Application of machine learning to associative scRNA-seq data gene expression and alternative polyadenylation sites clustering

Jiongsong Hu[1], Chao Ren[2], Wenjie Shu[3] and Gangqiao Zhou[4,*]

[1]University of South China, Hengyang Hunan 421001, China

[2]Institute of Health Service and Transfusion Medicine, Beijing, 100850, China

[3]Beijing Institute of Microbiology and Epidemiology, Beijing, 100850, China

[4]Department of Genetics & Integrative Omics, State Key Laboratory of Proteomics, National Center for Protein Sciences, Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing, 100850, P. R. China

Abstract. Cell type identification is a vital step in the analysis of scRNA-seq data. Transcriptome subtype pivotal information such as alternative polyadenylation (APA) obtained from standard scRNA-seq data can also provide valid clues for cell type identification with no alteration of experimental techniques or increased experimental costs. Furthermore, using multimodal analysis techniques and their methods, more confident cell type identification results can be obtained. For that purpose, we constructed a workflow framework: On five different scRNA-seq datasets, 18 methods based on machine learning that have not yet been applied to identify cell types by association APA and single-cell gene expression fusion were compared with three single-cell clustering methods, and compared these method against the advanced method scLAPA based on similarity network fusion (SNF). In our experiments, we used the adjusted Rand index (ARI) as a metric. We found that unsupervised methods like WMSC and supervised methods like MOGONET have more robust and excellent results in associating APA with single-cell gene expression clustering than methods based only on single-cell gene expression clustering and advanced scLAPA methods.

1. Introduction

In recent years, many innovative multimodal methods applied in the fields of image and natural language processing have been transferred to the analysis of biomedical research by many researchers. To mine the expression of a gene in a specific environment and action pathway, standard scRNA-seq data are frequently combined with other genomic data for association analysis. However, in the actual scientific research process, due to the relationship between funds, materials, platforms, and technology, many standard single-cell data often do not have corresponding multi-group data such as methylation data, proteomics data, or metabolome data.

Recent bioinformatics studies suggest that alternative polyadenation (APA) [1] expression profiles can be used to identify cell types by capturing key transcriptomic information about APA sites from standard scRNA-seq data and revealing intercellular dynamics between cell types. The use of the APA expression to probe intercellular dynamics not only allows the discovery of alternative patterns to gene expression profiles from standard scRNA-seq data without changing experimental techniques, but also offers great potential for the development of efficient methods to resolve cell types. Thus, APA site expression data can be used as transcriptome isoform information instead of DNA methylation, open chromatin, or proteome information with gene expression data for multi-view analysis. In the state-of-the-art studies, scLAPA [1] based on similarity network fusion (SNF) [2] is the most advanced approach. The three proposed methodological strategies applied to fusion clustering of single-cell gene expression and APA sites in this paper, with the exception of the method based on unsupervised autoencoder fusion embedding, show improved results compared to existing state-of-the-art methods in all five datasets.

The organization of this paper is as follow： the methods are presented in section 2.Section 3 is the results. Section 4 is the discussion. Section 5 is the conclusion.

2. Methods

2.1 Identify APA sites and Data processing

Two methods, including SCAPTURE (Guo et al., 2021)

* E-mail: zhougq114@126.com

[3] and scAPAtrap (Tao et al., 2020) [4], were used in this study to identify APA sites and to quantify the transcriptome level of these sites. SCAPTURE identifies peaks using the findPeaks command from HOMOER and trains the sequence shifts of peaks using an embedded deep learning network, DeepPASS, after which it evaluates the predicted high-confidence peaks and validates them with the collected APA sites. scAPAtrap uses the region matrix function in the R packages named derfinder [5] to identify potential peaks at the whole genome level. Afterwards, peaks with widths >1000 bps were used as the threshold for cleavage based on the size of the peak area, using a quarter digit of the total read coverages, and this step was repeated until all peaks were not broad peaks.

We collected three gold standard datasets [1] and two silver standard datasets [6] as benchmarks to test the methods we needed to validate.We used scAPAtrap to identify APA sites in the three datasets of the gold standard dataset and used SCAPTURE to identify APA sites in other two datasets of the silver standard dataset. We quantified the APA sites at the transcriptome level for each of the five datasets to obtain the PA matrix, and the initial data for the gene expression were quantified as the GE matrix. We used the FindVariousFeatures function from Seurat [7] to extract the top 2000 highly expressed genes in the GE matrix and the top 2000 highly expressed APA sites in the PA matrix.

