Computational and Structural Biotechnology Journal 20 (2022) 2181–2197

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

scEFSC: Accurate single-cell RNA-seq data analysis via ensemble consensus clustering based on multiple feature selections

Chuang Bian[a], Xubin Wang[a], Yanchi Su[a], Yunhe Wang[b,*], Ka-chun Wong[c], Xiangtao Li[a,*]

[a] School of Artificial Intelligence, Jilin University, Changchun, 130000, Jilin, China
[b] School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China
[c] Department of Computer science, City University of Hong Kong, Hong Kong Special Administrative Region

A R T I C L E   I N F O

Article history:
Received 5 December 2021
Received in revised form 9 April 2022
Accepted 17 April 2022
Available online 27 April 2022

Keywords:
scEFSC
scRNA-seq
Feature selection
Consensus clustering

A B S T R A C T

With the development of next-generation sequencing technologies, single-cell RNA sequencing (scRNA-seq) has become one indispensable tool to reveal the wide heterogeneity between cells. Clustering is a fundamental task in this analysis to disclose the transcriptomic profiles of single cells and is one of the key computational problems that has received widespread attention. Recently, many clustering algorithms have been developed for the scRNA-seq data. Nevertheless, the computational models often suffer from realistic restrictions such as numerical instability, high dimensionality and computational scalability. Moreover, the accumulating cell numbers and high dropout rates bring a huge computational challenge to the analysis. To address these limitations, we first provide a systematic and extensive performance evaluation of four feature selection methods and nine scRNA-seq clustering algorithms on fourteen real single-cell RNA-seq datasets. Based on this, we then propose an accurate single-cell data analysis via Ensemble Feature Selection based Clustering, called scEFSC. Indeed, the algorithm employs several unsupervised feature selections to remove genes that do not contribute significantly to the scRNA-seq data. After that, different single-cell RNA-seq clustering algorithms are proposed to cluster the data filtered by multiple unsupervised feature selections, and then the clustering results are combined using weighted-based meta-clustering. We applied scEFSC to the fourteen real single-cell RNA-seq datasets and the experimental results demonstrated that our proposed scEFSC outperformed the other scRNA-seq clustering algorithms with several evaluation metrics. In addition, we established the biological interpretability of scEFSC by carrying out differential gene expression analysis, gene ontology enrichment and KEGG analysis. scEFSC is available at https://github.com/Conan-Bian/scEFSC.

1. Introduction

Advances in next-generation sequencing technologies brought single-cell RNA sequencing (scRNA-seq), which is rapidly generating large amounts of the single-cell RNA sequencing data. The main improvement achieved by scRNA-seq is that it overcomes the drawback of traditional bulk RNA sequencing in averaging gene expression across all cells in a sample [1]. On the contrary, scRNA-seq measures the transcriptome of individual cells, which makes high throughput investigations of tissue samples contributing to reveal heterogeneity of the cells and the molecular basis of phenotypic variation between them. This technology can provide new opportunities to characterize and understand the investigation of complex diseases and the dynamics of cell developmental cycles at the single cell resolution [2]. Therefore, the accurate identification of cell types and their underlying profiles has become an important step in single-cell RNA-seq analysis [3].

Clustering has been proven to be a critical computational method in grouping cells according to transcriptomic characteristics and thereby annotating cell types [4]. Recently, a plethora of clustering algorithms have been used to address the complexities of the scRNA-seq data [5]; for instance, Lin et al. [6] proposed CIDR, a clustering algorithm that reduces data dropouts through an implicit interpolation method using principal coordinate analysis (PCoA) to reduce the dimensionality. Qiu et al. [7] presented the Monocle model for clustering using t-SNE and density peak clustering. Yau et al. [8] proposed the pcaReduce algorithm, which integrates principal component analysis (PCA) and hierar-

* Corresponding author.
E-mail addresses: conanbian486@gmail.com (C. Bian), wangxb19@mails.jlu.edu.cn (X. Wang), suyanchi@gmail.com (Y. Su), wangyh082@hebut.edu.cn (Y. Wang), kc.w@cityu.edu.hk (K.-c. Wong), lixt314@jlu.edu.cn (X. Li).

https://doi.org/10.1016/j.csbj.2022.04.023
2001-0370/

chical clustering to generate a cell state hierarchy in which each clustering branch is associated with a variant principal component. Levine et al. [9] developed the PhenoGraph algorithm for the analysis of high-dimensional single-cell data, by creating a graph representing the phenotypic similarity between cells. Satija et al. [10] proposed the Seurat algorithm for clustering based on PCA, Shared Nearest Neighbour (SNN) graphs and Louvain's algorithm. Guo et al. [11] proposed the SINCERA algorithm to optimize cell clusters by hierarchical clustering for detecting differentially expressed genes. Grün et al. [12] developed RaceID for identifying rare cell types in complex single cell populations. It is difficult to imagine, however, that each clustering method can be equally effective across all scRNA-seq datasets. Indeed, each clustering algorithm has its own strengths and weaknesses and performances exhibit specific characteristics on particular scRNA-seq datasets [13]. Therefore, this is a challenge for users to decide which clustering algorithm is the most appropriate choice for the scRNA-seq data in hand.

Clustering ensembles have emerged as an effective method for combining solutions from multiple individual clustering algorithms into a consensus result. A number of clustering ensemble models have been developed for the scRNA-seq data; for example, Kiselev et al. [14] proposed the SC3 algorithm to combine the base clustering results obtained using spectral transformation and k-means algorithms into a consensus matrix, and then employing a hierarchical clustering to achieve the final clustering result. Wan et al. [15] presented the SHARP algorithm to cluster scRNA-seq data, using meta-clustering algorithms to obtain the final clustering result. Yang et al. [16] proposed the SAFE-clustering algorithm to first perform an independent clustering using four methods, SC3, CIDR, Seurat, and t-SNE + k -means, and then integrating solutions using three hypergraph-based partitioning algorithms. Geddes et al. [17] developed an autoencoder-based cluster integration framework to obtain random subspace projections from the data and the reducing dimensionality using an autoencoder. Zhu et al. [18] developed the Sc-GPE algorithm by combining five clustering methods based on single-cell graph partitioning and calculating the probability of cell pairs being classified into the same cluster. Huh et al. [19] presented the SAME-clustering algorithm to generate clustering solutions from multiple methods and selected a subset of maximum diversity to produce improved integrated solutions. However, we find that most clustering algorithms suppose that all features are equally significant in the clustering while in reality, diverse features have different effects on clustering [20]. This is one of the reasons why most clustering algorithms often perform poorly when faced with high-dimensional scRNA-seq data [21]. Therefore, efficient feature selection methods need to be developed for optimizing scRNA-seq data analysis.

To address these challenges, we first carry out a performance evaluation of multiple feature selection methods and nine scRNA-seq clustering algorithms on fourteen real single-cell RNA-seq datasets. Then, we present an accurate single-cell data analysis via Ensemble Feature Selection based Clustering, called scEFSC. The first module proposes removing genes that do not add significantly to the analysis of the scRNA-seq data using multiple unsupervised feature selections. Then, the second module implements different scRNA-seq clustering algorithms on the data generated in the first module to cluster the data, and then combines all the clustering results using a weighting-based meta-clustering method to obtain the final result. We applied scEFSC to the fourteen tested real scRNA-seq datasets compared clustering performance using several evaluation metrics to other scRNA-seq clustering algorithms. Results showed that our proposed scEFSC outperformed the other clustering algorithms. In addition, we realized differential gene expression analysis, gene ontology enrichment and KEGG analysis, demonstrating the biological interpretability of scEFSC.

2. Methods

2.1. Overview of the scEFSC algorithm

In our study, we developed an accurate single-cell data analysis via Ensemble Feature Selection based Clustering, called scEFSC. The input of the proposed scEFSC was the scRNA-seq data, which is composed of an $n \times d$ gene expression matrix $X = \{x_1, x_2, \cdots, x_n\}; x_i = \{x_i^1, x_i^2, \cdots, x_i^d\}$, where n is the number of cells and d is the number of genes. To begin, data pre-processing was performed using a log2 transformation to normalize the data and then genes detected in the normalized data in less than 2% of the cells were removed to filter out the low-level expressed genes from the scRNA-seq data. The overall framework of our proposed scEFSC is summarized in Fig. 1. As depicted in this figure, scEFSC consists of three important phases.

In phase A, we first employed a non-negative kernel autoencoder [22] to pre-select 5000 genes to remove insignificant genes. After that, we proposed multiple unsupervised feature selections including Low Variance [23], Laplacian Score [24], SPEC [25], and MCFS [26] to remove genes that do not contribute significantly to the analysis of the scRNA-seq data. We then fed the derived feature subsets into the clustering algorithms. Among them, we can observe that Low Variance is based on statistics; Laplacian Score and SPEC are based on similarity and MCFS is based on sparse learning. The last three methods are extensions of spectral model previously used for scRNA-seq data analysis. Unlike feature extraction methods, these feature selection methods do not change the original representation of the data and are considered to provide better readability and interpretability [27].

In phase B, we applied several different scRNA-seq clustering algorithms to cluster the feature subsets obtained by the multiple feature selection models. Various scRNA-seq clustering methods exist to run in our model. These methods are based on different underlying mathematical formulations as described above, including SC3 [14], CIDR [6], monocle [7], pcaReduce [8], Rphenograph [9], Seurat [10], SHARP [15], SINCERA [11], and RaceID [12]. For each feature subset derived from the different feature selection models, we applied the stated clustering methods to generate cluster labels to finally yield a set of individual cluster labels. In addition, to enhance the diversity of the individual cluster labels in the set, the pairwise Adjusted Rand Index (ARI) was employed to measure the similarity between any two individual clustering labels and then remove the method having similarity with the lowest variance [19]. In phase C, a weighted-ensemble clustering method called wMetaC was used to obtain the final clustering result of the individual cluster labels [15].

2.2. Non-negative kernel autoencoder

Since non-negative kernel autoencoders have been proven to be successful for screening single-cell sequencing data with highly expressed genes [22], in our study we first use non-negative nuclear autoencoders to preselect genes. Indeed, the normalized data is fed to a single-layer autoencoder to remove the insignificant genes, which consists of two important parts: the encoder and the decoder. The autoencoder can be formulated as follows.

$$e = f_E(x), f_E(x) = xW_E + b_E \tag{1}$$

$$\bar{x} = f_D(e), f_D(e) = eW_D + b_D \tag{2}$$

where f_E and f_D represent the conversion function of the encoder and decoder layers, x is the input of the encoder, $\bar{x}$ is a restruction of x, e is the reduced dimensional data of x, W is the weight matrix, and b is the bias vector. The encoder aims to represent the data in a

