Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A systematic benchmark and practical guide for rare-cell detection in single-cell RNA-seq data

Ke Liu

keliu.iluke@hotmail.com

ShanDong University https://orcid.org/0000-0003-3190-1656

Xiuyuan Jin

Shandong University

Chunyang Fu

Shandong University https://orcid.org/0009-0008-4775-667X

Article

Keywords:

Posted Date: September 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10583472/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

A systematic benchmark and practical guide for rare cell detection in single-cell RNA-seq data

Xiuyuan Jin[1], Chunyang Fu[1], Ke Liu[1,*]

Affiliation:

1. Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China.

* Correspondence: keliu.iluke@email.sdu.edu.cn

Abstract

Discovery of rare cell populations from single-cell RNA sequencing (scRNA-seq) data remains challenging. Despite the development of multiple methods, their performance and practical utility remain incompletely characterized. Here, we systematically benchmark eleven published methods using large-scale benchmarking datasets generated from real scRNA-seq datasets through cell-type-specific downsampling. We find that aKNNO and RareQ exhibit consistently better performance across diverse conditions. The intrinsic characteristics of rare cell populations, including abundance and transcriptional separability, are important factors associated with detection performance, although their effects vary across methods. Notably, rare-cell abundance should be carefully considered when selecting parameter values for scCAD. We identify substantial differences in computational efficiency and scalability among methods. Finally, we demonstrate that intersecting predictions from top-performing methods can improve the precision of rare cell identification and facilitate marker gene discovery in biological applications. Together, our findings provide practical guidance for rare cell detection in scRNA-seq analysis.

Introduction

Single-cell RNA sequencing (scRNA-seq) has enabled the characterization of cellular heterogeneity and diversity at single-cell resolution[1, 2]. The increasing availability of large-scale scRNA-seq datasets provides unprecedented opportunities to discover previously uncharacterized cell types from complex tissues, including rare cell types. Although present at low abundance, rare cell types can play pivotal roles in a wide range of biological processes, including embryonic development, immune responses, and cancer progression[3, 4, 5, 6, 7].

Rare cell types are typically represented by very few cells in scRNA-seq datasets, and their transcriptional differences from abundant cell types can be obscured by the high dimensionality, sparsity, and noise inherent to scRNA-seq data[8]. Consequently, rare cell detection remains challenging. To address this problem, numerous computational methods have been developed[3, 9, 10]. Most methods adopt a conventional clustering framework, in which cells are first represented in a low-dimensional space and cell–cell similarities (or distances) are subsequently used to construct clusters[11, 12, 13]. Rare cell populations are then identified from the resulting clusters based on cluster size and, in some cases, additional criteria. Rather than clustering all cells indiscriminately, Jindal *et al.* proposed the FiRE[14] score to quantify the rarity of individual cells and subsequently applied clustering to cells with high scores. In addition, methods such as EDGE[15] and SCA[16] were primarily developed to generate low-dimensional representations of single cells, and their embeddings can be combined with downstream clustering for rare cell detection. Thus, despite substantial differences in their specific strategies, many existing rare cell detection methods are clustering-based or incorporate clustering as a key component of their workflows.

Despite the availability of multiple methods, several fundamental questions remain unresolved. First, their performance across diverse biological and technical conditions has not been comprehensively and systematically evaluated, largely owing to the

limited availability of large-scale scRNA-seq datasets with reliably annotated rare cell populations. Second, how the intrinsic characteristics of rare cell populations influence detection performance remains poorly understood. Third, from a practical perspective, there is a lack of strategies for integrating the outputs of multiple detection methods to improve the reliability of rare cell identification in real-world biological applications.

In this study, we constructed 720 benchmarking scenarios from real scRNA-seq datasets and used them to systematically evaluate 11 rare cell detection methods **(Supplementary Table 1)**. We identified the methods that performed consistently well across diverse scenarios and uncovered intrinsic characteristics of rare cell populations that influenced detection performance. In addition, we demonstrated the utility of ensemble strategies for improving the identification of high-confidence rare cell populations, thereby facilitating downstream marker gene discovery in biological applications. Our study provides a comprehensive assessment of the strengths and limitations of current rare cell detection methods and offers practical guidance for rare cell detection in scRNA-seq analysis.

Results

Construction of benchmarking datasets

We utilized a cell-type-specific downsampling approach to construct benchmarking datasets with known rare-cell ground truth from existing annotated scRNA-seq datasets **(Fig. 1a)**. In detail, given an annotated dataset D, suppose that a cell type C constitutes a proportion q of all cells ($0 < q < 1$). For any target proportion r satisfying $0 < r < q$, the proportion of C can be reduced to r by randomly removing a subset of cells belonging to C while retaining all cells from other cell populations. When r is sufficiently small, C effectively becomes a rare cell population in the perturbed dataset D', which can therefore serve as a benchmarking dataset with known ground-truth labels. Unlike fully simulated datasets, our approach largely preserves the native transcriptional characteristics, cellular heterogeneity, and technical noise of real scRNA-seq data while enabling precise control over rare-cell abundance.

