Computational and Structural Biotechnology Journal 27 (2025) 451–460

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Improving doublet cell removal efficiency through multiple algorithm runs

Yong She, Chaoye Wang, Qi Zhao *

State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong 510060, China

ARTICLE INFO

Keywords:
Single-cell RNA sequencing
Doublet removal
Multi-round doublet removal strategy
Synthetic dataset

ABSTRACT

Doublets are a key confounding factor in the analysis of scRNA-seq data, as they can interfere with differential expression analysis and disrupt developmental trajectories. However, due to the randomness of the algorithms, most doublet removal methods still leave a certain proportion of doublets after application. In this study, we proposed a multi-round doublet removal (MRDR) strategy, that ran the algorithm in cycles multiple times to effectively reduce randomness while enhancing the effectiveness of doublet removal. We evaluated the MRDR strategy in 14 real-world datasets, 29 barcoded scRNA-seq datasets, and 106 synthetic datasets with four popular doublet detection tools, including DoubletFinder, cxds, bcds, and hybrid. We found that in real-world datasets, the DoubletFinder had a better performance in MRDR strategy compared to a single removal of doublets and the recall rate improved by 50 % for two rounds of doublet removal compared to one round, and the performance of the other three doublet algorithms improved the ROC by about 0.04. In barcoded scRNA-seq datasets, we found that using cxds for two rounds of doublet removal yielded the best results. Subsequently, in simulated datasets, we proved that the multi-round removal strategy was more effective in removing doublets than a single removal, with cxds showing the best results when applied twice, and the ROC of the four methods during the two rounds of removal improved by at least 0.05 compared to single removal. Finally, compared to running the algorithm once, we found that the MRDR strategy was more beneficial for differential gene expression analysis and cell trajectory inference when using default analysis parameters. Overall, we proved that the MRDR strategy was more effective in removing doublets and advantageous for downstream analyses, and the strategy could be incorporated into the standard analysis pipeline for scRNA-seq experiments and recommend using cxds to remove doublets through two rounds of algorithm iteration.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) is a rapidly developing method of biological research in recent years, capable of comprehensive analysis of gene expression in individual cells [1]. However, during the distribution steps of scRNA-seq experiments, a droplet may encapsulate multiple cells, leading to the formation of doublets that appear as single cells [2,3]. The presence of doublets can potentially confound downstream analysis. These doublets can form false cell clusters, interfere with differential gene expression analysis, and obscure the inference of cell developmental trajectories. Doublets formed by different cell types were hard to annotate and can lead to false biological conclusions. Therefore, removing doublet was crucial for scRNA-seq data analysis.

So far, more than ten doublet detection methods have been developed based on different algorithms [4–10]. However, the performance of different methods can vary significantly. To assist in determining the most suitable method, a benchmark study conducted by Jingyi Jessica Li's team evaluated nine doublet detection methods. They found that DoubletFinder exhibited the highest detection accuracy, while cxds demonstrated both a high level of accuracy and the greatest computational efficiency [11]. However, It's worth noting that these evaluations show that even when using these tools, a portion of doublets may still be retained. We observed that the principle of DoubletFinder was as follows. DoubletFinder generated artificial doublets by averaging the gene expression profiles of two randomly selected droplets [4]. The doublet score for each droplet was defined as the proportion of artificial doublets in its k-nearest neighbors in PC space, with the dimension specified by the user. The number of neighbors (k) was chosen by maximizing the mean-variance-normalized doublet score distribution. The principle of doublets involved randomness, which led to the incomplete removal of doublets in a single attempt.

Several studies have shown that after a single round of doublet

* Corresponding author.
E-mail address: zhaoqi@sysucc.org.cn (Q. Zhao).

https://doi.org/10.1016/j.csbj.2025.01.009
Received 29 November 2024; Received in revised form 12 January 2025; Accepted 14 January 2025
Available online 15 January 2025

removal, multiple rounds of manual curation may be needed to achieve a more refined analysis by eliminating atypical doublets [12,13]. Inspired by this, we propose a strategy for multiple rounds of doublet removal, where each subsequent round builds upon the previous one, continuing until multiple rounds are done. This MRDR strategy removal minimizes the impact of randomization.To evaluate the strategy, We collected 14 real scRNA-seq datasets with doublet annotation [4,8,11, 14–19], 29 barcoded scRNA-seq datasets, and created 106 synthetic datasets. Using a doublet rate of cells/500 * 0.004 determined from 10x single-cell sequencing, We applied four different algorithms, including DoubletFinder, cxds, bcds, and hybrid, to evaluate the performance of the strategy for removing doublets.