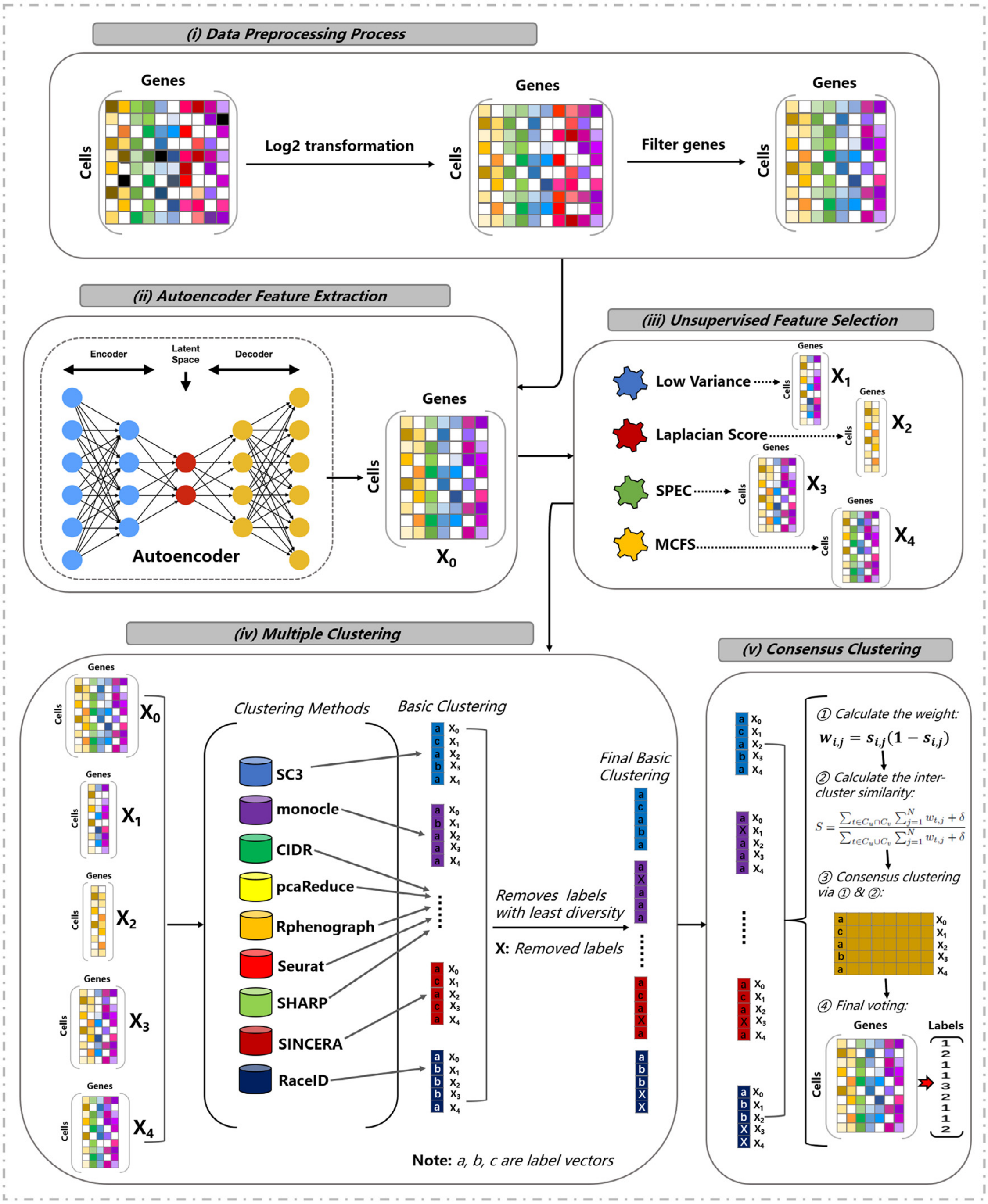
(i) Data Preprocessing Process
Genes
Cells
Log2 transformation
Filter genes
(ii) Autoencoder Feature Extraction
Encoder
Latent Space
Decoder
Autoencoder
X_0
(iii) Unsupervised Feature Selection
Low Variance
Laplacian Score
SPEC
MCFS
X_1
X_2
X_3
X_4
(iv) Multiple Clustering
Clustering Methods
Basic Clustering
SC3
monocle
CIDR
pcaReduce
Rphenograph
Seurat
SHARP
SINCERA
RaceID
Final Basic Clustering
Removes labels with least diversity
X: Removed labels
Note: a, b, c are label vectors
(v) Consensus Clustering
① Calculate the weight:
$w_{i,j} = s_{i,j}(1 - s_{i,j})$
② Calculate the inter-cluster similarity:
$S = \frac{\sum_{t \in C_u \cap C_v} \sum_{j=1}^{N} w_{t,j} + \delta}{\sum_{t \in C_u \cup C_v} \sum_{j=1}^{N} w_{t,j} + \delta}$
③ Consensus clustering via ① & ②:
④ Final voting:
Labels
1 2 1 1 3 2 1 1 2

Fig. 1. Diagram of the framework of our proposed scEFSC algorithm. It consists of data pre-processing and three important phases. In Phase A, a non-negative kernel autoencoder to pre-select a portion of genes to remove genes that were insignificant and multiple unsupervised feature selections are proposed to remove genes that do not contribute significantly to the analysis of the scRNA-seq data. In Phase B, several different scRNA-seq clustering algorithms are employed to cluster the feature subsets obtained by the multiple feature selection models. In Phase C, a weighted-ensemble clustering method called wMetaC was used to obtain the final clustering result of the individual cluster labels.

low-dimensional space, while the decoder tries to rebuild the original input from the reduced data. After that, the encoder weights are imposed as non-negative so that each latent variable is an additive representation of the original features. Then, the non-negative factors of less important features are reduced to zero. Based on the calculated weights, the method retains only those genes with large weight variances that can be considered as important features to characterize the original data. Following the reference [22], we also pre-selected 5000 genes for subsequent algorithm evaluation and analysis.

2.3. Unsupervised Feature Selection Algorithms

After data pre-processing, we employed four unsupervised feature selection algorithms, Low Variance [23], Laplacian Score [24], SPEC [25] and MCFS [26] to further select genes with relevant information, and then obtained four different feature subsets. The detailed section is summarized in the Supplementary Section S1.

Low Variance is a simple and effective unsupervised algorithm for choosing features, consisting of a typical filtering method that evaluates features by their variance which can be formularized as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}, \tag{3}$$

where x_i is the value of the i-th sample, μ is the mean of samples, and n is the number of samples.

Laplacian Score is a filter method with Laplacian Eigenmaps and Locality Preserving Projection. The Laplacian Score first builds a graph and connects the adjacent points. The incidence matrix S can be defined as follows:

$$S_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{t}}, & \text{if nodes } i \text{ and } j \text{ are connected}, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where t is a constant. Then, it establishes a diagonal matrix D on the basis of S, where the diagonal element D_{ii} of D is the sum of the elements in the i-th row of S, generates a Laplacian matrix $L = D - S$ according to S and D. The Laplacian Score of the i-th feature is calculated as follows:

$$L_i = \frac{\hat{f}_i^T L \hat{f}_i}{\hat{f}_i^T D \hat{f}_i} \tag{5}$$

$$\hat{f}_i = f_i - \frac{f_i^T D1}{1^T D1} 1, \tag{6}$$

where f_i is the vector formed by the i-th feature of the data and 1 is the vector formed by the number 1.

SPEC is a feature selection framework based on spectral graph theory, which unifies the algorithmic selection of the supervised and unsupervised features. SPEC uses the RBF kernel function, which is a popular similarity measure to construct similarity matrix S, which can be calculated as follows:

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}. \tag{7}$$

SPEC, induces the undirected graph G from the data and obtains its adjacency matrix $W = S$. The degree matrx D of the graph G, is established by W. D is a diagonal matrix and its diagonal element D_{ii} is the sum of the elements in the i-th row of W. The Laplace matrix L is defined as $L = D - W$. The normalized Laplacian matrix $\widehat{L}$ is formulated as $\widehat{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. The relevance of the features is evaluated by using the spectrum of the graph in SPEC. The scoring functions is calculated as follows:

$$\hat{\varphi}_1(F_i) = \hat{f}_i^T \gamma\left(\widehat{L}\right) \hat{f}_i \tag{8}$$

$$\hat{\varphi}_2(F_i) = \frac{\hat{f}_i^T \gamma\left(\widehat{L}\right) \hat{f}_i}{1 - \hat{f}_i^T \xi_0} \tag{9}$$

$$\hat{f}_i = \frac{D^{\frac{1}{2}} f_i}{\|D^{\frac{1}{2}} f_i\|}, \tag{10}$$

where F_i is the i-th feature of data, f_i is the vector formed by the i-th feature of data, $\gamma\left(\widehat{L}\right)$ is the Fourier transform of $\widehat{L}$, and $\xi_0 = D^{\frac{1}{2}} e$.

MCFS is based on manifold learning and L1-regularized models for feature selection so that the multi-cluster structure of the data can be well preserved. In MCFS, the representation of the data in the embedding space is obtained by solving the generalized feature problem $Ly = \lambda Dy$. Let $Y = [y_1, \ldots y_k]; y_k$ be the eigenvector corresponding to the smallest eigenvalue, and K be the dimension of the embedding space. The coefficients of linear regression are obtained by addressing the following problems:

$$\min_{a_k} \|y_k - X^T a_k\|^2 + \beta |a_k| n, \tag{11}$$

where a_k is the regression coefficient vector, $|a_k|$ is the L1-norm of a_k. The importance of each feature is determined by the MCFS score. The MCFS score is obtained as follows:

$$MCFS(j) = \max_k |a_{k,j}|, \tag{12}$$

where $a_{k,j}$ is the j-th element of the vector a_k.

2.4. Implementation of multiple scRNA-seq clustering methods

After implementing the unsupervised feature selection methods, nine scRNA-seq clustering algorithms, including SC3[14], CIDR[6], monocle[7], pcaReduce[8], Rphenograph[9], Seurat[10], SHARP[15], SINCERA[11] and RaceID[12], were employed for clustering under those four different feature subsets generated by the above four unsupervised feature selection algorithms. We describe this briefly below, with detailed sections in the Supplementary Section S2.

Table 1 summarizes these single-cell clustering algorithms. For SC3, we used the version 1.18.0 from Bioconductor and employed the number of cell types as the number of clusters k in our proposed model. For CIDR, we used the version 0.1.5 from GitHub (github.com/VCCRI/CIDR). The Monocle R package was the version 2.18.0 from GitHub (github.com/cole-trapnell-lab/monocle-release). For pcaReduce, we used the version 1.0 from GitHub (github.com/JustinaZ/pcaReduce) and performed merging based on largest probability. The Rphenograph R package was the version 0.99.1 from GitHub (github.com/JinmiaoChenLab/Rphenograph). For Seurat, we used the version 4.0.0 from CRAN and the dimension of reduction to use as input was 10. For SHARP, we used the version 1.1.0 from GitHub (github.com/shibiaowan/SHARP). The SINCERA R package was the version 0.99.0 from GitHub (github.com/xu-lab/SINCERA) and the agglomeration method to be used was the group average method in the hierarchical clustering. For RaceID, we used the version 0.2.2 from CRAN and the maximum number of clusters for the derivation of the cluster number by the saturation of mean within-cluster-dispersion was 20.

2.5. scEFSC: Ensemble Consensus Clustering Based on Multiple Feature Selections

After data preprocessing and non-negative kernel autoencoder, multiple unsupervised feature selection was proposed to select different feature subsets to remove genes that do not contribute significantly to the scRNA-seq data. On this basis, the data generated by the four unsupervised feature selections were clustered by exe-

Table 1
Summary of the compared single-cell clustering algorithms on the scRNA-seq data.

Algorithm	Type	Version	Published
SC3	k-means	1.18.0	*Nature methods* [14]
CIDR	hierarchical	0.1.5	*Genome biology* [6]
monocle	density peaks clustering	2.18.0	*Nature methods* [7]
pcaReduce	k-means + hierarchical	1.0	*BMC bioinformatics* [8]
Rphenograph	graph-based	0.99.1	*BMC bioinformatics* [9]
Seurat	graph-based	4.0.0	*Nature biotechnology* [10]
SHARP	hierarchical	1.1.0	*Genome research* [15]
SINCERA	hierarchical	0.99.0	*PLoS computational biology* [11]
RaceID	k-means	0.2.2	*Nature* [12]

cuting nine single-cell RNA-seq clustering algorithms separately to finally obtain the underlying set of clustering results.

2.5.1. Ensemble consensus clustering

After obtaining the basic clustering results and removing labels with least diversity, we proposed employing weighted-based meta-clustering (wMetaC) [15] to combine the clustering results from the multiple individual clustering methods. In contrast to traditional cluster-based similarity partitioning algorithms that give equal importance to each instance and each cluster, wMetaC assigns different weights to different instances or pairs of instances and different clusters to enhance the clustering capabilities. With wMetaC, the individual clustering results were transformed into a clustering similarity matrix S whose elements $s_{i,j}$ represented the similarity between the ith and jth cells. Since the weight of each cell pair is determined by the degree of agreement of the colocation clustering results of the two cells, the similarity matrix S is transformed into a weight matrix W as follows:

$$w_{i,j} = s_{i,j}(1 - s_{i,j}), \tag{13}$$

where $w_{i,j}$ is the element of the ith row and jth column in W. When $s_{i,j} = 1$ or $s_{i,j} = 0, w_{i,j}$ reaches a minimum value of 0; when $s_{i,j} = 0.5, w_{i,j}$ reaches a maximum value of 0.25. Zero weights were assigned to cell pairs that were easiest to cluster, while the highest weights were assigned to the most difficult pairs to cluster. The weight associated with each cell was then calculated as the cumulative sum of all the intercellular weights associated with the corresponding cell.

