Computational and Structural Biotechnology Journal 23 (2024) 589–600

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

scSID: A lightweight algorithm for identifying rare cell types by capturing differential expression from single-cell sequencing data

Shudong Wang [a], Hengxiao Li [a], Kuijie Zhang [a], Hao Wu [b,c], Shanchen Pang [a,*], Wenhao Wu [a], Lan Ye [d], Jionglong Su [e], Yulin Zhang [f,*]

[a] Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, 266580, China
[b] College of Information Engineering, Northwest A&F University, 712100, Yangling, China
[c] School of Software, Shandong University, 250100, Jinan, China
[d] Cancer Center, the Second Hospital of Shandong University, Jinan, 250033, China
[e] School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Suzhou, 215123, Jiangsu, China
[f] College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, 266590, China

ARTICLE INFO

Dataset link: https://support.10xgenomics.com/single-cell-gene-expression/datasets

Dataset link: https://github.com/lgmzfl/scSID

Keywords:
Single-cell RNA sequencing
Rare cell types
Similarity analysis
Scalability

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) is currently an important technology for identifying cell types and studying diseases at the genetic level. Identifying rare cell types is biologically important as one of the downstream data analyses of single-cell RNA sequencing. Although rare cell identification methods have been developed, most of these suffer from insufficient mining of intercellular similarities, low scalability, and being time-consuming. In this paper, we propose a single-cell similarity division algorithm (scSID) for identifying rare cells. It takes cell-to-cell similarity into consideration by analyzing both inter-cluster and intra-cluster similarities, and discovers rare cell types based on the similarity differences. We show that scSID outperforms other existing methods by benchmarking it on different experimental datasets. Application of scSID to multiple datasets, including 68K PBMC and intestine, highlights its exceptional scalability and remarkable ability to identify rare cell populations.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has become a powerful technology revealing differences between different cell types and states [1]. It has greatly contributed to the development of transcriptomics, which can help us study cellular heterogeneity to gain insights into complex and rare cell populations and reveal the regulatory relationships between genes. By analyzing the transcriptional profiles of these rare cells, we can discover new disease-causing drivers and biomarkers, advancing the development of precision medicine and personalized therapy [2–4]. Despite their scarcity, rare cells play a pivotal role in a variety of key processes, such as immune responses to cancer and other diseases, cancer pathogenesis, and angiogenesis. For instance, invariant natural killer T cells have provided valuable insights into defense mechanisms against Mycobacterium tuberculosis [5]. Tumor stem cells have emerged as significant factors in tumorigenesis, tumor recurrence, and metastasis [6,7]. Moreover, endothelial progenitor cells from the bone marrow have demonstrated reliability as biomarkers for tumor angiogenesis [8,9]. In short, advancements in cell analysis techniques have opened up exciting possibilities for the discovery of new rare cell types, enriching our understanding of cellular diversity and function.

Single-cell clustering is a key step in understanding cell populations through single-cell analysis [10,11]. However, the small proportion of rare cells in scRNA-seq data poses a significant cluster-based cell type identification challenge. Most of the existing clustering methods cater to major cell types [12–14], such as Seurat [15] and Scater [16]. Whereas rare cell-specific genes often do not have a major impact in single-cell downstream analyses, traditional clustering methods may fail to identify cell populations that occur at low frequencies, i.e., rare cells.

Numerous attempts have been undertaken to identify rare cells and develop algorithms for detecting their transcriptomes. Notably, several methods have been proposed for this purpose, including rare cell type

* Corresponding authors.
E-mail addresses: wangsd@upc.edu.cn (S. Wang), z22070033@s.upc.edu.cn (H. Li), z20070009@gmail.com (K. Zhang), haowu@sdu.edu.cn (H. Wu), pangsc@upc.edu.cn (S. Pang), wenhaowu_1024@163.com (W. Wu), sdeyyelan@email.sdu.edu.cn (L. Ye), Jionglong.Su@xjtlu.edu.cn (J. Su), zhangyulin@sdust.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.csbj.2023.12.043
Received 2 November 2023; Received in revised form 27 December 2023; Accepted 27 December 2023
Available online 3 January 2024

identification (RaceID) [17], GiniClust [18], cell subtype identification from up-regulated gene sets (CellSIUS) [19], finder of rare entities (FiRE) [20], and novel deep generative model for leveraging the small samples of cells (scLDS2) [21]. RaceID employs k-mean clustering to calculate count probabilities for individual cells, enabling the identification of abnormal cells. To mitigate the impact of noise on clustering results, Herman et al. introduced RaceID3 by incorporating feature selection techniques [22]. GiniClust utilizes Gini coefficients as a criterion for gene selection prior to density clustering, filtering out cells corresponding to genes expressed only in rare cell populations. Daphne et al. later proposed GiniClust2, which utilizes Fano factor-based clustering combined with GiniClust to identify rare versus important cell populations through weighted integrated clustering [23,24]. CellSIUS adopts a two-step approach, first clustering major cell populations and then screening for marker genes exhibiting a bimodal distribution in each cluster. Subsequently, one-dimensional clustering is performed based on the identified bimodal distribution of marker genes to detect specific subpopulations [19]. FiRE leverages a sketching technique [25] to assign hash codes to cells multiple times, thereby calculating the rarity score. Cells surpassing a specified rarity threshold are considered rare. ScLDS2 distinguishes rare and non-rare cells through adversarial learning, transforming the rare cell type detection problem into a classification task [21]. These different approaches contribute to the ongoing advances in rare cell identification and transcriptome analysis.

Despite significant progress in detecting rare cell types, analysis of single-cell sequencing data on a large scale remains challenging. Several existing methods have limitations when dealing with such extensive datasets. RaceID3, while effective in identifying rare cell types, takes substantial time to process when thousands of cell counts are involved [23,25]. GiniClust2, although capable of handling large data, requires a considerable amount of memory for single-cell analysis [23]. CellSIUS relies on pre-existing clustering information obtained for major cell types, leading to potential drawbacks in its performance [19]. FiRE presents an improvement in both time and memory consumption when analyzing large datasets, but it necessitates a single clustering of the results to differentiate between rare and abnormal cells [20]. Moreover, scLDS2 is unable to better infer cell-cell interactions from changes in gene expression patterns so there are limitations in effectively utilizing indirect relationships between cells and genes [21]. In short, while various rare cell detection methods have shown promise, addressing the challenges of processing large-scale single-cell sequencing data remains crucial for further advancement in this field.

In this research, we introduce a novel similarity partitioning model termed scSID (single-cell similarity division). scSID's design is motivated by the observation that cells within the same cluster of cells exhibit significantly higher intercellular similarity compared to cells from neighboring clusters, mainly due to their intrinsic structural similarities. Based on this, we propose a method that utilizes the differences in similarity between cells to delineate rare cell populations. The effectiveness of scSID is rigorously evaluated through several simulation experiments using various single-cell sequencing datasets. Results indicate that scSID exceeds current methods for efficiency in detecting rare cells. Furthermore, when scSID is applied to multiple datasets, it demonstrates excellent scalability and memory efficiency in rare cell detection [26].

2. Methods

2.1. Overview of scSID

Traditional methods for identifying rare cells often rely on bimodal distributions of specific genes or preliminary clustering. This may lead to overlooked rare cell populations, potential biases and increased computational costs, especially in cases of low differential gene expression or misclassifications from preliminary clustering. To address these limitations, we propose a novel approach, scSID, for rare cell identification.

Our proposed scSID method is divided into two main steps: cell division based on individual similarity and rare cell detection based on population similarity. An overview of scSID is given in Fig. 1. The first step is as a result of the number of rare cells being low, when calculating the K nearest neighbor (KNN) [27] of a rare cell, cells within the KNN may cross over into the distribution of other cell types if K is large enough. This change in distribution leads to a steep change in similarity. We used the Euclidean distance in the gene expression space of individual cells as a measure of similarity between cells. We then characterize the similarity of cells to their K nearest neighbors, and classify cells with minimal characteristic differences into the same group.

In the second step, the primary objective is to address the potential impact of noise and outliers on the results obtained in the first step. Therefore, it is essential to further partition the similarity among cell clusters to reveal cellular heterogeneity. To achieve this, scSID employs a step-by-step clustering synthesis approach that aims to explore the hierarchical relationships between the cells within the identified clusters and their nearest neighbors outside the clusters.

2.2. Cell division based on individual similarity

In single-cell data, cells of the same class tend to have similar features in function and phenotype. This means that their gene expression patterns will be more similar, showing up as coordinate points in closer proximity in the dimensionally reduced data space. Conversely, cells have different functions and phenotypes from other classes of cells, which means that their gene expression patterns may be more different, giving different distributions in the dimensionally reduced data space. As a result, the difference in similarity between cells from the same category to different categories of cells will show significant changes due to the changes in distribution in the feature space.