The HY-dataset is a scRNA-seq dataset of mouse hypothalamus composed of 727 single cell data contains 7 categories. The EPI-dataset is a scRNA-seq dataset of mouse mammary epithelial cells composed of 2127 single cell contains 5 categories. The TAIR-dataset is a scRNA-seq dataset of Arabidopsis roots composed of 1473 single-cell data contains 7 categories. PBMC-4K is a scRNA-seq dataset of 4K human peripheral blood cells composed of 4292 single cell data. It contain 11 categories. PBMC-8K is a scRNA-seq dataset of 8K human peripheral composed of 8352 single cells and contains 11 categories. The specific cell type of five datasets is shown as Table 1.

Table 1. The count of cells and Categories among datasets.

Datasets	Categories	Counts of cells	Cell type
HY	7	727	Astrocytes,Endothelial cells,Ependymal cells,Microglia,Neurons, Oligodendrocyte progenitor cells,Oligodendrocytes
EPI	5	2127	Avp,Bsl,Hsd,Hsp,Lp
TAIR	7	1473	Cortex,Endocortex,Endodermis,Epidermis(H),Epidermis(N),Root Cap,Stele
PBMC-4K	11	4292	CD14+ Monocyte,CD19+ B,CD34+,CD4+ T Helper2, CD4+/CD25 T Reg,CD4+/CD45RA+/CD25- Naive T,CD4+/CD45RO+ Memory,CD56+ NK CD8+ Cytotoxic T,CD8+/CD45RA+ Naive Cytotoxic, Dendritic
PBMC-8K	11	8352	CD14+ Monocyte,CD19+ B,CD34+, CD4+ T Helper2,CD4+/CD25 T Reg,CD4+/CD45RA+/CD25- Naive T,CD4+/CD45RO+ Memory, CD56+ NK,CD8+ Cytotoxic T,CD8+/CD45RA+ Naive Cytotoxic, Dendritic

2.2 Algorithm scheme

We validated the effectiveness of multiple machine learning and deep learning approaches for the optimization of clustering of single-cell gene expression data associated with APA sites. Recent studies have found that supervised learning has the advantage of fast training, but has some limitations in accuracy, compared to unsupervised learning, which can substantially compensate for the shortcomings of supervised learning. Therefore, we designed three schemes to build the fusion clustering part of the workflow.Adjusted Rand Index (ARI) [18] are used to evaluate the performance of the proposed algorithm and the baselines. To avoid the randomness, we run all the algorithms 5 times and report their average values.

2.2.1Based on unsupervised spectral clustering algorithms

There is a consensus on the ground truth of the Laplacian matrix among all the views. Typically, the consensus Laplacian matrix is unknown. However, it can be approximated by a weighted combination of Laplacian matrices for each view [8].

With a given data set $X = \{X^{(1)}, X^{(2)}, \ldots, X^{(n_{view})}\}$ with n_{view} views , where $X^{(a)} = \{x_1{}^{(a)}, x_2{}^{(a)}, \ldots, x_n{}^{(a)}\}$, $a \in \{1,2,\ldots,n_{view}\}$, n is the number of data points, $L^{(a)} \in R^{n\times n}$ is the Laplacian of the a-th view, $L^* \in R^{n\times n}$ is the consensus Laplacian matrix. The objective function of calculating L^* is shown as follows:

$$L^* = \sum_{a=1}^{n_{view}} \mu_a L^{(a)} \ \ s.t \sum_{a=1}^{n_{view}} \mu_a = 1, \mu_a \geq 0 \quad (1)$$

where μ_a is the weight of the a -th view. Afterwards, L^* can be applied to spectral clustering.

Therefore,in this scheme, we validated six methods based on multi-view spectral clustering that have been optimized . These include co-regularized spectral clustering(CoRegSC) [9] , affinity aggregation for spectral clustering(AASC) [10], robust multi-view spectral clustering(RMSC) [11], multiview consensus graph clustering(MCGC) [12], multi-view clustering via adaptively weighted procrustes (AWP) [13] and weighted multi-view spectral clustering based on spectral perturbation(WMSC) [14].