We applied this approach to 18 annotated scRNA-seq datasets spanning multiple tissues, species (human and mouse), and sequencing protocols **(Supplementary Table 2)**. For each dataset, we selected cell types comprising more than 5% of the total cells (termed abundant cell types) and generated benchmarking datasets at eight rare-cell proportions (0.125%, 0.25%, 0.5%, 1%, 2%, 3%, 4%, and 5%). Each combination of dataset D_i, target cell type $C_{i,j}$, and rare-cell proportion r_k defined a unique benchmarking scenario $S_{i,j,k} = (D_i, C_{i,j}, r_k)$, where $i = 1, 2, \ldots, 18$; $j = 1, 2, \ldots, N_i$; and $k = 1, 2, \ldots, 8$. Here, N_i denotes the number of abundant cell types in D_i, and $C_{i,j}$ denotes the jth abundant cell type in D_i. To account for stochastic variation introduced by random sampling, ten benchmarking datasets were generated for each benchmarking scenario. Overall, this procedure produced 7,200 benchmarking datasets representing 720 unique benchmarking scenarios, providing a comprehensive benchmarking framework covering diverse biological contexts and degrees of rarity **(Fig. 1b)**.

Because all evaluated methods ultimately identify candidate rare-cell clusters, we

adopted a unified evaluation strategy independent of each algorithm's implementation. For each benchmarking dataset, each method was run using its default parameters unless otherwise specified. For evaluation, among all clusters identified by each method, the cluster containing the largest number of ground-truth rare cells was designated as the predicted rare-cell cluster. All cells within this cluster were then considered predicted rare cells, and performance was quantified using the F1 score **(Fig. 1c)**. This evaluation measures each method's ability to isolate the target rare cell population into a distinct cluster while jointly accounting for both precision and recall. Unless otherwise specified, the F1 score for each benc-hmarking scenario was summarized as the median F1 score across its ten benchmarking datasets. An overview of all analyses performed in this study is shown in **Fig. 1d**.

Comparison of performance across benchmarking scenarios

Rather than directly comparing methods using raw F1 scores, we defined a scenario-specific relative F1 score (ssr-F1 score) to facilitate comparisons across benchmarking scenarios with different levels of intrinsic difficulty. For each benchmarking dataset, the relative F1 score of a method was calculated by subtracting the median F1 score across all evaluated methods for that dataset from its raw F1 score. For each benchmarking scenario, the ssr-F1 score was then defined as the median relative F1 score across the ten benchmarking datasets corresponding to that scenario. It is worth noting that this normalization mitigates differences in the overall performance baseline among benchmarking datasets while preserving the relative performance differences among methods. Applying this procedure yielded an 11 × 720 ssr-F1 score matrix, in which rows represent evaluated methods and columns represent benchmarking scenarios **(Fig. 2a)**. Ranking methods according to their median ssr-F1 scores identified aKNNO, RareQ, and scCAD as the three best-performing methods **(Fig. 2a)**. Notably, these methods are also among the most recently developed rare cell detection approaches **(Supplementary Table 1)**. We compared ssr-F1 scores between adjacent methods in the ranking and identified significant differences for eight of the ten comparisons. The comparisons between scCAD and CellSIUS and between SCA and

GiniClust3 were not statistically significant, and these results remained unchanged after Benjamini–Hochberg correction.

We next investigated whether method rankings depended on the abundance of the target rare cell population. For each of the eight predefined rare-cell proportions, we repeated the ranking procedure using only benchmarking scenarios corresponding to that proportion. Rankings derived from individual rare-cell proportions were highly concordant with the overall ranking, with Spearman correlation coefficients ranging from 0.73 to 0.96 **(Fig. 2b)**, indicating that the overall ranking was generally robust. Nevertheless, the relative ordering of methods varied across rare-cell proportions. While rankings obtained at neighboring proportions were highly similar, agreement gradually decreased as the difference in rare-cell proportion increased. For example, the ranking at 0.125% showed only modest concordance with those at 4% and 5% (ρ = 0.34 and 0.29, respectively), indicating that the best-performing method may depend, at least in part, on the abundance of the target population. In particular, scCAD consistently ranked first at extremely low rare-cell proportions (0.125%–1%), with a clear performance advantage over the second-ranked method; however, its rank decreased to eighth at rare-cell proportions of 4% and 5%. Notably, aKNNO ranked first at higher rare-cell proportions (2%–5%) while remaining among the top three methods across all eight rare-cell proportions **(Supplementary Fig. 1a–h)**.