Given a sufficient value of K, we believe that by computing the K nearest neighbors [27] of each cell, we can capture differences in similarity between cells from the same class and different classes. The determination of the K value is closely related to the number of rare cells. We found that most clustering methods were able to effectively identify groups that accounted for more than 2% of the total number of cells when the K value was greater than the number of rare cells detected [28]. Meanwhile, in order to balance the computational performance and accuracy, we generally set the K value to no more than 2% of the total number of cells in large sample datasets. For datasets with a data volume of about 5000 or less, a K value of 100 is selected by default. Considering the small proportion of rare cells compared to common cells, this differential variation can be observed among the first few neighbors of a cell, thus reducing the need to compute an excessive number of neighbors for each cell.

In the implementation process, we identified cells with differential gene expression by selecting genes with high expression levels [29,30]. Subsequently, we applied principal component analysis (PCA) [31] to reduce the dimensionality of the features to n dimensions (with a default value of 50 dimensions). Next, in single-cell gene expression analysis, each cell in the downscaled data space can be regarded as a point in the space, and the expression level of each gene corresponds to the position of that point on a particular axis. Thus, the Euclidean distance between two cells can be interpreted as a combined measure of the difference in their positions in gene expression space. For this purpose, we performed a KNN analysis based on the Euclidean distance for each cell to identify the top k most similar neighboring cells and their corresponding distances. Let $x_{<j,p>}$ represent the p_{th} principal component of cell j, and $x_{<k,p>}$ be the p_{th} principal component of the k_{th} nearest neighbor of cell j. The Euclidean distance $D_{<j,k>}$ between cell j and its k_{th} neighbor is defined as follows:

$$D_{<j,k>} = \sqrt{\sum_{p=1}^{n} \left(x_{<j,p>} - x_{<k,p>}\right)^2}. \quad (1)$$

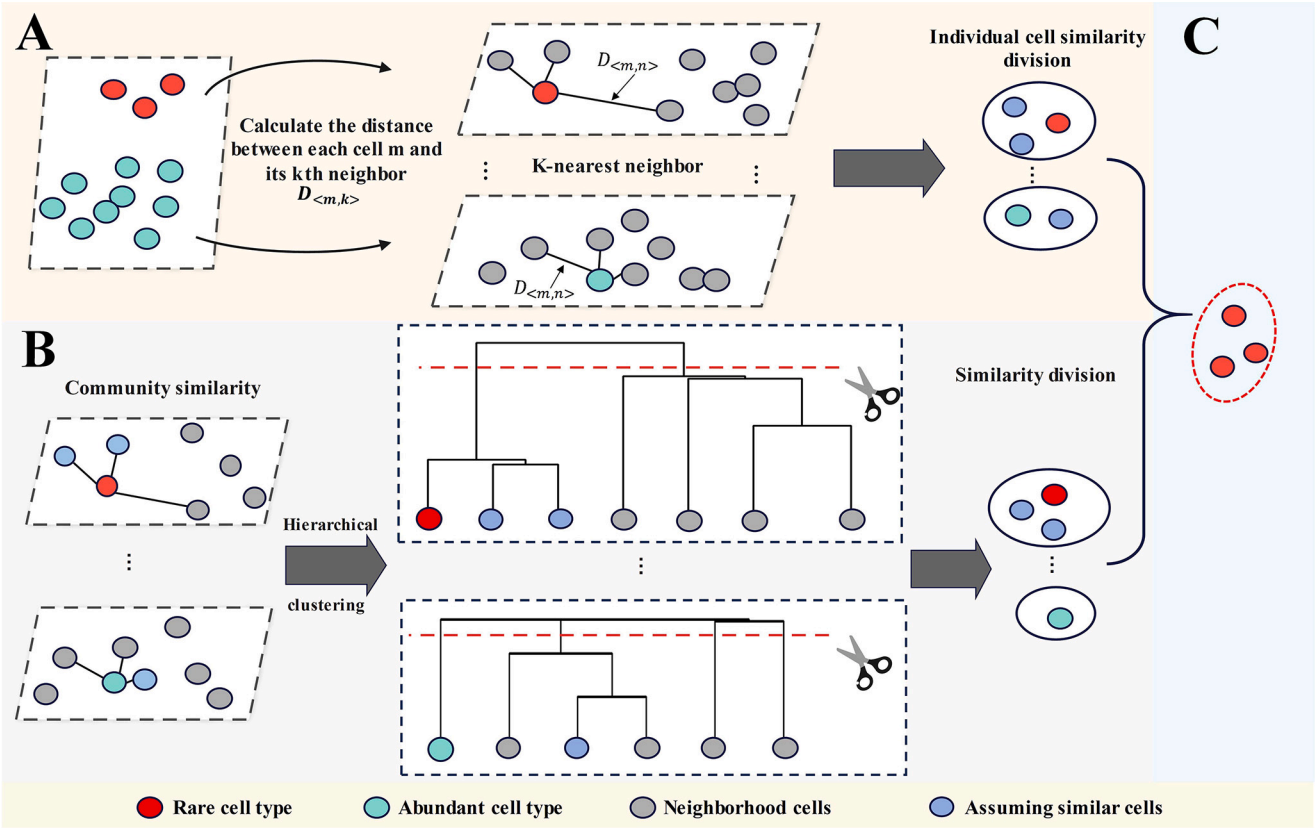

Fig. 1. Overview of scSID. (A) For each cell, the number of its first K neighbors was counted, and $D_{<m,k>}$, was calculated using the provided formula. Subsequently, based on the maximum value of $D_{<m,k>}$, the cell was initially assigned to a specific cluster. (B) For each cell, the single-cell cluster it was assigned to in the first step was further subdivided using hierarchical clustering based on similarity with the neighboring cells obtained from KNN. (C) A comparison was made between the results obtained from the first and second steps in order to determine the final rare cell type.

For rare cells, their scarcity leads to less variation in intercellular distance among the smaller k neighbors. Therefore, we can classify a cell as a rare cell cluster if its first k neighbors exhibit higher similarity. Conversely, when rare cells are compared to other cells beyond the k neighbors, the distance increases significantly. To capture this change in difference, we utilize the similarity between a cell and its K nearest neighbors as a feature and represent it using the first-order difference. The first-order difference refers to the difference between two consecutive neighbor terms in a discrete function, which helps to smooth out irregular fluctuations in the data and effectively captures the variation in differences. More specifically, we compute the similarity characteristics $D_{<j,k>}$ between each cell and its neighbors and define the first-order difference $\Delta D_{<j,k>}$ as:

$$\Delta D_{<j,k>} = D_{<j,k>} - D_{<j,k-1>}. \quad (2)$$

Intercellular feature differences become more prominent when a cell belongs to a different class than its neighbors, resulting in numerically larger values. To analyze these feature differences, we examined the difference in similarity between each cell c_j and its neighboring cells ΔD_j to identify the location s_j where the difference increased steeply, represented by:

$$s_j = \max\left(\Delta D_j\right). \quad (3)$$

For each $s_j \in \{2, 3, \ldots, K-1\}$, we focus on the first s_j neighbors of the cell, which demonstrate the highest individual similarity. We provisionally classify these neighbors as rare cell clusters C_j, while the remaining neighbors are assigned to the nearest neighbor clusters C_{near}. With this approach, we can accurately identify and distinguish rare cell clusters and nearest neighbor clusters based on the significant similarity differences observed among cells.

2.3. Rare cell detection based on population similarity

Cells possess diverse biological characteristics, and solely relying on similarities based on individual cells may not fully explain the intricate structures and relationships within cell clusters. Consequently, further partitioning of group similarity is warranted to unveil the heterogeneity between cells. Additionally, similarity partitioning based solely on individual cells can be susceptible to the influence of noise and outliers, potentially impacting interpretability. To address this problem, it is essential to perform similarity analyses among cell populations to distinguish between noise and outliers, so as to enhance the interpretability of the results. As a result, performing rare cell group similarity detection on the obtained division results becomes necessary.

To analyze rare cell clusters, we chose a hierarchical clustering approach that incorporates both similarity considerations between cell clusters and their nearest neighbors. The choice of hierarchical clustering was motivated by its ability to efficiently deal with the multilevel structure in the data, which often stems from differences between cell types. Compared to other clustering techniques, hierarchical clustering performs much better in capturing such complex relationships, providing a more flexible analytical framework for accurately identifying cell clusters. In the initial step, we treat each cell as a separate cluster. Following this, based on the distance matrix, we implemented a minimum distance merging strategy, i.e., the distance between the nearest cell pairs in two different clusters was selected as the distance between clusters. This strategy synthesizes the similarity between cells and aims to merge the cells with the closest distance in the feature space. It helps to maintain sensitivity to the true association between cells during the merging process and ensures that the most similar cells are merged. Consequently, we selected the minimum distance between two clusters, merged them into a new cluster, and concurrently updated the

distance matrix. This process is repeated iteratively until all the clusters are eventually merged into a large cluster. The minimum distance between cluster m and cluster n is denoted as follows:

$$L(m,k) = \min D(c_p, c_q) \tag{4}$$

where c_p, c_q are the two most similar cells between cluster m and cluster k. After we perform hierarchical clustering [32] based on this idea, these neighbors belong to a set of similarity samples if they clearly show different clusters. After applying hierarchical clustering based on this concept, if the neighbors do not belong to the same group of similar samples, they will be present in separate clusters, as shown in the first clustering plot in Fig. 1B. This observation indicates the presence of different biological properties and inter-cellular heterogeneity within the data. Conversely, when the similarity among these neighbors is high, the clustering structure shows smaller classes, as shown in the second clustering tree, and the hierarchical clustering tree shows a denser distribution. To this end, based on the above results, we set a threshold h (with height = 0.85 by default) to cut the resulting clustering tree (Fig. 2), and for the resulting clusters, we calculate whether the number of clusters in which individual cells are located matches the s_j result, and if so, the cluster is verified as a rare cell cluster with close similarity.

