Single-cell RNA sequencing data imputation using similarity preserving network

Duc Tran
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
duct@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

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

2021 13th International Conference on Knowledge and Systems Engineering (KSE) | 978-1-6654-9975-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/KSE53942.2021.9648794

***Abstract*—Recent advancements in single-cell RNA sequencing (scRNA-seq) technologies have allowed us to monitor the gene expression of individual cells. This level of detail in monitoring and characterization enables the research of cells in rapidly changing and heterogeneous environments such as early stage embryo or tumor tissue. However, the current scRNA-seq technologies are still facing many outstanding challenges. Due to the low amount of starting material, a large portion of expression values in scRNA-seq data is missing and reported as zeros. Moreover, scRNA-seq platforms are trending toward prioritizing high throughput over sequencing depth, which makes the problem become more serious in large datasets. These missing values can greatly affect the accuracy of downstream analyses. Here we introduce scINN, 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 scINN with state-of-the-art imputation methods using 10 scRNA-seq datasets with a total of more than 100,000 cells. In an extensive analysis, we demonstrate that scINN outperforms existing imputation methods in improving the identification of cell sub-populations and the quality of transcriptome landscape visualization.**

***Index Terms*—single cell, scRNA-seq, imputation, neural network, gene expression, 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 tumors [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, 4].

Current single-cell RNA sequencing (scRNA-seq) technologies still need to overcome significant challenges to ensure the accurate measurement of gene expression [5, 6]. One notable challenge of scRNA-seq is the dropout events, which happen when a gene that generally has high expression values but does not express in some cells [7]. The source 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 [8–10]. 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 also be due to biological variability. Since most 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 events would greatly improve the quality of downstream analyses.

Thus far, numerous methods have been developed to infer the missing values caused by dropout events [11–18]. 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 [11], SAVER [12], scImpute [13], scRecover [19], and RIA [15]. 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 [12] 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. More recent methods, RIA [15] and scIRN [18], assume that highly expressed genes follow a normal distribution and apply hypothesis testing method to identify true dropouts. Next, they impute missing values by using a linear regression model. All of these methods assume the gene expression data follows a specific distribution, which does not always hold true in reality. In addition, exiting methods involve the estimation of many parameters for genes across the whole genome. This

978-1-6654-9975-0/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: UNIVERSITY OF IDAHO. Downloaded on May 27,2022 at 19:44:28 UTC from IEEE Xplore. Restrictions apply.

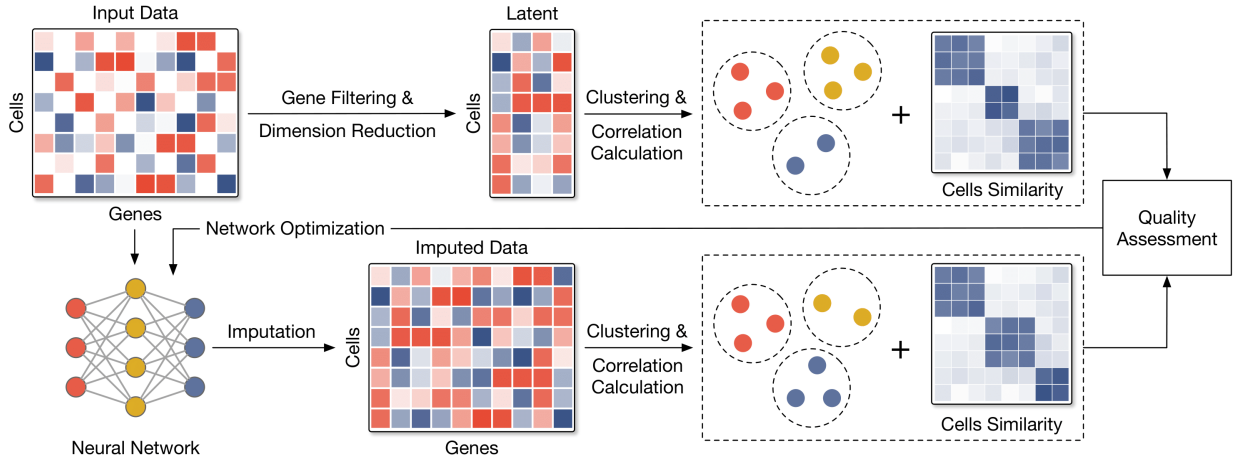

Fig. 1. The workflow of single-cell Imputation using Residual Network (scINN). The first module (similarity module, upper part) generates an accurate clustering result of the data, and calculates the similarity between all pairs of samples. The input data is first filtered using a one-layer, non-negative kernel autoencoder to remove genes that have insignificant contribution to the global structure of the data. Next, the data is projected onto a low-dimensional space to obtain a compressed data matrix (latent data). Using this latent data, we cluster the samples into groups and compute the similarity matrix for all samples. In the second module (imputation module, lower part), 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. The parameters of the neural network are repeatedly adjusted so that the clustering assignments and similarity matrix inferred from the imputed data is as similar to the output of the first module as possible.

