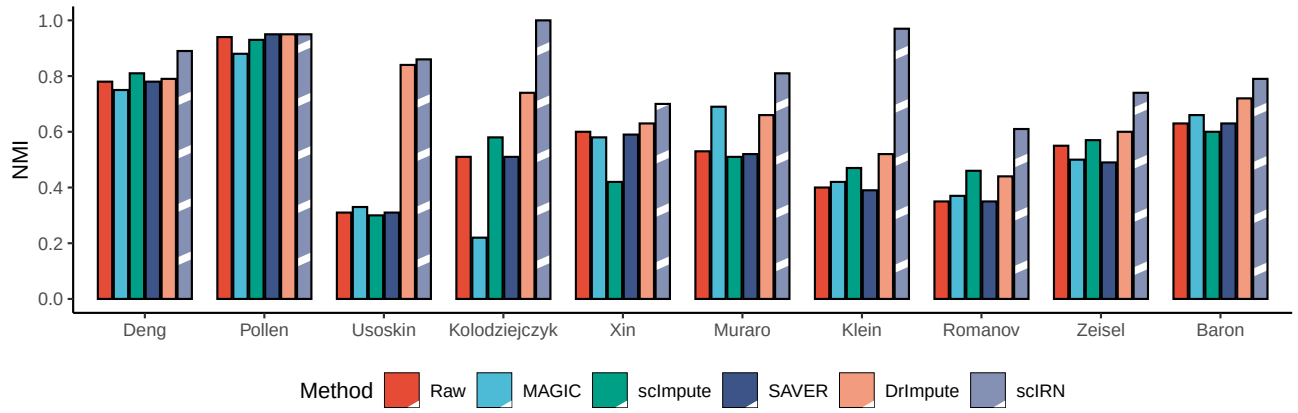

Fig. 3. Normalized mutual information (NMI) obtained from clustering on raw data and data imputed by MAGIC, SAVER, scImpute, DrImpute, and scIRN. The x-axis shows the names of the datasets while the y-axis shows NMI value of each method. scIRN outperforms other methods in all datasets.

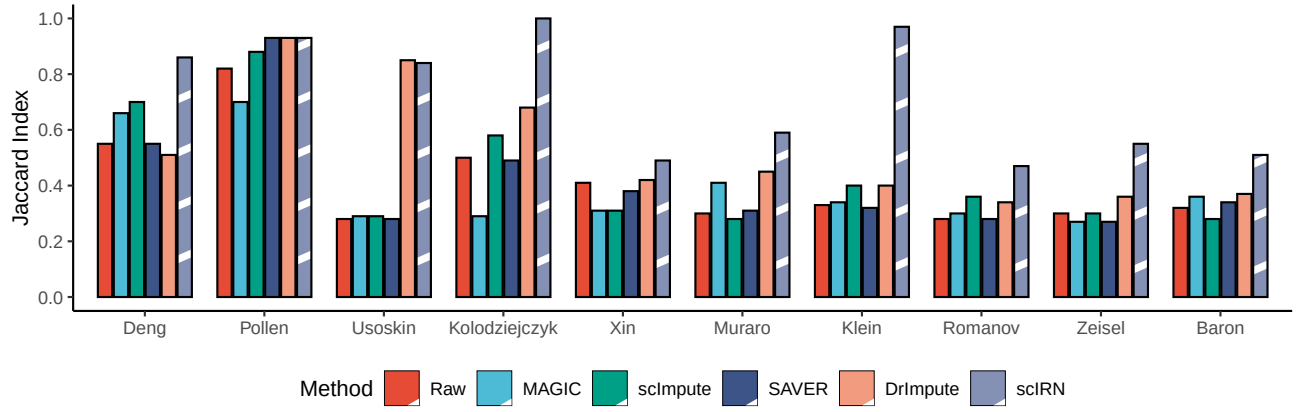

Fig. 4. Jaccard index (JI) obtained from clustering on raw data and data imputed by MAGIC, SAVER, scImpute, DrImpute, and scIRN. The x-axis shows the names of the datasets while the y-axis shows JI value of each method. scIRN outperforms other methods in all datasets except Usoskin.