3. Materials

3.1. Datasets and preprocessing

In our analysis, we employed a variety of publicly accessible single-cell RNA sequencing (scRNA-seq) datasets. One dataset was particularly noteworthy, comprising expression profiles from approximately 68,579 healthy recipient donor cells, specifically peripheral blood mononuclear cells (PBMC) [33]. Reference data was derived from single-cell expression profiles of 11 distinct purified PBMC subpopulations to categorize the cell types. Our downstream analysis focused on genes with read counts exceeding 2 in a minimum of 3 cells, ensuring their retention for further examination. We standardized each scRNA-seq dataset by applying median normalization and log2 transformations.

To simulate experiments involving rare cells, we utilized 293T and Jurkat cell data comprising roughly 3200 cells, which underwent filtration and normalization processes. The intestine dataset of 4500 cells was processed further [34]. We used scanpy [35] to exclude genes expressed in less than 3 cells and retained 15,172 genes for further analysis. In addition, we excluded cells with fewer than 200 expressed genes, resulting in a dataset of 4,301 cells, which were subsequently subjected to normalization.

In the high-fat diet dataset, we retained genes expressed in a minimum of two cells for normalization. Subsequently, we applied a logarithmic transformation to each element of the input matrix x_{ij}, accompanied by the addition of a pseudo-count of one.

3.2. Performance assessment

In order to evaluate the results of rare cell detection in the simulated experiments, we chose F_1 score and $Sensitivity$ to reflect the degree of rare cell detection [36–38]. The F_1 score and $Sensitivity$ are calculated as follows.

$$F_1 score = \frac{2 * TP}{2 * TP + FP + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

Where TP stands for true positive, indicating the precise identification of rare cells. FP denoting false positive refers to the misclassification of ordinary cells as rare. On the other hand, FN representing false negatives occurs when rare cells are mistakenly categorized as non-rare. The F_1 score and $Sensitivity$ metrics lie within the [0,1] interval. Within this range, higher values denote a more accurate correspondence in the identification of rare cells, indicating enhanced precision in cell detection.

3.3. The 68K PBMC data subsampling

To systematically simulate the rare cell phenomenon, we downsampled the full 68K PBMC dataset in order to evaluate the model. Our analysis focused on three distinct cell types that exhibited differential transcriptional profiles (Fig. 3A). These cell types included CD14+ monocytes, CD19+ B cells, and CD56+ NK cells. To create the rare cell datasets, we employed a sampling strategy wherein we randomly selected 2 to 100 CD14+ monocytes from the CD14+ monocyte population and added them to a population of 500 cells randomly selected from the CD56+ NK and CD19+ B cell populations. By applying this approach, we generated a total of 99 rare cell datasets, where the proportion of rare cells ranged from 0.2% to 10%.

To assess the sensitivity of scSID to cell type characterization, we used the DBSCAN clustering method on the basis of a 68K PBMC dataset in CD14+ monocyte and CD19+ B cell populations, respectively. Then, in each largest cluster, we randomly selected 2 CD14+ monocytes and 500 CD19+ B cells for subsequent combinatorial analysis.

3.4. The 293T-Jurkat data subsampling

We created ten subsample datasets using the approach of Jindal et al. [20]: a cell cluster of 1540 cells was randomly selected from the 293T cell cluster as the main cell cluster. Ten datasets were then generated by randomly selecting different numbers of cells from the Jurkat population. The percentage of Jurkat cells ranged from 0.5% to 5%.

3.5. Differential expression analysis

We employed a standard Wilcoxon's rank sum test to identify differentially expressed genes, with a false discovery rate (FDR) threshold of 0.05 and an inter-group absolute fold-change threshold of 1. Fold-change values were calculated based on mean expression levels between groups for each gene. A gene was classified as cell type-specific if it exhibited higher expression in a specific cluster compared to all other clusters.

3.6. Hyper-parameter

Through subsampling the 68K PBMC dataset, we acquired a rare cell dataset comprising a 2% rarity proportion. We then conducted a comparative analysis using varying scSID thresholds h (0.75, 0.80, 0.85, 0.90, 0.95) and distinct PCA dimensions (50, 100, 150). Our analysis revealed that the algorithm reached its peak performance when the scSID threshold h was set at 0.85. This decision was informed by the observation that the performance of scSID's dimensionality reduction was remarkably similar across 50, 100, and 150 dimensions, as evidenced in Fig. 2. Considering the reduced memory requirement for the 50 dimensions, we ultimately opted to set the PCA dimension at 50.

3.7. Methods of comparison

To further validate the performance of scSID algorithm, we selected popular and high-performing methods for comparison, including RaceID3 [22], GiniClust2 [23], CellSIUS [19], FiRE [20], and scLDS2 [21]. Specifically, RaceID3, GiniClust2 and CellSIUS are traditional clustering-based methods, and FiRE is a fast and novel method for rare cell identification based on sketching techniques. scLDS2 is a deep clustering algorithm based on generative adversarial learning. All of the above methods are available on GitHub. RaceID3 package is applied directly to the preprocessed matrices with all parameters at default values except for the initial clustering based on the enrichment of cells.

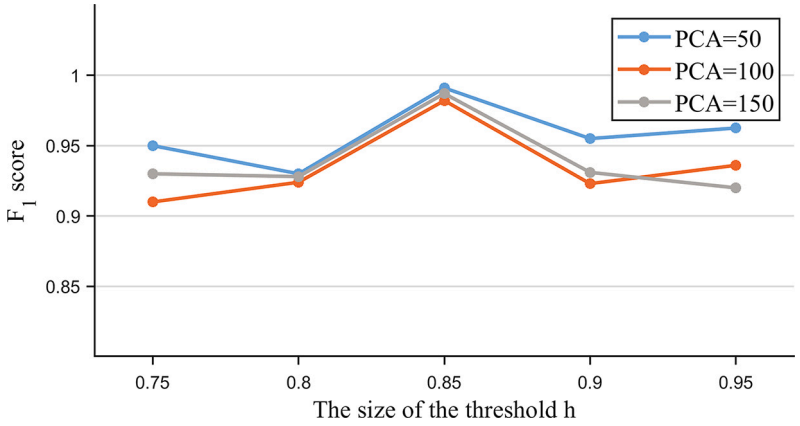

Fig. 2. Performance comparison (F_1 score) of scSID on 68K PBMC subsampling dataset with different threshold h and PCA parameters.

GiniClust2 package runs the analysis with default parameters. CellSIUS package parameters were set to default values. All FiRE package parameters were set to default values. scLDS2 package clustering clusters were changed according to the number of enriched cell clusters, and all other parameters were set to default values.

4. Results

4.1. scSID accurately identifies different proportions of rare cells under data subsampling

To evaluate the performance of scSID in the analysis of real scRNA-seq datasets, we compared scSID capabilities with RaceID3, GiniClust2, CellSIUS, FiRE and scLDS2 for rare cell detection. We utilized a publicly accessible scRNA-seq dataset comprising the transcriptome of around 68,000 peripheral blood mononuclear cells (68K PBMC) covering 11 subtypes. There is an extremely high similarity between several of these subpopulations, which is difficult to distinguish. To construct the rare cell dataset, our analysis focused on three distinct cell types characterized by differing transcriptional profiles: CD14+ monocytes, CD19+ B cells, and CD56+ NK cells (Fig. 3A). We employed a sampling strategy, as described in the Materials section, to generate 99 rare cell datasets, spanning proportions ranging from 0.2% to 10%.

We assessed the detection of rare cell types using F_1 scores, which offer a balanced measure of accuracy and sensitivity. Across datasets with varying proportions of rare cells, scSID consistently outperformed RaceID3, GiniClust2, CellSIUS, FiRE, and scLDS2 in identifying rare monocytes (Fig. 3B). Notably, the F_1 scores of the other methods were close to zero for low rare cell proportions. As the percentage of rare cells increased, RaceID3 and CellSIUS showed significant improvements in F_1 scores, although they did not achieve perfect detection results in most cases. FiRE gave more scattered scores. GiniClust2 shows a relatively stable growth trend, but the overall F_1 score is not high. The mean F_1 score of the scLDS2 is proportional to the percentage of rare cells, but the score is not stable. Furthermore, we further evaluated the six methods using sensitivity (Fig. 3C). Both RaceID3 and scSID consistently maintained high sensitivities across all datasets. The sensitivity of CellSIUS increased with the proportion of rare cells, whereas FiRE, GiniClust2, and scLDS2 exhibited lower average sensitivity scores compared to the other three methods.