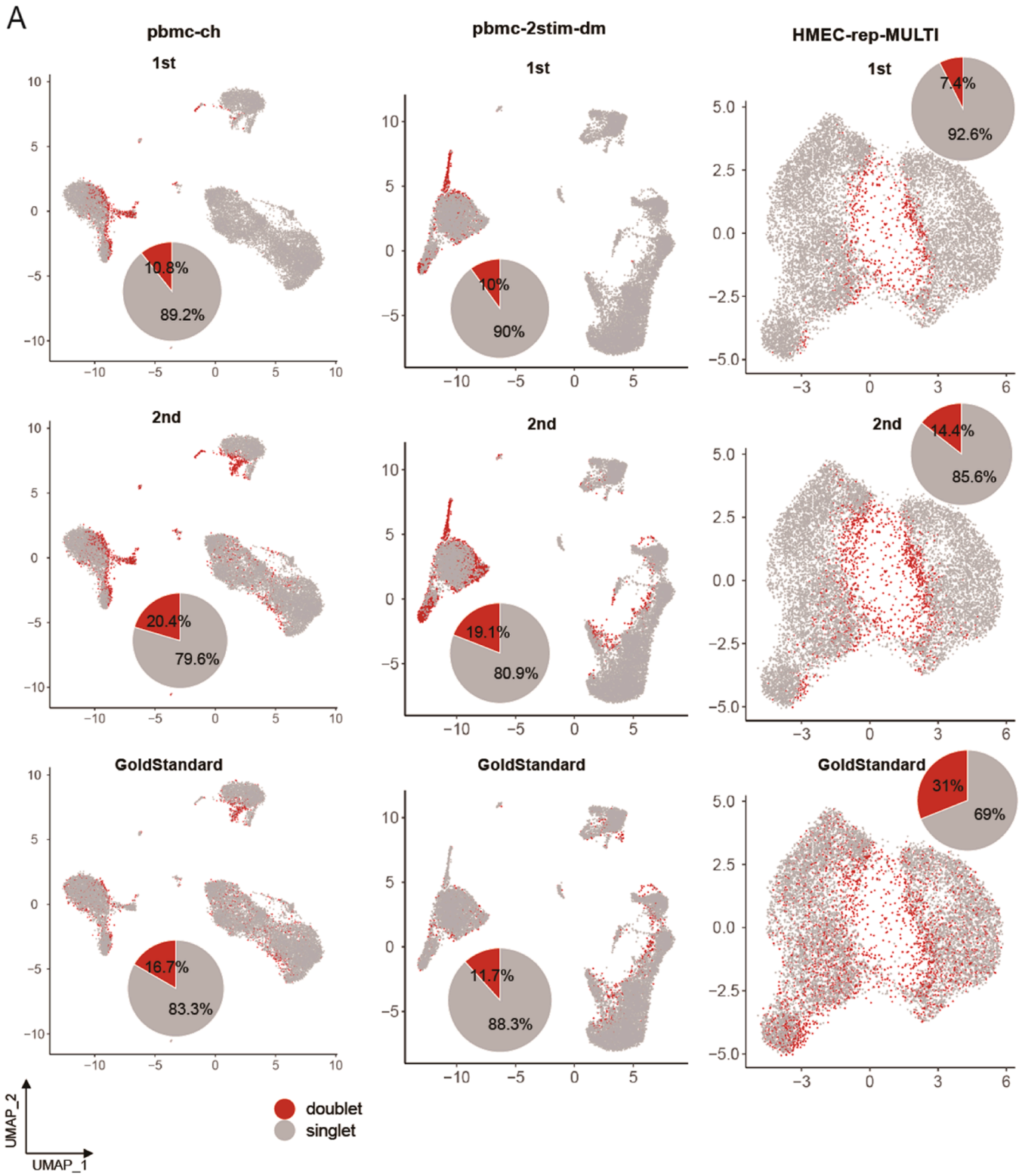

Fig. 1. Residual doublets were a widespread phenomenon. (A) Results of one and two removal using DoubletFinder in datasets with true doublet labels.

2. Result

2.1. Residual doublets evaluation after single-round removal

To initially explore the retention of doublets after using common doublet detection tools, we collected 16 public scRNA-seq datasets that had doublets annotated through experimental methods. In the two datasets, the doublets were identified as droplets containing cells from two different species, and all annotated doublets were easily recognizable. As a result, we removed these two datasets. Our collection included a variety of doublet rates, cell types, gene numbers, and sequencing depths (Table S1), thus indicating different degrees of challenge in identifying doublets from scRNA-seq data.

First, we examined the residual presence of double cells after a single removal attempt in the collected real dataset. In the dataset with doublet annotation (pbmc-ch), we observed that even after a single round of removal, a considerable proportion of doublet cells persisted in pbmc-ch (Fig. 1A). Some of these non-removed doublets formed clusters, while