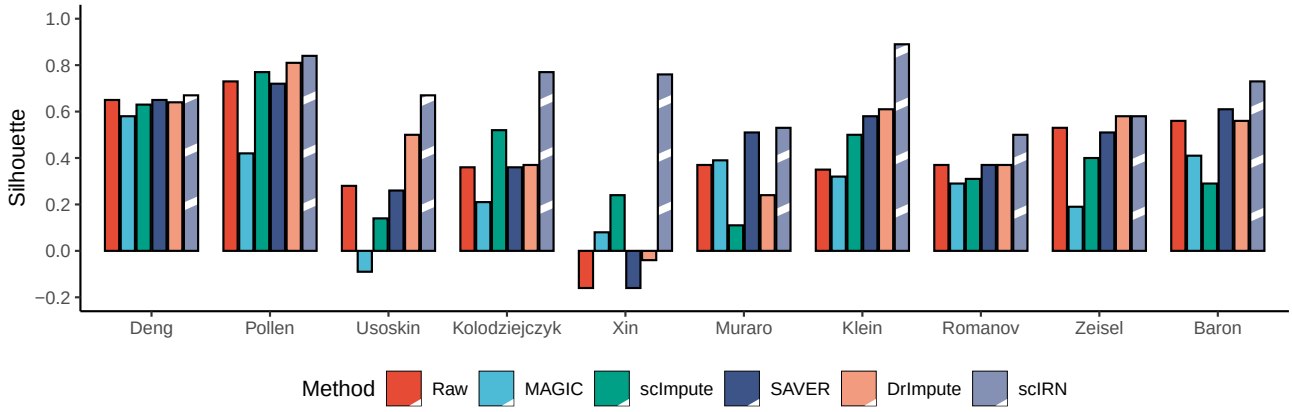

Fig. 5. Visualization quality using raw and imputed data, measured by silhouette index (SI). The x-axis shows the names of the datasets while the y-axis shows SI value of each method. scIRN outperforms other methods in all datasets.

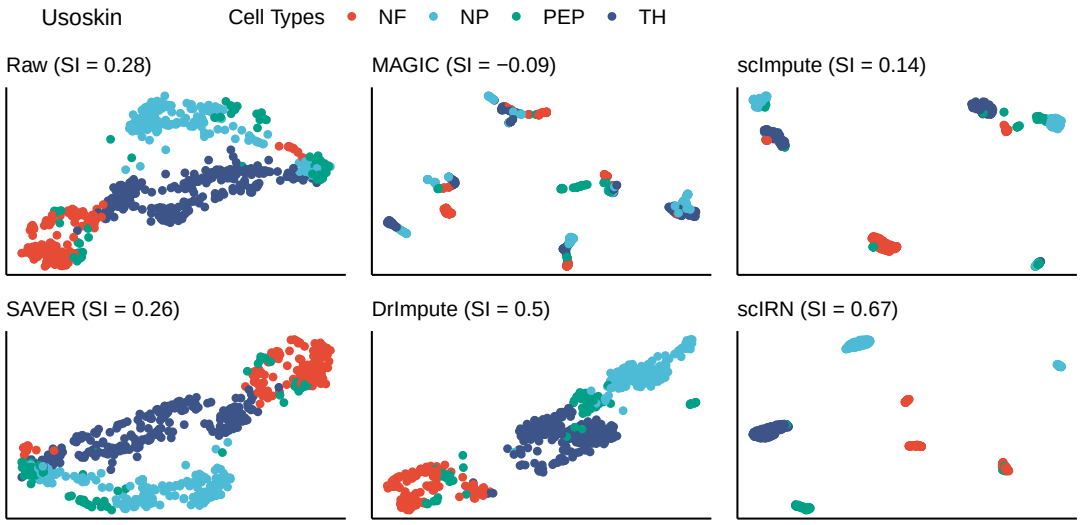

Fig. 6. Transcriptome landscape of the Usoskin dataset. The scatter plot shows the first two principal components calculated by UMAP. Different colors represent different cell types. The 2D representation generated by scIRN has a clear structure, where cells from different groups are separated from one other.