2.2.2Based on unsupervised autoencoder fusion embedding

Autoencoder is a deep neural network. It mainly consists of an encoder and a decoder, both of which are multilayer neural networks and can be represented by Equation (2) and Equation(3):

$$z = f_{encoder}(x) \quad (2)$$

$$z = f_{decoder}(x) \quad (3)$$

There are two types of model structures based on autoencoders including early fusion and late fusion.

The n_{view} views vectors are into a feature vector X . It's the early fusion way. Therefore, the encoder and the decoder can be represented as $z = f_{encoder}(X)$ and $X' = f_{decoder}(z)$. And the other late fusion way is that n_{view} autoencoder used to perform feature extraction on the n_{view} views vectors. The encoder and decoder can be expressed in Equation(4) and Equation(5), respectively.

$$z_i = f_{encoder(i)}(x_i), i = 1,2,\dots,n_{view} \quad (4)$$

$$x_i' = f_{decoder(i)}(z_i), i = 1,2,\dots,n_{view} \quad (5)$$

Finally, the latent features z_i of each views were concatenated as multi-views fusion features z_{fusion}.

Therefore, in this scheme, we validated 10 methods based on unsupervised autoencoder s fusion embedding. These methods were built by (Leng et al, 2022) [15] for the validation of fusion clustering of biological multimodal data. These methods, which consist of the combined five autoencoders and varied with two fusion strategies is shown as Table2.

Table 2. The models based on autoencoder

	early fusion	late_fusion
Autoencoder	efAE	lfAE
Denoising AutoEncoder	efDAE	lfDAE
Variational AutoEncoder	efVAE	lfVAE
Stacked Variational AutoEncoder	efSVAE	lfSVAE
Autoencoder with maximum mean discrepancy function	efmmdVAE	lfmmdVAE

2.2.3Based on supervised deep learning model

The third scheme is inspired by MOGONET **错误!未找到引用源。**. It combines graph convolutional networks (GCN) for multi-omics-specific learning and the VCDN [17] for multi-omics integration. It is mainly divided into two parts: 1. Initial feature prediction of individual classes of each omics dataset in GCNs 2. Using the results of the initial prediction, a cross-omics discovery tensor is constructed and sent to VCDN for training. The specific application to association single-cell gene expression and APA sites association clustering is shown in Figure1.

Besides, we also combine fully connected fully connected neural network (NN) for multi-omics-specific learning and the VCDN for multi-omics integration to compare against MOGONET (GCN-VCDN).

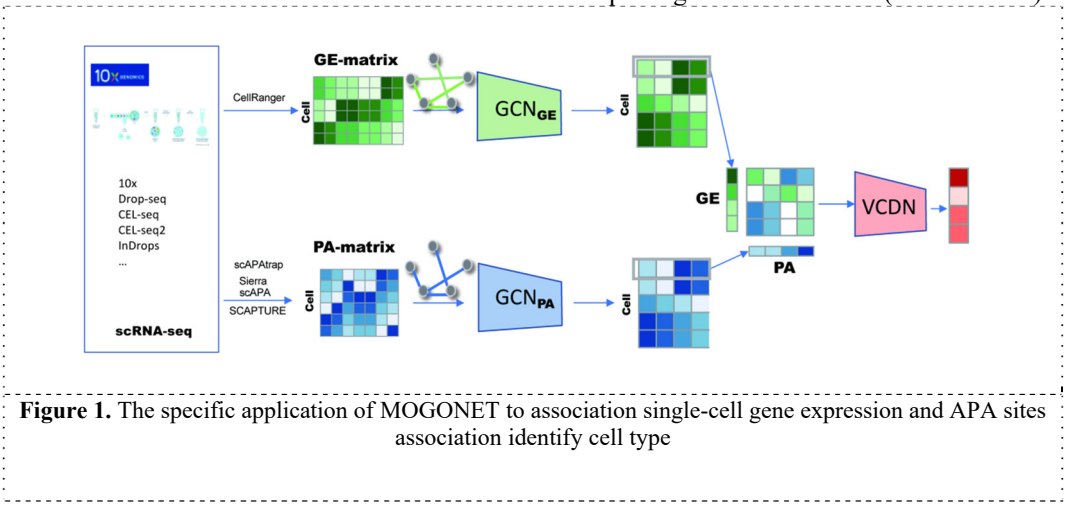

Figure 1. The specific application of MOGONET to association single-cell gene expression and APA sites association identify cell type