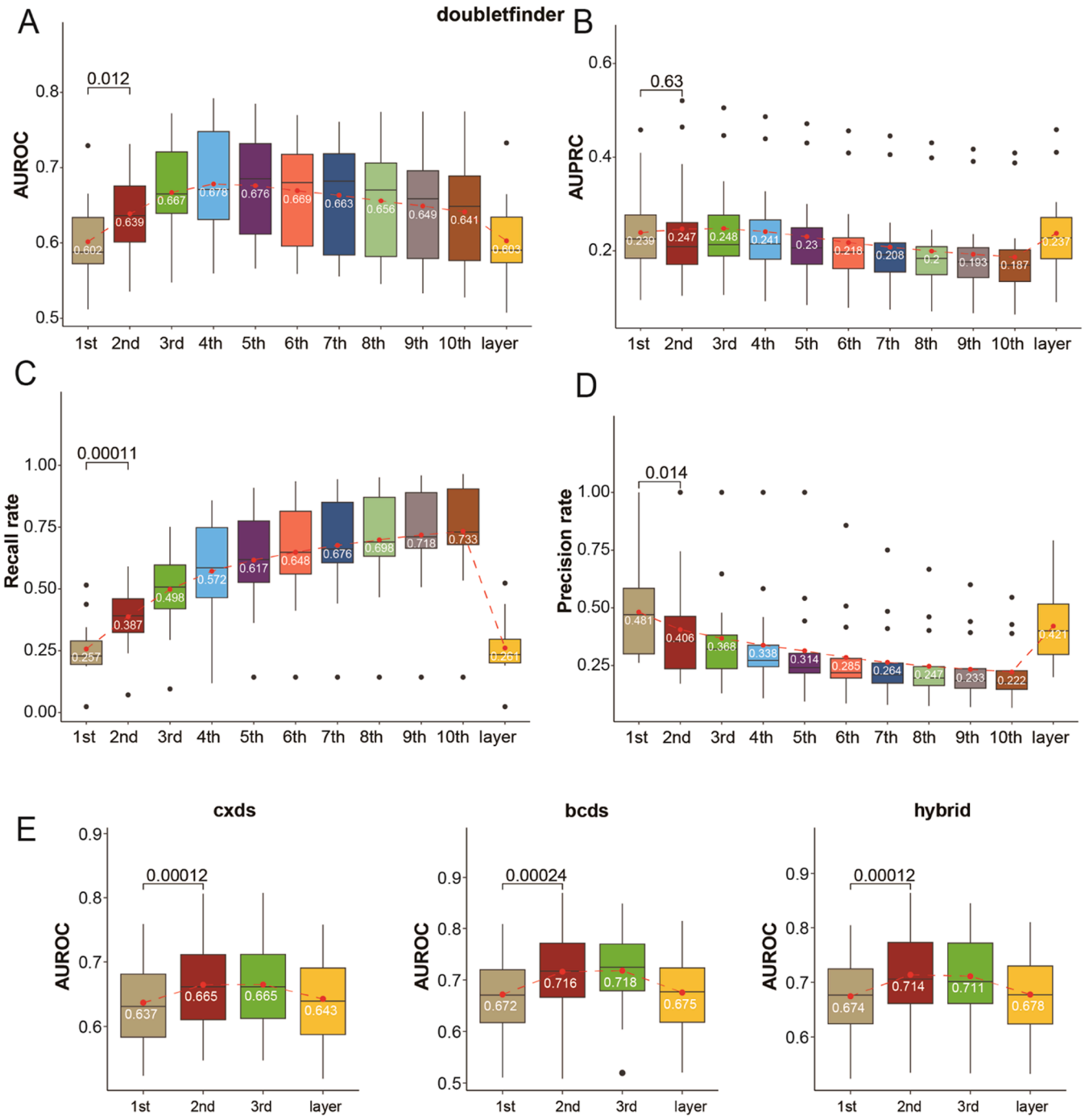

Fig. 2. Evaluation of the MRDR strategy using 14 Benchmark real scRNA-seq datasets. (A) Comparison of AUROC values from one removal to 10 doublet removals and hierarchical removals using the DoubletFinder. (B) Comparison of AUPRC values from one removal to 10 removals and hierarchical removal using the DoubletFinder. (C) Comparison of recall rates from one removal to 10 removals and hierarchical removal using the DoubletFinder. (D) Comparison of precision rates from one removal to 10 doublet removals and hierarchical removal using the DoubletFinder (E) Comparison of AUROC for one removal and two removals of the other three methods.

others were scattered. Surprisingly, the second round of removal eliminated many doublets that were not removed in the first round (Fig. 1A). Similar phenomena were observed in pbmc-2stim-dm and HMEC-rep-MULTI (Fig. 1A). These findings confirmed that the persistence of doublets after a single removal attempt is common, and that a second round of removal could eliminate many of the doublets overlooked initially.

2.2. The MRDR strategy was effective for removing doublets on real scRNA-Seq datasets

Since DoubletFinder is a widely used method in doublet detection and has shown the highest accuracy in the benchmark evaluation, we initially tested the impact of applying DoubletFinder for single removal up to ten times, as well as hierarchical doublet removal. We assessed their detection accuracy based on their AUPRC and AUROC values. A significant performance improvement in AUROC with multiple removals was obtained, especially at the second and third iterations (Fig. 2A-D, Figure S1A). As the number of removals increased, the detection accuracy gradually decreased, but the recall rate increased. We noted that the recall rate of the two-round removal improved by 13 % compared to the single round, and the ROC increased by 3 % These results confirmed that using the MRDR strategy was effective in DoubletFinder.

2.3. Using multiple tools to evaluate the performance of the MRDR strategy on real scRNA-Seq datasets

Based on the previous tests, we examined whether the MRDR strategy had a similar effect to other methods. We then evaluated the impact of multiple removal using algorithms including cxds, bcds, and hybrid methods. To note, these methods removed doublets bearing on different principles. cxds utilized binarized (absence/presence) gene expression data and, employing a binomial model for the co-expression of pairs of genes, yields interpretable doublet annotations. bcds used a binary classification approach to discriminate artificial doublets from original data. The hybrid normalized the doublet scores of cxds and bcds to values between 0 and 1, and the doublet score of each droplet was defined as the sum of the two normalized doublet scores [5]. As a result, the AUROC indicated that two-round removals were significantly more effective than single removal (Fig. 2E, Figure S1B-D, Figure S2). These results likewise suggested that using the MRDR strategy was effective.

To further explore whether hierarchical doublet removal (group layer) could improve performance, we analyzed its effect with four different tools across multiple datasets. The strategy of hierarchical doublet removal is that the first run removal of doublets was followed by the second doublet removal based on cluster information. However, we found that hierarchical doublet removal didn't significantly enhance performance (Fig. 2, Figure S2).

2.4. Evaluating the performance of the MRDR strategy on barcoded scRNA-seq datasets

Although we collected all scRNA-Seq datasets containing doublet annotations (removing two mixed-species datasets), none of the datasets had completely accurate annotations due to experimental design reasons. The six Demuxlet datasets only labeled doublets formed by two individuals, and many homogeneous doublets were not annotated in the fourteen real datasets. Consequently, the incompleteness of doublet annotations likely reduced the accuracy of our benchmarking. To overcome the issue of incomplete doublet annotations in real datasets, we employed a previous study to identify singlets from barcoded scRNA-seq datasets and obtained real-world scRNA-seq datasets that exclusively included singlets [20] (Table S2). Based on doublet rates of 0.05 and 0.1, we randomly selected two cells, averaged their gene expression profiles to generate simulated doublets, and removed the singlets that were used to generate the doublets. We then used these barcoded scRNA-seq datasets to evaluate the performance of the MRDR strategy.

Fig. 3 demonstrated that the MRDR strategy could identify some true doublets that were overlooked by the single-round detection on barcoded scRNA-seq datasets. We noticed that the performance of bcds and bybrid using the MRDR strategy showed little improvement, while DoubltFinder and cxds demonstrated significant performance enhancements with the MRDR strategy. Among these, cxds achieved the best results through two rounds of algorithm iteration. Overall, our test results indicated that the MRDR strategy was more effective than single-round doublet removal on barcoded scRNA-seq datasets, with cxds yielding the best outcome after two rounds of doublet removal.

2.5. The MRDR strategy improved the accuracy of doublet removal on synthetic scRNA-Seq datasets

To thoroughly evaluate the performance of the MRDR strategy under various biological conditions, we generated 100 synthetic scRNA-seq datasets using scDesign [21], featuring different doublet rates, sequencing depths, cell types, and levels of inter-cell type heterogeneity. Using synthetic data was advantageous for benchmarking the MRDR strategy, as we could flexibly and comprehensively alter the experimental settings and biological conditions, while also knowing the true doublets.

Fig. 4 demonstrated that the two-round doublet removal strategy could identify many true doublets that were overlooked by the single-round detection on synthetic scRNA-Seq datasets. Fig. 4A and B showed how the performance of the MRDR strategy varied under different doublet rates, sequencing depths, numbers of cell types, or levels of heterogeneity among cell types. First, as the doublet rate increased, AUROC gradually decreased while AUPRC first increased, then decreased, and finally rose again. Consistent with expectations, we found that using four methods with two rounds of removing double cells performed significantly better than a single round of removal and the two-round removal using cxds yielded the best performance. We then found that different sequencing depths, cell types, and levels of inter-cell type heterogeneity did not have a significant impact on the performance of doublet removal. Surprisingly, the AUROC and AUPRC of these methods also indicated that two-round removal was significantly more effective than one-round removal (Fig. 4A-B). We observed that the ROC of simulated datasets under almost all biological conditions improved by at least 5 % during two-round removal compared to single-round removal. Notably, cxds showed a 10 % improvement in ROC during two-round removal compared to single removal. In addition, we noted that as the number of cell types increased, the performance of DoubletFinder improved in both one-round and two-round removal processes (Fig. 4A-B, Figure S4). This observation is quite understandable; with a greater diversity of cell types, the simulated doublets generated by DoubletFinder began to closely resemble the actual doublets found in the dataset, thereby enhancing its ability to detect these doublets. Meanwhile, in practical analyses of single-cell RNA sequencing data, the number of cell types is typically quite high. Overall, Our testing results suggested that two rounds of doublet removal were more effective than a single round of double cell removal on the synthetic datasets.

