**SHDC:** **A Method of Similarity Measurement Using Heat Kernel based on Denoising for Clustering scRNA-seq Data**

Hai-Yun Wang1※, Jian-ping Zhao1※,2,\*, Chun-Hou Zheng1,3,\*

1College of Mathematics and System Sciences, Xinjiang University, Urumqi, China

2Institute of Mathematics and Physics, Xinjiang University, Urumqi, China

3 College of Computer Science and Technology, Anhui University, Hefei, China

※These authors contributed equally to this work

**\*Correspondence to:** Jian-ping Zhao(jpzhao@xju.edu.cn) **and** Chun-Hou Zheng(zhengch99@126.com)

**Abstract:** Identifying cell types is one of the most important goals of single cell analysis. The similarity between cells is the principal basis for dividing cell subpopulations. Kernel method is a powerful tool to measure similarity. Because of the high noise of scRNA-seq data, it is very common to measure the similarity between cells by using multi-kernel. However, it usually brings higher computational complexity. In this paper, a new method called “A Method of Similarity Measurement Using Heat Kernel based on Denoising for Clustering scRNA-seq Data” (SHDC) was proposed for single cell clustering. SHDC denoised scRNA-seq datasets to obtain a more stable data structure and only a heat kernel was used to measure similarity between cells. It consists of the following steps. First, scRNA-seq datasets were denoised by deep count autoencoder network (DCA). Second, data standardization after denoised data. Third, similarity matrix was constructed between cells by heat kernel. Fourth, extracted k-nearest neighbors (k-NN) graph from similarity matrix to obtain adjacency matrix. Final, The Uniform Manifold Approximation and Projection (UMAP) algorithm is used to project the adjacency matrix into a low dimensional space to perform clustering. We used several real scRNA-seq datasets to test the performance SHDC. We also compared SHDC with previous single cell clustering methods and obtained comparable or superior results.

**Keywords:** Clustering, scRNA-seq, Heat kernel, Deep count autoencoder network, Uniform Manifold Approximation and Projection

# Introduction

Cell is the basic unit of life. In recent years, single cell sequencing technology has developed rapidly, so especially is single-cell RNA sequencing. It helps researchers explore the nature and laws of life activities [1]. Because single cell sequencing technology takes into account the heterogeneity between cells [2]. More importantly, it can help us explore the causes of complex diseases at a cell resolution. But these are closely related to cell types [3]. Therefore, identifying cell types is an important goal.

Nowadays, a large number of single cell sequencing data have been generated, but only a small part of the data contains the corresponding cell types [4]. This makes it necessary to use clustering methods to identify cell types. For example, traditional machine learning methods, k-means [5], and spectral clustering [6] are the representatives. In most cases, these methods will not be used to cluster single cell RNA sequencing data directly, but they will still appear as an important step in many algorithms.

High dimension and high noise are two characteristics of single cell RNA sequencing data [7]. They are also two difficulties in the process of single cell clustering. So far, many single cell clustering methods and tools taking them into account have been proposed. These include single cell consensus clustering (SC3) [8], single-cell interpretation via multikernel learning (SIMLR) [9], Spearman subsampling-clustering-classification (SSCC) [10], Phenotyping by Accelerated Refined Community-partitioning (PARC) [11], Tools for Single Cell Genomics (Seurat) [12] Clustering through Imputation and Dimensionality Reduction (CIDR) [13], Tools for Single Cell Analysis (TSCAN) [14], GiniClust [15], Scater [16] and so on.

Seurat, TSCAN and Scater are three R language toolkits for single cell analysis. They include a series of single cell analysis processes. SC3 uses different parameters in distance measurement and data transformation which leads to different clustering results. In other words, SC3 firstly obtains different clustering results in different parameter subspaces. Then the consistency strategy is used to integrate the different results to get a consistency matrix. Finally, hierarchical clustering [17] is performed on the consistency matrix to get the final clustering result. SC3 has good performance in processing scRNA-seq data with high dropout rate. SSCC is a new single cell clustering framework based on subsampling-clustering-classifying (SCC) [10]. It also contains random projection and feature construction. The novelty of SSCC is that the gene expression matrix of single cell RNA is randomly divided into two sub matrices for clustering and classification respectively. This reduces the computational complexity to a certain extent. CIDR considers the influence of dropouts on clustering. In order to reduce the impact of dropouts on clustering results, it imputes the missing values before clustering. High-throughput single cell sequencing technology provides an opportunity to discover new cell types. It also brings challenges to discover rare cell types from large a large population simultaneously [15]. PARC integrates hierarchical graph construction and data-driven graph pruning with community detection algorithm. The biggest advantage of PARC is that it can process large-scale data in less time. The similarity between cells is the main basis for dividing cell subpopulations. SIMLR uses multiple Gaussian kernels with different parameters to measure the similarity between cells. Then the final similarity matrix is obtained by weighted integration of the precious results. Finally, the similarity matrix is used for clustering analysis and cell visualization.