Then, the weighted inter-cluster similarity was calculated by dividing the sum of the weights of their overlapping elements by their combined weights. Given two clusters C_u and C_v, the inter-cluster similarity in wMetaC was calculated as:

$$S = \frac{\sum\limits_{t \in C_u \cap C_v} \sum\limits_{j=1}^{N} w_{t,j} + \delta}{\sum\limits_{t \in C_u \cup C_v} \sum\limits_{j=1}^{N} w_{t,j} + \delta}, \tag{14}$$

where $w_{t,j}$ is the colocation weight of the tth cell and the jth cell obtained above, N is the number of cells, and δ is a very small positive number, with $\delta = 0.01$ used to avoid a zero denominator. Since $\sum_{j=1}^{N} w_{t,j}$ summarizes all possible colocation weights between the tth cell and all cells, it is set to the overall colocation weight of the tth cell. The numerator of the inter-cluster similarity formula represents the sum of the colocation weights of the cells that appear in both the *Cu* and *Cv* groups, while the denominator represents the sum of the colocation weights of the cells that appear in either the *Cu* or *Cv* group.

With this, the similarity matrix obtained was clustered using the hierarchical clustering method with the ”ward.D”. The correspondence between the clusters was obtained after clustering. Finally a voting scheme [15] was applied to the clustering results obtained in the previous step and the individual cells were assigned to the clusters to which they fit in the highest proportion, resulting in the final clustering result.

2.5.2. Removing labels with least diversity

Since the diversity of clustering results is beneficial to enhance the performance of the ensemble solution [28], we computed pairwise ARI between clustering labels to quantify the similarity between the results of individual clustering methods to evaluate the diversity. After that, a similarity matrix was constructed by calculating the pairwise similarity between all individual clustering labels, including the clustering labels with a self-similarity value of 1. Then the variance of the similarity vector was evaluated for each clustering labels. Since high pairwise similarity has a high ARI value and a self-similarity value of 1, the similarity vector with the most similar cluster label to the other cluster labels has the smallest variance. Because the clustering labels with the lowest variance contribute least in terms of diversity, we removed clustering labels obtained from the few clustering methods with the lowest variance to obtain the set of clustering labels.

2.6. Evaluation metrics

Comparing the agreement of cell clustering results with the previously published labels is a common method for evaluating the performance of clustering methods for scRNA-seq data analysis. In this study, we used two evaluation metrics: Normalised Mutual Information (NMI) and Adjusted Rand Index (ARI), as summarized in Supplementary Section S3.

3. Results and Discussion

3.1. Data source

We collected fourteen publicly available scRNA-seq datasets containing cell type annotations and gene expression values from various scRNA-seq platforms, which can be downloaded from the Gene Expression Omnibus and Broad Institute Single Cell Portal. All the datasets are from different species, including mouse and human, as well as from different organs, such as brain, lung and kidney. The detailed information on the datasets are summarized in Table 2. Of note, we removed cells labeled ambiguously as ”abandoned” and then performed a logarithmic transformation (base 2) to rescale the data if its gene expression values were higher than 100. Indeed, for the 14 scRNA-seq datasets in Table 2, we can see that the number of cells varies from 90 to 13316, the number of genes from 19020 to 55186 and the number of clusters from 4 to 14. To further remove redundant and irrelevant genes and reduce the computational cost of the algorithm, we used a non-negative kernel autoencoder [22] to pre-select 5000 genes for downstream analysis.

3.2. scEFSC is the most accurate among computational methods on the scRNA-seq data

To investigate the advantages of clustering using scEFSC, we benchmarked our proposed scEFSC method alongside the nine scRNA-seq clustering methods on the fourteen published scRNA-seq datasets (Table 2). The clustering results measured by NMI and ARI are summarized in Fig. 2.

The results of the NMI evaluation, are summarized in Fig. 2a. We observe the following: it appears that of the ten single-cell clustering algorithms, scEFSC provided the best clustering results on eleven of the fourteen datasets, and ranked second, slightly below SC3 on the E-MTAB-5061 dataset, and third and fourth on the GSE36552 and GSE84133 datasets; CIDR had the worst performance. Compared to Rphenograph and Seurat that based on the community detection method Louvain, scEFSC outperformed both. In addition, compared to the consensus clustering algorithms, SC3 and SHARP, scEFSC was superior to both indicating that multiple clustering is more effective for base clusters in consensus clustering.

In terms of ARI evaluation, as shown in Fig. 2b, we observe broadly similar results to those using the NMI. Our proposed scEFSC performed better than the other nine clustering algorithms, while CIDR and SHARP simultaneously performed the worst. scEFSC clustered best on ten datasets, being second only to SC3 on the GSE71585 and E-MTAB-5061 datasets, second only to SIN-

Table 2
Summary of the 14 real scRNA-seq datasets.

Source	Organism	cell	gene	class	platform	Ref
GSE36552	Human embryo	90	20214	6	Tang	*Nature structural & molecular biology* [29]
GSE83139	Human pancreas	457	19950	7	SMARTer	*Diabetes* [30]
GSE81861	Human tissues	561	55186	9	SMARTer	*Nature genetics* [31]
GSE59739	Mouse brain	622	25334	4	STRT-Seq	*Nature neuroscience* [32]
GSE81252	Human liver	777	19020	7	SMARTer	*Nature* [33]
GSE81608	Human pancreas	1600	39851	8	SMARTer	*Cell metabolism* [34]
GSE71585	Mouse brain	1679	24150	18	SMARTer	*Nature neuroscience* [35]
GSE85241	Human pancreas	2126	19140	10	CEL-Seq2	*Cell systems* [36]
E-MTAB-5061	Human pancreas	2209	25525	14	Smart-Seq2	*Cell metabolism* [37]
GSE65525	Mouse embryo	2717	24175	4	inDrop	*Cell* [38]
GSE60361	Mouse brain	3005	19972	9	STRT-Seq	*Science* [39]
phs000833.v3.p1	Human brain	3042	25123	16	Fluidigm C1	*Science* [40]
GSE84133	Human pancreas	8569	20125	14	inDrop	*Cell systems* [41]
SCP345	Human pancreas	13316	21813	8	10X Genomics	*The reference* [42]

CERA on the SCP345 dataset and fourth on the GSE84133 dataset. Although CIDR showed a higher performance than most of the other algorithms on the GSE83139 and GSE81861 datasets, it performed particularly poorly on GSE85241, E-MTAB-5061, phs000833.v3.p1, GSE84133 datasets. SHARP performed the worst on the GSE81861 and GSE71585 datasets, and did not perform well on the remaining datasets. SC3 achieved better performance than the others on most scRNA-seq datasets, while scEFSC outperformed SC3 on most datasets. scEFSC was better than community detection method Louvain algorithms, Rphenograph and Seurat. The higher ARI values of scEFSC clustering results over consensus clustering algorithms, SC3 and SHARP, suggest that multiple clustering is more effective in base clustering for consensus clustering.

In summary, our proposed scEFSC algorithm performed well and demonstrated superior clustering performance on fourteen scRNA-seq datasets, improving the clustering performance on the scRNA-seq data. The results also reflect that the combination of multiple clustering models can enhance the performance of consensus clustering.

3.3. scEFSC outperforms other ensemble clustering algorithms

We compared scEFSC, SAME[19], scCCESS[43] and RSEC[44] on the fourteen published scRNA-seq datasets as in the previous section. SAME, scCCESS and RSEC are three ensemble clustering algorithms for single-cell RNA-seq data. The clustering results measured by NMI and ARI are summarized in Fig. 3.

Fig. 3a summarizes the clustering results evaluated by NMI. Compared with SAME and RSEC, our proposed scEFSC had the best NMI values on all 14 datasets. Of the fourteen datasets, our proposed scEFSC had better NMI values on twelve datasets compared with scCCESS, they provided the similar results on the GSE84133 dataset, and on the GSE36552 datasets scCCESS had better NMI values than our proposed scEFSC. Our proposed scEFSC performed the best among the four algorithms, and RSEC performed the worst. On the GSE59739, GSE81608, and SCP345 datasets, our proposed scEFSC was superior to the other three algorithms by more than 0.1.

Fig. 3b summarizes the clustering results evaluated by ARI. The results using ARI are roughly similar to the results using NMI. Compared with SAME and RSEC, our proposed scEFSC had the best ARI values on all 14 datasets. Of the fourteen datasets, our proposed scEFSC had better ARI values on twelve datasets compared with scCCESS, both algorithms had the same result on the GSE84133 dataset, and on the GSE36552 datasets scCCESS had better ARI values than our proposed scEFSC. Our proposed scEFSC performed the best among the four algorithms, and RSEC performed the worst. On the GSE59739, GSE81608, GSE85241, GSE65525, GSE60361, phs000833.v3.p1, and SCP345 datasets, our proposed scEFSC was superior to the other three algorithms by more than 0.1.

Overall, the clustering results showed that our proposed scEFSC outperformed SAME, scCCESS and RSEC on most scRNA-seq datasets, and our proposed scEFSC exhibited excellent clustering performance in comparison with other ensemble clustering algorithms.

3.4. The performance of scEFSC outperforms the scDHA algorithm recently reported in Nature Communications

We compared scEFSC and scDHA [22] on the fourteen published scRNA-seq datasets as in the previous section. scDHA is a scRNA-seq clustering method, which uses a hierarchical autoencoder published in Nature Communications. The scDHA algorithm framework consists of two main components. The first component is a non-negative kernel autoencoder that aims to filter genes that contribute little to the data representation. The second part is a stacked Bayesian autoencoder for data dimensionality reduction. Fig. 4 depicts the clustering results of scEFSC and scDHA as measured by NMI and ARI.

The clustering results of scEFSC and scDHA assessed by NMI are summarized in Fig. 4a. Of the fourteen datasets, our proposed scEFSC had better NMI values on ten datasets compared with scDHA, both algorithms had the same result on the GSE71585 dataset, and on the GSE36552, E-MTAB-5061 and GSE65525 datasets scDHA had better NMI values than our proposed scEFSC. Our proposed scEFSC significantly outperformed scDHA on the GSE59739 dataset, with an improvement in clustering performance of 0.11. Our proposed scEFSC was superior to scDHA by more than or equal to 0.03 on the GSE81252, GSE81608, GSE85241, GSE60361 and GSE84133 datasets. On the E-MTAB-5061 and GSE65525 datasets, our proposed scEFSC was worse than scDHA.

Fig. 4b summarizes the clustering results for scEFSC and scDHA assessed by ARI. For the fourteen datasets our proposed scEFSC had better ARI values than scDHA on ten datasets, scEFSC and scDHA gave the similar ARI value on the GSE71585 dataset, and on the GSE36552, E-MTAB-5061 and GSE65525 datasets scDHA had better ARI values than scEFSC. scEFSC significantly outperformed scDHA on the GSE81252, GSE81608, GSE85241, GSE60361 and phs000833.v3.p1 datasets, with a clustering performance improvement of no less than 0.10. Meanwhile, on GSE83139, GSE59739, GSE84133 and SCP345 datasets, our proposed scEFSC outperformed scDHA by more than or equal to 0.6. On the dataset GSE65525, scDHA outperformed our proposed method.