2.6. The impact of doublet detection on DE

It was well known that the presence of doublets in scRNA-seq datasets was anticipated to complicate the downstream differential expression gene analysis. Consequently, if the MRDR strategy proved to be effective, its removal of doublets should have enhanced the accuracy of differential expression gene analysis.

To evaluate the effectiveness of the MRDR strategy from this perspective, we used scDesign to generate four synthetic scRNA-seq datasets. These datasets included two cell types and 1154 between-cell-type DE genes (10 % of a total of 11,543 genes) [21]. We labeled this dataset as the "clean data." Afterward, we blended each cell type with randomly formed doublets, targeting doublet rates of 5 %, 10 %,

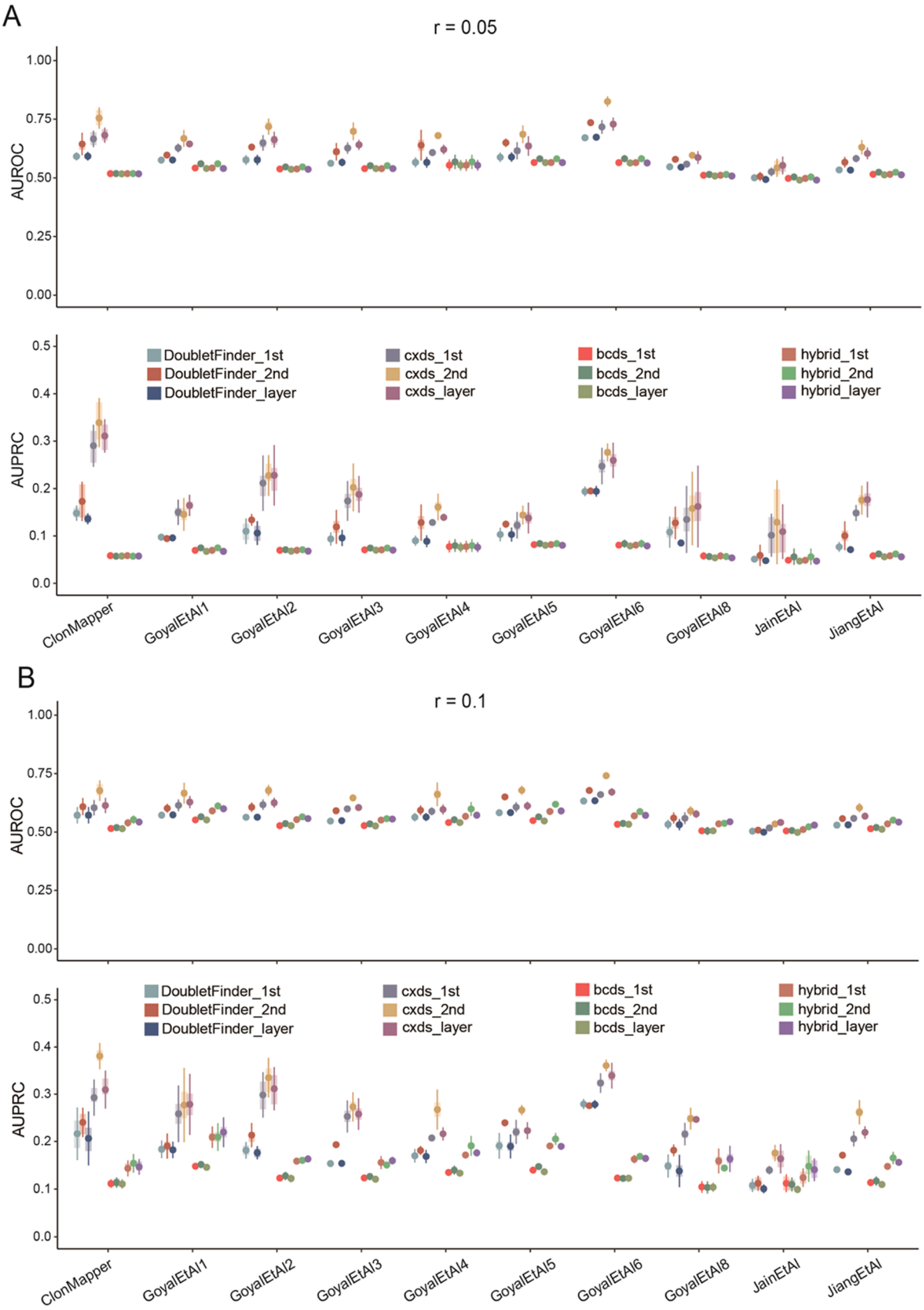

Fig. 3. Evaluation of the MRDR strategy using 29 barcode scRNA-seq datasets. (A) Each barcode scRNA-seq dataset constructed doublets with a doublet rate of 0.05. (B) Each barcode scRNA-seq dataset constructed doublets with a doublet rate of 0.1.