To evaluate whether species influenced method ranking, we repeated the above analysis separately using benchmarking scenarios derived from human and mouse datasets. The human- and mouse-specific rankings were highly concordant with each other (ρ = 0.81) and with the overall ranking (ρ = 0.99 and 0.83 for human and mouse datasets, respectively; **Supplementary Fig. 1i–k**), suggesting that the relative performance of the evaluated methods was generally consistent across species.

To obtain a global view of method performance, we performed hierarchical clustering using the complete ssr-F1 score matrix. Interestingly, methods with similar overall

performance tended to cluster together, indicating that they also exhibited similar performance profiles across benchmarking scenarios **(Fig. 2c)**. Based on the clustering results, the eleven methods could be grouped into three performance tiers. The first tier comprised the four highest-performing methods (aKNNO, RareQ, scCAD, and CellSIUS), the second comprised four intermediate-performing methods (SCISSORS, CIARA, SCA, and GiniClust3), and the third comprised the remaining methods (EDGE, RaceID2, and RaceID3). Repeating this analysis separately for each rare-cell proportion yielded broadly similar clustering patterns **(Supplementary Fig. 2a–h)**. A notable exception was scCAD, which clustered together with aKNNO and RareQ at extremely low rare-cell proportions but shifted toward other groups as the rare-cell proportion increased. Notably, RaceID2 and RaceID3 consistently clustered together across all analyses, suggesting that they retained highly similar performance profiles despite the methodological updates introduced in RaceID3.

Collectively, these analyses provide a comprehensive evaluation of eleven representative rare cell detection methods across 720 benchmarking scenarios. Overall, aKNNO and RareQ exhibited better performance across diverse conditions, whereas scCAD showed a distinct advantage for identifying extremely rare cell populations. These findings indicate that no single method is universally optimal and highlight the importance of considering the expected abundance of the target population when selecting a rare cell detection method.

Rare-cell abundance influences detection performance

We next investigated how rare-cell abundance influenced the performance of the evaluated methods. For each method, we aggregated all benchmarking scenarios corresponding to the same rare-cell proportion and calculated the median of the scenario-level F1 scores, hereafter referred to as the rarity-specific F1 score (rs-F1 score). Unlike the ssr-F1 score, the rs-F1 score preserves absolute detection performance and is therefore more suitable for characterizing how performance changes with rare-cell abundance. All methods except scCAD exhibited an overall

positive relationship between rs-F1 score and rare-cell proportion **(Fig. 3a)**. Notably, most methods had rs-F1 scores below 0.5 when the rare-cell proportion was 0.5% or lower, indicating that the detection of extremely rare cell populations remains challenging.

Interestingly, we noticed that several top-performing methods, including aKNNO and RareQ, exhibited a characteristic two-phase pattern: rs-F1 scores increased rapidly as the rare-cell proportion rose from 0.125% to approximately 2%, followed by a plateau at higher abundances. To further characterize these trends, for each method except scCAD, we performed segmented linear regression analysis using scenario-level F1 scores as the response variable and the log10-transformed rare-cell proportion as the predictor. aKNNO and RareQ exhibited markedly decreased slopes after the breakpoint, consistent with a transition from rapid improvement to a plateau, whereas the other methods showed increased slopes after the breakpoint **(Supplementary Fig. 3a–b)**.

To determine the factors underlying the positive relationship between rs-F1 score and rare-cell proportion, we separately examined rarity-specific precision (rs-precision) and rarity-specific recall (rs-recall) across rare-cell proportions **(Fig. 3b–c)**. The rs-precision exhibited a consistently positive association with rare-cell proportion (median Spearman correlation coefficient, $\rho = 0.69$), whereas rs-recall showed a substantially weaker and generally negative association (median $\rho = -0.15$) **(Supplementary Fig. 3c)**. These findings indicate that the improvement in F1 score with increasing rare-cell proportion was driven primarily by increased precision rather than improved recall.

We next investigated why scCAD exhibited a distinct non-monotonic pattern, with performance increasing from 0.125% to 1% but declining thereafter. Although rs-precision increased with rare-cell proportion similarly to the other methods, rs-recall exhibited a striking biphasic pattern **(Fig. 3b–c)**. The rs-recall remained relatively high at low rare-cell proportions but declined abruptly as the rare-cell proportion increased beyond approximately 1%. This pattern suggested that the non-monotonic rs-F1 curve

of scCAD was largely attributable to the sharp reduction in rs-recall at higher rare-cell proportions. We hypothesized that this behavior was related to the default setting of scCAD's key parameter, *rare_h*, which has a default value of 1%. Notably, the inflection point in the rs-recall curve closely coincided with this threshold. To test this hypothesis, we repeated the analysis using different values of *rare_h*. While the overall relationship between rs-precision and rare-cell proportion remained largely unchanged, increasing *rare_h* progressively shifted the inflection point in the rs-recall curve toward higher rare-cell proportions **(Fig. 3d–e)**. Correspondingly, the peak of the rs-F1 curve also shifted toward higher rare-cell proportions, substantially attenuating the performance decline observed with the default setting **(Fig. 3f)**.