Overall, the clustering results showed that our proposed scEFSC outperformed scDHA on most of the scRNA-seq datasets and was

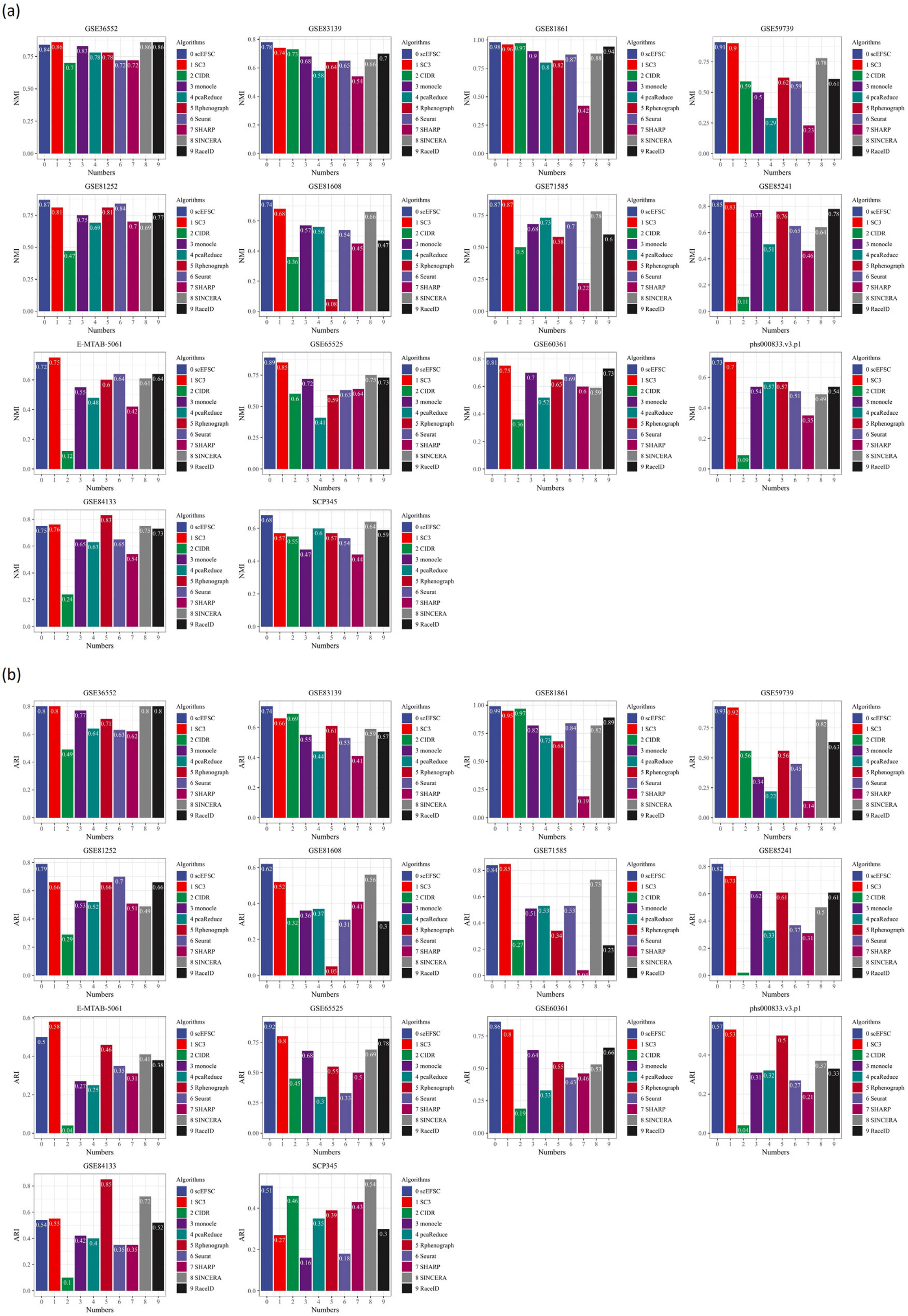

Fig. 2. Clustering performance comparison of scEFSC, SC3, CIDR, monocle, pcaReduce, Rphenograph, Seurat, SHARP, SINCERA and RaceID on the fourteen published scRNA-seq datasets. (a) Clustering performance evaluated by NMI. (b) Clustering performance evaluated by ARI.

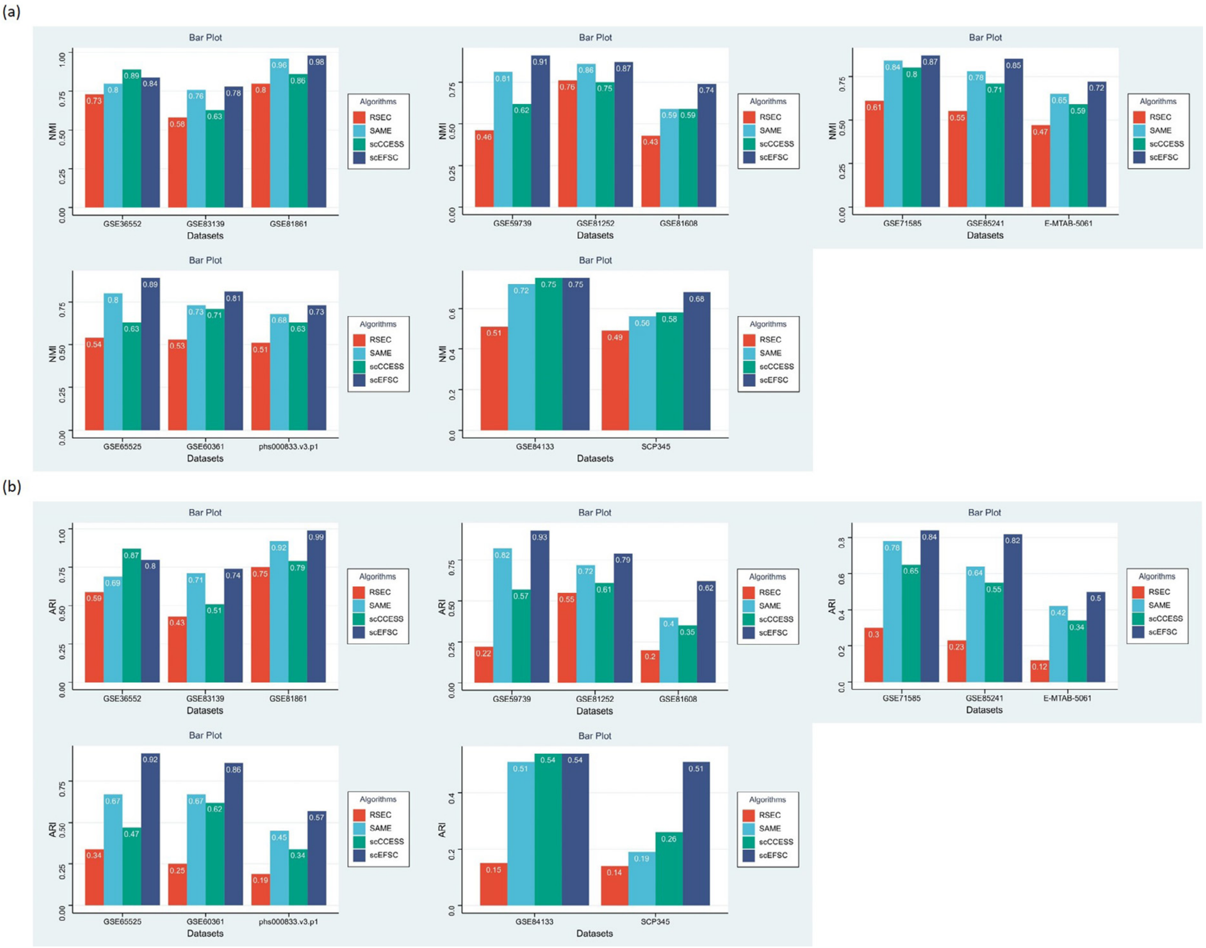

Fig. 3. Comparison of the clustering performance of scEFSC, SAME, scCCESS and RSEC on the fourteen published scRNA-seq datasets. (a) Clustering performance evaluated by NMI. (b) Clustering performance evaluated by ARI.

inferior to scDHA on only a few scRNA-seq datasets. Our proposed scEFSC demonstrated an excellent clustering performance therefore, compared to scDHA.

3.5. Integration of multiple feature selection methods can improve downstream functional analysis

We compared our proposed scEFSC with the scEFSC algorithms with one of the four feature selections on the fourteen scRNA-seq datasets shown in Table 2 using the clustering evaluation metrics NMI and ARI to evaluate the clustering results and summarize the results in Fig. 5.

With the NMI evaluation, the full scEFSC outperforms the other four single-feature selection algorithms on six scRNA-seq datasets including GSE59739, GSE81608, GSE85241, GSE65525, GSE60361, and phs000833.v3.p1. On GSE83139, GSE81861, GSE81252 and GSE71585 datasets, our proposed full scEFSC performed equally well as some of the other four algorithms. On GSE36552 and E-MTAB-5061 datasets, scEFSC performed next to scEFSC containing only SPEC. On the GSE84133 dataset, performance of our proposed scEFSC was second to that of scEFSC containing only MCFS. Finally, on the SCP345 dataset, the performance of full scEFSC is second to that of scEFSC with only Laplacian Score. Overall, our proposed scEFSC performed best, and scEFSC including SPEC was the second best, scEFSC having only Laplacian Score or only MCFS had the worst performances.

With ARI evaluation, our proposed scEFSC similarly outperformed the other four algorithms on six datasets including GSE83139, GSE81861, GSE81608, GSE65525, GSE60361, and phs000833.v3.p1. On the GSE36552 and GSE81252 datasets, full scEFSC performed just as well as some of the other four algorithms. On GSE85241, the performance of our proposed scEFSC was only lower than scEFSC having only Low Variance. The performance of full scEFSC on the GSE59739, GSE71585, E-MTAB-5061, GSE84133 and SCP345 datasets, however was not good. In summary, our proposed scEFSC had the best performance, scEFSC with only Low Variance was second best and scEFSC with only MCFS the worst.

Altogether, our proposed scEFSC algorithm exhibited superior clustering performance on the fourteen scRNA-seq datasets, indicating that using multiple unsupervised feature selection algorithms can strengthen the clustering ability of consensus clustering over a single unsupervised feature selection algorithm.

3.6. The clustering performance of scEFSC is affected by different parameters

In terms of parameters, our proposed sEFSC default selects 5000 genes by a non-negative kernel autoencoder and 2000 genes by unsupervised feature selections, using nine clustering algorithms. To verify the superiority of the default parameters, we conducted two sets of experiments on the fourteen scRNA-seq datasets testing these parameters. The first experiment compared the effect