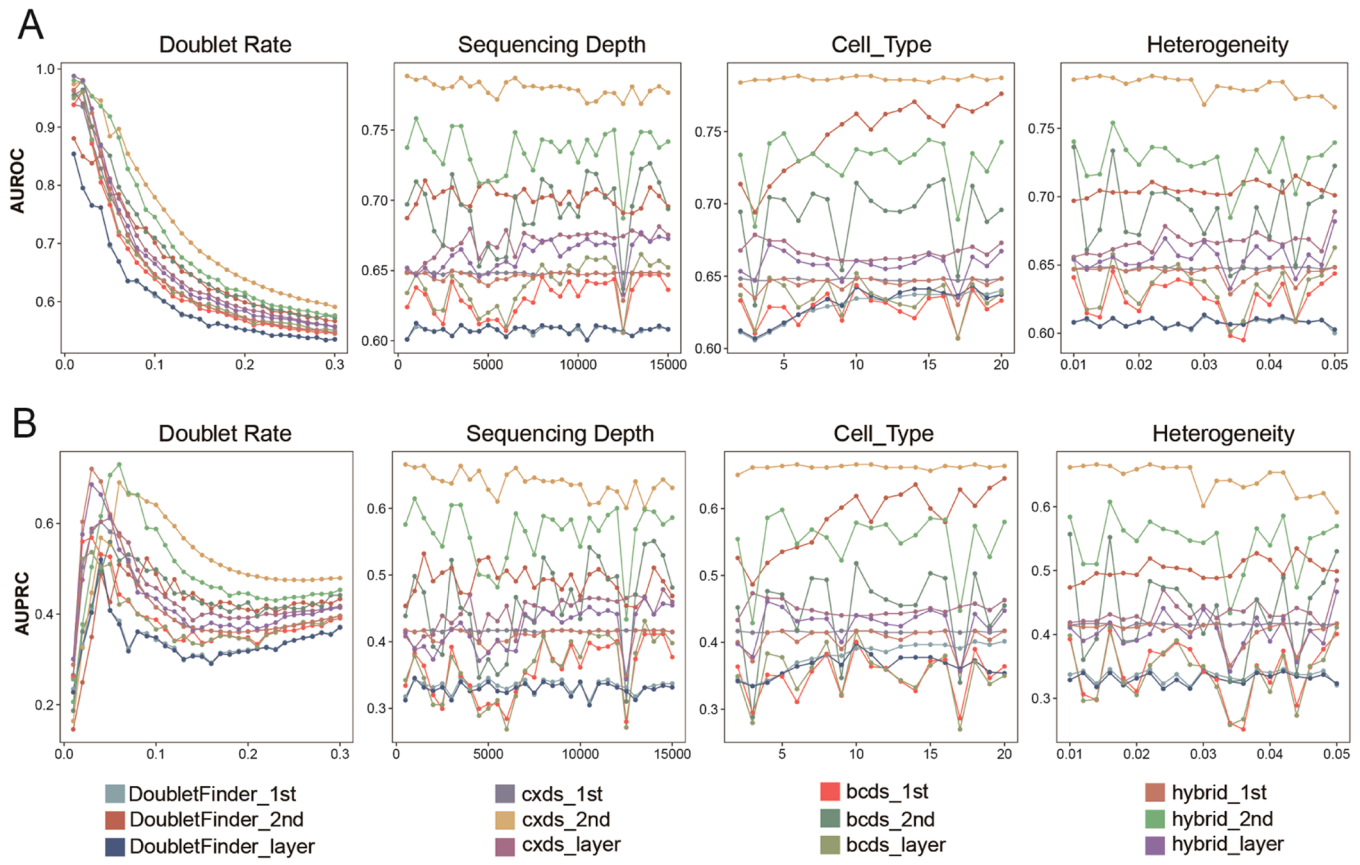

Fig. 4. Evaluation of MRDR strategy using 100 synthetic scRNA-seq datasets based on four simulation settings. (A) The AUROC values of MRDR strategy for each doublet detection method were evaluated across four distinct simulation settings: varying doublet rates (from 1 % to 30 % with a step size of 1 %), varying sequencing depths (from 500 to 15,000 UMI counts per cell with a step size of 500 counts), varying numbers of cell types (from 2 to 20 with a step size of 1), and 21 heterogeneity levels which specified the degree of differentiation of genes between the two cell types. (B)The AUPRC values of MRDR strategy for each doublet detection method were evaluated across four distinct simulation settings: varying doublet rates (from 1 % to 30 % with a step size of 1 %), varying sequencing depths (from 500 to 15,000 UMI counts per cell with a step size of 500 counts), varying numbers of cell types (from 2 to 20 with a step size of 1), and 21 heterogeneity levels which specified the degree of differentiation of genes between the two cell types.

15 %, and 20 %, and the resulting dataset was referred to as the "contaminated data". Then, we applied the MRDR strategy to these four datasets to eliminate doublets and subsequently used the widely used method Wilcox to calculate DE genes [22]. We found that the positive control group had the highest precision, while the negative control exhibited the lowest precision. Furthermore, we discovered that the MRDR strategy was effective for each doublet detection method. Specifically, the precision of DE genes for each algorithm showed significant improvement with two rounds of doublet removal compared to a single round (Fig. 5A). Overall, these results demonstrated that the MRDR strategy was more effective in eliminating doublets and was advantageous for differential gene expression analysis.

2.7. The impact of doublet detection on trajectory inference

Cell trajectory inference relies on the similarity of gene expression profiles to estimate cell trajectories, often referred to as pseudotime. The accuracy of trajectory inference depended both on the quality of scRNA-seq data and the method used for inference. Therefore, the presence of doublets in the dataset could introduce bias into trajectory inference, particularly atypical doublets that lead to false branches [23–26]. Previous results confirmed that the MRDR strategy significantly improved overall performance. It could be reasonably assumed that the cell trajectory might have become more accurate after the MRDR strategy.

To evaluate the effectiveness of the MRDR strategy from this perspective, we generated two scRNA-seq datasets with cell trajectories using the splatSimulatePaths function from Splatter [27]. Both datasets exhibited branching trajectories. Then, doublets were introduced based on different doublet rates (0.07 and 0.12), and the single cells used to create doublet were removed. Similar to DE gene analysis, we used the MRDR strategy to remove doublet from each contaminated dataset. We then used Slingshot to infer cell trajectories [28], as it had demonstrated top performance in previous studies. As expected, we found that removing doublets once using each tool resulted in a significant deviation between the inferred trajectories and the true trajectories. However, after two rounds of doublet removal, this deviation was reduced or eliminated for some of the tools (Fig. 5B-C). These results indicated that the MRDR strategy was more effective in eliminating doublets and facilitating cell trajectory inference.