In scCAD, cells are first partitioned into initial clusters, after which clusters exceeding a predefined size threshold undergo iterative decomposition to generate smaller decomposed clusters. Highly similar decomposed clusters are subsequently merged to generate the final clustering results. The *rare_h* parameter defines the cluster-size threshold above which a cluster undergoes further decomposition. Thus, a lower *rare_h* allows a greater number of intermediate clusters to undergo additional decomposition. Our results suggest that when the target rare-cell population constitutes a proportion greater than the specified *rare_h*, clusters containing these cells may be more likely to undergo further decomposition into multiple smaller clusters. If these clusters are not fully re-merged during the subsequent merging step, fewer rare cells may be retained within a single predicted cluster, thereby reducing recall while leaving precision largely unaffected. These results suggest that the default *rare_h* value may be well suited to detecting extremely rare cell populations but may promote over-partitioning when the target population is more abundant.

Together, these results demonstrate that rare-cell abundance influences detection performance.

Rare-cell separability influences detection performance

We next investigated whether the transcriptional separability of rare cell populations was associated with detection performance. For each benchmarking dataset, we quantified rare-cell separability using the mean silhouette score of cells in the rare cell population, with higher scores indicating greater separation of the rare cell population from other cells in the dataset. Given that rare-cell abundance was an important determinant of detection performance, we examined the relationship between F1 score and silhouette score at each rare-cell proportion **(Fig. 4a; Supplementary Fig. 4)**. For all methods except EDGE, F1 score was generally positively correlated with silhouette score, although the strength of this association varied across rare-cell proportions. Taking SCISSORS as an example, we divided benchmarking scenarios into low- and high-silhouette groups according to their silhouette scores at each rare-cell proportion. At a rare-cell proportion of 0.125%, the difference in F1 score between the two groups was only 0.02, whereas at 5% the difference increased to 0.28 **(Fig. 4b)**.

To test whether the association between rare-cell separability and detection performance depended on rare-cell abundance, we focused on the two extreme rare-cell proportions (0.125% and 5%) and fitted a linear regression model including an interaction term between silhouette score and rare-cell proportion. Among the ten methods other than EDGE, four methods (SCISSORS, CIARA, GiniClust3, and SCA) showed a significant positive interaction coefficient, indicating that the positive association between silhouette score and F1 score was stronger at the higher rare-cell proportion **(Fig. 4c)**. Interaction coefficients were positive for most methods, whereas scCAD showed a negative coefficient. Consistent with this pattern, at a rare-cell proportion of 5%, the Spearman correlation between silhouette score and F1 score for scCAD was weak and not statistically significant ($\rho = 0.11$; **Fig. 4a; Supplementary Fig. 4**). For scCAD, precision remained positively associated with silhouette score at a rare-cell proportion of 5% ($\rho = 0.43$), whereas recall showed little association ($\rho = -0.08$), which may partly explain the weak relationship between silhouette score and F1 score at this abundance **(Fig. 4d; Supplementary Fig. 5–6)**.

Together, these findings indicate that rare-cell separability is an additional factor associated with detection performance beyond rare-cell abundance, with the strength of this association varying across methods and rare-cell proportions.

Runtime performance of rare cell detection methods

We next evaluated the computational efficiency of the eleven methods. For each benchmarking scenario, runtime was defined as the median runtime across the ten corresponding benchmarking datasets. Ranking methods by their median runtime across all benchmarking scenarios identified GiniClust3, EDGE, and aKNNO as the three fastest methods **(Fig. 5a)**. We compared runtime between adjacent methods in the ranking and identified significant differences for all ten comparisons. Notably, computational costs differed by more than two orders of magnitude across methods. GiniClust3 required a median runtime of less than 1 minute per benchmarking dataset, whereas the slowest method, CIARA, required nearly 100 minutes.

As expected, runtime generally increased with dataset size for all methods **(Fig. 5b)**. To quantify runtime scalability, we fitted a linear regression model for each method relating log10-transformed runtime to log10-transformed total cell number. We ranked methods according to the fitted slope and found that RareQ, CellSIUS, and aKNNO showed the three lowest values, indicating comparatively favorable scalability **(Fig. 5c)**. Given that the relationship between runtime and total cell number may not be purely linear, we additionally assessed the variability of runtime across benchmarking scenarios using the interquartile range (IQR), with lower IQR suggesting higher scalability. GiniClust3 showed the smallest runtime IQR, followed by aKNNO and RareQ, whereas CIARA showed the largest **(Supplementary Fig. 7a)**. Rankings based on runtime IQR and fitted slope showed a positive association **(Supplementary Fig. 7b)**, indicating that the two measures yielded broadly consistent rankings of method scalability.