can potentially lead to overfitting and high time complexity.

Methods in the second category include DrImpute [14], MAGIC [16], and kNN-smoothing [17]. MAGIC imputes zero expression values using a heat diffusion algorithm [20]. It constructs the affinity matrix between cells using a 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 [14] that is based on the cluster ensemble and consensus clustering. 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. 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 Neural Network (scINN), that can reliably impute missing values from single-cell data. The method consists of two steps. The first step is to generate an accurate clustering result of the original data, and calculate the similarity between all pairs of samples. The second step is to estimate the missing values using a neural network and the similarity information generated in the first module. The approach is evaluated using 10 single-cell datasets in comparison with four other methods. We demonstrate that scINN outperforms existing imputation methods (DrImpute [14], MAGIC [16], scImpute [13], and SAVER [12]) in improving the identification of cell subpopulations and the quality of biological landscape.

II. Methods

The input of scINN is an expression matrix, in which rows represent cells and columns represent genes or transcripts. The overall workflow of scINN is described in Figure 1, which consists of two modules: (i) generating an accurate clustering results of the original data, and calculating the similarity between all samples, and (ii) imputing the dropout values. The purpose of the first module is to learn the similarity information between each pair of samples. The output of the first module is the clustering assignments for samples in the dataset, and a similarity matrix with Pearson correlations for all pairs of samples. These information are used as the target for the second module. In the second module, we impute the original data using a neural network. The parameters of the neural network are repeatedly adjusted so that the clustering assignments and similarity matrix inferred from the imputed data is as similar to the outputs of the first module as possible. The details of each step are described in the following subsections.

A. Generating similarity information

To generate a compressed, low-dimensional representation of original data, we apply our previously developed method, called scDHA [21]. scDHA consists of two core modules. The first module is a non-negative kernel autoencoder that can filter out genes or components that have insignificant contributions to the representation. The second module is a Stacked Bayesian Self-learning Network that is built upon the Variational Autoencoder [22] to project the filtered data onto

Authorized licensed use limited to: UNIVERSITY OF IDAHO. Downloaded on May 27,2022 at 19:44:28 UTC from IEEE Xplore. Restrictions apply.

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

Dataset	Accession ID	Tissue	Sequencing Protocol	Drop. Rate	Class	Size
1. Yan	GSE36552	Human Embryo	Tang	0.456	6	90
2. Goolam	E-MTAB-3321	Mouse Embryo	Smart-Seq2	0.685	5	124
3. Deng	GSE45719	Mouse Embryo	Smart-Seq	0.605	6	268
4. Camp	GSE75140	Human Brain	SMARTer	0.801	7	734
5. Klein	GSE65525	Mouse Embryo	inDrop	0.658	4	2,717
6. Romanov	GSE74672	Human Brain	SMARTer	0.878	7	2,881
7. Baron	GSE84133	Human Pancreas	inDrop	0.906	14	8,569
8. Tasic	GSE115746	Mouse Visual Cortex	SMART-Seq	0.798	6	23,178
9. Zilionis	GSE127465	Human Lung	inDrop	0.982	9	34,558
10. Hrvatin	GSE102827	Mouse Visual Cortex	inDrop	0.942	8	48,266

a much lower-dimensional space. The output of scDHA is a low-dimensional matrix that preserves the global structure of the original data. Using this representation, scDHA can cluster the samples into groups with high accuracy. We also generate the similarity matrix for all samples in the dataset. The similarity between two samples is measured by Pearson correlation. We use the similarity information between samples in the dataset to optimize our imputation module so the same information can be inferred from imputed data using a network with simpler structure.

B. Imputing dropout data using neural 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 infer the clusters of input cells and the Pearson correlations between them. By minimizing the difference between the inferred results and the results from the first module, the imputed values are ensured to have high accuracy.

