scCAN: A Single cell clustering method using autoencoder and network fusion

Bang Tran, Duc Tran, Ha Nguyen, 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

04/30/2021

Background

Single-cell RNA sequencing technologies (scRNA-seq) have allowed us to monitor biological systems at single-cell resolution [1]. Defining cell types through unsupervised learning is considered the most powerful application of scRNA-seq data. This has led to the creation of a number of atlas projects that aim to build the references of all cell types [2,3].

Objectives

The ever-increasing number of cells, the high-dimensionality of scRNA-seg data, technical noise, and high dropout rate pose significant computational challenges in cell segregation. The goal here is to develop a novel method able to accurately separate different cell types in scRNA-seq data.

Results

Data: 26 datasets with more than a million cells and simulation.

Metric: Adjusted Rand Index (ARI) [4], Adjusted Mutual Information [5] and V-measure [6].

Methods: CIDR [7], SEURAT3 [8], Monocle3 [9], SHARP [10], SCANPY [11].

Results: scCAN outperforms other methods by having the highest ARI, AMI and V-measure values (panels A, B, C in Figure 1). scCAN is also the most scalable (panels D and E in Figure 1).

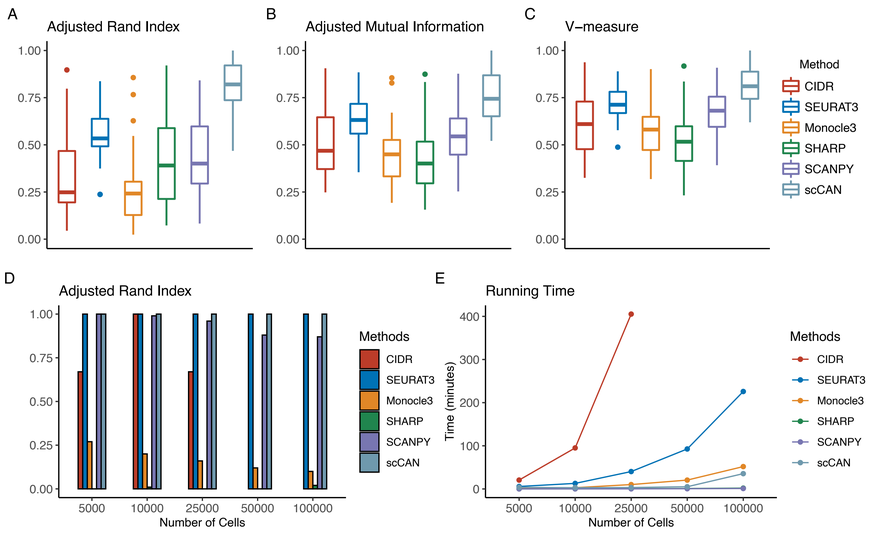

Figure 1: *scCAN outperforms state-of-the-art methods real data analysis and simulation studies.*

Methodology

Genes filtering and generation of larent variables: scCAN uses non-negative autoencoder to keep the 5,000 most informative genes. Then, the denoised data is passed to Bayesian stack autoencoder to obtain multiple latent variables (Figure 2A).

Network fusion based scRNA-seq clustering:

- For small datasets (less than 2,000 cells), scCAN converts latent variables to networks. The obtained networks are combined to a single fused network. scCAN applies spectral clustering algorithm on fused network to group the cells (Figure 2B)
- For big datasets (more than 2,000 cells), scCAN uses sub-sampling strategy to get 2,000 cells for training. Then, scCAN uses the approach mentioned in Figure 2B to partition the training cells. scCAN uses k-NN to map remaining cells to the training data clusters (Figure 2C).