For a more comprehensive analysis, we also report the assessment using normalized mutual information (NMI) and Jaccard index (JI) in Figures 3 and 4, respectively. Regardless of the assessment metrics, scIRN outperforms other methods by having the highest NMI (10/10 datasets) and JI (9/10 datasets) values. These results demonstrate that cluster analysis using scIRN-imputed data leads to a better accuracy than using the raw data or data imputed by other imputation methods.

B. scIRN improves transcriptome landscape visualization

In this section, we demonstrate that scIRN improves the visualization of the single-cell data. We used UMAP [38] to generate the transcriptome landscapes from raw and data imputed by MAGIC, scImpute, SAVER, DrImpute, and scIRN. We performed data visualization and calculated the silhouette index for each of the 10 datasets. Figure 5 shows the SI values obtained for the raw data and data imputed by the five imputation methods. The figure shows that scIRN can improve the quality of data visualization in all datasets. scIRN also has the highest SI in each of these datasets. These results demonstrate that data imputation using scIRN would lead to a much better visualization of transcriptome landscapes compared to using raw data or data imputed by other methods.

Figure 6 shows the transcriptome landscapes of the Usoskin dataset. Using scIRN imputed data, UMAP was able to generate a clear representation, where cells from different groups are well-separated. When using data imputed by other methods, cells are usually mixed together. scIRN outperformed other imputation methods by having the highest SI value (0.67 compared to 0.28, -0.09, 0.14, 0.26, 0.5 of raw data, MAGIC, scImpute, SAVER, and DrImpute, respectively).

Figure 7 shows the transcriptome landscapes of the Klein dataset. The 2D representation of scIRN-imputed data is the only one that has four separable groups, corresponding to the

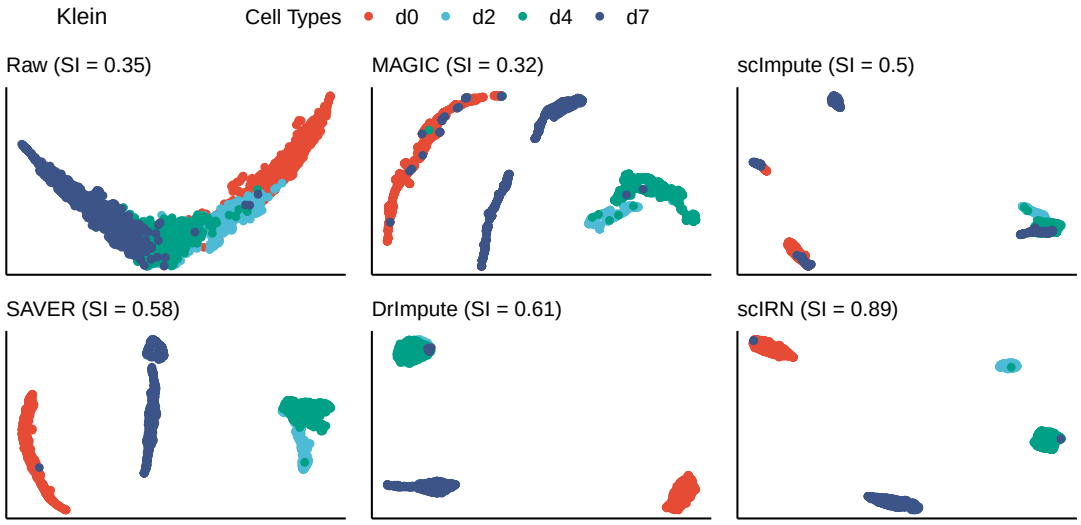

Fig. 7. Transcriptomics landscape of the Klein dataset. The scatter plot shows the first two principal components calculated by UMAP for raw and imputed data. The 2D representation generated from scIRN has a clear structure, where cells from different groups are separate from each other.

four real cell types. The landscapes generated using raw and data imputed by other methods have different cell types mixed together. The data imputed by scIRN has the highest SI value (0.89 compared to 0.61 of the second best).