The formulation of the neural network can be written as:

$$\begin{aligned} X_I &= f_I(X) \\ C + S &= f_P(X_I) \end{aligned}$$

where $X \in R_+^n$ is the input of the model (X is simply the original data), f_I and f_P represent the transformation by the two sub-networks, f_I imputes the zero values in the data, f_P predicts the clusters of the input cells and the correlations between them, C is the clustering results, and S is the similarity matrix between all input cells. The network is optimized by minimizing: (i) the binary cross entropy loss between the inferred clusters and the clustering result from the first module, and (ii) the mean square error loss between the inferred similarity matrix and the similarity matrix calculated using the representations from the first module.

III. RESULTS

We compare our method with four state-of-the-art imputation methods: DrImpute [14], MAGIC [16], scImpute [13], and SAVER [12]. Each of these methods represents a distinct strategy to single-cell data imputation: DrImpute integrates clustering result from other software, MAGIC is a Markov-based technique, while scImpute and SAVER use statistical models. Table I shows the 10 datasets used in our data analysis. These scRNA-seq datasets are available on NCBI [23], and ArrayExpress [24]. The processed data of the first 7 datasets are 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) [25]. For visualization, we used UMAP [26] to generate the 2D representation and then calculated the silhouette index (SI) [27] of the 2D representation. SI measures the cohesion among cells of the same type, as well as the separation between different cell types.

A. scINN 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 DrImpute, MAGIC, scImpute, SAVER, and scINN). 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 where k is the true number of cell types. 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, SAVER has higher ARIs than the raw data for the Goolam, Camp, Klein, Romanov, Baron, and Zilionis but has lower ARIs in the remaining 4 datasets. scINN is the only method able to improve the clustering performance compared to raw data in every dataset. Moreover, scINN has the highest ARIs in all but Zilionis datasets. The average ARI of scINN-imputed data is 0.72, which is higher than those obtained from raw data and

Authorized licensed use limited to: UNIVERSITY OF IDAHO. Downloaded on May 27,2022 at 19:44:28 UTC from IEEE Xplore. Restrictions apply.

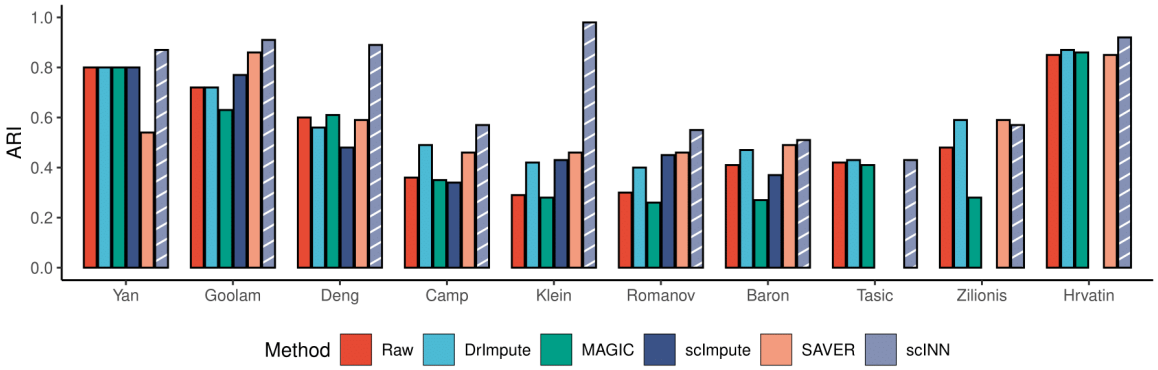

Fig. 2. Adjusted Rand index (ARI) obtained from clustering on raw data and data imputed by DrImpute, MAGIC, scImpute, SAVER, and scINN. The x-axis shows the names of the datasets while the y-axis shows ARI value of each method. scINN outperforms other methods in all datasets except Zilionis.