3. Results

When we tested the GCN-VCDN model [16], we found that one of the hyper-parameters, k, has a significant impact on the experimental results. In this model, the hyper-parameter k represents the average number of edges of the sample in the similarity network. If k is too large, the similarity network is too dense, which will lead to noise generation. On the contrary, if k is too small, the correlation between samples in the similarity network may be lost. Therefore, we tested the performance of different hyper-parameters on each data set and took the best score as the score for GCN-VCDN in the later data analysis, as shown in Figure 2.

In addition to the designed strategy approach, we compared three single-cell gene expression matrix clustering methods, SINCERA [19], SNN-Clip [20], and dynamic Tree Out [21]. The clustering visualization comparison diagram of different algorithms on the five datasets is shown in Table 3.

From Table 3, it can be seen that the algorithms based on supervised deep learning models are the best in terms of ARI. The average ARI of the method based on the supervised deep learning model was improved by 70.93% compared to the optimal method of single-cell gene expression matrix clustering and by 37.1% compared to scLAPA. Based on unsupervised spectral clustering algorithms, they also achieved better results, and their average ARI improved by 32.58% compared with the optimal method of single-cell gene expression matrix clustering and 6.3% compared with scLAPA. Nevertheless, the method based on unsupervised autoencoder fusion embedding improved the average ARI by 11.16% compared to the optimal method of single-cell gene expression matrix clustering. However, it decreased by 10.87% compared to scLAPA.

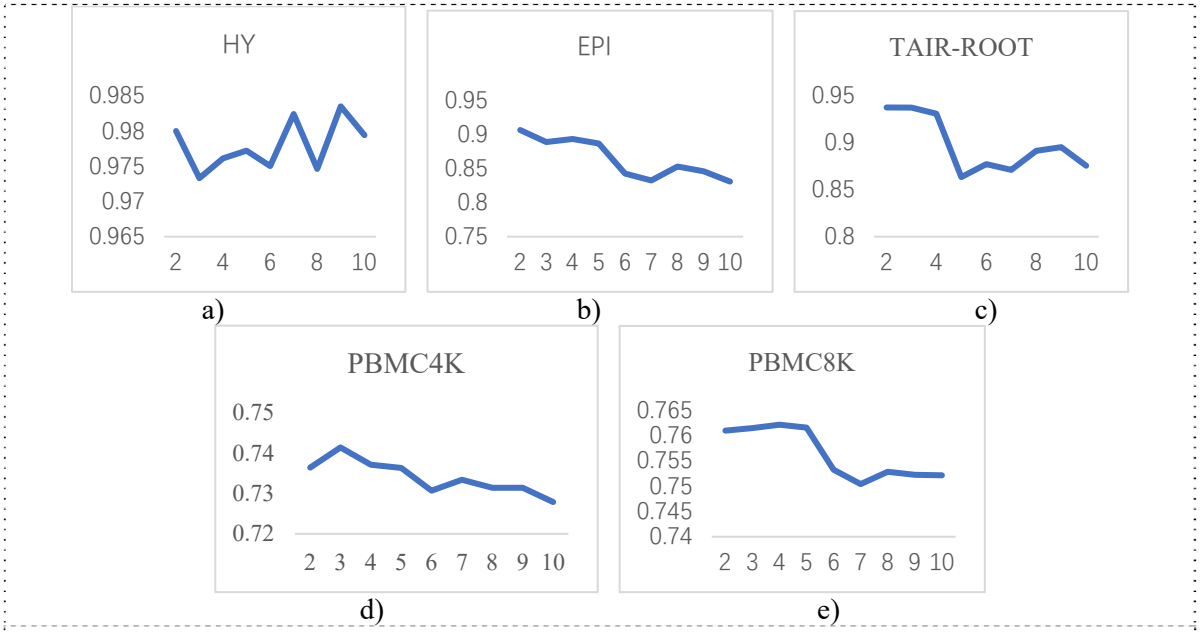

Figure 2. (a-e) show the performance of the datasets HY, EPI, TAIR, PBMC-4K, and PBMC-8K under different values of k.