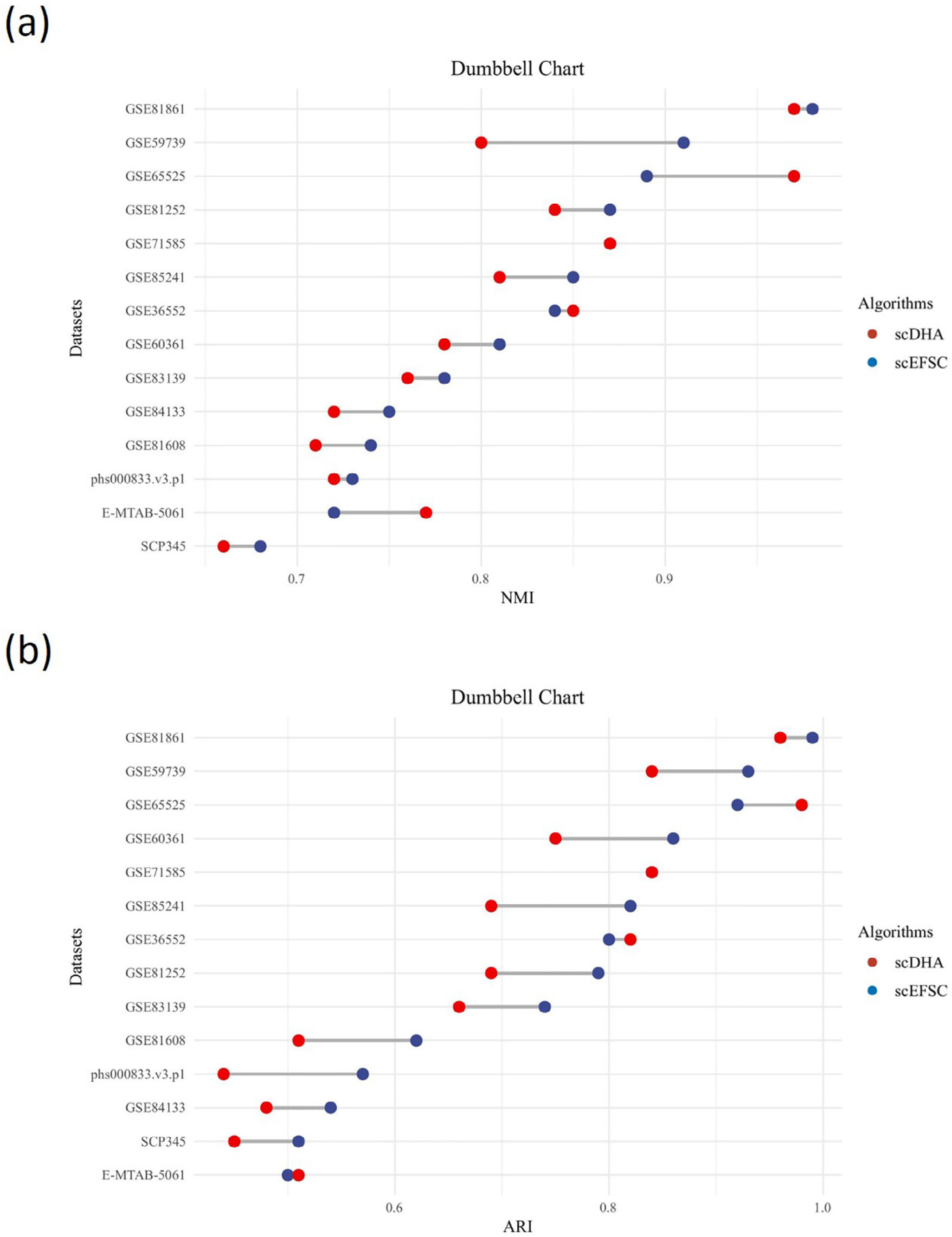

Fig. 4. Comparison of the clustering performance of scEFSC and scDHA on the fourteen published scRNA-seq datasets. (a) Clustering performance evaluated by NMI. (b) Clustering performance evaluated by ARI.

of the number of features selected on scEFSC. The second experiment compared the effect of the number of clustering algorithms in the multiple clustering on scEFSC.

For the first parametric experiment, we compared scEFSC with the proposed default parameters to scEFSC with 10,000 or 5,000 genes selected, and scEFSC with 5,000 or 3,000 genes selected, respectively and the clustering results assessed by NMI and ARI are summarized in Fig. 6a and 6b. NMI and ARI values were broadly similar. scEFSC with default parameters outperformed scEFSC with two feature selections of 10,000 and 5,000 genes on thirteen datasets and was the same on the GSE60361 dataset. Therefore, our proposed scEFSC using a non-negative encoder for 5,000 genes feature selection was effective. In addition, scEFSC with default parameters performed better on eight datasets, the same on three datasets and worse on three datasets, compared to the scEFSC with two feature selections of 5,000 and 3,000 genes. This shows that our proposed scEFSC was effective in selecting 2000 genes using unsupervised feature selection. In conclusion, our proposed scEFSC default setting of 5000 and 2000 genes for feature selection effectively enhanced the clustering performance.

In the second parametric experiment, on the fourteen datasets, we compared scEFSC with our proposed default parameters to scEFSC with eight clustering algorithms and scEFSC with six clustering algorithms. Fig. 6c and 6d summarize the clustering results,

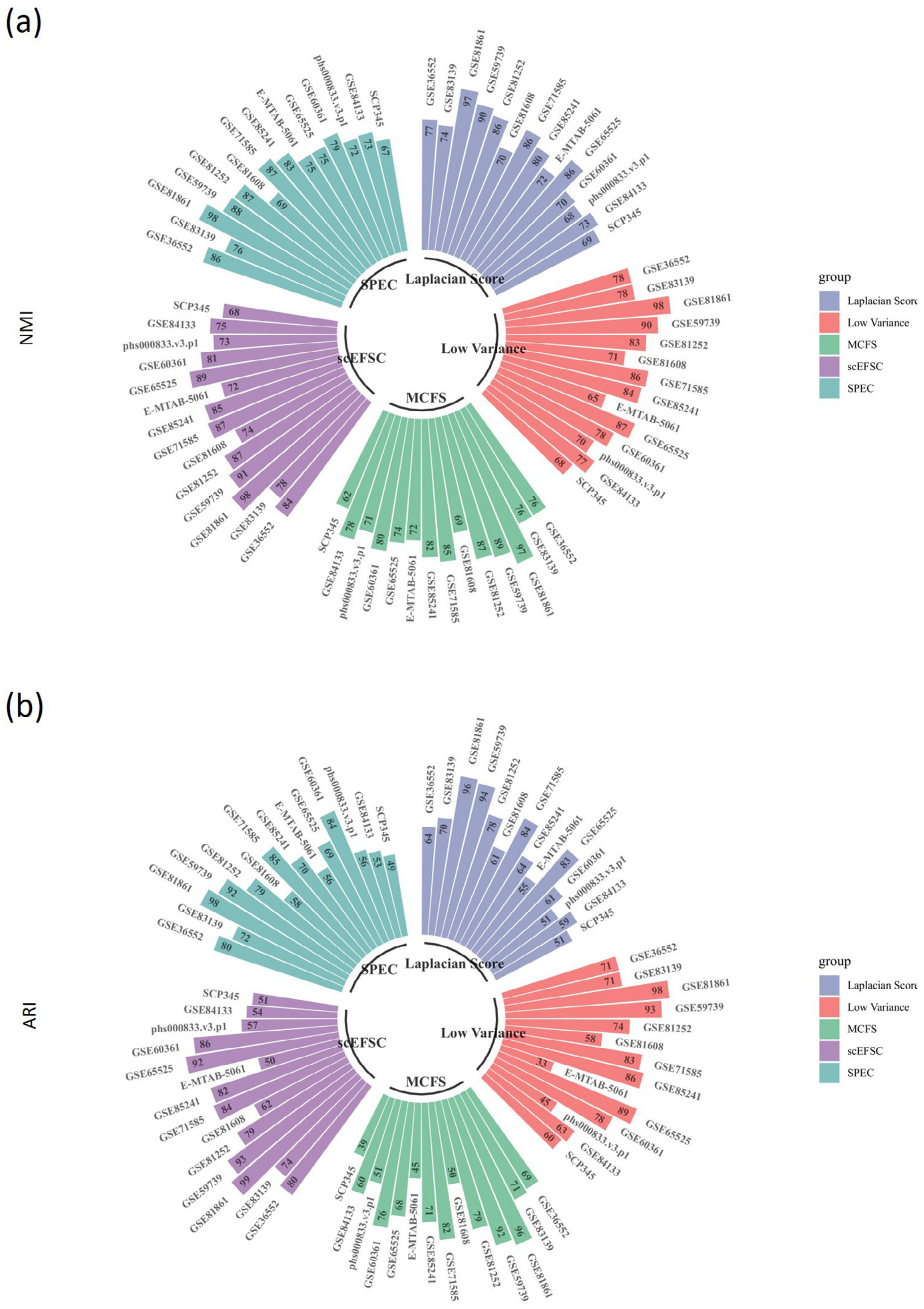

Fig. 5. Performance comparisons of the different feature selections. NMI and ARI values were multiplied by 100 for comparison purposes. (a) Clustering performance evaluated by NMI evaluation metric. (b) Clustering performance evaluated by ARI evaluation metric.

showing the largely consistent results evaluated by NMI and ARI. In the previous comparison of our proposed scEFSC and nine single-cell clustering algorithms, SC3 was second to scEFSC in terms of average clustering results across the fourteen datasets and was better than the other nine single-cell clustering algorithms, while CIDR, SHARP and pcaReduce were the three worst algorithms. The scEFSC using eight clustering algorithms eliminated the SC3 algorithm and the scEFSC using six clustering algorithms excluded CIDR, SHARP and pcaReduce. Our proposed scEFSC clustered better on ten datasets, equally well on the

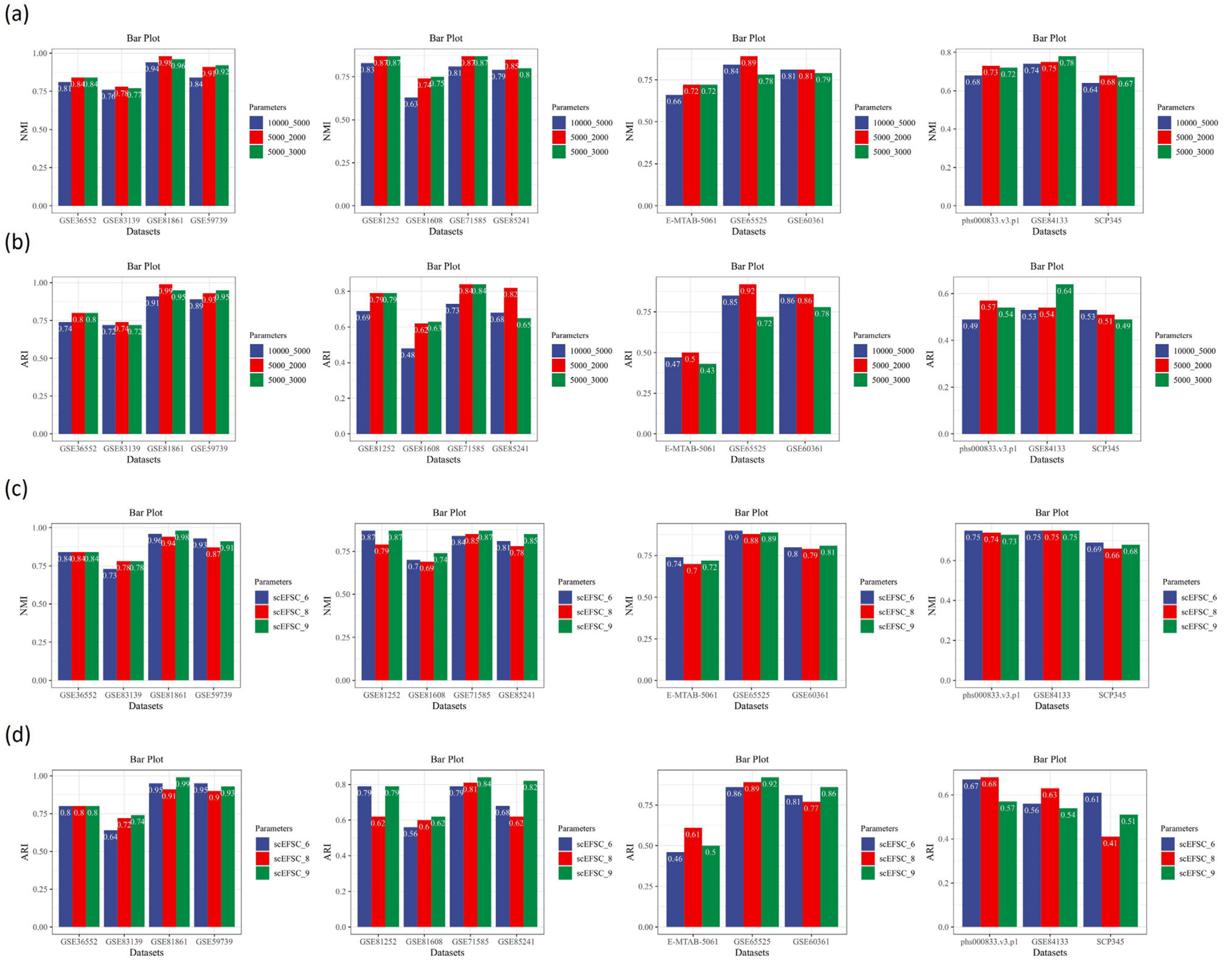

Fig. 6. Comparison of the clustering performance with different parameters of scEFSC on the fourteen published scRNA-seq datasets. (a) Comparison of the clustering performance of scEFSC with two default feature selections of 5,000 and 2,000 genes to scEFSC with two feature selections of 10,000 and 5,000 genes and scEFSC with two feature selections of 5,000 and 3,000 genes. Clustering performance evaluated by NMI. (b) The same as a. with clustering performance evaluated by ARI. (c) Comparison of the clustering performance of scEFSC with default multiple clustering using nine clustering algorithms to scEFSC with multiple clustering using eight clustering algorithms and scEFSC with multiple clustering using six clustering algorithms. Clustering performance evaluated by NMI. (d) The same as c. with clustering performance evaluated by ARI.