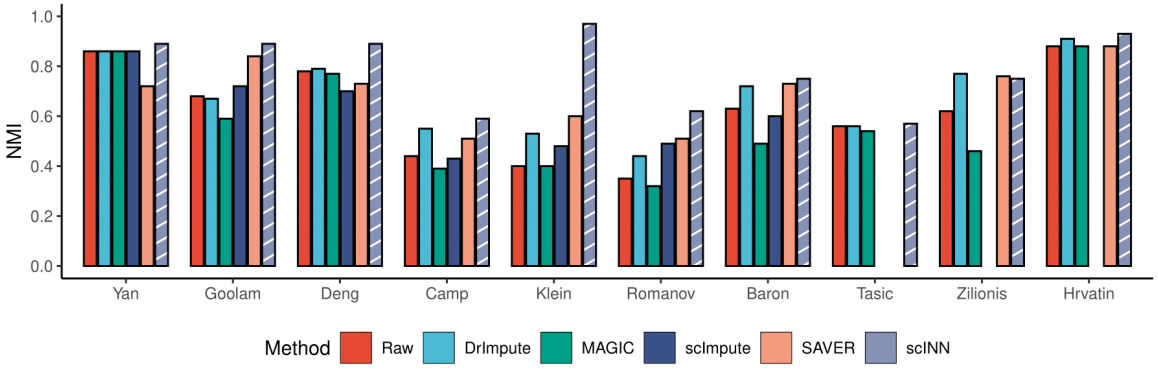

Fig. 3. Normalized mutual information (NMI) obtained from clustering on raw data and data imputed by DrImpute, MAGIC, scImpute, SAVER, and scINN. The y-axis shows NMI value of each method. scINN outperforms other methods in all datasets except Zilionis.

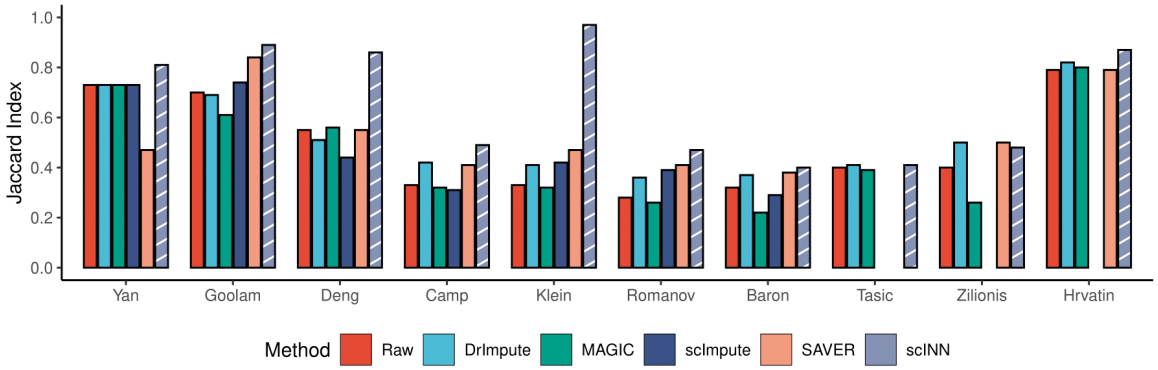

Fig. 4. Jaccard index (JI) obtained from clustering on raw data and data imputed by DrImpute, MAGIC, scImpute, SAVER, and scINN. The y-axis shows JI value of each method. scINN outperforms other methods in all datasets except Zilionis.

data imputed by DrImpute, MAGIC, scImpute, SAVER (0.52, 0.58, 0.48, 0.36, 0.53, respectively).

For a more comprehensive analysis, we also report the assessment using normalized mutual information (NMI) and Jaccard index (JI) [28] in Figures 3 and 4, respectively. Regardless of the assessment metrics, scINN outperforms other methods by having the highest NMI (9/10 datasets) and JI (9/10 datasets) values. These results demonstrate that cluster analysis using scINN-imputed data leads to a better accuracy than using the raw data or data imputed by other imputation methods.

B. scINN improves transcriptome landscape visualization

In this subsection, we demonstrate that scINN improves the visualization of the single-cell data. We used UMAP [26] to generate the transcriptome landscapes from raw and data

Authorized licensed use limited to: UNIVERSITY OF IDAHO. Downloaded on May 27,2022 at 19:44:28 UTC from IEEE Xplore. Restrictions apply.