To validate the robustness of the scSID algorithm, we implemented an experimental design proposed by the authors of the FiRE algorithm [20]. Following their methodology, we employed Jurkat cells as rare cells and mixed them with 239T cells, which were extracted in different proportions, resulting in 10 datasets with rare cell proportions ranging from 0.5% to 5% (Materials). Notably, since CellSIUS relies on initial clustering and is not applicable to scenarios with only one major cell type, we excluded its evaluation from this experiment. However, for the remaining five methods, scSID consistently outperformed RaceID3, GiniClust2, FiRE, and scLDS2 across different dataset proportions (Fig. 4A). As the proportion of rare cells increased, scSID exhibited slightly superior performance compared to RaceID3, while the scores of FiRE and scLDS2 also showed improvement. In line with the findings reported by Jindal [20], GiniClust2 [23] failed to identify rare cells in any portion of the dataset. Furthermore, we assessed the sensitivity of these five methods using the sample dataset, and the results demonstrated that scSID maintained a high sensitivity across all datasets (Fig. 4B). To visually showcase the detection outcomes of the different methods, we highlighted and presented in Fig. 4C the rare cells detected by each algorithm at a rare cell percentage of 2.5%. Notably, scSID consistently and accurately identified rare cells in agreement with the known annotations. These results serve to further demonstrate the robustness of the scSID algorithm and highlight its advantages for rare cell detection.

4.2. scSID is sensitive to cell type identity

To evaluate the robustness and sensitivity of scSID under different numbers of differentially expressed genes, as well as its performance at low rare cell ratios, we conducted secondary sampling of the CD19+ B cell population and the CD14+ monocyte population from the 68K PBMC dataset. This allowed us to generate a rare cell data set with a rare cell ratio of 0.4%. Through a rigorous screening process [39], we identified a total of 114 differentially expressed genes between these two cell types (Materials).

In each iteration of the experiment, we systematically replaced the same amount of differentially expressed genes with non-differentially expressed genes, thus altering the count of differentially expressed genes to assess the sensitivity of scSID in detecting rare cell populations (Fig. 5A). For each dataset with varying numbers of differentially expressed genes, we calculated the average area under the curve (AUC) relative to the secondary population. This process was repeated 100 times to obtain its mean values.

In each iteration of the experiment, we also compared the performance of RaceID3, GiniClust2, FiRE, and scLDS2. However, the CellSIUS method was not applicable in this scenario as the dataset consisted of only one dominant cell population and no prior knowledge was available [19,40]. At smaller proportions of differentially expressed genes, all methods struggled to effectively detect rare cells. However, as the proportion of differentially expressed genes increased, scSID reported substantial improvement over the other four methods (Fig. 5B). With fewer than 70 differentially expressed genes, scSID achieved an AUC value of above 0.9. RaceID3 also exhibited noticeable improvement in performance. On the other hand, the results of GiniClust2, FiRE and scLDS2 showed high concordance, which was due to noise interference caused by cell type-specific expression, resulting in poor detection performance in the presence of a small number of differentially expressed

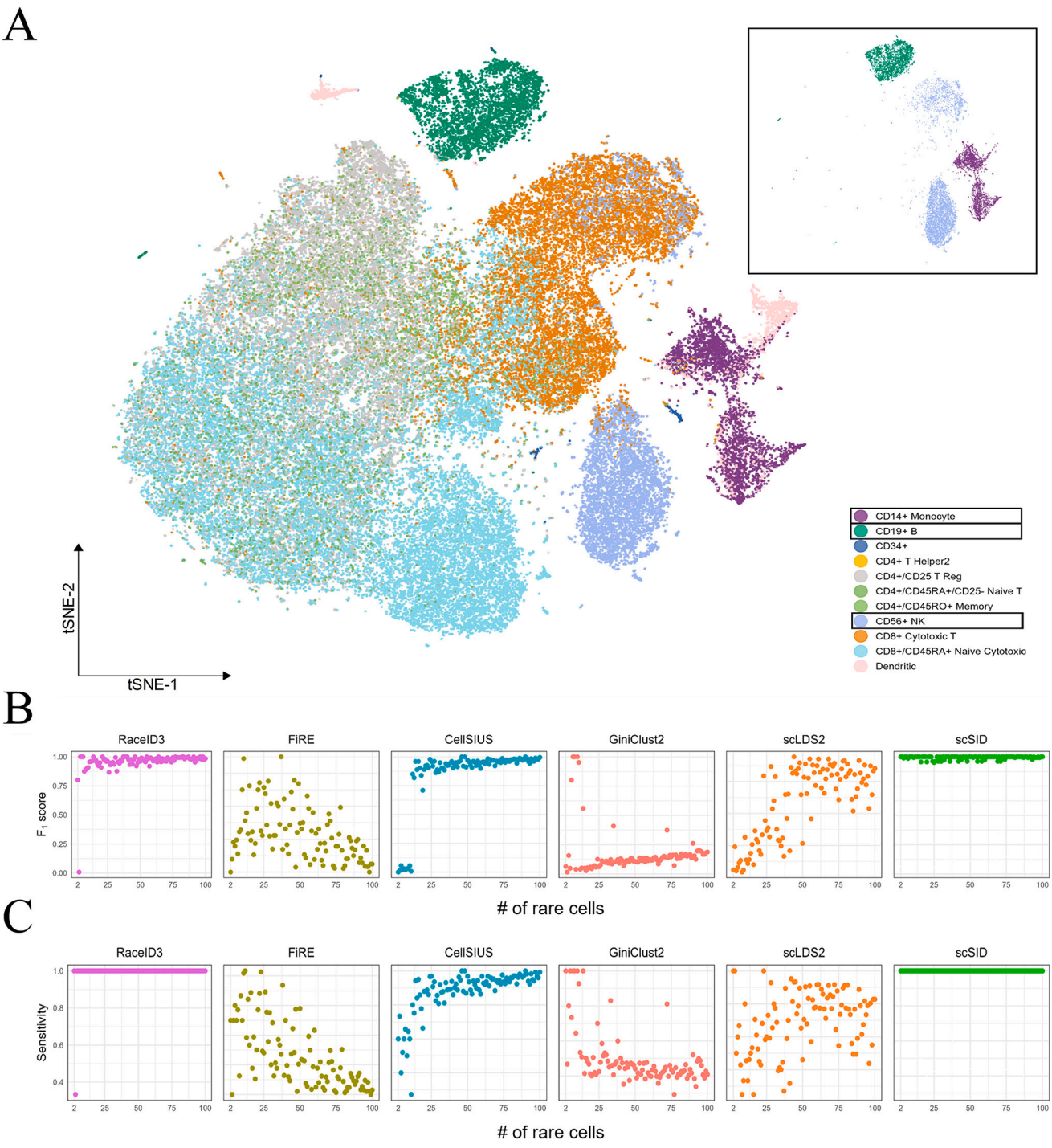

Fig. 3. scSID analysis of different degrees of rare cells after subsampling the 68K PBMC dataset. (A) A tSNE plot was generated for the 68K PBMC dataset showing cell types and three cell types selected for further analysis. (B) Evaluating the performance of different rare cell type detecting methods based on their F_1 score. (C) Evaluating the performance of different rare cell type detecting methods based on their sensitivity.

genes Scores were all less than 0.6 and their results overlapped and were overlaid in Fig. 5B.

4.3. scSID is scalable and efficient

The advancement of scRNA-seq technology has greatly facilitated the study of cell types. As the field progresses towards large-scale single-cell sequencing, the computational efficiency of algorithms becomes a prominent concern in scRNA-seq research. To assess the computational efficiency of RaceID3, GiniClust2, scLDS2, FiRE, and CellSIUS, we implemented these methods on a single machine equipped with four Intel Xeon E5-2620v4 CPUs and 376.33 GB of memory. We recorded the runtimes of these methods using different input data sizes. To facilitate comparison, we performed subsampling from the 68K PBMC dataset, which comprises of expression profiles ranging from 1,000 cells to approximately 68,000 cells. Our findings indicate that RaceID3, GiniClust2, CellSIUS, and scLDS2 exhibit longer runtimes (Fig. 6A). Among them, RaceID3 has the longest runtime compared to the other methods. For instance, while scSID processes 50,000 cells within seconds, RaceID3 requires 8,338.03 minutes to process slightly over 50,000 cells. In contrast, FiRE, leveraging sketching technology [25], demonstrates rapid processing of cells, completing the analysis of the entire 68K