The gaussian kernel used in SIMLR has two parameters and . Different values of and correspond to different Gaussian kernel functions. They take different values to combine 55 different kernel functions, and then there are 55 similarity matrices. This leads to higher computational complexity, not to mention the weighted integration of these 55 similarity matrices. Single cell RNA sequencing data usually contain a lot of noise. It has been proved that denoising can improve the stability of data structure. It is useful for subsequent cluster analysis. We used real scRNA-seq data for experiments and came to the conclusion. It was that the performance of single gaussian kernel is similar to that of multiple gaussian kernels for being denoised data. In addition, compared with gaussian kernel, heat kernel had better performance and less computational complexity.

So, we proposed a new method combining denoising and kernel method. Deep count autoencoder network (DCA) [18] was used to denoise data. Only a heat kernel [19] that has less computational complexity compare to gaussian kernel [20] was used to measure similarity between cells. We extensively evaluated our approach with competing methods using several real scRNA-seq datasets. Experimental results showed that our method has better performance.

# Method

## Framework overview

SHDC consists of the following six steps (Fig 1).

1. The DCA is used to denoise and reconstruct scRNA-seq datasets.
2. Data normalization and this contained two processes. Firstly, the output data of DCA model need to be log-transformed. Secondly, data should be standardized according to the formula (1):

. (1)

Where f is a feature, is the corresponding feature after normalization. And and represent sample means and sample variance of *f*.

1. The similarity matrix S is obtained by using heat kernel (formula 2) to measure the similarity between cells.

. (2)

Where |||| is 2-norm, represents the distance between cell *i* and cell *j*. And *t* is a constant, in this article, *t* is 100 by default. represents the similarity between cell *i* and cell *j* and is in the range of 0 to 1.

1. Extract k-nearest neighbors (k-NN) graph from similarity matrix *S* to obtain adjacency matrix *W* according to the formula (3):

. (3)

Where is the set of the first *k* nearest neighbors of cell *i*, represents the *jth* nearest neighbor of cell *i*.

This process is also used to sparse the similarity matrix *S*.

1. Project adjacency matrix *W* to a lower dimensional space using the Uniform Manifold Approximation and Projection (UMAP) [21] algorithm.
2. The k-means was used to obtain final clustering result in this low dimensional space.