Together, these results reveal substantial differences in runtime and scalability among

existing rare cell detection methods. Notably, aKNNO and RareQ combined ideal performance with efficient computation and good scalability, supporting their use in large-scale scRNA-seq analyses.

Practical evaluation of ensemble strategies for rare cell detection

We next sought to translate our benchmarking results into a practical workflow for biological applications. In a typical rare-cell analysis, computational identification of rare cell populations is usually followed by marker gene discovery and experimental validation. Unlike benchmarking, however, real-world studies typically lack ground-truth labels, making it difficult to directly determine whether an algorithm-identified rare-cell cluster represents a genuine rare cell population or instead arises from imperfect clustering. Consequently, marker genes derived from false-positive clusters could mislead downstream experimental validation. We therefore reasoned that, for marker discovery, prioritizing prediction precision over maximal rare-cell recovery may provide a more reliable route to experimental validation.

Given the favorable overall performance of aKNNO, RareQ, and CellSIUS across our benchmarking scenarios, we next investigated whether combining these methods could improve the precision of rare cell identification using an independent, well-annotated fetal lung scRNA-seq dataset (GSE264407) comprising 20 samples[17]. Here, we evaluated whether cells were rare irrespective of their specific cell-type identity. For each sample, clusters identified by each method containing no more than 1% of the total cells and at least three cells were defined as candidate rare-cell clusters, and all cells within these clusters were considered predicted rare cells. We then compared the three individual methods with pairwise intersection and union strategies, as well as three-method intersection, union, and majority-voting strategies. Among the three individual methods, aKNNO achieved the highest precision but the lowest recall **(Fig. 6a; Supplementary Fig. 8a)**. Notably, all intersection strategies involving aKNNO achieved significantly higher precision than aKNNO alone. In contrast, the RareQ + CellSIUS intersection and the three-method majority-voting strategy showed precision

comparable to that of aKNNO. As expected, intersection-based strategies generally exhibited lower recall than aKNNO, reflecting the increased stringency of requiring cells to be consistently identified by multiple methods. Importantly, repeating the analysis using a less stringent threshold of 5% to define candidate rare-cell clusters yielded consistent results **(Supplementary Fig. 8b–c)**, indicating that the precision improvements observed for several intersection-based strategies were robust across the two rarity thresholds.

These results motivated a two-step workflow for high-confidence rare-cell marker discovery **(Fig. 6b)**. Rather than directly using the output of a single rare cell detection algorithm, we first used the intersection of multiple top-performing methods to enrich for cells with consistent rare-cell predictions. The resulting high-confidence rare cells were then subjected to clustering and marker gene identification, followed by cell type annotation. Although this strategy inevitably sacrifices a fraction of true rare cells, the rapidly increasing scale of modern single-cell sequencing datasets can provide sufficient cellular coverage to accommodate this loss while substantially reducing false-positive predictions.

To demonstrate the feasibility of this strategy, we applied the two-step workflow to sample GSM8217713 from dataset GSE264407, which contained 12,427 cells, including 1,494 cells representing 28 rare cell types. The three-method intersection retained 348 high-confidence rare cells. We then applied Louvain clustering to these enriched cells, yielding seven clusters that showed substantial correspondence with the original cell-type annotations **(Fig. 6c–d)**. Except for cluster 2, each cluster was dominated by a single rare cell type that accounted for more than 50% of the cells within the cluster. For each cluster, we next compared the expression of marker genes corresponding to its dominant cell type between cells within the cluster and the remaining cells in the sample. As expected, the corresponding marker genes exhibited significantly higher expression in the cells within the cluster. For example, cluster 0 consisted entirely of originally annotated aM-like cells and exhibited *CCL3* expression

comparable to that of the originally annotated aM-like cell population, while both groups showed substantially higher expression than other cells **(Fig. 6e; Supplementary Fig. 8d)**.

Together, these results demonstrate that intersection-based integration of the outputs of multiple top-performing methods can improve prediction precision and provide a practical two-step pipeline for high-confidence rare-cell enrichment and marker gene discovery.

Discussion

Accurate detection of rare cell types remains a challenging task in scRNA-seq analysis, particularly under conditions of high dimensionality and severe class imbalance[18]. Although many computational methods have been developed, their performance and practical utility in biological applications remain incompletely characterized. Our study provides a comprehensive assessment of their strengths, limitations, and practical applicability.