3. Discussion

With the rapid advancement of scRNA-seq technology, there has been a proliferation of computational methods for various scRNA-seq data analyses. In the distribution step of scRNA-seq experiments, it's possible for a single droplet to encapsulate two or more cells, resulting in doublets [1–3]. To date, more than ten tools have been developed to remove doublets [4,5,11,28]. However, it has been observed in practical single-cell data analysis that even after the initial removal of doublets, some cell clusters express both T-cell and myeloid markers or markers from other major cell groups, indicating the presence of residual doublets. This suggested that residual doublets were a widespread

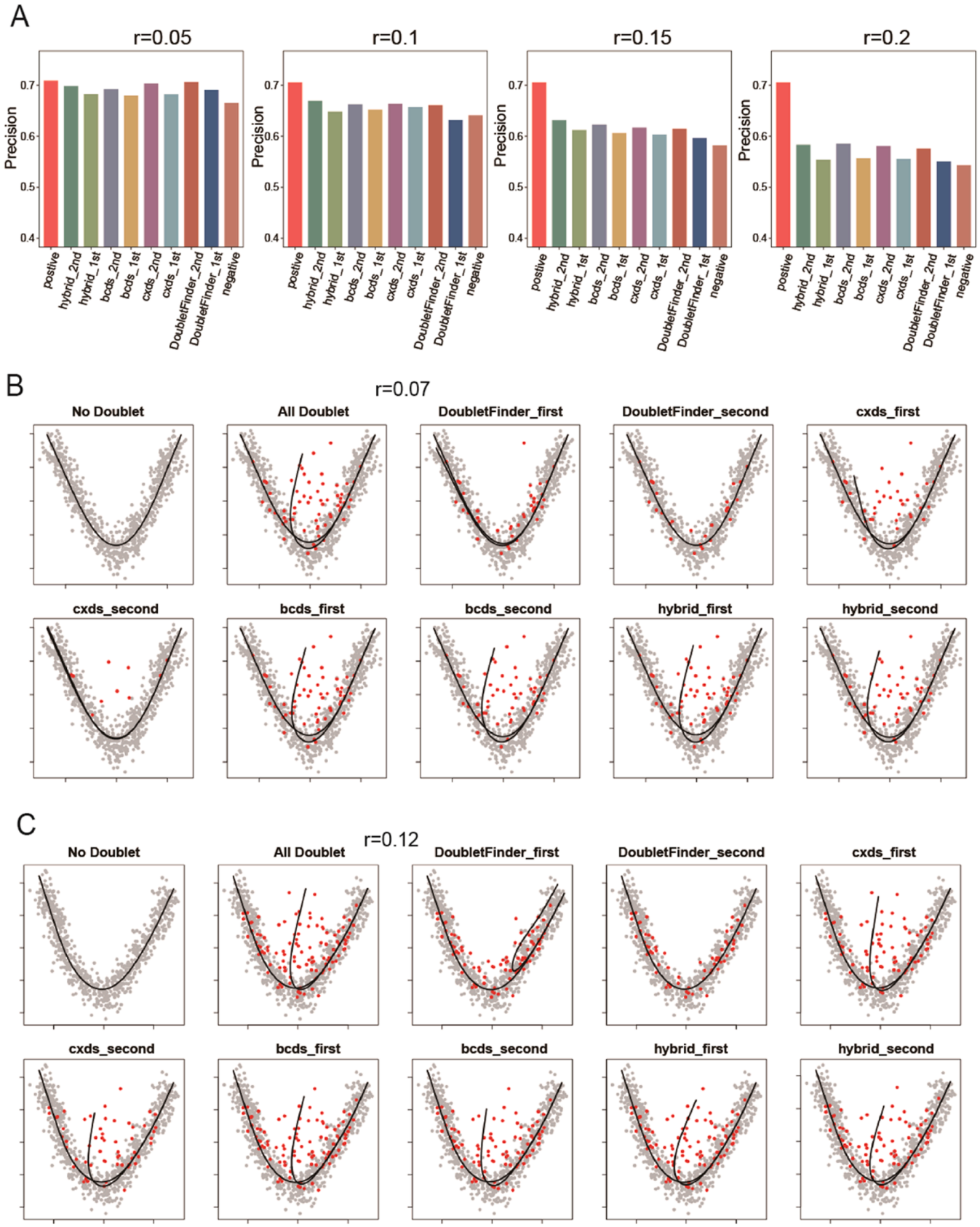

Fig. 5. MRDR strategy influenced the computation of differentially expressed genes and the inference of cell trajectory. (A) Comparison of precision of differentially expressed genes after removing doublets using MRDR strategy across four distinct doublet rates. (B) Trajectories constructed by Slingshot after MRDR strategy were applied to remove identified doublets in a synthetic dataset with 7 %. The true cell trajectory was branched. (C) Trajectories constructed by Slingshot after MRDR strategy were applied to remove identified doublets in a synthetic dataset with 12 %. The true cell trajectory was branched.

phenomenon even after the initial doublet removal. Some single-cell studies have proposed strategies that involve using DoubletFinder to remove doublets and then multiple rounds of manual annotation to further eliminate doublets [12,13,29]. Inspired by this, we proposed the MRDR strategy, where each subsequent round built upon the previous one.

To provide the first comprehensive benchmarking of MRDR strategy, we evaluated it across four doublet detection algorithms using 14 real scRNA-seq datasets, 29 barcoded scRNA-seq datasets, and 106 synthetic scRNA-seq datasets, assessing its overall detection accuracy and impact on downstream analyses. Our tests demonstrated that the MRDR strategy was more effective in eliminating doublets and was advantageous for differential gene expression analysis and trajectory inference. This may be attributed to the algorithms themselves. Methods like DoubletFinder often construct artificial doublets by summing or averaging real droplet counts [4]. The simulated artificial doublets are unlikely to cover all real doublets, so after the first round of doublet removal, the second round of simulated doublets may encompass some that were not removed in the first round. Hence, multiple rounds of removal naturally lead to the elimination of more real doublets, resulting in better overall performance when compared to a single round of removal. The iterative process of the multi-round algorithm effectively reduced the impact of randomness.

Based on our evaluation of the MRDR strategy in barcoded scRNA-seq datasets and synthetic scRNA-seq datasets, the MRDR strategy significantly improved the efficiency of doublet removal. Notably, cxds showed the best performance when using the MRDR strategy, with both AUROC and AUPRC increasing by more than 5 %. At the same time, the computational speed of cxds was also particularly fast (completing in under two minutes). This also enhanced the feasibility of cxds in practical applications.

During our benchmarking process, we identified a critical unresolved issue: how to accurately estimate the doublet rate in scRNA-seq datasets. Essentially, each doublet detection algorithm required us to provide an expected doublet rate, which would determine how many doublets were removed. If the expected doublet rate was higher than the actual rate, it could lead to the removal of some rare cells, resulting in the loss of biologically significant cells. To address this, we needed to obtain prior knowledge from previously assessed sequencing data to estimate the doublet rate in our dataset.