IV. Conclusion

In this article, we introduce a new method, scIRN, to recover the missing data caused by dropout events in scRNA-seq. We assess the performance of our approach using 10 single-cell datasets in a comparison with four current state-of-the-art imputation methods. Our analysis shows that scIRN outperforms existing approaches in improving the identification of cell sub-populations. scIRN also improves the quality of transcriptome landscapes generated by UMAP. A potential improvement of this research is to investigate the scalability of scIRN by analyzing datasets with higher number of cells. Another direction is to investigate the imputation method in other research applications, including pseudo-time trajectory inference and supervised learning. For future work, we will combine scIRN with current methods to improve the quality of downstream data analysis in the context of gene networks [40–47] and multi-omics integration [48–53].

V. Acknowledgments

This work was partially supported by NASA under grant number NNX15AI02H (subaward no. 21-02), by NIH NIGMS under grant number GM103440, and by NSF under grant numbers 2001385 and 2019609. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

References

[1] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev, and B. E. Bernstein, "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, no. 6190, pp. 1396–1401, 2014.

[2] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells," *Science*, vol. 343, no. 6167, pp. 193–196, 2014.

[3] P. A. Darrah, J. J. Zeppa, P. Maiello, J. A. Hackney, M. H. Wadsworth, T. K. Hughes, S. Pokkali, P. A. Swanson, N. L. Grant, M. A. Rodgers, M. Kamath, C. M. Causgrove, D. J. Laddy, A. Bonavia, D. Casimiro, P. L. Lin, E. Klein, A. G. White, C. A. Scanga, A. K. Shalek, M. Roederer, J. L. Flynn, and R. A. Seder, "Prevention of tuberculosis in macaques after intravenous BCG immunization," *Nature*, vol. 577, no. 7788, pp. 95–102, 2020.

[4] L. D. Orozco, H.-H. Chen, C. Cox, K. J. Katschke Jr, R. Arceo, C. Espiritu, P. Caplazi, S. S. Nghiem, Y.-J. Chen, Z. Modrusan, A. Dressen, L. D. Goldstein, C. Clarke, T. Bhangale, B. Yaspan, M. Jeanne, M. J. Townsend, M. v. L. Campagne, and J. A. Hackney, "Integration of eQTL and a single-cell atlas in the human eye identifies causal genes for age-related macular degeneration," *Cell Reports*, vol. 30, no. 4, pp. 1246–1259, 2020.

[5] V. Kozareva, C. Martin, T. Osorno, S. Rudolph, C. Guo, C. Vanderburg, N. M. Nadaf, A. Regev, W. Regehr, and E. Macosko, "A transcriptomic atlas of the mouse cerebellum reveals regional specializations and novel cell types," *bioRxiv*, 2020.

[6] P. Brennecke, S. Anders, J. K. Kim, A. A. Kolodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Te-

ichmann, J. C. Marioni, and M. G. Heisler, "Accounting for technical noise in single-cell RNA-seq experiments," *Nature Methods*, vol. 10, no. 11, pp. 1093–1095, 2013.

[7] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells," *Nature Biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[8] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature Methods*, vol. 11, no. 7, pp. 740–742, 2014.

[9] S. Rizzetto, A. A. Eltahla, P. Lin, R. Bull, A. R. Lloyd, J. W. Ho, V. Venturi, and F. Luciani, "Impact of sequencing depth and read length on single cell RNA sequencing data of T cells," *Scientific Reports*, vol. 7, p. 12781, 2017.

[10] S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, and I. Hellmann, "The impact of amplification on differential expression analyses by RNA-seq," *Scientific Reports*, vol. 6, p. 25533, 2016.

[11] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications," *Genome Medicine*, vol. 9, no. 1, p. 75, 2017.

[12] W. Tang, F. Bertaux, P. Thomas, C. Stefanelli, M. Saint, S. Marguerat, and V. Shahrezaei, "baynorm: Bayesian gene expression recovery, imputation and normalization for single-cell rna-sequencing data," *Bioinformatics*, vol. 36, no. 4, pp. 1174–1181, 2020.

[13] M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang, "SAVER: gene expression recovery for single-cell RNA sequencing," *Nature Methods*, vol. 15, no. 7, pp. 539–542, 2018.