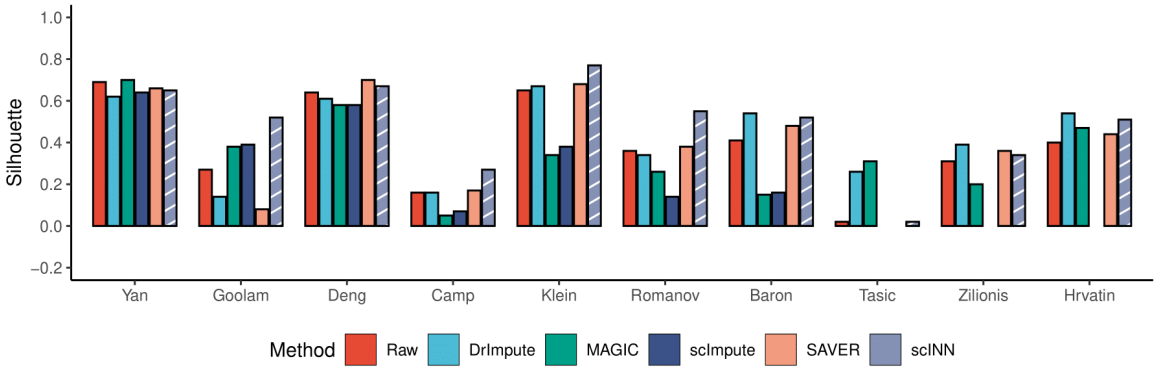

Fig. 5. Visualization quality using raw and imputed data, measured by silhouette index (SI). The y-axis shows SI value of each method.

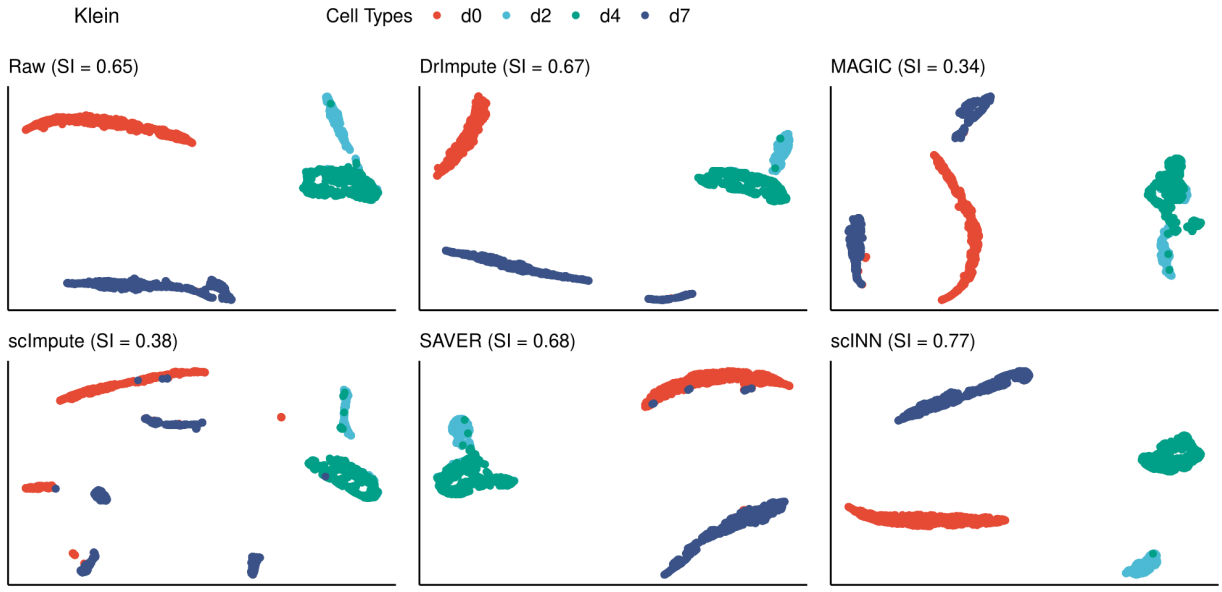

Fig. 6. Transcriptome landscape of the Klein dataset. The scatter plot shows the first two principal components calculated by UMAP. Different colors represent different cell types. The 2D representation generated by scINN has a clear structure, where cells from different groups are separated from one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

imputed by DrImpute, MAGIC, scImpute, SAVER, and scINN. 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 scINN can improve the quality of data visualization in most of the datasets (8/10 datasets). These results demonstrate that data imputation using scINN 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 Klein dataset. The 2D representation of scINN-imputed data is the only one that has four separable groups, corresponding to the 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 scINN has the highest SI value (0.77 compared to 0.68 of the second best).