In summary, doublet removal was crucial for the analysis of scRNA-seq data. We then proposed a strategy for multi-round doublet removal and benchmarked it using four doublet detection tools. Our tests indicated that using the MRDR strategy was effective, and especially the cxds using the MRDR strategy had the best performance. So, we propose that this strategy should be incorporated into standard analysis workflows for scRNA-seq, and recommend using cxds to remove doublets through two rounds of algorithm iteration.

4. Materials and methods

4.1. Doublet detection methods description

Several doublet-detection methods were implemented in the study. The parameters for the software used in these methods were set to their recommended values or to default values when no recommended values were available. A random seed was saved in our code to ensure reproducibility. Below are the detailed configurations for each software.

4.1.1. Doubletfinder

The method was carried out according to the instructions found at https://github.com/chris-mcginnis-ucsf/DoubletFinder. The parameters were set to their default values, and the doublet scores were obtained using the function doubletFinder_v3() in the R package DoubletFinder (v 2.0.3).

4.1.2. scds

The R package scds (v1.16.0) included three methods: cxds, bcds, and hybrid. These methods were executed by following the instructions provided at https://github.com/kostkalab/scds. The doublet scores were obtained separately using the functions cxds(), bcds(), and cxds_bcds_hybrid() in the R package scds.

4.2. Measures of double detection accuracy

Computational doublet detection methods used binary classification algorithms to differentiate between two categories: doublet and singlet. To assess the accuracy of the MRDR strategy, four measures were calculated in the study.

4.2.1. AUROC

The area under the receiver operation characteristic(AUROC) is a performance metric used to evaluate classification models. The AUROC is calculated as the area under the ROC curve and lies between 0 and 1. The bigger the AUROC is, the higher accuracy the model has. The measures are calculated using function *roc.curve ()* in R package PRROC (v 1.3.1).

4.2.2. AUPRC

The area under the precision-recall curve (AUPRC) is a performance metric for imbalanced data in a problem setting where care about finding the positive examples. If the model achieves a perfect AUPRC, it means models find all of the positive examples without marketing any negative examples as positive. The measure could be calculated using function *pr.curve ()* in R package PRROC (v 1.3.1).

4.2.3. Recall rate

The recall rate is a vector to evaluate the proportion of number of true positive samples predicted by classification models to the number of actual positive samples. It is defined as follows:

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN})$$

TP (true positive): the samples are actually positive and predicted as positive.

FN (false negative): the samples are actually positive but predicted as negative.

FP (false positive): the samples are actually negative but predicted as positive.

4.2.4. Precision rate

The precision rate is a vector to evaluate the proportion of number of true positive samples predicted by classification models to the number of all samples predicted as positive. It is defined as follows:

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$$

4.3. The collection of real scRNA-seq datasets with doublet label

In our research, we used some publicly available datasets from *Xi and Li* [4,8,11,14–17], which contained 14 real scRNA-seq datasets with experimentally annotated doublets. These datasets can be available at https://zenodo.org/record/4062232#.X3YR9Hn0kuU%E3%80%82.

4.4. Identification of singlets and construction of doublets in barcode scRNA-seq datasets

To test the MRDR strategy on more real-world scRNA-seq datasets, we utilized a previous study to identify singlets from barcoded scRNA-seq datasets and obtained real-world scRNA-seq datasets that contained only singlets [20]. Based on doublet rates of 0.05 and 0.1, we

randomly selected two cells, averaged them to generate simulated doublets, and removed the singlets used to generate the doublets.

4.5. Synthetic scRNA-seq dataset containing doublets

All synthetic scRNA-seq datasets used in this study were produced in two steps. In Step 1, singlets in each dataset were produced by scDesign [21]. In Step 2, with the given number of singlets and a predetermined doublet rate, the corresponding number of doublets was generated by random pairing of singlets. Specifically, the gene expression profiles (in UMI counts) of two randomly selected singlets were averaged by gene, and this averaged profile was referred to as a prototype doublet. For each of the 14 real scRNA-seq datasets, a doublet-to-singlet size ratio, which was defined as (average doublet library size)/(average singlet library size), was computed. Finally, the singlets used to generate doublets were removed, excluding the DE gene section. In mathematical terms, if N singlets had been produced in Step 1 and the doublet rate had been R (a value between 0 and 1), then after Step 2, the quantities of singlets and doublets were calculated as N(1-R)/(1 +R) and NR/(1 +R), respectively, both rounded to the nearest integers. For instance, if 1500 singlets were generated in Step 1 and the doublet rate was 10 %, the numbers of doublets and singlets in the final dataset were 136 and 1227, respectively, resulting in a total of 1363 droplets.

4.6. Simulated scRNA-seq datasets in benchmarking simulations

100 scRNA-seq datasets were generated by scDesign to benchmark MRDR strategy across four aspects: varying doublet rates, sequencing depths, cell types, and levels of heterogeneity between cell types [21]. (1) 30 synthetic datasets were created with doublet rates ranging from 1 % to 30 %, increasing in steps of 1 %. Each dataset included two cell types. The per-cell library size was set to 5000 UMI counts. 2000 singlets were produced for each cell type. In Step 2, doublets were introduced according to each specified doublet rate, and the singlets that were used to create the doublets were removed. (2) 30 synthetic datasets were created with per-cell library sizes ranging from 500 to 15,000 UMI counts, increasing in increments of 500 counts. Each dataset consisted of two cell types. According to the data generation scheme outlined in the previous subsection 2, 000 singlets were produced for each cell type in Step 1. In Step 2, doublets were introduced at a rate of 10 %, and the singlets used to form the doublets were subsequently removed. (3) A total of 19 synthetic datasets were produced, with the number of cell types fluctuating between 2 and 20 in increments of 1. The per-cell library size was established at 5000 UMI counts. According to the data generation scheme outlined in the previous subsection 2000 singlets were produced for each cell type in Step 1. In Step 2, doublets were introduced with a 10 % doublet rate, and the singlets utilized to generate the doublets were eliminated. (4) A total of 21 synthetic datasets were produced, exhibiting varying degrees of heterogeneity between two cell types. The heterogeneity level was controlled by four parameters (pUp, pDown, fU, and fL) in scDesign. In particular, pUp and pDown represented the proportions of up-regulated and down-regulated genes, whereas fU and fL specified the upper and lower bounds of fold changes in the expression levels of differentially expressed (DE) genes. The parameter combinations described below were implemented to create the 21 levels of heterogeneity:

Level 1: pDown = 0.010, pUp = 0.010, fL = 0.5, and fU = 1.0;
Level 2: pDown = 0.012, pUp = 0.012, fL = 0.6, and fU = 1.2;
Level 3: pDown = 0.014, pUp = 0.014, fL = 0.7, and fU = 1.4;
…
Level 21: pDown = 0.050, pUp = 0.050, fL = 2.5, and fU = 5.0.