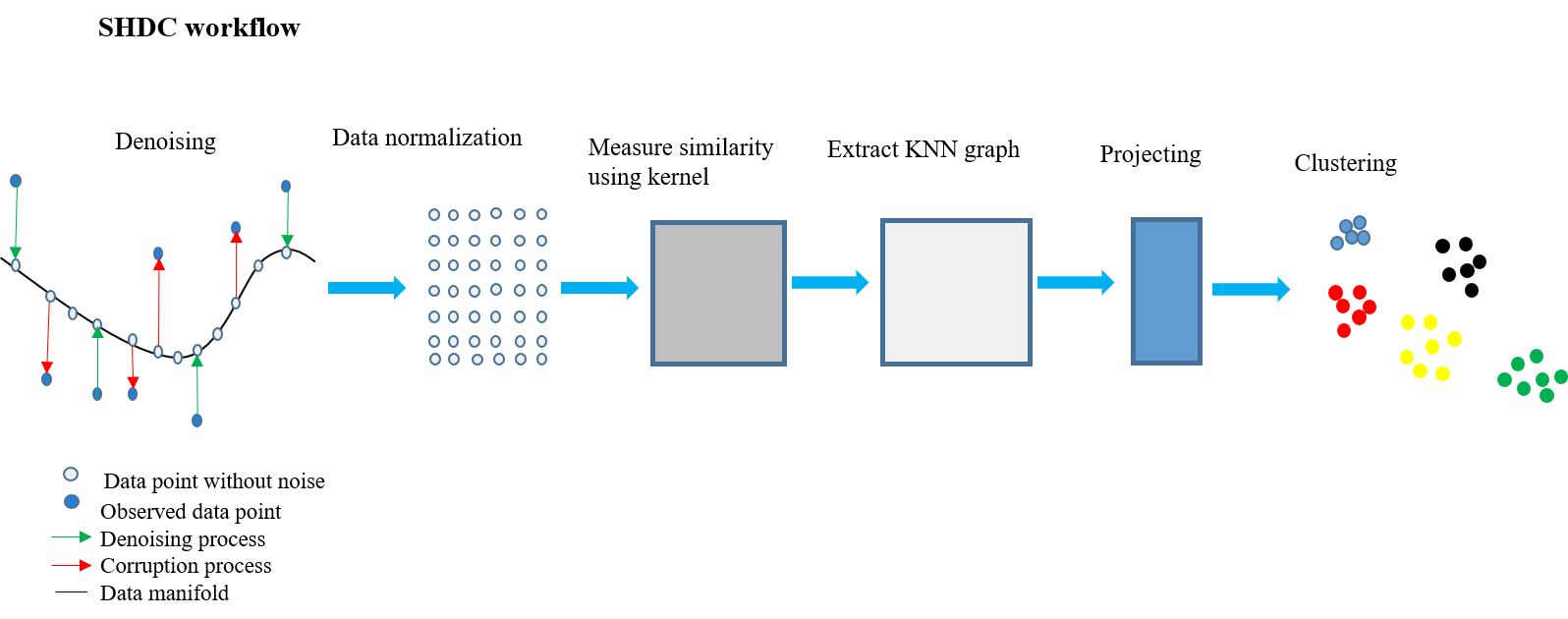


Figure 1 SHDC workflow

## Denoise using DCA

Noise has always been a major challenge for single-cell clustering [22]. There are also dropouts in noise, including “true zeros” and “false zero” values. We all know that the lower the ratio of dropouts in single cell RNA sequencing data, the better for single cell clustering [22]. Therefore, data denoising is also an effective means to improve the accuracy of clustering.

Deep count autoencoder network (DCA) [18] has been widely used in scRNA-seq data denoising. DCA uses the framework of autoencoder to denoise single cell RNA data by learning the manifold of noisy data. DCA is a multi-layer automatic encoder model. It has one input layer, three hidden layers (also called bottleneck layer) and three output layers. DCA uses zero-inflation- negative-binomial (ZINB) distribution to fit real manifold of noisy scRNA-seq data.

The ZINB distribution is parameterized with negative binomial component (μ and θ) and the mixture coefficient. is a mixture coefficient that represents the weight of the point mass (π):

, (4)

. (5)

The framework of autoencoder was used to estimate three parameters in formula (5). The formulation of the architecture is given below:

. (6)

Where *E*, *B* and *D* represent the encoder layer, bottleneck layer and decoding layer respectively. represents the standardized gene expression matrix, with rows representing cells and columns representing genes. is defined as:

. (7)

where X and "zscore" represent the raw count matrix and z-score normalization, is size factor of cell *i*.

The ZINB distribution is applied to the loss function of DCA model, and the corresponding optimization objective is:

. (8)

Where represents negative log likelihood function of ZINB distribution.

## Use single kernel instead of multi-kernel

A conclusion can be drawn from SIMLR [9]. It is that multiple gaussian kernels are an effective method to measure cell similarity for high-noise single-cell RNA data with unstable data structure. But it takes 2000 to 3000 seconds for SIMLR to process a dataset of more than 2000 cells. In addition, the same denoised scRNA-seq dataset was used to measure performance of single gaussian kernel only needed less than 40 seconds. We used six pairs of parameters corresponding to single gaussian kernels to construct the similarity matrix between cells. We also compared the final clustering results with those obtained by SIMLR (Table 1). We found that the performance of using single Gaussian kernel for denoised data is similar to that of using multiple Gaussian kernels for original data. The detailed results were shown in Table 1.