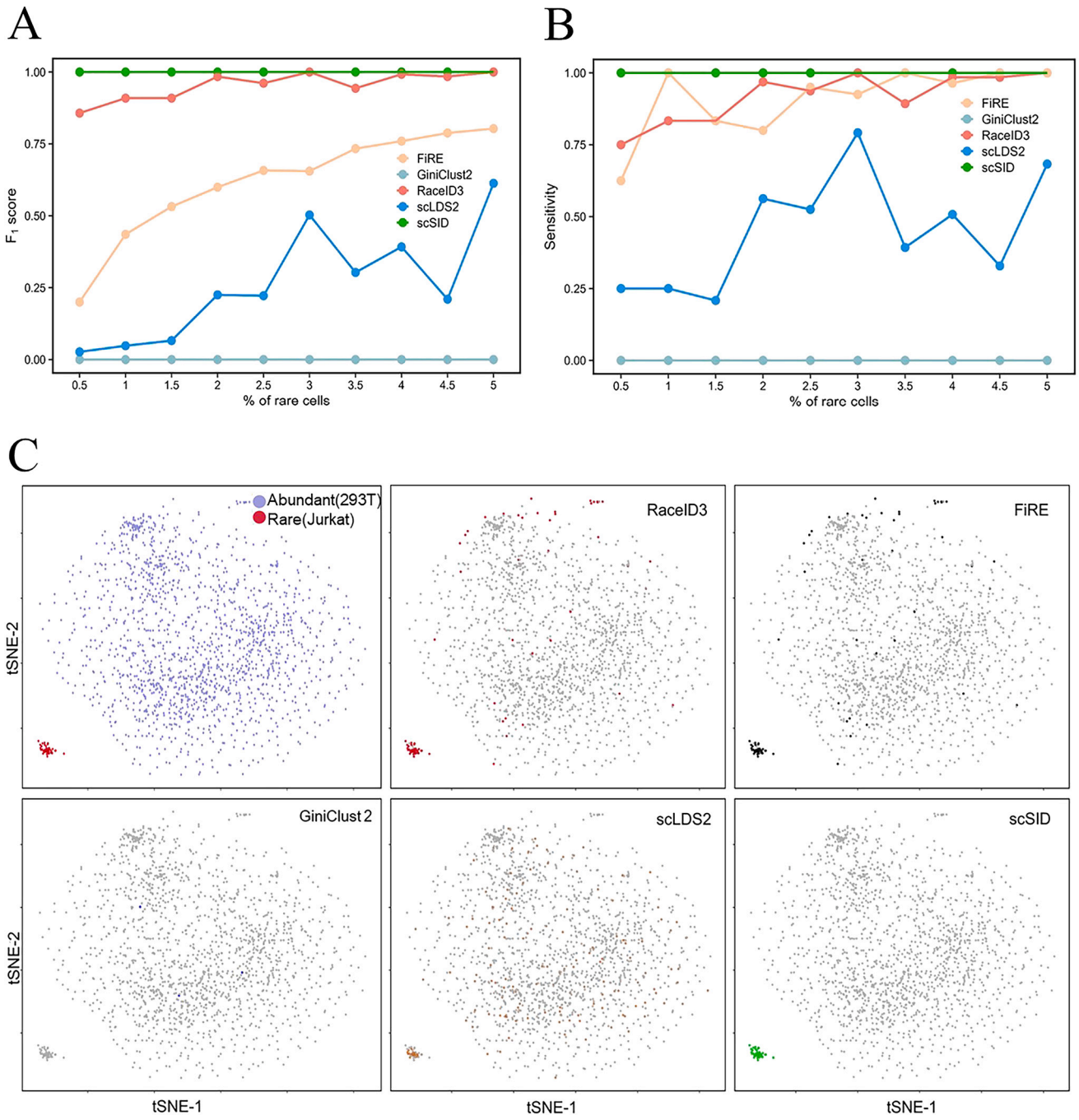

Fig. 4. scSID analysis unveiled differences in the detection of rare cells within a simulated dataset that included Jurkat and 293T cells. (A) F_1 scores were computed for the rare (Jurkat) populations to evaluate the performance of various methods in identifying these rare cell types. (B) The sensitivity of rare (Jurkat) populations were calculated to assess how well various approaches for identifying rare cell types performed. (C) Rare cells detected by various algorithms were plotted on a 2D plot based on t-SNE, specifically focusing on the 2.5% rare cell concentration.

PBMC dataset in just a few seconds, which is approximately twice the runtime of scSID.

In the current landscape of large-scale single-cell transcriptomics, the spatial complexity of algorithms poses a significant challenge to their scalability. When studying rare cells, it is crucial to consider not only the accuracy of the algorithms, but also their memory utilization during runtime [20,23]. In our assessment of six algorithms on the 68K PBMC dataset, we observed that GiniClust2 exhibited the highest memory usage, exceeding 250 GB. This finding aligns with the results reported in GiniClust2 by its authors [23] (see Fig. 6B). CellSIUS and RaceID3 demonstrated relatively lower memory usage, although both exceeded 100 GB. scLDS2 showed an improved memory footprint, utilizing less than 50 GB. The FiRE algorithm achieved a significant reduction in memory usage, requiring less than 20 GB. Similarly, scSID also demonstrated improved memory utilization, consuming less than 3 GB. By experimental comparison of methods, we found that scSID offers faster processing speed and lower memory utilization when dealing with large datasets.

4.4. scSID identifies rare cells from the single-cell transcriptome of the intestine and in a large 68K PBMC dataset

To evaluate the performance of the model, we utilized two publicly available datasets: the single-cell transcriptome of the intestine [34,41]

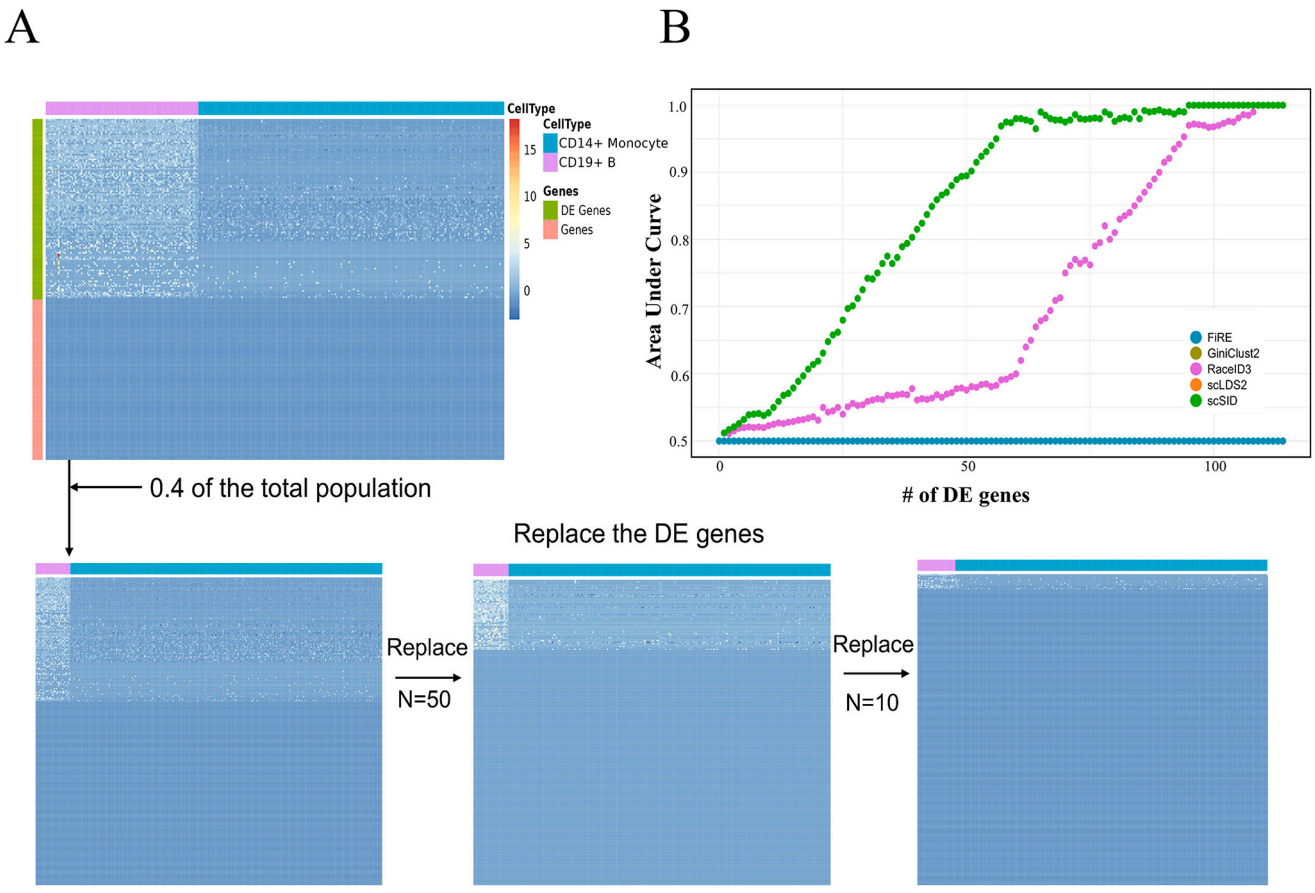

Fig. 5. Comparison of sensitivity in different methods for cell type identification with differential gene expression (DE) analysis. (A) The schematic overview illustrates perturbations in the dataset. Starting with the 68K PBMC dataset, where CD19+ B cells constitute the predominant cell type and CD14+ monocytes are rare, we created a dataset with a rarity of approximately 0.4% by employing secondary sampling. Pairwise screening was conducted to identify DE genes between the two cell types. Non-DE genes were then replaced with a predetermined number of DE genes. The number of replaced genes ranged from 1 to the total number of DE genes. (B) The performance of different methods for cell type detection was compared using sensitivity as the evaluation criterion. The number of replaced DE genes was varied to assess the impact on sensitivity.