The per-cell library size was established at 5000 UMI counts. In Step 2, doublets were introduced with a 10 % doublet rate, and the singlets utilized to generate the doublets were eliminated.

4.7. DE gene analysis

scDesign generated four synthetic scRNA-seq datasets comprising two cell types [21]. Each cell library had a size of 5000 UMI counts. In scDesign, both the pUp and pDown parameters were adjusted to 0.05, indicating that 10 % of the genes were differentially expressed (5 % were upregulated and 5 % were downregulated) between the two cell types. The fL and fU parameters (lower and upper limits of fold change for DE genes) were set to 1.5 and 3, respectively. For each cell type, 1500 single cells were created in Step 1. In Step 2, the four datasets had introduced doublets at various rates (0.05, 0.1, 0.15, and 0.2). In this instance, the singlets employed to introduce the doublets were retained, as their removal would have impacted the true differentially expressed genes between the two cell types. The Wilcoxon rank-sum test was applied to this dataset (the “contaminated dataset,” which contained both singlets and doublets), a clean version without doublets (the “clean dataset,” which contained only singlets), and the post-doublet detection versions after applying MRDR strategy. After the MRDR strategy had been applied to each dataset, genes with Bonferroni-corrected p-values not greater than 0.05 were recognized as DE genes. Precision was computed as a performance metric for each group of identified DE genes. The results from the contaminated dataset were used as a negative control, whereas the clean dataset served as a positive control.

4.8. Cell trajectory inference

In Step 1, we generated two scRNA-seq datasets with cell trajectories using the splatSimulatePaths function from Splatter [27]. Each dataset contained 1000 genes. In Step 2, doublets were introduced based on different doublet rates (0.07 and 0.12), and singlets used to create doublets were removed. Within the parameters of Splatter, de.prob and de.facLoc were set to 0.5 and 0.2, respectively, while all other parameters were kept at their default values. Each dataset was expanded into a suite that included its original version (the “contaminated dataset”), a clean version without doublets (the “clean dataset”), and the post-doublet detection versions after applying each doublet detection method once and twice. For each group of datasets, cell trajectories were inferred using Slingshot [28].

Code availability

We uploaded the specific code for the MRDR strategy to GitHub, which could provide some reference for users (https://github.com/sheyong111/Doublet_Guidance).

CRediT authorship contribution statement

Qi Zhao: Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **Yong She:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Chaoye Wang:** Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Data curation.

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 was supported by the National Key Research and Development Program of China (2021YFA1302100); the National Natural Science Foundation of China (82172861); Young Talents Program of Sun

Yat-sen University Cancer Center (YTP-SYSUCC-0033).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.01.009.

References

[1] Tang X, Huang Y, Lei J, Luo H, Zhu X. The single-cell sequencing: new developments and medical applications. Cell Biosci 2019;9:53.
[2] Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med 2017;9:75.
[3] Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018;50:1–14.
[4] McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst 2019;8:329–337.e4.
[5] Bais AS, Kostka D. scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics 2020;36:1150–8.
[6] Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst 2019;8:281–291.e9.
[7] Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research 2016;5:2122.
[8] Bernstein NJ, et al. Solo: doublet identification in single-cell RNA-seq via semi-supervised deep learning. Cell Syst 2020;11:95–101.e5.
[9] DePasquale EAK, et al. DoubletDecon: deconvoluting doublets from single-cell RNA-sequencing data. Cell Rep 2019;29:1718–1727.e8.
[10] Germain P-L, Lun A, Garcia Meixide C, Macnair W, Robinson MD. Doublet identification in single-cell sequencing data using scDblFinder. F1000Research 2022;10:979.
[11] Xi NM, Li JJ. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell Syst 2021;12:176–194.e6.
[12] Wang R, et al. Evolution of immune and stromal cell states and ecotypes during gastric adenocarcinoma progression. Cancer Cell 2023. https://doi.org/10.1016/j.ccell.2023.06.005.
[13] Dang M, et al. Single cell clonotypic and transcriptional evolution of multiple myeloma precursor disease. Cancer Cell 2023;41:1032–1047.e4.
[14] MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices | Nature Methods. ⟨https://www.nature.com/articles/s41592-019-0433-8⟩.
[15] Stoeckius M, et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol 2018;19:224.
[16] Massively parallel digital transcriptional profiling of single cells | Nature Communications. ⟨https://www.nature.com/articles/ncomms14049⟩.
[17] Kang HM, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol 2018;36:89–94.
[18] Pan X, et al. Identifying a confused cell identity for esophageal squamous cell carcinoma. Signal Transduct Target Ther 2022;7:1–11.
[19] Pai JA, et al. Lineage tracing reveals clonal progenitors and long-term persistence of tumor-specific T cells during immune checkpoint blockade. Cancer Cell 2023;41:776–790.e7.
[20] Zhang Z, et al. Synthetic DNA barcodes identify singlets in scRNA-seq datasets and evaluate doublet algorithms. Cell Genom 2024;4:100592.
[21] Li WV, Li JJ. A statistical simulator scDesign for rational scRNA-seq experimental design. Bioinformatics 2019;35:i41–50.
[22] Fay, M.P. & Proschan, M.A. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. (2010).
[23] Trapnell C, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381–6.
[24] Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments | Nature Methods. ⟨https://www.nature.com/articles/s41592-019-0425-8⟩.
[25] Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics | BMC Genomics | Full Text. ⟨https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4772-0⟩.
[26] Wang K, et al. PhyloVelo enhances transcriptomic velocity field mapping using monotonically expressed genes. Nat Biotechnol 2023:1–12. https://doi.org/10.1038/s41587-023-01887-5.
[27] Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol 2017;18:174.
[28] Street K, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genom 2018;19:477.
[29] Ghaddar B, De S. Reconstructing physical cell interaction networks from single-cell data using neighbor-seq. Nucleic Acids Res 2022;50:e82. –e82.