[14] W. V. Li and J. J. Li, "An accurate and robust imputation method scImpute for single-cell RNA-seq data," *Nature Communications*, vol. 9, p. 997, 2018.

[15] Z. Miao, J. Li, and X. Zhang, "screcover: Discriminating true and false zeros in single-cell rna-seq data for imputation," *bioRxiv*, p. 665323, 2019.

[16] W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry, "DrImpute: imputing dropout events in single cell RNA sequencing data," *BMC Bioinformatics*, vol. 19, p. 220, 2018.

[17] B. Tran, D. Tran, H. Nguyen, N. S. Vo, and T. Nguyen, "Ria: a novel regression-based imputation approach for single-cell rna sequencing," in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2019, pp. 1–9.

[18] D. Van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er, "Recovering gene interactions from single-cell data using data diffusion," *Cell*, vol. 174, no. 3, pp. 716–729, 2018.

[19] F. Wagner, Y. Yan, and I. Yanai, "K-nearest neighbor smoothing for high-throughput single-cell rna-seq data," *BioRxiv*, p. 217737, 2017.

[20] A. M. Garay, E. M. Hashimoto, E. M. Ortega, and V. H. Lachos, "On estimation and influence diagnostics for zero-inflated negative binomial regression models," *Computational Statistics & Data Analysis*, vol. 55, no. 3, pp. 1304–1318, 2011.

[21] Z. I. Botev, J. F. Grotowski, D. P. Kroese *et al.*, "Kernel density estimation via diffusion," *The Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.

[22] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[23] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. 1-2, pp. 91–118, 2003.

[24] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, and M. Hamberg, "SC3: consensus clustering of single-cell RNA-seq data," *Nature Methods*, vol. 14, no. 5, pp. 483–486, 2017.

[25] A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, D. W. Kemp Ii, M. Wong, B. Clerkson, B. N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L. S. Weaver, A. P. May, R. C. Jones, M. A. Unger, A. R. Kriegstein, and J. A. A. West, "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex," *Nature Biotechnology*, vol. 32, no. 10, pp. 1053–1058, 2014.

[26] D. Usoskin, A. Furlan, S. Islam, H. Abdo, P. Lönnerberg, D. Lou, J. Hjerling-Leffler, J. Haeggström, O. Kharchenko, P. V. Kharchenko, S. Linnarson, and P. Ernfors, "Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing," *Nature Neuroscience*, vol. 18, no. 1, pp. 145–153, 2015.

[27] A. A. Kolodziejczyk, J. K. Kim, J. C. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Bühler, P. Liu, J. C. Marioni, and S. A. Teichmann, "Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation," *Cell Stem Cell*, vol. 17, no. 4, pp. 471–485, 2015.

[28] Y. Xin, J. Kim, H. Okamoto, M. Ni, Y. Wei, C. Adler, A. J. Murphy, G. D. Yancopoulos, C. Lin, and J. Gromada, "RNA sequencing of single human islet cells reveals type 2 diabetes genes," *Cell Metabolism*, vol. 24, no. 4, pp. 608–615, 2016.

[29] M. J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti, E. J. de Koning, and A. van Oudenaarden, "A

Single-Cell Transcriptome Atlas of the Human Pancreas," *Cell Systems*, vol. 3, no. 4, pp. 385–394.e3, 2016.

[30] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, 2015.

[31] R. A. Romanov, A. Zeisel, J. Bakker, F. Girach, A. Hellysaz, R. Tomer, A. Alpár, J. Mulder, F. Clotman, E. Keimpema, B. Hsueh, A. K. Crow, H. Martens, C. Schwindling, D. Calvigioni, J. S. Bains, Z. Máté, G. Szabó, Y. Yanagawa, M.-D. Zhang, A. Rendeiro, M. Farlik, M. Uhlén, P. Wulff, C. Bock, C. Broberger, K. Deisseroth, T. Hökfelt, S. Linnarsson, T. L. Horvath, and T. Harkany, "Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes," *Nature Neuroscience*, vol. 20, no. 2, pp. 176–188, 2017.

