Number of sIngle-cell Clusters Estimation (NICE): a robust and priori knowledge independent algorithm for single-cell RNA-seq data analysis

Xin Zou, Jie Hao, Ze-Guang Han
Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine,
Shanghai Jiao Tong University, 800 Dongchuan Road,
Shanghai 200240, China.
x.zou@sjtu.edu.cn

***Abstract*—Interpretation of single-cell transcriptomic data is usually achieved by dissecting cells into a few clusters. A successful interpretation performance is highly dependent on how well the number of cell clusters matches the intrinsic structure of the dataset. However, the existing estimation methods for number of clusters either require priori information which is not usually available for single-cell data, or are subject to interferences, such as artificial bias and technological noise. To tackle these issues, we propose a novel algorithm, Number of sIngle-cell Clusters Estimation (NICE), to estimate the number of cell clusters based on single-cell RNA-seq data. The new algorithm can effectively discriminate significant variations from subtle perturbations without requiring any priori information about the datasets, and therefore, it is highly robust. Furthermore, the output of algorithm guarantees each cell cluster is significantly distinctive from the others, and thereby, each cell cluster has specific biological meaning.**

Keywords—single-cell analysis; clustering; number of clusters estimation

I. INTRODUCTION

Single-cell RNA-seq transcriptome analysis measures the amount of transcripts within individual cells and can reveal cell-to-cell variations; therefore, it has wide applications in many areas, such as tumor and cell lineage commitment. The interpretations of single-cell datasets are usually achieved by decomposing the cells into a few clusters based on their gene expression profiles. In most scenarios, the bunch of cells to be investigated are either considered as homogeneous in traditional bulk experiments or the cell classification information is unknown [1]. Therefore, unsupervised clustering methods have been widely applied, which usually need to estimate the number of clusters.

The estimation methods can be categorized into two types. The first type is visual inspection. The clustering patterns of cells can be made visible by using projection methods, e.g., principle component analysis (PCA) [2] and t-distribution stochastic neighbor embedding (tSNE) [3]. Such type of methods is usually subject to artificial bias and noise interference, which may blur the cell clustering patterns.

The other type is based on statistical metrics. Gap statistic is one of the most popular criteria for cluster numbers evaluation [4] and has been successfully applied in single-cell data analyses [5]. The method maximizes the differences between the real gap statistic and the one obtained by permutation test. However, the method do not guarantee statistical differences between clutters.

To tackle the issues mentioned above, we propose a novel algorithm, Number of sIngle-cell Clusters Estimation (NICE), to automatically estimate the number of cell clusters. The algorithm consists of two main steps, hybrid clustering for feature selection and critical number of cell clusters identification. The hybrid clustering aims to identify the features with high contribution to the discrimination of different cell clusters. In the identification step of critical number of clusters, subtle perturbations are discriminated from significant variations by modelling the bimodality property of variable genes. As a result, the proposed identification criterion is completely objective and robust to random perturbations.

We exemplified the proposed algorithm on several simulated and experimental single-cell RNA-seq datasets containing hundreds to over three thousands of cells. The proposed algorithm has demonstrated the capability in reliably revealing heterogeneity of single-cell datasets and their corresponding biological meanings.

II. PRINCIPLES OF NICE

Here, we propose a novel algorithm, NICE, to guide the generation of statistically optimized cells clustering results. The NICE algorithm consists of two main steps: i) Identification of variable genes by a hybrid clustering scheme for a given number of clusters, N. We adopt a similar principle to SHOCSY [6], which combines a supervised (OPLS-DA) and an unsupervised (k-means) clustering methods. ii) Identification of critical number of clusters. In principle, the gene expression patterns of each cell cluster are modelled in a binary manner, i.e., using 1/0 to represent high/low repression. Then, the similarities between different clusters are evaluated to detect which value of N best reflects the intrinsic characteristics of data. Within this step, binaryzation process is performed by modelling the bimodality property of variable genes using auto-clustering and enrichment test methods. By doing so, trivial cell-to-cell and gene-to-gene perturbations can be precluded. A schematic diagram of the proposed algorithm has been given in Fig. 1.

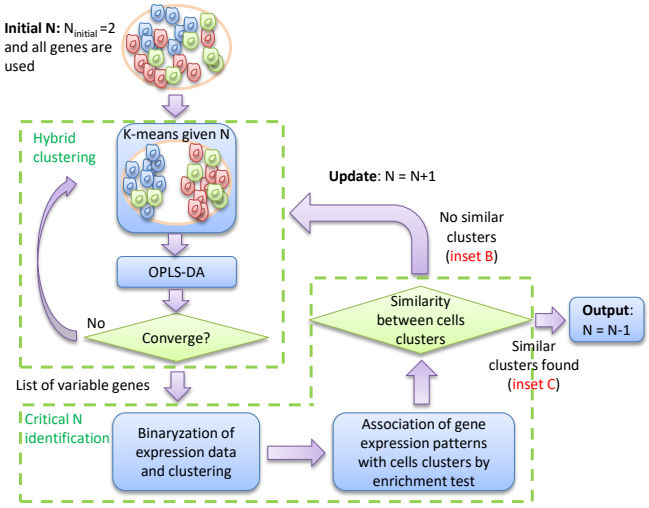

Figure 1. A demonstration of NICE being applied in analyzing single-cell RNA-seq data.