Table 1 Comparison results of single gaussian kernel on denoised data and multi gaussian kernel on raw data for clustering.

|  |  |  |  |
| --- | --- | --- | --- |
| Single gaussian kernel | | | |
| *k* |  | NMI | ARI |
| 10 | 1.5 | 0.759 | 0.743 |
| 12 | 2.0 | 0.761 | 0.753 |
| 14 | 2.0 | 0.760 | 0.744 |
| 16 | 1.75 | 0.763 | 0.743 |
| 18 | 1.5 | 0.762 | 0.743 |
| 20 | 1.75 | 0.761 | 0.742 |
| multi gaussian kernel (SIMLR) | | 0.772 | 0.753 |

We used the 10X\_PBMC dataset [23] and randomly selected 2100 cells from it. Where *k* and *sigma* were two parameters in gaussian kernel. For a given value of *k*, it may correspond to a unique m, so that the performance of a single Gaussian kernel was the best.

Of course, we also selected the other three real scRNA-seq datasets, which are mouse\_bladder\_cell dataset [24], mouse\_ES\_cell dataset [25] and worm\_neuron\_cell dataset [26]. The datasets were denoised by DCA with the same process. Then, single gaussian kernel and multi gaussian kernel (SIMLR) were used to measure similarity between cells. We compared final results obtained by two different means. We can find that denoising can improve the structural stability of scRNA-seq data. Single gaussian kernel and multiple gaussian kernels have the comparable performance on the denoised single cell RNA data. Obviously, single Gaussian kernel takes less time than multiple Gaussian kernels. See Table 2 for details.

Table 2 Comparison results of single gaussian kernel and multi gaussian kernel on raw data for clustering.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | NMI | | Running time (s) | |
| Single gaussian kernel | Multi gaussian kernel (SIMLR) | Single gaussian kernel | Multi gaussian kernel (SIMLR) |
| 10X\_PBMC\_DCA | 0.793 | 0.775 | 419.58 | 2877.47 |
| mouse\_bladder\_DCA | 0.722 | 0.696 | 175.64 | 2898.50 |
| mouse\_ES\_DCA | 0.966 | 0.829 | 172.37 | 1892.61 |
| worm\_neuron\_DCA | 0.697 | 0.684 | 401.29 | 1728.67 |

## Select heat kernel instead of gaussian kernel

Single kernel function can save computation time to some extent, but the computation of gaussian kernel is still complex. In order to speed up the calculation, the heat kernel was used to replace the gaussian kernel. It can be seen from formulas (9-11) and (12) that the computational complexity of heat kernel is smaller than that of gaussian kernel.

. (9)

. (10)

. (11)

. (12)

Where the formula (9) represents gaussian kernel. The formula (10) and (11) represent the parameters used in formula (9). Where formula (12) represents heat kernel. We found that the heat core not only greatly reduces the calculation time (Fig 2), but also improved the clustering accuracy (Fig 3) after finished numerical experiments.

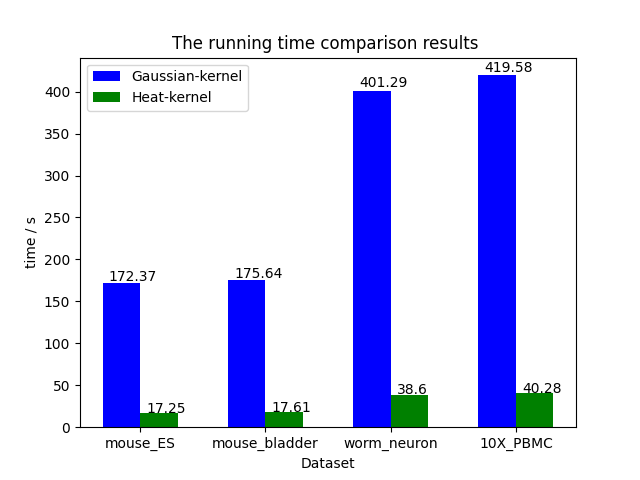


Figure 2 The running time comparison results between heat kernel and gaussian kernel

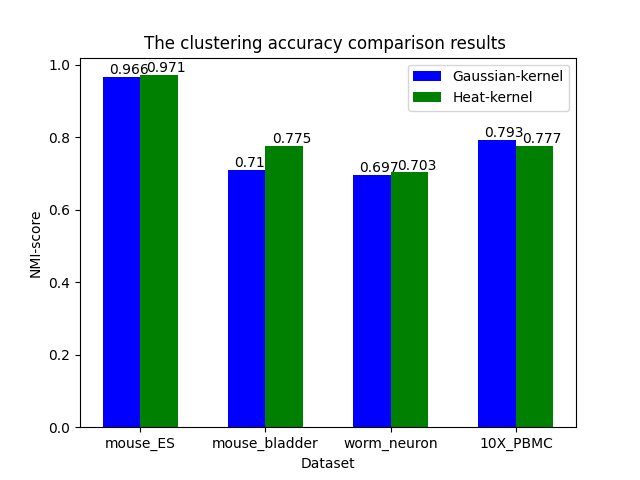


Figure 3 The clustering accuracy comparison results between heat kernel and gaussian kernel