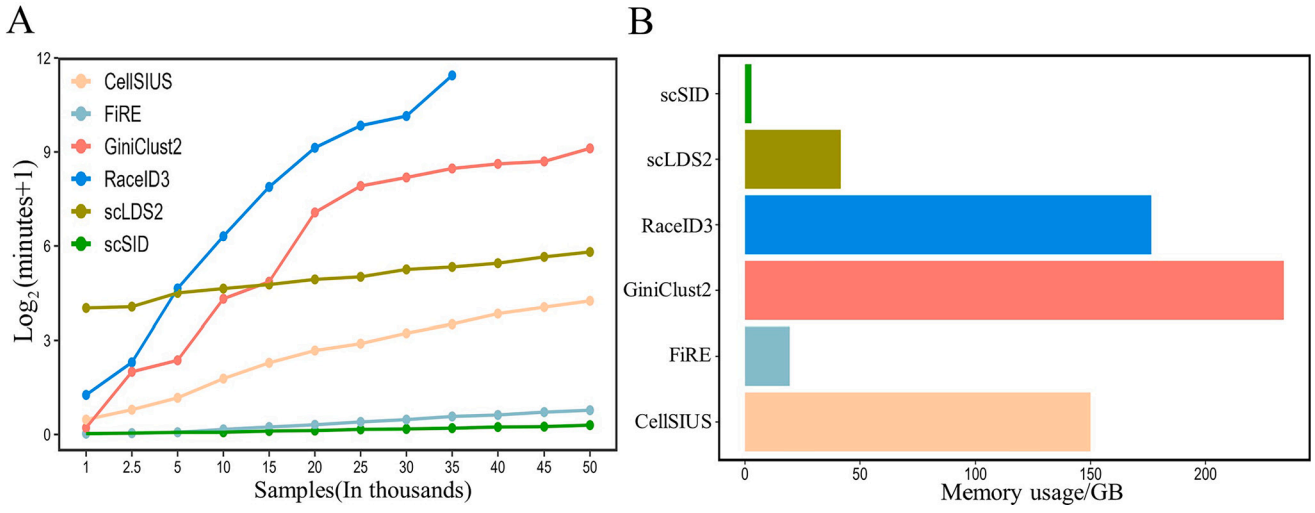

Fig. 6. scSID achieves the balance of efficiency and scalability. (A) The execution time of six methods, RaceID3, GiniClust2, CellSIUS, FiRE, scLDS2, and scSID, with cell counts varying from 1 to approximately 50,000. (B) The memory usage of the six methods was recorded when applied to a dataset containing approximately 68,000 cells.

and the 68K PBMC dataset (Fig. 7, A and B). The 68K PBMC dataset is widely used as a large-scale dataset in various computational algorithms. Previous analyses using GiniClust2 and FiRE have identified a rare cell type, specifically (CD34+) megakaryocytes, with cell proportions below 0.4%. Our interest lies in investigating whether scSID can detect the rare cells in these datasets.

We utilized scSID to analyze the single-cell transcriptome of the irradiated mouse intestine. This approach yielded 11 enteroendocrine

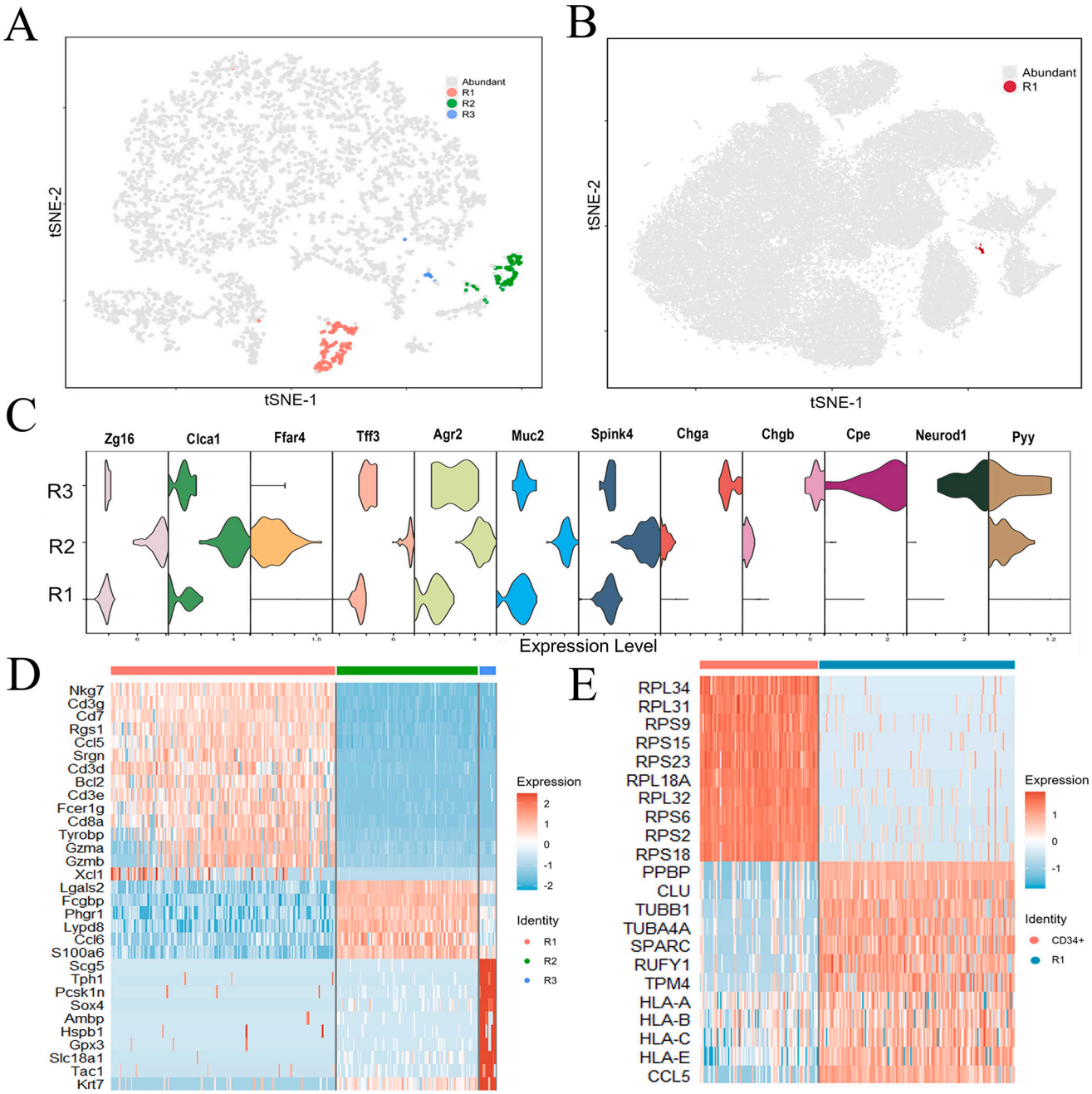

Fig. 7. Analysis of the intestinal dataset and a large 68K PBMC dataset. (A-B) Rare cell types within the intestinal dataset and the 68K PBMC dataset were visually highlighted using different colors in the 2D embedded plots. (C) Marker gene expression within all rare cell types was analyzed, providing insights into the specific genetic signatures associated with these rare cell populations. (D-E) Heatmaps were generated to display the rare cell clusters detected within the intestinal dataset and the 68K PBMC dataset, showcasing their gene expression patterns. Additionally, the top differentially expressed genes within each sub-cluster were identified and depicted in the heatmaps.

CD3+CD8+ T cells and 95 goblet cells. Additionally, 150 enteroendocrine cells were identified, corresponding to three distinct rare cell clusters, designated as R1, R2, and R3. We performed differential analyses to identify marker genes to distinguish each cell cluster from other cell types (Fig. 7C). Specifically, we found that marker genes such as Cd3d and Gzmb showed expression specifically in subpopulation R1 (Fig. 7D), indicating that this group of 11 cells is among the enteroendocrine CD3+CD8+ T cells [42]. Markers such as ZG16, CLCA1, FFAR4, TFF3 and SPINK4 genes were specifically expressed in R2, suggesting that the 95 cells in R2 represent thrush cells. Similarly, marker genes, including CHGA1, CHGA2, and CHGA3, were also specifically expressed in R2, further confirming the presence of thrush cells in this cell population [43]. In addition, we identified marker genes such as CHGA, CHGB, CPE, NEUROD1, and PYY, which are specifically expressed in R3, suggesting that these 150 cells belong to enteroendocrine cells.