We identified rare-cell proportion as an important determinant of detection performance. In practical applications, however, the true abundance of a rare cell population is generally unknown, making it difficult to anticipate whether a given method will perform reliably. At extreme rarity, current approaches showed substantial limitations. For example, at rare-cell proportions of 0.125% and 0.5%, even the best-performing method, scCAD, achieved a precision below 0.25, indicating that most cells identified as rare cell candidates did not correspond to the target population. It is worth noting, however, that extremely low abundance should not be interpreted as evidence that a cell population is biologically unimportant or merely represents technical noise. Biologically meaningful cell populations can occur at exceptionally low frequencies; for example, circulating tumor cells can be detected at very low abundance in peripheral blood and are closely associated with cancer progression and clinical outcomes[19, 20]. Thus, our results suggest that future rare cell detection methods should be designed and optimized specifically for conditions of extreme rarity.

One potential avenue for improvement lies in increasing the resolution of graph-based clustering. Graph-based approaches, including Louvain[21] and Leiden[22] clustering, are widely used in scRNA-seq data analysis, but their utility for resolving extremely rare populations may be limited by the resolution of the initial clustering. Notably, scCAD, which performed best at extremely low rare-cell proportions, employs an iterative cluster-decomposition and re-merging strategy. Compared with a single round of

clustering, decomposition may provide greater opportunities to isolate purer rare-cell populations. Although such a strategy may increase the risk of over-partitioning, improved cluster re-merging strategies could help mitigate the limitation.

Effective representation of single cells in low-dimensional space is another important consideration for rare cell detection. For many of the methods evaluated here, cell–cell distances were calculated as Euclidean distances in PCA space. Although the diversity of clustering strategies across methods prevents us from concluding that PCA-based Euclidean distance is intrinsically optimal, our benchmarking suggests that this representation may be advantageous for rare cell detection. Among the methods evaluated, RaceID2 and RaceID3 used Pearson correlation-based similarity[23, 24], whereas EDGE calculated Euclidean distances in its own embedding space[15]. By contrast, the four highest-performing methods (aKNNO, RareQ, scCAD, and CellSIUS) used PCA-based representations to quantify cell–cell distances. Despite their favorable performance, current PCA-based representations leave considerable room for improvement. PCA is typically performed using highly variable genes (HVGs), which are selected to capture major sources of transcriptional variation across the entire dataset. Consequently, genes that are highly informative for a rare population but contribute little to global variance may be underrepresented in the resulting embedding. One possible strategy is to introduce dynamic embedding in the algorithm; for example, during the cluster-decomposition stage, PCA could be re-performed using only cells within the same cluster. Such a strategy could better capture subtle transcriptional differences within heterogeneous clusters and potentially improve the detection of rare cell populations.

We also highlight a gap between the objectives of algorithm development and those of biological applications and emphasize that precision may be particularly important when rare cell detection serves as the first step of a biological discovery pipeline. Importantly, our results showed that combining multiple high-performing methods can substantially increase the precision of recovered rare cell populations, although this

gain is accompanied by reduced recall. Beyond increasing precision, such joint analysis may provide an additional advantage. When clustering is performed across the entire dataset, abundant cell populations may dominate the representation of cellular heterogeneity and potentially obscure subtle transcriptional differences within rare populations. In contrast, re-analysis of high-confidence rare-cell candidates reduces the influence of abundant populations and may enable finer resolution of heterogeneity within the rare-cell compartment.

Cell type annotation is a fundamental step in scRNA-seq analysis and relies heavily on predefined marker genes[25]. However, existing annotation resources are predominantly designed around well-characterized cell populations[26, 27], whereas comprehensive resources specifically covering rare cell types remain limited. Our pipeline could be applied to large-scale scRNA-seq datasets to facilitate the systematic discovery of rare cell types and their defining molecular features, ultimately supporting the construction of a comprehensive reference that complements existing cell-type annotation resources.

Taken together, we comprehensively benchmarked eleven rare cell detection methods and identified key factors associated with their performance. Our findings provide practical guidance for rare cell detection in biological applications and offer insights into the development of more robust and effective methods for rare cell identification.

Methods

scRNA-seq datasets

We collected a total of 18 published scRNA-seq datasets for benchmarking rare cell detection methods. The PBMC3k dataset was obtained from the SeuratData R package[28], while all other original benchmarking datasets were retrieved from the CZ CELLxGENE Discover database (https://cellxgene.cziscience.com/). Only datasets with clearly annotated cell types were included to enable the construction of benchmarking scenarios with known ground-truth labels. Cell type annotations provided by the original studies were retained and used as reference labels for benchmarking dataset generation and performance evaluation. Detailed information for all original benchmarking datasets, including tissue source, species, and sequencing platform, is provided in Supplementary Table 2.

For the practical evaluation of ensemble strategies, we further used an independent fetal lung scRNA-seq dataset from Quach *et al*., comprising 20 samples[17]. This dataset was obtained from the CZ CELLxGENE Discover database. Its original cell type annotations were retained for retrospective evaluation of ensemble predictions.

