RIA: A NOVEL REGRESSION-BASED IMPUTATION APPROACH FOR SINGLE-CELL RNA SEQUENCING

Bang Tran, Duc Tran, Hung Nguyen, Nam Sy Vo and Tin Nguyen*
Department of Computer Science and Engineering, University of Nevada, Reno
Contact: tinn@unr.edu , Website: https://bioinformatics.cse.unr.edu/

NEVADA NASA SPACE GRANT CONSORTIUM

UNIVERSITY OF NEVADA, RENO

05/08/2020

Background

Advances in single-cell technologies have shifted genomics research from the analysis of bulk tissues toward a comprehensive characterization of individual cells. This holds enormous opportunities for both basic biology and clinical research. However, low amount of mRNA available within individual cells leads to the excess amount of zero counts caused by dropout events

Objectives

Develop an imputation method, RIA, that can reliably impute missing values from single-cell data. RIA consists of two modules. The first module performs a hypothesis testing to identify the values that are likely to be impacted by the dropout events. The second module estimates the missing value using a robust regression approach.

Results

Data: 5 datasets with a total of 3,535 cells.

Metric: Adjusted Rand Index (ARI) [8], Jaccard Index [9] and Purity Index [10].

Methods: scImpute [15], MAGIC [16], t-SNE [17].

Results: RIA produces the best ARI values, preserve the transcriptomics landscape and significantly elucidates the cell lineage identification.

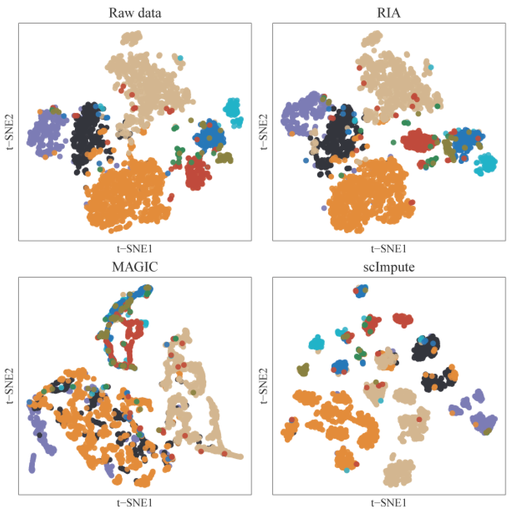

Fig. 2. *RIA preserves the transcriptomics landscape for Zeisel [14] dataset.*

Methodology

Hypothesis Testing and Identification of Dropout : to determine genes that are likely to be impacted by dropouts. Genes that are not impacted by dropouts, the log-transformed expression values are normally distributed. We use z-test to determine whether a zero is impacted by the dropout events. Original data is divided into two sets of genes: a set *G* that include genes affected by dropout (imputable set), and a set *M* that have high confidence of not being affected by dropout . (training set)

Raw Data

Genes \ Cells	S_1	S_2	...	S_m
g_1	2	8	...	13
g_2	1	1	...	5
g_3	18	0	...	2
g_4	5	5	...	1
...	...	...	...	...
g_n	3	6	...	0

Hypothesis Testing

Genes \ Cells	S_1	S_2	...	S_m
g_1	2	8	...	13
g_2	1	1	...	5
g_3	18	0	...	2
g_4	5	5	...	1
...	...	...	...	...
g_n	3	6	...	0

Regression-based Imputation:

- We select genes from the training set that are highly correlated with the gene we need to impute.
- We train the linear model using these highly-correlated genes and then estimate the missing values

Training Data

Genes \ Cells	S_1	S_2	...	S_p
g_1	2	8	...	13
g_2	1	1	...	5
...	...	...	...	...
g_i	3	6	...	15

Imputable Data

Genes \ Cells	S_1	S_2	...	S_i
g_3	18	0	...	2
g_4	5	5	...	1
g_5	0	1		0
...	...	...	...	...
g_j	0	2	...	0

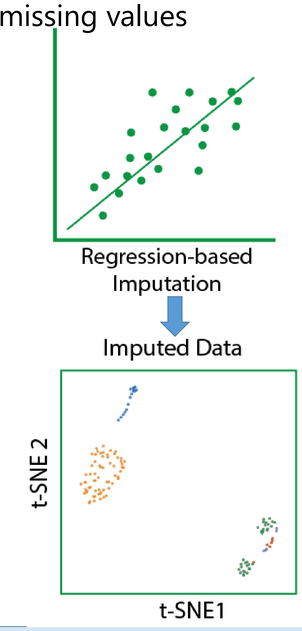

Fig. 1. *The overall pipeline of RIA.*

Conclusion

- Outperforms existing state-of-the-art approaches in cell group identification.
- Recover temporal trajectories in embryonic development stages
- RIA is fast and is able to impute thousands of cells with tens of thousands of genes in minutes

Future work

We plan to utilize the perturbation clustering [3],[4],[6].

Acknowledgement

This material is based upon work supported by the National Aeronautics and Space Administration under Grant No. 80NSSC19M0170

References

1. Nguyen et al. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biology*, 20(1), 1-15.
2. Nguyen et al. (2017). DANUBE: data-driven meta-ANalysis using UnBiased empirical distributions—applied to biological pathway analysis. *Proceedings of the IEEE, 105*(3), 496-515.
3. Nguyen et al. (2017). A novel approach for data integration and disease subtyping. *Genome Research*, 27(12), 2025-2039.
4. Nguyen et al. (2016). A novel bi-level meta-analysis approach: applied to biological pathway analysis. *Bioinformatics*, 32(3), 409-416.
5. Nguyen et al. (2020). NBIA: a network-based integrative analysis framework–applied to pathway analysis. *Scientific Reports*, 10(1), 1-11.
6. Nguyen et al. (2019). PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, 35(16), 2843-2846.
7. Tran et al. (2019). Fast and precise single-cell data analysis using hierarchical autoencoder. *bioRxiv*, 799817.
8. Hubert et al. (1985). Comparing partitions. *Journal of Classification*, vol. 2, no. 1, pp. 193–218.
9. Jaccard et al. (1901). Etude comparative de la distribution florale dans une portion des Alpes et des jura. *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579.
10. Manning et al. (2010). Introduction to information retrieval. *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103.
11. Tran et al. (2019). RIA: a novel Regression-based Imputation Approach for single-cell RNA sequencing. In 2019 *11th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 1-9). IEEE.
12. Tanevski et al. (2019). Predicting cellular position in the Drosophila embryo from Single-Cell Transcriptomics data. *bioRxiv*, 796029.
13. Nguyen et al. (2019). A comprehensive survey of tools and software for active subnetwork identification. *Frontiers in genetics*, 10, 155.
14. Zeisel et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq, *Science*, vol. 347, no. 6226, pp. 1138–1142.
15. Li et al. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data, *Nature Communications*, vol. 9, no. 1, p. 997.
16. Van Dijk et al. (2018). Recovering gene interactions from single-cell data using data diffusion, *Cell*, vol. 174, no. 3, pp. 716–729.
17. Krijthe, et al. (2015). Rtsne: T-distributed stochastic neighbor embedding using barnes-hut implementation, R package version 0.13, URL https://github.com/jkrijthe/Rtsne.