With the rapid advances in single-cell technology, it is now possible to analyze tens of thousands or even millions of transcriptomes at the single-cell level. The 68K PBMC dataset has become a benchmark for large-scale data analysis, and we employed scSID to analyze the entire 68K PBMC dataset [25] in order to uncover rare cell types. Using scSID, we successfully detected 164 CD34+ megakaryocytes, while excluding the remaining 77 CD34+ cells in PBMC, leading to the differential analysis of the two clusters (Fig. 7E). In cluster R1, we observed high expression levels of HLA-A, HLA-B, and HLA-C genes, which belong to the human leukocyte antigen (HLA) gene family [44,45]. HLA genes

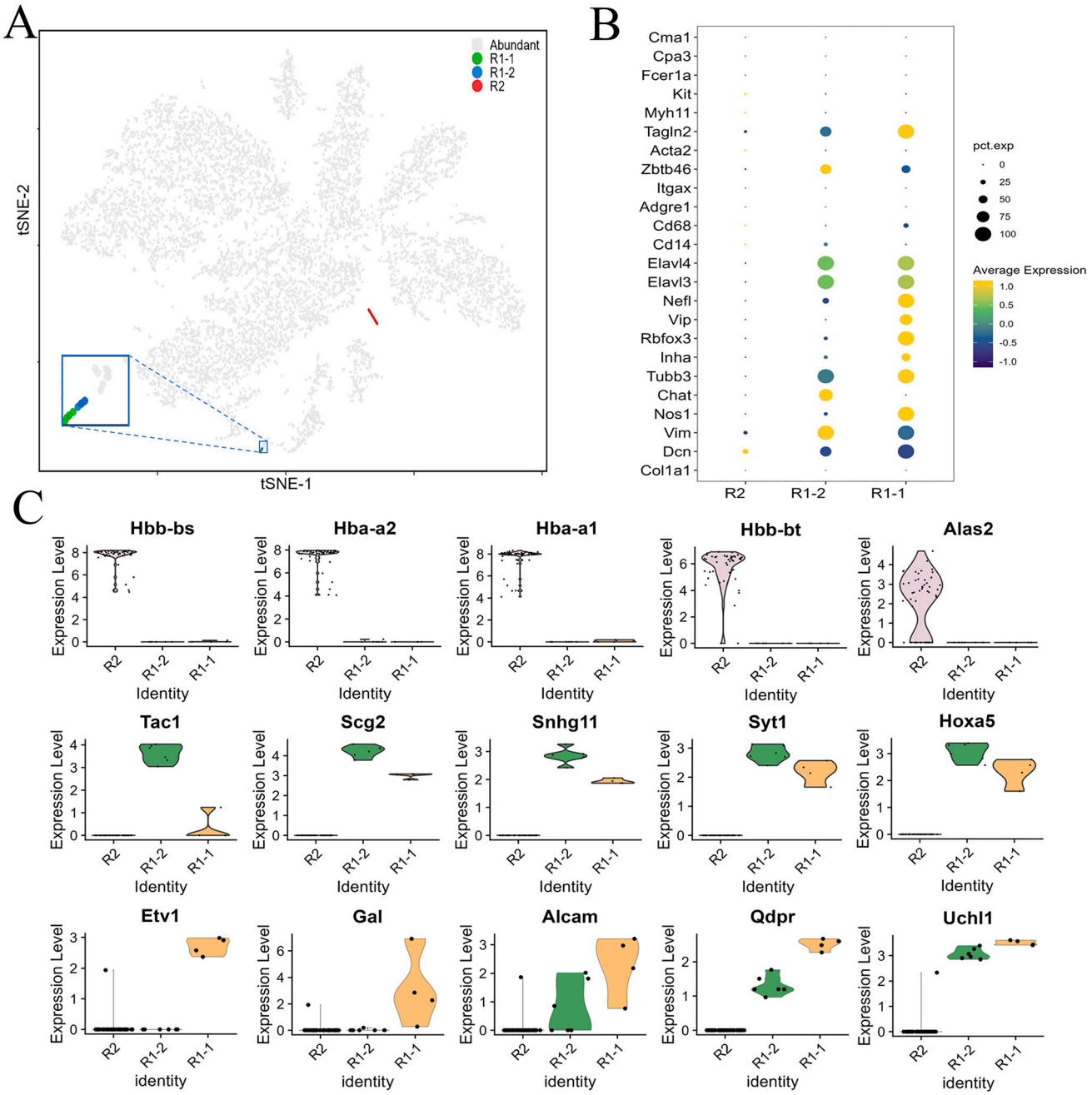

Fig. 8. Analysis of the High-Fat diet dataset. (A) Rare cell types within the high-fat diet dataset were visually highlighted using different colors in the 2D embedded plot. (B) Marker gene expression in all rare cell types was examined. (C) Violin plots were generated to depict the expression of differentially expressed genes within each cluster.

play a crucial role in cell recognition, immune response regulation, and are widely studied in immune-related diseases, organ and bone marrow transplantation, vaccine and drug target population screening, tumor immunology research, and more. In addition, we found that ribosomal protein genes such as RPL34, RPL31, RPS9 and RPL32, as well as inflammatory chemokine genes, were highly expressed in the remaining CD34+ cells. This suggests the presence of different isoforms within the CD34+ cell population that were captured by scSID.

4.5. scSID identifies nitrergic neurons in high-fat diet dataset

To validate the detection ability of our model, we applied it to the high-fat diet dataset (Fig. 8A). Our objective was to determine whether the model could successfully detect rare cells in this dataset.

We applied scSID to analyze 9,139 cells from the high-fat diet single-cell transcriptome dataset. In our experimental investigations, we discerned two distinct rare cell clusters, characterized as neuronal and erythroid clusters, labeled R1 and R2 in Fig. 8A. The cell counts for these clusters were 10 for R1 and 47 for R2, respectively. Notably, cluster R1 is subdivided into two subgroups, R1-1 and R1-2, consisting of six neurons and four nitrergic neurons. We conducted differential analyses on these cell populations (R1-1, R1-2 and R2) to ascertain their respective cell types accurately.

In R1-1 and R1-2, we observed specific expression of marker genes such as Tubb3, Elavl3, and Elavl4 (Fig. 8B). Based on this expression pattern, we concluded that a total of 10 cells in the R1 subpopulation are neuronal cells. Furthermore, significant expression of the Nos1 gene was observed specifically in R1-2, indicating the presence of nitrergic neurons, which are a specific subtype of neurons [46,47].

For the R2 cluster, comprising a total of 47 cells, all of which were identified as erythrocytes based on the presence of contamination markers such as Hbb-bs, Hba-a2, Hbb-a1, and Hbb-bt, that are associated with hemoglobin genes (Fig. 8C). Additionally, significant expression of erythrocyte markers, including Alas, was detected [48].

Through analysis of the dataset via screening for high expression of neuronal cell marker genes, we observed that neuronal populations constituted approximately 0.3% of the total cell count within the high-fat diet dataset. Therefore, scSID successfully identified rare neuronal cells and further classified them into subtypes, specifically nitrergic neurons, which represented approximately 0.043% of the total cells in the dataset.

5. Discussion

The advancement of transcriptomics in single cells has dramatically improved cell type identification. Furthermore, the detection of rare cell types remains challenging due to the presence of a large number of other cell types. In this study, we aim to address the need for an algorithm that combines efficiency, scalability, and accuracy for identifying rare cells in single-cell data. To achieve this, we propose a novel algorithm based on cell similarity division, which identifies rare cells by analyzing the variation of similarity differences between cells of the same class and those of different classes.

By comparing scSID with RaceID3, GiniClust2, CellSIUS, FiRE, and scLDS2 methods on 68K PBMC and 293T-Jurkat datasets, we observed that scSID has the ability to detect subtypes that may not be recognized using other single-cell analysis methods. Utilization of single-cell data demonstrates that scSID achieves high sensitivity and specificity in detecting subtypes within single-cell populations, thereby facilitating downstream classification of single cells. Notably, the application of scSID to the high-fat diet dataset uncovered previously unrecognized subtypes, highlighting its potential in identifying rare cell types. The high sensitivity of scSID contributes to the effective fractionation of single cells in downstream analyses. Measuring the similarity between cells from different perspectives should be considered in future work. In addition, scSID was designed for single-cell RNA sequencing data, and future studies should extend its application to multi-omics data.

CRediT authorship contribution statement

Shudong Wang: Conceptualization, Methodology, Validation. **Hengxiao Li:** Conceptualization, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Kuijie Zhang:** Conceptualization, Methodology, Writing – review & editing. **Hao Wu:** Writing – review & editing. **Shanchen Pang:** Funding acquisition, Supervision. **Wenhao Wu:** Writing – review & editing. **Lan Ye:** Writing – review & editing. **Jionglong Su:** Writing – review & editing. **Yulin Zhang:** 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.

Data availability

The study used several publicly available scRNA-seq datasets and one private dataset. 68K PBMC and 293T-Jurkat cell datasets are available at single-cell-gene-expression. The intestine dataset can be assessed at the GEO under accession code GSE123516. Moreover, for the high-fat diet dataset, due to the nature of this study, participants in this study did not agree to share their data publicly, and therefore supporting data could not be provided.

Code availability

The scSID source code is available at https://github.com/lgmzfl/scSID.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2021YFA1000102, 2021YFA1000103), Shandong Province Natural Science Foundation (ZR2020MH208, ZR2021MH104), the Young Taishan Scholars Program (tsqn201909178) and the Key Program State Fund in XJTLU (KSF-A-22).