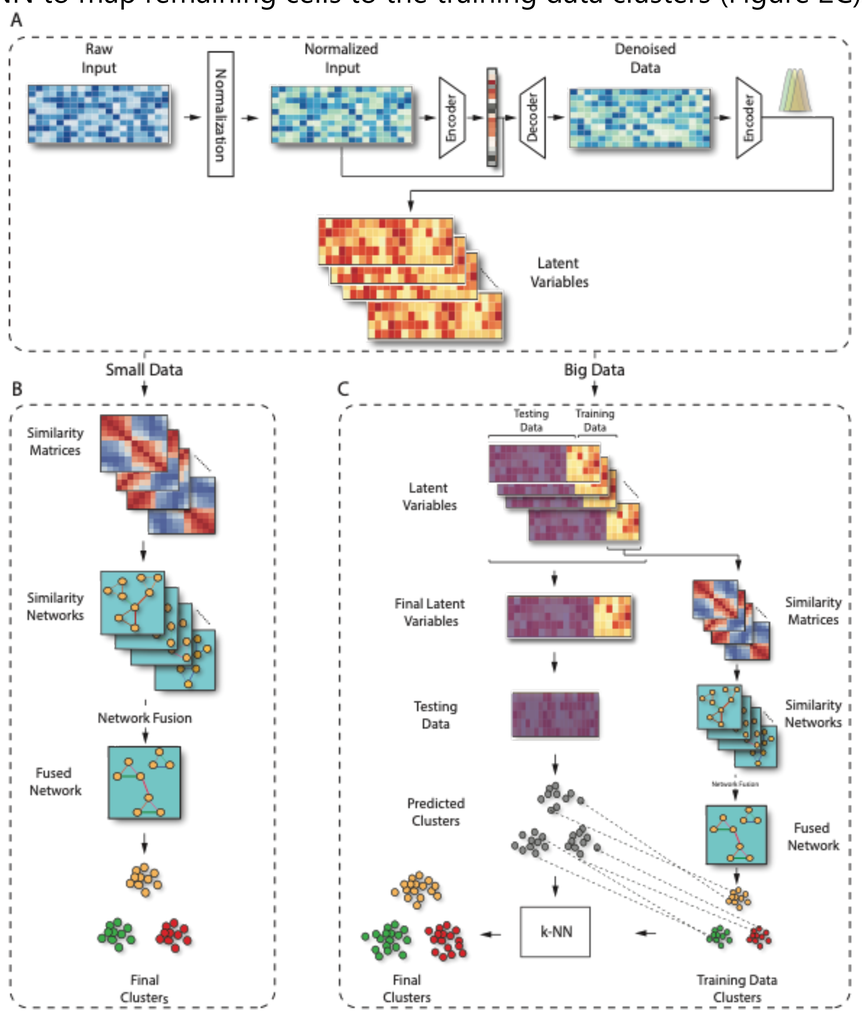

Figure 2: *The overall analysis pipeline of scCAN.*

Conclusion

Here we introduce a new method scCAN for single-cell analysis. Our analysis results demonstrate that the method:

- Outperforms existing state-of-the-art approaches for cell segregation using scRNA-seq.
- is the fast method for big data.
- is robust against dropout events.
- is the best method in predicting the true number of cell types.

Acknowledgement

This work was partially supported by NASA under grant numbers 80NSSC19M0170 and NNX15AI02H (subaward no. 21-02), and by NIH NIGMS under grant number GM103440.

References

1. Lähnemann et al. (2020). Eleven grand challenges in single-cell data science. Genome Biology, 21(1), 1–35.
2. Davie et al. (2018). A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. Cell 174, 982–998.
3. Rozenblatt-Rosen et al. (2017). The Human Cell Atlas: From vision to reality. Nature 550, 451–453.
4. Hubert et al. (1985). Comparing partitions. Journal of Classification, vol. 2, no. 1, pp. 193–218.
5. Vinh et al. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. The Journal of Machine Learning Research, 11, 2837–2854.
6. Rosenberg et al.(2007). V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pages 410–420.
7. Lin et al.(2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biology, 18(1), 59.
8. Stuart et al. (2019). Comprehensive integration of single-cell data. Cell, 177(7), 1888–1902.
9. Cao et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature, 566(7745), 496–502.
10. Wan et al. (2020). SHARP: hyperfast and accurate processing of single-cell RNA-seq data via ensemble random projection. Genome Research, 30(2), 205–213.
11. Wolf et al. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 19(1), 15.