Table 3.ARI on all the datasets.

Method	HY	EPI	TAIR	PBMC-4K	PBMC-8K
SINCERA	0.9027(0.0000)	0.1248(0.0000)	0.2964(0.0000)	0.2395(0.0000)	0.2109(.0.0000)
SNN-Clip	**0.9248(0.0000)**	**0.3944(0.0000)**	**0.7011(0.0000)**	**0.2650(0.0000)**	**0.2650(0.0000)**
dynamic-Tree-out	0.9244(0.0000)	0.2433(0.0000)	0.3834(0.0000)	0.1732(0.0000)	0.2427(0.0000)
scLAPA	**0.9315(0.0000)**	**0.6166(0.0000)**	**0.7686(0.0000)**	**0.4490(0.0000)**	**0.4823(0.0000)**
CoeReg	0.9112(0.0071)	0.8371(0.0002)	0.6761(0.0000)	0.4292(0.0034)	**0.4030(0.0308)**
AASC	0.9652(0.0000)	0.8780(0.0000)	0.6850(0.0000)	0.3572(0.0003)	0.3544(0.0000)
RMSC	0.6153(0.0000)	**0.9076(0.0000)**	0.0653(0.0002)	**0.3794(0.0368)**	0.3381(0.0163)
MCGC	0.1083(0.0000)	0.3606(0.0000)	0.0012(0.0000)	0.1957(0.0000)	0.0007(0.0000)
AWP	**0.9845(0.0000)**	0.7478(0.0000)	**0.6880(0.0000)**	0.4069(0.0000)	0.3225(0.0000)
WMSC	**0.9795(0.0000)**	**0.8222(0.0000)**	**0.7305(0.0000)**	**0.4318(0.0014)**	**0.3918(0.0367)**
efAE	0.5555(0.2239)	**0.8444(0.0387)**	**0.5111(0.2114)**	**0.4537(0.0237)**	**0.4497(0.0414)**
lfAE	0.6569(0.1206)	**0.6624(0.1512)**	0.2053(0.0971)	0.3313(0.1062)	0.2042(0.1956)
efDAE	0.6108(0.2004)	0.6360(0.0424)	0.4473(0.1979)	**0.4503(0.0286)**	0.4270(0.0451)
lfDAE	**0.7884(0.0714)**	0.6123(0.1617)	0.1728(0.1122)	0.3111(0.0915)	0.1314(0.0652)
efMMDVAE	0.7812(0.0000)	0.6326(0.0000)	0.2238(0.0000)	0.4473(0.0000)	0.4880(0.0000)
lfMMDVAE	0.4579(0.1746)	0.6329(0.0000)	0.2521(0.0000)	0.2277(0.0000)	0.3180(0.0000)

lfVAE	0.7676(0.1075)	0.5855(0.0358)	0.4746(0.1267)	0.4347(0.0359)	**0.4644(0.0364)**
efVAE	**0.8106(0.0535)**	0.6789(0.0604)	**0.5972(0.2161)**	0.4007(0.0095)	0.4099(0.0648)
efSVAE	0.0175(0.0107)	0.0130(0.0050)	0.0025(0.0066)	0.0453(0.0128)	0.1332(0.0316)
lfSVAE	0.0142(0.0236)	0.0101(0.0069)	0.0014(0.0019)	0.0670(0.0253)	0.1227(0.0165)
NN-VCDN	0.9714(0.0143)	0.8632(0.0093)	0.9291(0.0026)	**0.7535(0.0112)**	**0.7771(0.0078)**
GCN-VCDN	**0.9835(0.0072)**	**0.9062(0.0173)**	**0.9369(0.0144)**	0.7414(0.0054)	0.7614(0.0028)

4. Discussion

The results demonstrate the validity of the APA sites information as other modal data for scRNA-seq data again . The approach based on supervised deep learning model have a significant improvement against clustering using gene expression data alone. This shows that the method of self-training single data using neural networks is effective before iterative training by constructing a cross-tensor, provided that labels are available. In the mean time, the scheme without any label annotations also has a significant improvement over using only gene expression profile clustering. However,the method based on unsupervised autoencoder fusion embedding didn't perform even though it exceeds the method using only gene expression profile clustering. Without any label to train, the method based on spectral clustering maybe perform better than deep learning methods that rely on training labels through optimization. In practical studies, there is not much data that has been benchmark corrected and has labels. The advanced unsupervised or semi-supervised methods are more widely used. With the development of deep learning, the models trained by getting rid of artificial labels are more valuable. The fusion embedding method based on self-encoder, although the performance is normal but reach the requirement of solving the problem. It also provides a research direction to the method researchers.