Comparison of rare cell identification methods

We benchmarked 11 rare cell detection methods. Unless otherwise specified, all methods were executed using their default or recommended parameters according to the original publications. The aKNNO[11] R package was obtained from GitHub (https://github.com/liuqivandy/aKNNO). The CellSIUS[29] R package was obtained from GitHub (https://github.com/Novartis/CellSIUS/). For CellSIUS analysis, initial clusters were generated using the FindClusters() function implemented in the Seurat R package with default parameters. The CIARA[30] R package was obtained from GitHub (https://github.com/ScialdoneLab/CIARA/). The EDGE[15] R package was obtained from GitHub (https://github.com/shawnstat/EDGE). Based on the two-dimensional embedding generated by EDGE, we constructed a k-nearest neighbor (kNN) graph (k

= 15) and applied Louvain clustering to obtain the global clustering results. The GiniClust3[31] Python implementation was obtained from GitHub (https://github.com/rdong08/GiniClust3/). The RaceID2[23] and RaceID3[24] R packages were obtained from GitHub (RaceID2: https://github.com/dgrun/StemID, RaceID3: https://github.com/dgrun/RaceID3_StemID2_package). The RareQ[13] R package was obtained from GitHub (https://github.com/fabotao/RareQ). The SCA[16] Python implementation was obtained from GitHub (https://github.com/bendemeo/shannonca). Following the recommendations of the original publication, we first computed a 50-dimensional SCA embedding. A kNN graph (k = 15) based on Euclidean distances was constructed based on the SCA embedding, and Leiden clustering with the default resolution of 1.0 was applied to obtain the final clustering results. The scCAD[12] Python implementation was obtained from GitHub (https://github.com/xuyp-csu/scCAD). The parameter *rare_h* defines the cluster-size threshold above which a cluster is subjected to further decomposition. We evaluated five *rare_h* settings, including the default value of 1% and additional settings of 2%, 3%, 4%, and 5%, denoted as scCAD (1%), scCAD (2%), scCAD (3%), scCAD (4%), and scCAD (5%), respectively. All other scCAD parameters were retained at their default settings. Unless otherwise specified, the default 1% setting was used in the primary cross-method comparisons, whereas the additional settings were used to evaluate the sensitivity of scCAD performance to the cluster-decomposition threshold across different rare-cell proportions. The SCISSORS[32] R package was obtained from GitHub (https://github.com/jr-leary7/SCISSORS). For SCISSORS analysis, clusters with an average silhouette coefficient lower than the overall average silhouette coefficient were reclustered to obtain the final clustering results.

Construction of benchmarking datasets

Given a well-annotated scRNA-seq dataset D_i, we selected each abundant cell type $C_{i,j}$ with an original proportion $q_{i,j} > 5\%$ as a candidate target population. For each selected cell type and each predefined rare-cell proportion r_k, a benchmarking scenario was

defined as

$$S_{i,j,k} = (D_i, C_{i,j}, r_k).$$

Eight rare-cell proportions were considered: 0.125%, 0.25%, 0.5%, 1%, 2%, 3%, 4%, and 5%.

For each benchmarking scenario, a perturbed dataset was generated by randomly removing cells from the target cell type without replacement while retaining all non-target cells. For simplicity, scenario indices are omitted in the following derivation. Let n denote the total number of cells in the original dataset, and c denote the number of cells belonging to the target cell type, such that

$$q = \frac{c}{n}.$$

If x cells of type C are removed, the resulting proportion of C in the perturbed dataset satisfies

$$\frac{c - x}{n - x} = r.$$

Solving this equation gives

$$x = \frac{q - r}{1 - r} \cdot n.$$

Thus, removing x target cells generates a benchmarking dataset with the predefined rare-cell proportion r, while preserving the expression profiles and cell numbers of all non-target populations. For each benchmarking scenario, the random sampling procedure was independently repeated ten times, generating ten benchmarking datasets per scenario. In total, 720 benchmarking scenarios and 7,200 benchmarking datasets were generated.

Performance evaluation metrics

To evaluate the performance of rare cell detection methods, we compared predicted rare-cell labels with ground-truth labels for each benchmarking dataset. Cells belonging to the target rare cell type were treated as ground-truth positives, whereas all other cells were treated as ground-truth negatives. True positives (TP), false positives (FP), true

negatives (TN), and false negatives (FN) were defined as follows: TP represents ground-truth rare cells correctly identified as rare; FP represents non-rare cells incorrectly identified as rare; TN represents non-rare cells correctly identified as non-rare; and FN represents ground-truth rare cells incorrectly identified as non-rare.

Precision, recall, and F1 score were calculated as:

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$\text{F1 score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Unless otherwise specified, performance for each benchmarking scenario was summarized as the median of each metric across its ten independently generated benchmarking datasets.