# Real scRNA-seq datasets in our method

In the part of numerical experiments, we selected six scRNA-seq datasets from different sequencing platforms to demonstrate the performance of SHDC. 10X\_PBMC dataset [23] was generated from 10X genomics platform. The mouse\_bladder\_cell dataset [24] containing 16 cell types was generated from Microwell-seq platform. The mouse\_ES cell dataset [25] was from droplet barcoding platform and the worm\_neuron\_cell dataset [26] was from sci-RNA-seq platform. These four are representative sequencing platforms. In addition to these four datasets, Zheng dataset [23] and Zeisel dataset [27] were also used to demonstrate the performance of our method. The Zheng dataset has only 500 cells, the Zeisel dataset has more than 3000 cells. The detailed information about these real scRNA-seq datasets were listed in Table 3.

Table 3 Basic information about scRNA-seq datasets

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Cell number | Gene number | Cluster number |
| Zheng | 500 | 32738 | 3 |
| Mouse\_ES\_cell | 2717 | 24046 | 4 |
| Mouse\_bladder\_cell | 2746 | 19079 | 16 |
| Zeisel | 3005 | 19972 | 9 |
| Worm\_neuron\_cell | 4186 | 11955 | 10 |
| 10X\_PBMC | 4271 | 16499 | 8 |

# Evaluate the accuracy of clustering

## Normalized mutual information (NMI)

Normalized mutual information (NMI) [28] was used to evaluate clustering accuracy in this paper. NMI is often used to measure the consistency of clustering results. If *A* and *B* are the clustering results corresponding to two different clustering mechanisms, where and . The normalized mutual information of two results A and B can be calculated by formula (13):

. (13)

The coincidence part between A and B can be represented by a confusion matrix C of size . Where represents the number of cells belonging to both cluster and cluster , *n* represents the total number of cells. is the amount of cells in subpopulation *i* in result A, is the amount of cells in subpopulation *j* in result *B*. If the result A and B are identical, . Therefore, the better the clustering performance is, the closer value is to 1.

## Adjusted rand index (ARI)

In order to test the performance of SHDC more comprehensively, the adjusted rand index (ARI) [28] was also used in this paper. Similarly, if there are two different clustering results *A* and *B*, where and . The adjusted rand index can be calculated by formula (14):

. (14)

Where *A* and *B* mean the real labels and clustering labels, respectively. In Equation (14), denotes the number of cells belonging to *i* group in result *A* and *j* group in result *B*, denotes the number of cells belonging to the *i* group in real labels while denotes the number of cells belonging to the *j* group in clustering labels. is in the range of 0 to 1. The closer is to 1, the closer the clustering result is to the real result.

# Results

In the part of numerical experiments, five popular single-cell clustering methods with default parameters are used to compare with our method. These five methods include SC3 [8], SSCC [10], PARC [11], Seurat [12] and “DCA+kmeans” [18]. It should be pointed out that “DCA+kmeans” is a method from Gökcen Eraslan et al. [18]. This method consists of three steps. First, DCA is used to denoise scRNA-seq data. Second, the denoised data is projected into a low dimensional space using principal component analysis (PCA) [29]. Finally, the first two principal components are used for k-means clustering to get the final clustering results.

We compared our method with these five methods on all datasets in Table 3, and use NMI and ARI to evaluate the model performance. Under NMI, only SC3 has the same performance as our method on Zheng dataset. Compared with other methods, our method performs better on all datasets. More detailed results were showed in Table 4.

We can see from Table 5 that, only on the Zeisel dataset SC3 is slightly better than our method and only on the Zheng dataset SC3 has the same performance as our method. Compared with other methods, our method performs better on all datasets.

Finally, we did cell visualization for all scRNA-seq datasets (Fig 4). After projecting adjacency matrix W to a lower dimensional space using UMAP algorithm, the first two-dimensional coordinates were directly taken for cell visualization. From the results of cell visualization, we can see that SHDC can well divide different types of cells into different clusters.