GSE36552 dataset and worse on three datasets compared to scEFSC using eight clustering algorithms, indicating that SC3 plays an important role in consensus clustering. Furthermore, scEFSC clustered better on eight datasets, the same on the GSE36552 and GSE81252 datasets and worse on four datasets compared to scEFSC using six clustering algorithms, suggesting that CIDR, SHARP and pcaReduce, the three clustering algorithms with the worst average clustering results, also play a positive role in the final clustering result for consensus clustering.

In summary, the selection of 5000 genes with a non-negative kernel autoencoder and 2000 genes with unsupervised feature selections and applying nine clustering algorithms in the multiple clustering are meaningful in enhancing the clustering advantages of the proposed scEFSC algorithm.

3.7. The running times of scEFSC with other single-cell clustering algorithms

We compared the running time of scEFSC with other single-cell clustering algorithms, including SC3, CIDR, monocle, pcaReduce, Rphenograph, Seurat, SHARP, SINCERA, and RaceID. All those algorithms were run on Ubuntu 20.04.3 LTS (GNU/Linux 5.4.0–90-generic x86_64). The running time comparisons are summarized in Supplementary Table S3 for fourteen scRNA-seq data. Since scEFSC performs multiple feature selection and multiple clustering, scEFSC has the longest running time compared to other single-cell clustering algorithms.

3.8. Evaluations on a large-scale scRNA-seq dataset GSE109774

Due to the development of scRNA-seq technology, an increasing number of cells are sequenced. Therefore, to demonstrate the scalability of our proposed scEFSC on large datasets, we conducted experiments on the GSE109774 dataset. The GSE109774 dataset is from Mouse tissues with 54,439 cells and 40 celltypes. Since computational resources were limited, we downsampled the large-scale GSE109774 dataset with 16,332 cells (about 30% of the cells). Then we compared scEFSC,SC3, CIDR, monocle, pcaReduce, Rphenograph, Seurat, SHARP, SINCERA, RaceID,scDHA,SAME, scCCESS and RSEC on GSE109774. Fig. 7 summarizes the performance comparisons under two evaluation metrics NMI and ARI. From the figure, it can be seen that our proposed scEFSC provided better performance than most other algorithms on the large dataset. Compared to SC3, CIDR, monocle, pcaReduce, Rphenograph, Seurat, SHARP, SINCERA and RaceID, we found that the clustering performance of scEFSC was worse than Rphenograph and SINCERA,

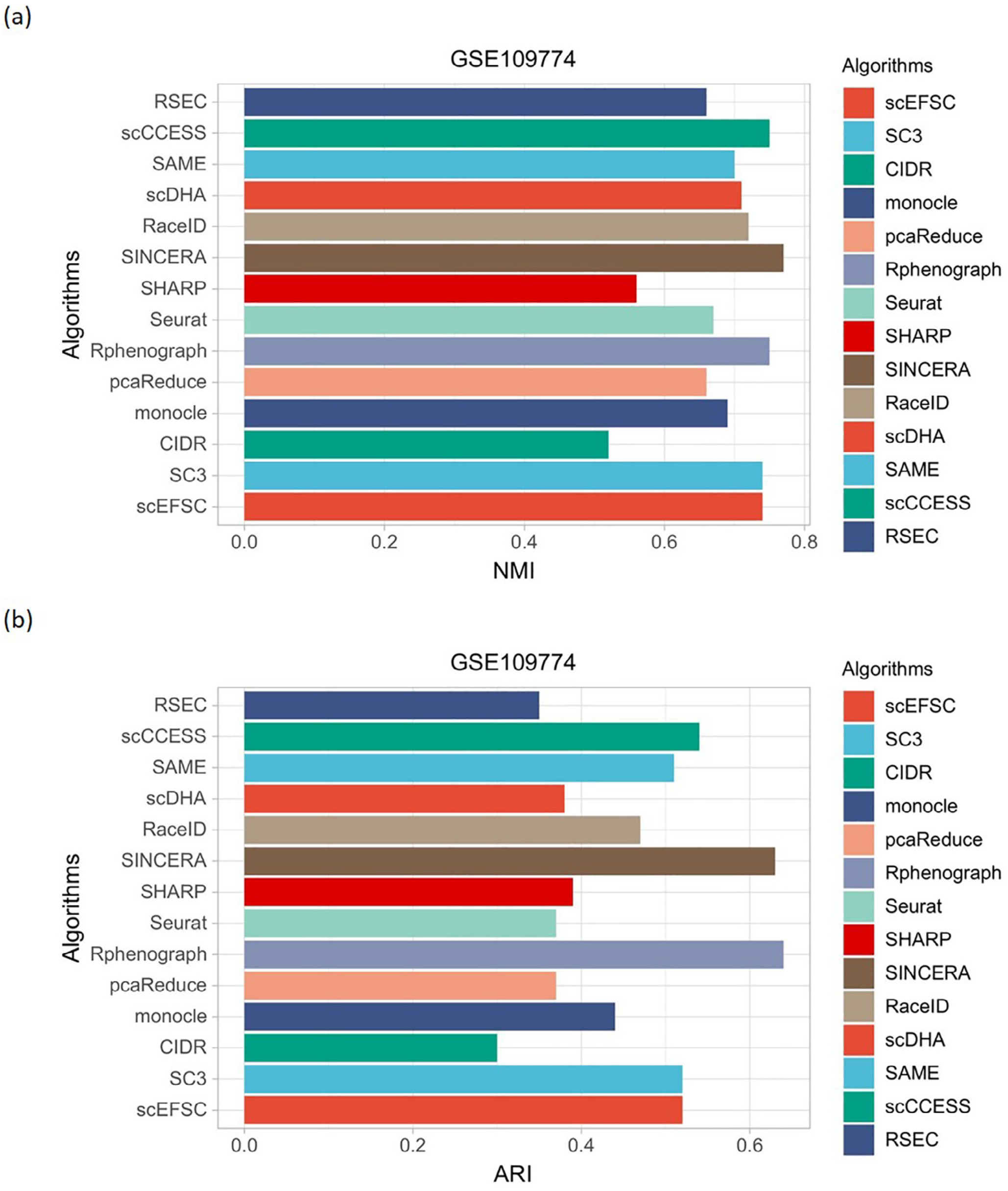

Fig. 7. Comparison of the clustering performance of scEFSC,SC3, CIDR, monocle, pcaReduce, Rphenograph, Seurat, SHARP, SINCERA, RaceID,scDHA,SAME, scCCESS and RSEC on GSE109774. (a) Clustering performance evaluated by NMI. (b) Clustering performance evaluated by ARI.

the same as SC3, and better than the other algorithms. SINCERA and Rphenograph had the best clustering performance, CIDR had the worst clustering performance, and SHARP had only better clustering performance than CIDR. Compared to scDHA, the clustering performance of scEFSC was better than scDHA. Compared to SAME, scCCESS and RSEC, the clustering performance of scEFSC was slightly worse than scCCESS, and better than SAME and RSEC. The clustering performance of RSEC was the worst among the three ensemble clustering algorithms, SAME, scCCESS and RSEC. Therefore, our proposed scEFSC is scalable for large-scale scRNA-seq data.

3.9. Functional genomic analysis

Although scEFSC shows superiority in the clustering performance analyses described above, the biological significance of its clustering results is particularly important for the understanding of the biological data. The biological analysis of differentially expressed genes, gene ontology enrichment and KEGG pathway analysis can be performed from the gene expression data and clustering labels obtained from scEFSC. Next, we elcidated the biological significance of scEFSC from the GSE81861 dataset of human colorectal tumour cells.

3.9.1. Differential expressed genes analysis

We used the gene expression matrix obtained from the GSE81861 dataset and the clustering labels from scEFSC to identify differentially expressed genes (DEGs) and marker genes for each cluster. Fig. 8 shows a volcano plot of each cluster analysed for differentially expressed genes, and the top twenty differentially expressed genes are labelled. The top twenty DEGs showed a high-fold change, indicating a greater difference in gene expression from the rest of the cluster and representing the marker genes for that cluster. In this analysis, we found the most differentially

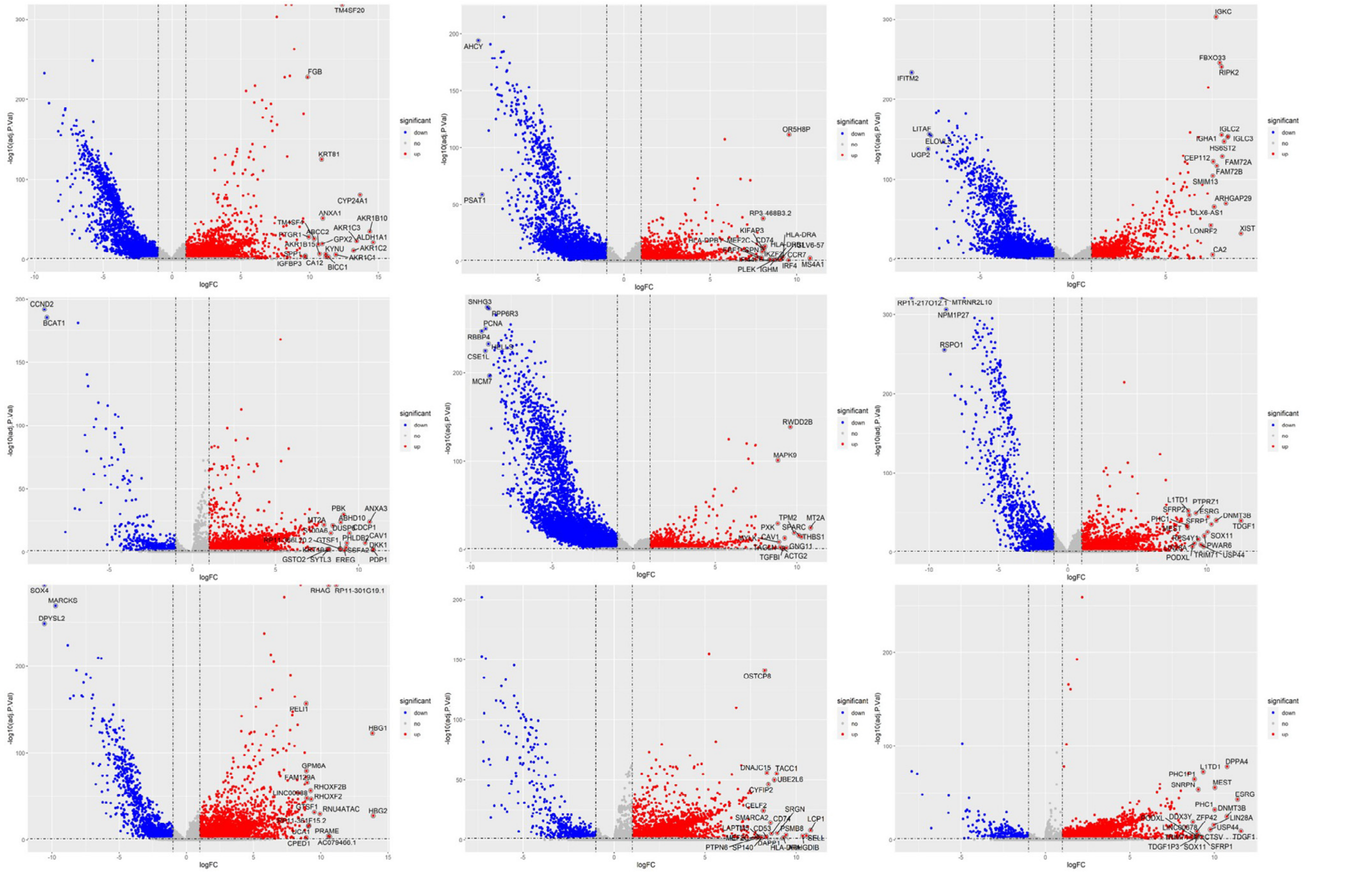

Fig. 8. Volcano plots for each identified cluster versus all other clusters, with the top twenty differentially expressed genes annotated. Selected differentially expressed genes are adjusted p-value < 0.05, fold change >1 or < -1. Up-regulated genes are in red and down-regulated genes in blue.