IV. Conclusion

In this article, we introduced a new method, scINN, to recover the missing data caused by dropout events in scRNA-seq data. We compared scINN with four state-of-the-art imputation methods using 10 scRNA-seq datasets. scINN outperformed existing approaches in improving the identification of cell sub-populations. scINN also improved the quality of transcriptome landscapes generated by UMAP. A potential improvement of this research is to investigate the scalability of scINN 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 and broader applications, we will apply scINN in conjunction with other analysis methods in the context of pathway analysis [29–36], meta-analysis [37–39], and multi-omics integration [40–43].

Authorized licensed use limited to: UNIVERSITY OF IDAHO. Downloaded on May 27,2022 at 19:44:28 UTC from IEEE Xplore. Restrictions apply.

V. Acknowledgments

This work was partially supported 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] P. Brennecke, S. Anders, J. K. Kim, A. A. Kolodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] D. Tran, F. C. Harris, B. Tran, N. S. Vo, H. Nguyen, and T. Nguyen, "Single-cell RNA sequencing data imputation using deep neural network," in *ITNG 2021 18th International Conference on Information Technology-New Generations*. Springer, 2021, pp. 403–410.

[19] 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.

[20] 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.

[21] 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," *Nature Communications*, vol. 12, p. 1029, 2021.

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

[23] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets–update," *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, 2013.

[24] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. P. Pereira, E. Pilicheva, J. Rung, A. Sharma, Y. A. Tang, T. Ternent, A. Tikhonov, D. Welter, E. Williams, A. Brazma, H. Parkinson, and U. Sarkans, "ArrayExpress update–trends in database growth and links to data analysis tools," *Nucleic Acids Research*, vol. 41, no. D1, pp. D987–D990, 2013.

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

[26] 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.

[27] 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.

[28] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[29] 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*, vol. 22, no. 3, pp. 1–15, 2021.

[30] T.-M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici, "Identifying significantly impacted pathways: a comprehensive review and assessment," *Genome Biology*, vol. 20, p. 203, 2019.

[31] 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.

[32] 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.

[33] 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.

[34] H. Nguyen, D. Tran, J. M. Galazka, S. V. Costes, A. Beheshti, S. Draghici, and T. Nguyen, "CPA: A web-based platform for consensus pathway analysis and interactive visualization," *Nucleic Acids Research*, vol. 49, no. W1, pp. W114–W124, 2021.

[35] 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.

[36] J. Tanevski, T. Nguyen, B. Truong, N. Karaiskos, M. E. Ahsen, X. Zhang, C. Shu, K. Xu, X. Liang, Y. Hu, H. V. Pham, L. Xiaomei, T. D. Le, A. L. Tarca, G. Bhatti, R. Romero, N. Karathanasis, P. Loher, Y. Chen, Z. Ouyang, D. Mao, Y. Zhang, M. Zand, J. Ruan, C. Hafemeister, P. Qiu, D. Tran, T. Nguyen, A. Gabor, T. Yu, J. Guinney, E. Glaab, R. Krause, P. Banda, DREAM SCTC Consortium, G. Stolovitzky, N. Rajewsky, J. Saez-Rodriguez, and P. Meyer, "Gene selection for optimal prediction of cell position in tissues from single-cell transcriptomics data," *Life Science Alliance*, vol. 3, no. 11, 2020.

[37] 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.

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

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

[40] 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.

[41] 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.

[42] D. Tran, H. Nguyen, U. Le, G. Bebis, H. N. Luu, and T. Nguyen, "A novel method for cancer subtyping and risk prediction using consensus factor analysis," *Frontiers in Oncology*, vol. 10, p. 1052, 2020.

[43] M. P. Menden, D. Wang, M. J. Mason, B. Szalai, K. C. Bulusu, Y. Guan, T. Yu, J. Kang, M. Jeon, R. Wolfinger, T. Nguyen, M. Zaslavskiy, AstraZeneca-Sanger Drug Combination DREAM Consortium, I. S. Jang, Z. Ghazoui, M. E. Ahsen, R. Vogel, E. C. Neto, T. Norman, E. K. Y. Tang, M. J. Garnett, G. Y. Di Veroli, C. Zwaan, S. Fawell, G. Stolovitzky, J. Guinney, J. R. Dry, and J. Saez-Rodriguez, "Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen," *Nature Communications*, vol. 10, p. 2674, 2019.

Authorized licensed use limited to: UNIVERSITY OF IDAHO. Downloaded on May 27,2022 at 19:44:28 UTC from IEEE Xplore. Restrictions apply.