Table 4 NMI score under different datasets

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Methods | SC3 | SSCC | PARC | Seurat | DCA+Kmeans | OUR |
| Zheng | 0.989 | 0.961 | 0.843 | 0.980 | 0.978 | 0.989 |
| 10X\_PBMC | 0.744 | 0.642 | 0.653 | 0.735 | 0.735 | 0.777 |
| Mouse\_bladder\_cell | 0.731 | 0.580 | 0.701 | 0.751 | 0.648 | 0.775 |
| Mouse\_ES\_cell | 0.858 | 0.741 | 0.560 | 0.771 | 0.856 | 0.971 |
| Worm\_neuron\_cell | 0.655 | 0.486 | 0.621 | 0.684 | 0.467 | 0.703 |
| Zeisel | 0.759 | 0.545 | 0.708 | 0.646 | 0.452 | 0.764 |

Table 5 ARI score under different dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Methods | SC3 | SSCC | PARC | Seurat | DCA+Kmeans | OUR |
| Zheng | 0.994 | 0.977 | 0.846 | 0.988 | 0.988 | 0.994 |
| 10X\_PBMC | 0.746 | 0.485 | 0.481 | 0.632 | 0.723 | 0.762 |
| Mouse\_bladder\_cell | 0.558 | 0.344 | 0.565 | 0.521 | 0.529 | 0.669 |
| Mouse\_ES\_cell | 0.803 | 0.705 | 0.433 | 0.597 | 0.852 | 0.983 |
| Worm\_neuron\_cell | 0.404 | 0.312 | 0.451 | 0.476 | 0.280 | 0.505 |
| Zeisel | 0.820 | 0.441 | 0.481 | 0.469 | 0.313 | 0.732 |