5. Conclusion

This paper examines the different forcing performance of 18 methods on five datasets from three schemes in the study of multimodal association analysis using only standard scRNA-seq data. All three schemes show greater enhancements compare to advance method . These three schemes offer other researchers wider ideas for multimodal data analysis in biomedicine.

Reference

1. Ji G, Xuan W, Zhuang Y, Ye L, Zhu S, Ye W, et al. 2021 Learning association for single-cell transcriptomics by integrating profiling of gene expression and alternative polyadenylation *BioRxiv*. 2021.01.04.425335
2. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. 2014 Similarity network fusion for aggregating data types on a genomic scale *Nature Methods*. **11(3)** 333-7
3. Li G-W, Nan F, Yuan G-H, Liu C-X, Liu X, Chen L-L, et al. 2021 SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3′ tag-based RNA-seq of single cells *Genome Biology*. **22(1)** 1-24
4. Wu X, Liu T, Ye C, Ye W, Ji G 2021 scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data *Briefings in Bioinformatics*. **22(4)** bbaa273
5. Collado-Torres L, Nellore A, Frazee AC, Wilks C, Love MI, Langmead B, et al. 2017 Flexible expressed region analysis for RNA-seq with derfinder *Nucleic Acids Research*. **45(2)** e9-e.
6. Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M 2018 Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data *F1000Research*. **7**
7. Hao Y, Hao S, Andersen-Nissen E, Mauck III WM, Zheng S, Butler A, et al. 2021 Integrated analysis of multimodal single-cell data *Cell*. **184(13)** 3573-87. e29
8. Zhou D, Burges CJ 2007 Spectral clustering and transductive learning with multiple views *Proceedings of the 24th international conference on Machine learning*. 1159-66
9. Kumar A, Rai P, Daume H 2011 Co-regularized multi-view spectral clustering *Advances in Neural Information Processing Systems*. **24**
10. Huang H-C, Chuang Y-Y, Chen C-S 2012 Affinity aggregation for spectral clustering *2012 IEEE Conference on computer vision and pattern recognition*. 773-80
11. Xia R, Pan Y, Du L, Yin J 2014 Robust multi-view spectral clustering via low-rank and sparse decomposition *Proceedings of the AAAI conference on artificial intelligence*. **28(1)**
12. Zhan K, Nie F, Wang J, Yang Y 2018 Multiview consensus graph clustering *IEEE Transactions on Image Processing*. **28(3)** 1261-70
13. Nie F, Tian L, Li X 2018 Multiview clustering via adaptively weighted procrustes *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2022-30.
14. Zong L, Zhang X, Liu X, Yu H 2018 Weighted multi-view spectral clustering based on spectral perturbation *Proceedings of the AAAI conference on artificial intelligence*. **32(1)**
15. Leng D, Zheng L, Wen Y, Zhang Y, Wu L, Wang J, et al. 2022 A benchmark study of deep learning-based

multi-omics data fusion methods for cancer *Genome Biology*. **23(1)** 1-32

16. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. 2021 MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification *Nature Communications*. **12(1)** 1-13
17. Wang L, Ding Z, Tao Z, Liu Y, Fu Y 2019 Generative multi-view human action recognition *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 6212-21
18. Steinley D 2004 Properties of the hubert-arable adjusted rand index *Psychological Methods.* **9(3)** 386
19. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y 2015 SINCERA: a pipeline for single-cell RNA-Seq profiling analysis *PLoS Computational Biology*. **11(11)** e1004575
20. Xu C, Su Z 2015 Identification of cell types from single-cell transcriptomes using a novel clustering method *Bioinformatics*. **31(12)** 1974-80
21. Langfelder P, Zhang B, Horvath S 2008 Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R *Bioinformatics*. **24(5)** 719-20