expressed gene was ALDH1A1, which is a member of the ALDH1 gene family. The ALDH1A1 gene is an important marker gene for human colorectal carcinoma stem cells and encodes an enzyme responsible for intracellular aldehyde oxidation, shown to have a function in the early differentiation of stem cells. Its expression profile shows potential as a universal marker for the identification and isolation of normal and cancer stem cells of multiple origins in a variety of tissue types. The remaining genes with large differences were AKR1B10, CYP24A1, and AKR1C3. AKR1B10 is a pathological diagnostic marker for tumours and is highly expressed in normal gastrointestinal epithelial tissues with low or no expression in other normal tissues, and highly significant expression in breast, liver and non-small cell lung cancers. CYP24A1 is a potential oncogene, which is aberrantly expressed in breast, lung and oesophageal cancer tissues and correlates with tumour malignancy and poor patient prognosis, based on its aberrant expression regulated by DNA methylation in some tumour types. AKR1C3 is a member of the AKR1 family and involved in synthesis of a steroid hormone closely associated with the development of many malignancies and also involved in regulating the sensitivity of many anti-tumour drugs.

3.9.2. Gene Ontology (GO) enrichment and KEGG analysis

From the differential gene expression analysis on the GSE81861 dataset, we selected the top twenty differentially expressed genes from each of the nine clusters and removed duplicate genes to obtain 162 genes. These 162 genes were queried by gene ID transformation to 153 genes for gene ontology (GO) enrichment analysis and KEGG pathway analysis to interpret the biological significance of the genes. For the GO enrichment analysis, we obtained 533 enriched GOs, of which 465 were biological processes (BP), 31 were cellular components (CC) and 37 were molecular functions (MF). Fig. 9a shows the top 20 categories of the three GO enrichments, ordered by adjusted p-value. The five most enriched GO biological processes were regulation of lymphocyte activation (GO:0051249), positive regulation of cell activation (GO:0050867), antigen receptor-mediated signaling pathway (GO:0050851), positive regulation of leukocyte activation (GO:0002696) and positive regulation of lymphocyte activation GO:0051251). We can see that most of the five enriched biological processes were related to signaling pathways and the positive regulation of cell activation, which are highly relevant to cancer.

The top five enriched GO cellular components are MHC class II protein complex (GO:0042613), blood microparticle (GO:0072562), external side of the plasma membrane (GO:0009897), MHC protein complex (GO:0042611) and integral component of lumenal side of endoplasmic reticulum membrane (GO:0071556). The MHC protein complex is closely associated with tumours [45] and there is a clear correlation between MHC-I gene composition and the genes that are mutated in the cancer of this person. The molecule MHC-II [46] may have a greater effect on neoplastic tumours than MHC-I. The remaining three enriched cellular components were all associated with the colorectum. The top five enriched GO molecular functions were alditol:NADP + 1-oxidoreductase activity (GO:0004032), alcohol dehydrogenase (NADP+) activity (GO:0008106), aldo–keto reductase (NADP) activity (GO:0004033), antigen binding (GO:0003823) and MHC class II protein complex binding (GO:0023026). Expression of MHC-II and related pathway components have been found in cancer cells from colorectal cancer. Tumor-specific MHC-II expression can increase

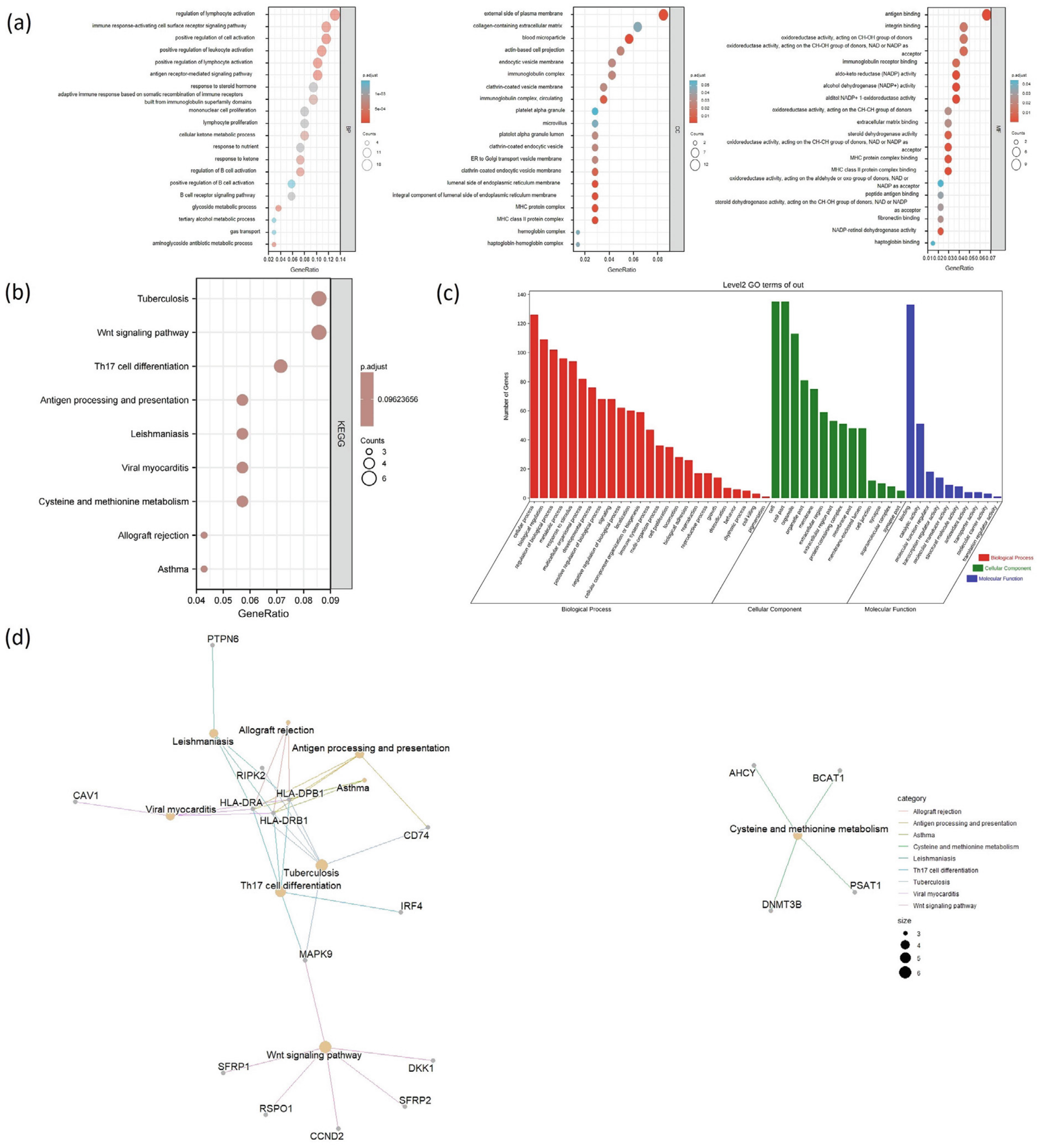

Fig. 9. (a) Top 20 classes of GO enrichment terms sorted by adjusted *p*-values. (b) Top 9 pathways of KEGG enrichment sorted by adjusted *p*-values. (c) The distribution of genes under the three GO enrichments. (d) The network diagram of the top 9 pathways and genes for the KEGG enrichment.

the recognition of tumors by the immune system and therefore could play an important role in immunotherapy. Fig. 9c depicts the distribution of genes of the three GO enrichments.

For the KEGG analysis, 9 pathways were matched and 70 genes were successfully annotated. Fig. 9b depicts the 9 KEGG pathways sorted by adjusted *p*-value. Fig. 9d depicts the network diagram of the top 9 KEGG pathways with genes. This plot shows that genes HLA-DRA, HLA-DRB1 and HLA-DPB1 were matched to most of the KEGG pathways. Human leukocyte antigens (HLA) are the expression products of the major histocompatibility complex (MHC) in humans. HLA plays a key role in the immune effect mechanism of the body against tumors, in which cellular immunity plays a leading role and is regulated by humoral immunity to synergistically kill tumor cells [47]. The top five KEGG pathways are Cysteine and methionine metabolism (hsa00270), Viral myocarditis (hsa05416), Th17 cell differentiation (hsa04659), Asthma (hsa05310) and Wnt signaling pathway (hsa04310). Fig. 10 shows Th17 cell differentiation (hsa04659) and Wnt signaling pathway (hsa04310) provided by the KEGG database. The Th17 cell differentiation belongs to the immune system of biological systems, has five related genes and is highly associated with tumors. Thl7 cells are a class of CD4 + effector T cells. Thl7 cells and their associated cytokines have been found to be presented in many tumors, taking an important role in inflammation-associated tumors [48]. The Wnt signalling pathway is the part of signal transduction in environmental information processing. This pathway has six related

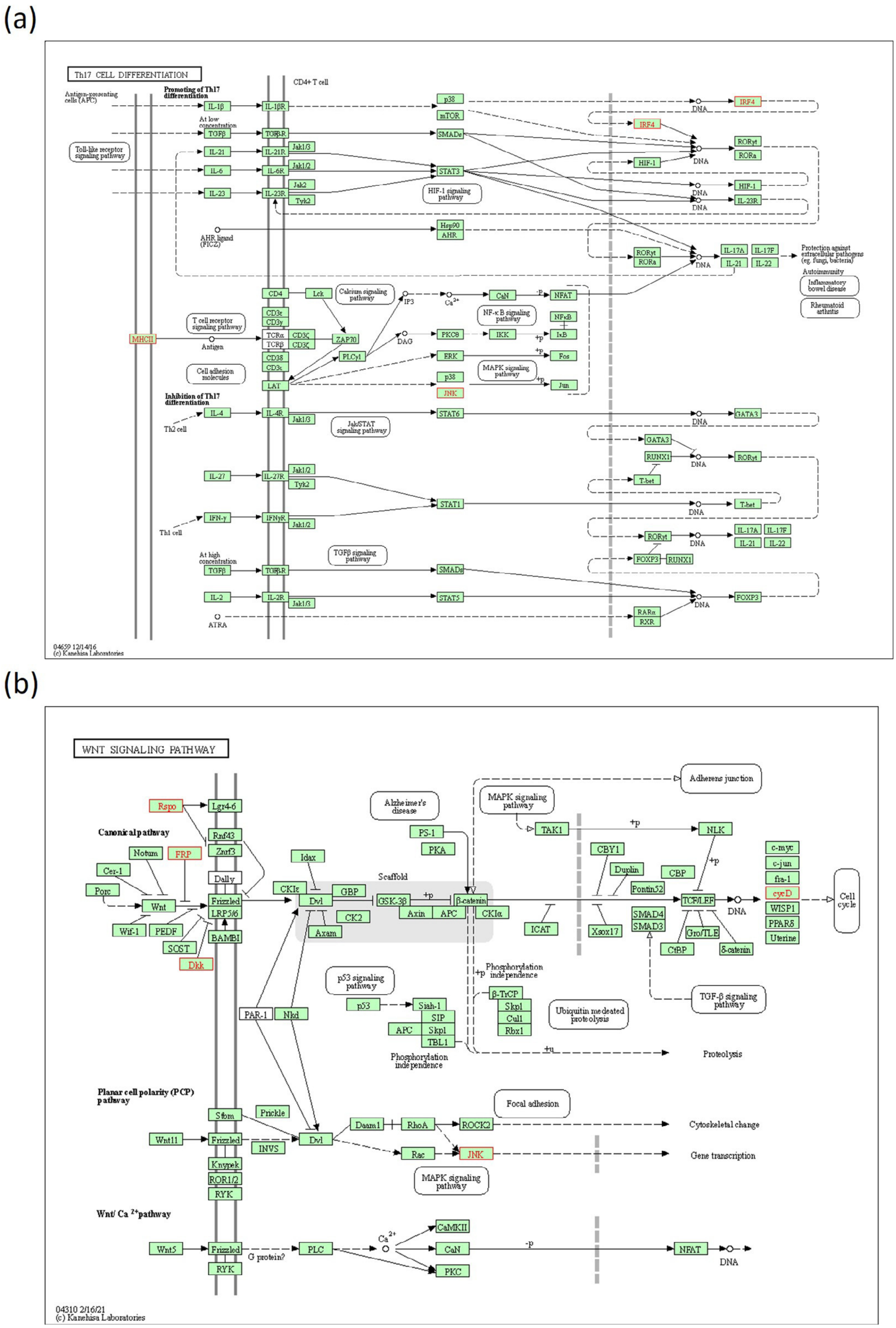

Fig. 10. (a) The KEGG database indicates Th17 cell differentiation (hsa04659). (b) The KEGG database indicates the Wnt signaling pathway (hsa04310).

genes and is closely associated with tumors. Indeed, the Wnt signalling pathway is not only associated with events related to tumor invasion and metastasis, but also plays an essential role in regulating the self-renewal, proliferation and differentiation of tumor stem cells [49].