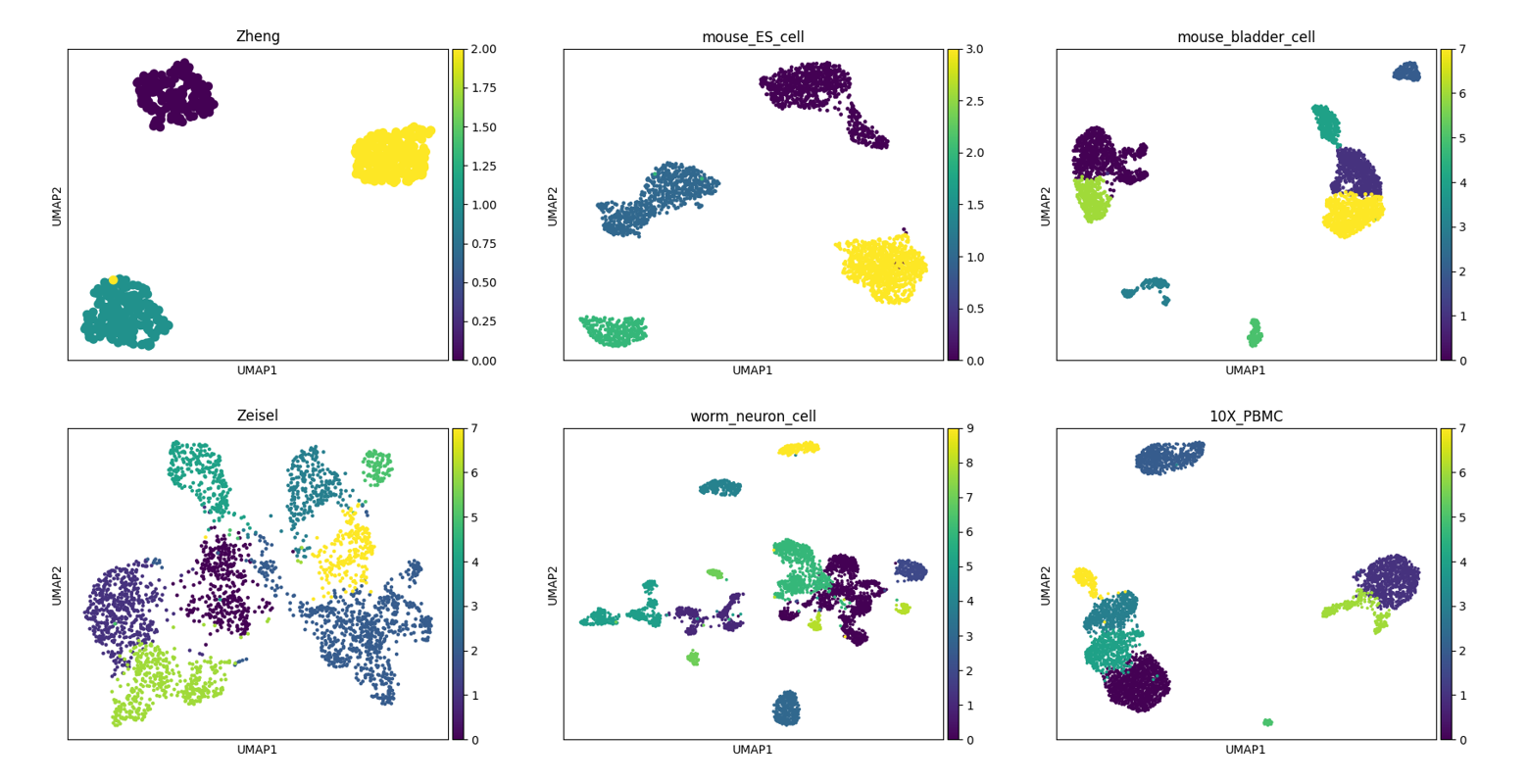


Figure 4 Cell visualization results for all scRNA-seq datasets

# Implementation

SHDC is implemented in Python 3 on HP Z840 workstation. In the implementation of SHDC, DCA module will be used. DCA is implemented in Python 3 using Keras [30] and its TensorFlow [31] backend. So before SHDC, make sure your computer has installed DCA module. In addition, SHDC uses Parallel Python (pp) module for CPU parallel computing, so it is necessary to ensure that pp module has been installed. The Python package scanpy [32] is used in the calculation, so SHDC supports h5ad-formatted HDF5 files and anndata data format.

# Discussion

The progress of single cell RNA sequencing technology has brought revolutionary changes to the research of transcriptomics. In recent years, many single-cell clustering methods have been proposed for scRNA-seq data. But many methods have some limitations in solving the impact of noise. In this paper, a new method called SHDC was proposed for single cell clustering. SHDC denoised scRNA-seq datasets to obtain a more stable data structure and only a heat kernel was used to measure similarity between cells. Experimental results show that SHDC has good performance. The progress of sequencing technology also makes the scale of scRNA-seq data increasing. For large-scale scRNA-seq datasets, there is still room for further improvement of clustering accuracy. These problems are the direction of our future efforts.

# Acknowledgements

This work was supported by grants from the Xinjiang Autonomous Region University Research Program (No. XJEDU2019Y002) and the National Natural Science Foundation of China (No. U19A2064, 61873001).