[32] A. Zeisel, A. B. Munoz-Manchado, S. Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson, "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq," *Science*, vol. 347, no. 6226, pp. 1138–1142, 2015.

[33] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, D. A. Melton, and I. Yanai, "A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure," *Cell Systems*, vol. 3, no. 4, pp. 346–360, 2016.

[34] D. Tran, H. Nguyen, B. Tran, C. La Vecchia, H. N. Luu, and T. Nguyen, "Fast and precise single-cell data analysis using hierarchical autoencoder," *bioRxiv*, p. 799817, 2019.

[35] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114 [cs, stat]*, Dec. 2013.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[37] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[38] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[39] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[40] H. Nguyen, D. Tran, B. Tran, B. Pehlivan, and T. Nguyen, "A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data," *Briefings in Bioinformatics*, p. bbaa190, 2020.

[41] T. Nguyen, A. Shafi, T.-M. Nguyen, A. G. Schissler, and S. Draghici, "NBIA: a network-based integrative analysis framework–applied to pathway analysis," *Nature Scientific Reports*, vol. 10, p. 4188, 2020.

[42] T.-M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici, "Identifying significantly impacted pathways: a comprehensive review and assessment," *Genome Biology*, vol. 20, no. 1, p. 203, 2019.

[43] H. Nguyen, S. Shrestha, D. Tran, A. Shafi, S. Draghici, and T. Nguyen, "A comprehensive survey of tools and software for active subnetwork identification," *Frontiers in Genetics*, vol. 10, p. 155, 2019.

[44] T. Nguyen, C. Mitrea, and S. Draghici, "Network-based approaches for pathway level analysis," *Current Protocols in Bioinformatics*, vol. 61, no. 1, pp. 8–25, 2018.

[45] T. Nguyen, C. Mitrea, R. Tagett, and S. Draghici, "DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions - applied to biological pathway analysis," *Proceedings of the IEEE*, vol. 105, no. 3, pp. 496–515, 2017.

[46] T. Nguyen, D. Diaz, R. Tagett, and S. Draghici, "Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data," *Nature Scientific Reports*, vol. 6, p. 29251, 2016.

[47] T. Nguyen, R. Tagett, M. Donato, C. Mitrea, and S. Draghici, "A novel bi-level meta-analysis approach-applied to biological pathway analysis," *Bioinformatics*, vol. 32, no. 3, pp. 409–416, 2016.

[48] H. Nguyen, S. Shrestha, S. Draghici, and T. Nguyen, "PINSPlus: A tool for tumor subtype discovery in integrated genomic data," *Bioinformatics*, vol. 35, no. 16, pp. 2843–2846, 2019.

[49] A. Shafi, T. Nguyen, A. Peyvandipour, and S. Draghici, "GSMA: an approach to identify robust global and test Gene Signatures using Meta-Analysis," *Bioinformatics*, vol. 36, no. 2, pp. 487–495, 2019.

[50] T. Nguyen, R. Tagett, D. Diaz, and S. Draghici, "A novel approach for data integration and disease subtyping," *Genome Research*, vol. 27, no. 12, pp. 2025–2039, 2017.

[51] A. Shafi, T. Nguyen, A. Peyvandipour, H. Nguyen, and S. Draghici, "A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures," *Frontiers in Genetics*, vol. 10, p. 159, 2019.

[52] A. Shafi, C. Mitrea, T. Nguyen, and S. Draghici, "A survey of the approaches for identifying differential methylation using bisulfite sequencing data," *Briefings in Bioinformatics*, vol. 19, no. 5, pp. 737–753, 2018.

[53] M. Menden, D. Wang, Y. Guan, M. Mason, B. Szalai, K. Bulusu, T. Yu, J. Kang, M. Jeon, R. Wolfinger, T. Nguyen, M. Zaslavskiy, S. Jang, Z. Ghazoui, M. Ahsen, R. Vogel, E. Neto, T. Norman, E. Tang, M. Garnett, G. Veroli, C. Zwaan, S. Fawell, G. Stolovitzky, J. Guinney, J. Dry, and J. Saez-Rodriguez, "Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen," *Nature Communications*, vol. 10, no. 1, p. 2674, 2019.