III. Results

The proposed algorithm has been evaluated using simulated and experimental single-cell RNA-seq datasets.

A. *Simulated datasets*

The simulated datasets, which consist of 2 to 8 cell clusters, have been adopted to justify the concept of the proposed method. The analysis results demonstrated that when the number of clusters were overestimated, the obtained minimum distances immediately dropped to zero as designed (Fig. 2). Therefore, the real cluster numbers can be reliably estimated by *N*-1, when the minimum distance reached strict zero.

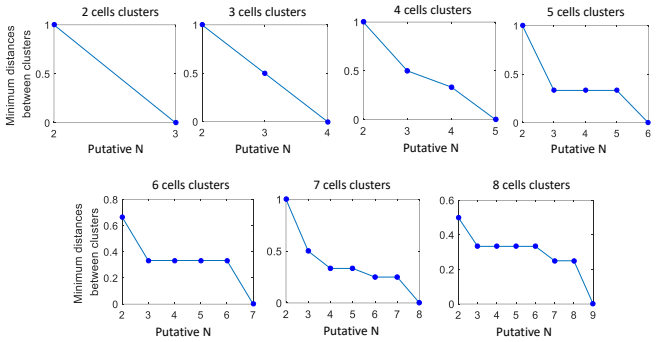

Figure 2. The minimum Hamming distances of different putative N for the simulated datasets containing various cell clusters.

B. *An experimental single-cell RNA-seq dataset*

The utility of NICE in dealing with large complex datasets was demonstrated on a single-cell RNA-seq dataset containing 3005 cells, which were collected from mouse cortex and hippocampus [7]. The new algorithm successfully dissected the major neuron types reported in the original paper in a hierarchical manner, Fig. 3. Furthermore, compared to a latest published method, NMF [8], our algorithm performed better in revealing neuron subtypes (data not shown).

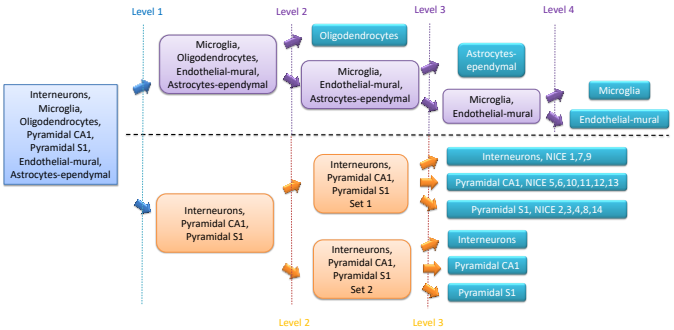

Figure 3. Summary of the NICE analysis results of the mouse neurons dataset (Zeisel, et al., 2015). NICE 1-14 denote the cell clusters identified by NICE.

References

[1] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nat. Rev. Genet.*, vol. 16, pp. 133-45, Mar 2015.

[2] P. Dalerba, T. Kalisky, D. Sahoo, P. S. Rajendran, M. E. Rothenberg, A. A. Leyrat, *et al.*, "Single-cell dissection of transcriptional heterogeneity in human colon tumors," *Nat Biotechnol*, vol. 29, pp. 1120-7, Nov 13 2011.

[3] A. Saadatpour, G. Guo, S. H. Orkin, and G. C. Yuan, "Characterizing heterogeneity in leukemic cells using single-cell gene expression analysis," *Genome Biol*, vol. 15, p. 525, Dec 03 2014.

[4] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, vol. 63, pp. 411-423, 2001.

[5] D. Grun, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, *et al.*, "Single-cell messenger RNA sequencing reveals rare intestinal cell types," *Nature*, vol. 525, pp. 251-+, Sep 10 2015.

[6] X. Zou, E. Holmes, J. K. Nicholson, and R. L. Loo, "Statistical HOmogeneous Cluster SpectroscopY (SHOCSY): an optimized statistical approach for clustering of (1)H NMR spectral data to reduce interference and enhance robust biomarkers selection," *Anal Chem*, vol. 86, pp. 5308-15, Jun 3 2014.

[7] A. Zeisel, A. B. Munoz-Manchado, S. Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, *et al.*, "Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq," *Science*, vol. 347, pp. 1138-42, Mar 06 2015.

[8] C. Shao and T. Hofer, "Robust classification of single-cell transcriptome data by nonnegative matrix factorization," *Bioinformatics*, vol. 33, pp. 235-242, Jan 15 2017.