# References

1. Na C, Menglu L, Le Z, Bo Z, Yuhua Y, Chun-Hou Z, Junfeng X (2020) Comparison and integration of computational methods for deleterious synonymous mutation prediction. Briefings in Bioinformatics 21:970-981. <https://doi.org/10.1093/bib/bbz047>
2. Tian T, Wan J, Song Q. et al. (2019) Clustering single-cell RNA-seq data with a model-based deep learning approach. Nat Mach Intell 1:191–198. [https://doi.org/10.1038/s42256-019-0037-0](%20https://doi.org/10.1038/s42256-019-0037-0)
3. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert, JP (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications 9(1):1- <https://doi.org/10.1038/s41467-017-02554-5>
4. Duò A, Robinson MD and Soneson C. (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 1; peer review: 2 approved with reservations]. *F1000Research* **7**:1141 <https://doi.org/10.12688/f1000research.15666.1>
5. Kuo RJ, Wang HS, Hu TL, Chou SH (2015) Application of ant K-means on clustering analysis. Computers & Mathematics with Applications 50(10):1709-1724. <https://doi.org/10.1016/j.camwa.2005.05.009>
6. Filippone M, Camastra F, Masulli F, Rovetta S (2008) A survey of kernel and spectral methods for clustering. PATTERN RECOGNITION 41(1):176-190. <https://doi.org/10.1016/j.patcog.2007.05.018>
7. Raphael P, Zhuliu L, Rui K (2020) Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Briefings in Bioinformatics 21(4):1209–1223. <https://doi.org/10.1093/bib/bbz063>
8. Kiselev V, Kirschner K, Schaub M et al. (2017) SC3: consensus clustering of single-cell RNA-seq data. Nature Methods 14:483-486. <https://doi.org/10.1038/nmeth.4236>
9. Wang B, Zhu J, Pierson E et al. (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nature Methods 14:414-416. <https://doi.org/10.1038/nmeth.4207>
10. Ren X, Zheng L, & Zhang Z (2019) SSCC: A Novel Computational Framework for Rapid and Accurate Clustering Large-scale Single Cell RNA-seq Data. Genomics, Proteomics & Bioinformatics 17(2):201-210. <https://doi.org/10.1016/j.gpb.2018.10.003>
11. Shobana V Stassen, Dickson M D Siu, Kelvin C M Lee, Joshua W K Ho, Hayden K H So, Kevin K Tsia. (2020) PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells. Bioinformatics 36(9):2778-2786. <https://doi.org/10.1093/bioinformatics/btaa042>
12. Rahul S, Farrell JA, David G, *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**(5): 495–502. <https://doi.org/10.1038/nbt.3192>
13. Lin, P., Troup, M. & Ho, J.W. (2017) CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* 18, 59. <https://doi.org/10.1186/s13059-017-1188-0>
14. Zhicheng Ji, Hongkai Ji. (2016) TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis, *Nucleic Acids Research* 14(13):117. <https://doi.org/10.1093/nar/gkw430>
15. Jiang, L., Chen, H., Pinello, L. *et al.* (2016) GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol* **17,**144 (2016). <https://doi.org/10.1186/s13059-016-1010-4>
16. Davis J McCarthy, Kieran R Campbell, Aaron T L Lun, Quin F Wills. (2017) Scater:pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R, *Bioinformatics* 33(8):1179–1186, <https://doi.org/10.1093/bioinformatics/btw777>
17. Sibson R (1973) SLINK: An optimally efficient algorithm for the single-link cluster method. The Computer Journal 16:30-34. <https://doi.org/10.1093/comjnl/16.1.30>
18. Eraslan, G., Simon, L.M., Mircea, M. *et al.* (2019) Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* **10,**390. <https://doi.org/10.1038/s41467-018-07931-2>
19. Hao Ding, Michael Sharpnack, Chao Wang, Kun Huang, Raghu Machiraju. (2019) Integrative cancer patient stratification via subspace merging. *Bioinformatics* 35(10):1653–1659. <https://doi.org/10.1093/bioinformatics/bty866>
20. Seyoung Park, Hongyu Zhao. (2018) Spectral clustering based on learning similarity matrix. *Bioinformatics* 34(12):2069–2076. <https://doi.org/10.1093/bioinformatics/bty050>
21. McInnes L, Healy J, & Melville J (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
22. Sinha D, Kumar A, Kumar H, Bandyopadhyay S, & Sengupta D (2018) dropClust: efficient clustering of ultra-large scRNA-seq data. Nucleic Acids Research 46(6):e36-e36.<https://doi.org/10.1093/nar/gky007>
23. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent Z. W, Wilson R, ... & Gregory MT (2017) Massively parallel digital transcriptional profiling of single cells. Nature Communications 8(1):1-12. <https://doi.org/10.1038/ncomms14049>
24. Han, X. et al. Mapping the mouse cell atlas by Microwell-seq. (2018) Cell 172:1091–1107. <https://doi.org/10.1016/j.cell.2018.02.001>
25. Klein, A. M. et al. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201. <https://doi.org/10.1016/j.cell.2015.04.044>
26. Angerer, P. et al. (2017) Single cells make big data: new challenges and opportunities in transcriptomics. Curr. Opin. Syst. Biol 4:85–91. <https://doi.org/10.1016/j.coisb.2017.07.004>
27. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, ... & Rolny C (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347(6226):1138-1142. <http://doi.org/10.1126/science.aaa1934>
28. Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X. & Garmire, L. (2018) DeepImpute: an accurate, fast and scalable deep neural network method to impute single-cell RNA-seq data. <https://doi.org/10.1101/353607>
29. Raphael P, Zhuliu L, Rui K (2020) Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Briefings in Bioinformatics 21(4):1209–1223. <https://doi.org/10.1093/bib/bbz063>
30. Chollet, F. (2015) Keras. Github. <https://github.com/fchollet/keras>
31. Martín, A. A. et al. (2015) TensorFlow: Large-Scale Machine Learning onHeterogeneous Systems. <https://www.tensorflow.org/>
32. Wolf F. A, Angerer P, & Theis F. J (2018) SCANPY: large-scale single-cell gene expression data analysis. Genome Biology 19(1):1-5. <https://doi.org/10.1186/s13059-017-1382-0>