The GO and KEGG enrichment results reflected that the clustering results of scEFSC can effectively guide the identification of cancer-related biological processes. We can conclude therefore, that the clustering results of scEFSC are biologically significant.

4. Conclusions

In this study, we propose a clustering method scEFSC based on ensemble feature selection. scEFSC is a method for clustering and analyzing the scRNA-seq data through feature selection and ensemble clustering. First, the feature selection is divided into two steps, the first step is a preliminary feature selection employing a non-negative autoencoder, and the second step uses four unsupervised feature selection algorithms to further select the results achieved by step one. For the high-dimensional sparse single-cell RNA-seq data, feature selection reduces the data dimensionality and selects effective features. Multiple clustering methods are performed on the data after feature selections. Multiple clustering methods are executed on multiple feature subsets to obtain multiple base clustering results. Then, multiple base clustering results are achieved for ensemble clustering with a high degree of variability, which facilitates the generation of better final consensus clustering results. Finally, the least diverse labels are removed and then consensus clustering is performed to obtain the final clustering results.

To demonstrate the effectiveness of the scEFSC clustering result, we tested fourteen publicly available scRNA-seq datasets. The experimental results revealed that our proposed scEFSC significantly outperformed other single-cell clustering algorithms in terms of ARI and NMI evaluation metrics. In addition, we performed differential gene expression analysis, gene ontology enrichment and KEGG analysis on the clustering results of scEFSC to validate the biological interpretability of the clustering results.

CRediT authorship contribution statement

Chuang Bian: Writing - original draft, Methodology, Investigation. **Xubin Wang:** Methodology, Investigation. **Yanchi Su:** Investigation. **Yunhe Wang:** Conceptualization, Writing - review & editing. **Ka-chun Wong:** Validation, Writing - review & editing. **Xiangtao Li:** Supervision, Conceptualization, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work described in this paper was substantially supported by the National Natural Science Foundation of China under Grant No. 62076109, and also funded by “the Fundamental Research Funds for the Central Universities”. The work described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 11200218], a grant from the Health and Medical Research Fund, of the Food and Health Bureau, The Government of the Hong Kong Special Administrative Region [07181426], and the funding from the Hong Kong Institute for Data Science (HKIDS) at the City University of Hong Kong. The work described in this paper was partially supported by two grants from the City University of Hong Kong (CityU 11202219, CityU 11203520). This research was substantially sponsored by a research project (Grant No. 32000464) from the National Natural Science Foundation of China and was substantially supported by the Shenzhen Research Institute, City University of Hong Kong.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2022.04.023.

References

[1] Xiangtao Li et al. Evolving Transcriptomic Profiles from Single-cell RNA-seq Data using Nature-Inspired Multiobjective Optimization. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2020;18(6):2445–58.
[2] Hedlund E, Deng Q. Single-cell rna sequencing: technical advancements and biological applications. Mol Aspects Med 2018;59:36–46.
[3] Xiangtao Li et al. Single-Cell RNA-seq Data Interpretation by Evolutionary Multiobjective Clustering. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2019;17(5):1773–84.
[4] Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science 2018;360(6392).
[5] Yunhe Wang et al. Multiobjective Deep Clustering and Its Applications in Single-cell RNA-seq Data. IEEE Transactions on Systems, Man, and Cybernetics: Systems 2021.
[6] Lin P, Troup M, Ho JW. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. Genome Biol 2017;18(1):1–11.
[7] Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. Nature Methods 2017;14(10):979.
[8] Yau C et al. pcareduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics 2016;17(1):1–11.
[9] Levine JH, Simonds EF, Bendall SC, Davis KL, El-ad DA, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. Cell 2015;162(1):184–97.
[10] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature Biotechnol 2015;33(5):495–502.
[11] Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. Sincera: a pipeline for single-cell rna-seq profiling analysis. PLoS Comput Biol 2015;11(11):e1004575.
[12] Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, Van Oudenaarden A. Single-cell messenger rna sequencing reveals rare intestinal cell types. Nature 2015;525(7568):251–5.
[13] Xiangtao Li et al. High-throughput Single-cell RNA-seq Data Imputation and Characterization with Surrogate-assisted Automated Deep Learning. Briefings in Bioinformatics 2021;23(1):bbab368.
[14] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al. Sc3: consensus clustering of single-cell rna-seq data. Nature methods 2017;14(5):483–6.
[15] Wan S, Kim J, Won KJ. Sharp: hyperfast and accurate processing of single-cell rna-seq data via ensemble random projection. Genome Res 2020;30(2):205–13.
[16] Yang Y, Huh R, Culpepper HW, Lin Y, Yun L. Safe-clustering: Single-cell aggregated (from ensemble) clustering for single-cell rna-seq data. Bioinformatics 2018;35(8).
[17] Geddes TA, Kim T, Nan L, Burchfield JG, Yang J, Tao D, Yang P. Autoencoder-based cluster ensembles for single-cell rna-seq data analysis. BMC Bioinformatics 2019;20(Suppl 19).
[18] Zhu X, Li J, Li HD, Xie M, Wang J. Sc-gpe: A graph partitioning-based cluster ensemble method for single-cell. Front Genetics 2020;11.
[19] Huh R, Yang Y, Jiang Y, Shen Y, Li Y. Same-clustering: S ingle-cell a ggregated clustering via m ixture model e nsemble. Nucleic Acids Res 2020;48(1):86–95.
[20] Zhuohan Yu et al. Elucidating Transcriptomic Profiles from Single-cell RNA sequencing Data using Nature-Inspired Compressed Sensing. Briefings in Bioinformatics 2021;22(5):bbab125.
[21] Xiangtao Li et al. Deep Embedded Clustering with Multiple Objectives on scRNA-seq Data. Briefings in Bioinformatics 2021;22(5):bbab090.
[22] Tran D, Nguyen H, Tran B, La Vecchia C, Luu HN, Nguyen T. Fast and precise single-cell data analysis using a hierarchical autoencoder. Nature Commun 2021;12(1):1–10.
[23] R. Silipo, I. Adae, A. Hart, M. Berthold, Seven techniques for dimensionality reduction, White Paper by KNIME. com AG (2014) 1–21..
[24] He X, Cai D, Niyogi P. Laplacian score for feature selection. Adv Neural Inform Processing Syst 2005;18:507–14.
[25] Zhao Z, Liu H. Spectral feature selection for supervised and unsupervised learning. In: Proceedings of the 24th international conference on Machine learning. p. 1151–7.
[26] Cai D, Zhang C, He X. Unsupervised feature selection for multi-cluster data. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. p. 333–42.
[27] Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: A data perspective. ACM computing surveys (CSUR) 2017;50(6):1–45.
[28] Hadjitodorov ST, Kuncheva LI, Todorova LP. Moderate diversity for better cluster ensembles. Information Fusion 2005;7(3):264–75.
[29] Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. Nature Struct Mol Biol 2013;20(9):1131.

[30] Wang YJ, Schug J, Won K-J, Liu C, Naji A, Avrahami D, Golson ML, Kaestner KH. Single-cell transcriptomics of the human endocrine pancreas. Diabetes 2016;65(10):3028–38.

[31] Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nature Genetics 2017;49(5):708–18.

[32] Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, Hjerling-Leffler J, Haeggström J, Kharchenko O, Kharchenko PV, et al. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. Nature Neurosci 2015;18(1):145–53.

[33] Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, Kanton S, Kageyama J, Damm G, Seehofer D, et al. Multilineage communication regulates human liver bud development from pluripotency. Nature 2017;546 (7659):533–8.

[34] Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, Murphy AJ, Yancopoulos GD, Lin C, Gromada J. Rna sequencing of single human islet cells reveals type 2 diabetes genes. Cell Metabolism 2016;24(4):608–15.

[35] Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature Neurosci 2016;19(2):335–46.

[36] Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, Van Gurp L, Engelse MA, Carlotti F, De Koning EJ, et al. A single-cell transcriptome atlas of the human pancreas. Cell Syst 2016;3(4):385–94.

[37] Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metabolism 2016;24(4):593–607.

[38] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 2015;161(5):1187–201.

[39] Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. Science 2015;347 (6226):1138–42.

[40] Lake BB, Ai R, Kaeser GE, Salathia NS, Yung YC, Liu R, Wildberg A, Gao D, Fung H-L, Chen S, et al. Neuronal subtypes and diversity revealed by single-nucleus rna sequencing of the human brain. Science 2016;352(6293):1586–90.

[41] Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell Systems 2016;3(4):346–60.

[42] M. Slyper, J. Waldman, D. Dionne, B. Li, Study: Ica: blood mononuclear cells (2 donors, 2 sites), https://singlecell. broadinstitute. org/single_cell/study/SCP345/ica-blood-mononuclear-cells-2-donors-2-sites..

[43] Geddes TA, Kim T, Nan L, Burchfield JG, Yang JY, Tao D, Yang P. Autoencoder-based cluster ensembles for single-cell rna-seq data analysis. BMC Bioinformatics 2019;20(19):1–11.

[44] Risso D, Purvis L, Fletcher RB, Das D, Ngai J, Dudoit S, Purdom E. clusterexperiment and rsec: A bioconductor package and framework for clustering of single-cell and other large gene expression datasets. PLoS Comput Biol 2018;14(9):e1006378.

[45] Barkal AA, Weiskopf K, Kao KS, Gordon SR, Rosental B, Yiu YY, George BM, Markovic M, Ring NG, Tsai JM, et al. Engagement of mhc class i by the inhibitory receptor lilrb1 suppresses macrophages and is a target of cancer immunotherapy. Nature Immunol 2018;19(1):76–84.

[46] Pyke RM, Thompson WK, Salem RM, Font-Burgada J, Zanetti M, Carter H. Evolutionary pressure against mhc class ii binding cancer mutations. Cell 2018;175(2):416–28.

[47] Zeestraten E, Reimers M, Saadatmand S, Dekker JT, Liefers G, Van Den Elsen P, Van De Velde C, Kuppen P. Combined analysis of hla class i, hla-e and hla-g predicts prognosis in colon cancer patients. British J Cancer 2014;110(2):459–68.

[48] Hertzen L, Joensuu H, Haahtela T. Microbial deprivation, inflammation and cancer. Cancer Metastasis Rev 2011;30(2):211–23.

[49] Bienz M, Clevers H. Linking colorectal cancer to wnt signaling. Cell 2000;103 (2):311–20.