References

[1] Papalexi Efthymia, Satija Rahul. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol 2018;18(1):35–45.
[2] Suvà Mario L, Tirosh Itay. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. Mol Cell 2019;75(1):7–12.
[3] Hwang Byungjin, Lee Ji Hyun, Bang Duhee. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018;50(8):1–14.
[4] Zhang Zilong, Cui Feifei, Lin Chen, Zhao Lingling, Wang Chunyu, Zou Quan. Critical downstream analysis steps for single-cell RNA sequencing data. Brief Bioinform 2021;22(5):bbab105.
[5] Rothchild Alissa C, Jayaraman Pushpa, Nunes-Alves Cláudio, Behar Samuel M. iNKT cell production of GM-CSF controls mycobacterium tuberculosis. PLoS Pathog 2014;10(1):e1003805.
[6] Cheng Qi, Zheng Hao, Li Ming, Wang Hongyi, Guo Xiaoxiao, Zheng Zhibo, et al. LGR4 cooperates with PrPc to endow the stemness of colorectal cancer stem cells contributing to tumorigenesis and liver metastasis. Cancer Lett 2022;540:215725.
[7] Lei Du, Cheng Qi, Zheng Hao, Liu Jinming, Liu Lei, Chen Quan. Targeting stemness of cancer stem cells to fight colorectal cancers. Seminars in cancer biology, vol. 82. Elsevier; 2022. p. 150–61.
[8] Kuo Yu-Hsuan, Lin Ching-Hung, Shau Wen-Yi, Chen Te-Jung, Yang Shih-Hung, Huang Shu-Min, et al. Dynamics of circulating endothelial cells and endothelial progenitor cells in breast cancer patients receiving cytotoxic chemotherapy. BMC Cancer 2012;12(1):1–9.
[9] Cima Igor, Kong Say Li, Sengupta Debarka, Tan Iain B, Phyo Wai Min, Lee Daniel, et al. Tumor-derived circulating endothelial cell clusters in colorectal cancer. Sci Transl Med 2016;8(345):345ra89.
[10] Grabski Isabella N, Kelly Street, Irizarry Rafael A. Significance analysis for clustering with single-cell RNA-sequencing data. Nat Methods 2023;20(8):1196–202.
[11] Wen Lu, Li Guoqiang, Huang Tao, Geng Wei, Pei Hao, Yang Jialiang, et al. Single-cell technologies: from research to application. Innovation 2022;3(6).
[12] Wang Hai-Yun, Zhao Jian-Ping, Zheng Chun-Hou, Su Yan-Sen. scGMAAE: Gaussian mixture adversarial autoencoders for diversification analysis of scRNA-seq data. Brief Bioinform 2023;24(1):bbac585.
[13] Jiang Jing, Xu Junlin, Liu Yuansheng, Song Bosheng, Guo Xiulan, Zeng Xiangxiang, et al. Dimensionality reduction and visualization of single-cell RNA-seq data with an improved deep variational autoencoder. Brief Bioinform 2023;24(3):bbad152.
[14] Wang HaiYun, Zhao JianPing, Zheng ChunHou, Su YanSen. scDSSC: deep sparse subspace clustering for scRNA-seq data. PLoS Comput Biol 2022;18(12):e1010772.
[15] Satija Rahul, Farrell Jeffrey A, Gennert David, Schier Alexander F, Regev Aviv. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015;33(5):495–502.
[16] McCarthy Davis J, Campbell Kieran R, Lun Aaron TL, Wills Quin F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 2017;33(8):1179–86.
[17] Grün Dominic, Lyubimova Anna, Kester Lennart, Wiebrands Kay, Basak Onur, Sasaki Nobuo, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 2015;525(7568):251–5.
[18] Jiang Lan, Chen Huidong, Pinello Luca, Yuan Guo-Cheng. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol 2016;17(1):1–13.
[19] Wegmann Rebekka, Neri Marilisa, Schuierer Sven, Bilican Bilada, Hartkopf Huyen, Nigsch Florian, et al. CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. Genome Biol 2019;20:1–21.
[20] Jindal Aashi, Gupta Prashant, Jayadeva, Sengupta Debarka. Discovery of rare cells from voluminous single cell expression data. Nat Commun 2018;9(1):4719.
[21] Wang Haiyue, Ma Xiaoke. Learning discriminative and structural samples for rare cell types with deep generative model. Brief Bioinform 2022;23(5):bbac317.
[22] Herman Josip S, Sagar, Gruen Dominic. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. Nat Methods 2018;15(5):379–86.
[23] Tsoucas Daphne, Yuan Guo-Cheng. GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. Genome Biol 2018;19:1–13.
[24] Fano Ugo. Ionization yield of radiations. II. The fluctuations of the number of ions. Phys Rev 1947;72(1):26.

[25] Lv Qin, Josephson William, Wang Zhe, Charikar Moses, Li Kai. Ferret: a toolkit for content-based similarity search of feature-rich data. In: Proceedings of the 1st ACM SIGOPS/EuroSys European conference on computer systems 2006; 2006. p. 317–30.
[26] Buettner Florian, Natarajan Kedar N, Casale F Paolo, Proserpio Valentina, Scialdone Antonio, Theis Fabian J, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol 2015;33(2):155–60.
[27] Peterson Leif E. K-nearest neighbor. Scholarpedia 2009;4(2):1883.
[28] Kiselev Vladimir Yu, Andrews Tallulah S, Hemberg Martin. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 2019;20(5):273–82.
[29] Van Der Maaten Laurens. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res 2014;15(1):3221–45.
[30] Wen Zhaoqi, Wang Xiaozhe, Hou Yanfang, Dong Yukun, Zhang Yulin. Nonlinear single-cell dimensionality reduction method based on scRNA-seq data. in Chinese, Math Model Appl 2023;12(3):33–44.
[31] Bro Rasmus, Smilde Age K. Principal component analysis. Anal Methods 2014;6(9):2812–31.
[32] Murtagh Fionn, Contreras Pedro. Algorithms for hierarchical clustering: an overview. Wiley Interdiscip Rev Data Min Knowl Discov 2012;2(1):86–97.
[33] Zheng Grace XY, Terry Jessica M, Belgrader Phillip, Ryvkin Paul, Bent Zachary W, Wilson Ryan, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8(1):14049.
[34] Ayyaz Arshad, Kumar Sandeep, Sangiorgi Bruno, Ghoshal Bibaswan, Gosio Jessica, Ouladan Shaida, et al. Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. Nature 2019;569(7754):121–5.
[35] Wolf F Alexander, Angerer Philipp, Theis Fabian J. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:1–5.
[36] Goutte Cyril, Gaussier Eric. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: European conference on information retrieval. Springer; 2005. p. 345–59.
[37] Altman Douglas G, Bland J Martin. Diagnostic tests. 1: sensitivity and specificity. BMJ, Br Med J 1994;308(6943):1552.
[38] Saito Takaya, Rehmsmeier Marc. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 2015;10(3):e0118432.
[39] Rosner Bernard, Glynn Robert J, Ting Lee Mei-Ling. Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. Biometrics 2003;59(4):1089–98.
[40] Fa Botao, Wei Ting, Zhou Yuan, Johnston Luke, Yuan Xin, Ma Yanran, et al. GapClust is a light-weight approach distinguishing rare cells from voluminous single cell expression profiles. Nat Commun 2021;12(1):4197.
[41] Ganesh Karuna, Basnet Harihar, Kaygusuz Yasemin, Laughney Ashley M, He Lan, Sharma Roshan, et al. L1CAM defines the regenerative origin of metastasis-initiating cells in colorectal cancer. Nat Cancer 2020;1(1):28–45.
[42] Wang Min, Windgassen Dirk, Papoutsakis Eleftherios T. Comparative analysis of transcriptional profiling of CD3+, CD4+ and CD8+ T cells identifies novel immune response players in T-cell activation. BMC Genomics 2008;9(1):1–16.
[43] Wang Yalong, Song Wanlu, Wang Jilian, Wang Ting, Xiong Xiaochen, Qi Zhen, et al. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. J Exp Med 2019;217(2):e20191130.
[44] Terasaki Paul I, Cai Junchao. Human leukocyte antigen antibodies and chronic rejection: from association to causation. Transplantation 2008;86(3):377–83.
[45] Mosaad YM. Clinical role of human leukocyte antigen in health and disease. Scand J Immunol 2015;82(4):283–306.
[46] Bódi Nikolett, Szalai Zita, Bagyánszki Mária. Nitrergic enteric neurons in health and disease—focus on animal models. Int J Mol Sci 2019;20(8):2003.
[47] McMenamin Caitlin A, Clyburn Courtney, Browning Kirsteen N. High-fat diet during the perinatal period induces loss of myenteric nitrergic neurons and increases enteric glial density, prior to the development of obesity. Neuroscience 2018;393:369–80.
[48] Bekri Soumeya, May Alison, Cotter Philip D, Al-Sabah Ala I, Guo Xiaojun, Masters Gillian S, et al. A promoter mutation in the erythroid-specific 5-aminolevulinate synthase (ALAS2) gene causes X-linked sideroblastic anemia. Blood 2003;102(2):698–704.