Scenario-specific relative performance analysis

To compare method performance under matched benchmarking scenarios, we calculated the scenario-specific relative F1 score (ssr-F1 score) for each method. Each benchmarking scenario S consisted of ten independently generated benchmarking datasets, indexed by $l = 1,2,\ldots,10$. For each benchmarking dataset, the median F1 score across all methods was first used as the dataset-specific reference:

$$F1_{\text{ref},S,l} = \text{median}_m\left(F1_{m,S,l}\right).$$

where $F1_{m,S,l}$ denotes the F1 score of method m for benchmarking dataset l within scenario S.

The relative F1 score of method m within each benchmarking dataset was then calculated as:

$$\Delta F1_{m,S,l} = F1_{m,S,l} - F1_{\text{ref},S,l}.$$

Finally, the ssr-F1 score of method m for scenario S was calculated as the median relative F1 score across the ten benchmarking datasets:

$$\text{ssr}-F1_{m,S} = \text{median}_{l=1,2,\ldots,10}(\Delta F1_{m,S,l}).$$

A positive ssr-F1 score therefore indicates that method m generally performed above the median of all evaluated methods within the benchmarking datasets of scenario S, whereas a negative value indicates relatively poorer performance.

Rarity-specific performance analysis

To characterize changes in detection performance across rare-cell proportions, we calculated rarity-specific F1 (rs-F1), precision (rs-precision), and recall (rs-recall). For each benchmarking scenario, F1 score, precision, and recall were first summarized as the median across the ten independently generated benchmarking datasets. For each method and rare-cell proportion, the corresponding rarity-specific metric was then calculated as the median of the scenario-level metric across all benchmarking scenarios at that proportion. Specifically,

$$\text{rs}-F1_{m,r} = \text{median}_{S\in S_r}(F1_{m,S}),$$

$$\text{rs}-\text{precision}_{m,r} = \text{median}_{S\in S_r}(\text{Precision}_{m,S}),$$

$$\text{rs}-\text{recall}_{m,r} = \text{median}_{S\in S_r}(\text{Recall}_{m,S}),$$

where m denotes the method, r denotes the rare-cell proportion, and S_r denotes the set of all benchmarking scenarios generated at rare-cell proportion r.

Segmented linear regression analysis

To characterize the relationship between rare-cell proportion and detection performance, we first summarized the F1 score for each benchmarking scenario as the median across its ten independently generated benchmarking datasets. For each method, rare-cell proportion was log10-transformed, and a segmented linear regression model with one unknown breakpoint was fitted to the scenario-level F1 scores:

$$F1 = \alpha + \beta_1 \log10(r) + \beta_2[\log_{10}(r) - \varphi]_+ + \varepsilon,$$

where r denotes the rare-cell proportion, φ denotes the estimated breakpoint on the

log10-transformed rare-cell proportion scale, and $[x]_+ = \max(x, 0)$. The slope before the breakpoint was β_1, whereas the slope after the breakpoint was $\beta_1 + \beta_2$. Segmented regression was performed using the segmented R package with one breakpoint.

Evidence for a change in slope at an unknown breakpoint was assessed using the Davies test.

Rare-cell silhouette score

For each benchmarking dataset, principal component analysis (PCA) was performed on the expression matrix, and the first 30 principal components (PCs) were used to represent the transcriptional space. Pairwise Euclidean distances were calculated in this 30-dimensional PCA space. For each cell *i* belonging to the target rare cell type *C*, we calculated its silhouette coefficient based on the reference cell-type annotations. Let $a(i)$ denote the average dissimilarity between cell *i* and all other cells belonging to *C*, and let $b(i)$ denote the minimum average dissimilarity between cell *i* and cells belonging to any other annotated cell type. The silhouette coefficient of cell *i* was calculated as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}.$$

The silhouette coefficient ranges from −1 to 1, with larger values indicating greater within-population cohesion and stronger separation from other cell populations. For each benchmarking dataset, the rare-cell silhouette score was defined as the mean silhouette coefficient across all cells belonging to the target rare cell type *C*:

$$\mathrm{Sil}_C = \frac{1}{N_C} \sum_{i \in C} s(i)$$

where N_C denotes the number of cells in the target rare cell type *C*.

Interaction analysis between rare-cell proportion and silhouette score

To assess whether the association between rare-cell separability and detection

performance varied with rare-cell proportion, we focused on the lowest and highest rare-cell proportions (0.125% and 5%). For each benchmarking scenario, F1 score and rare-cell silhouette score were summarized as the median values across the ten independently generated benchmarking datasets. A linear model was then fitted separately for each method:

$$F1_i = b + \alpha \mathrm{Sil}_i + \beta G_i + \gamma(\mathrm{Sil}_i \times G_i) + \varepsilon_i,$$

where i indexes the benchmarking scenario, Sil_i denotes the rare-cell silhouette score, and G_i denotes the rare-cell proportion group, coded as 0 for 0.125% and 1 for 5%. Here, b represents the intercept, α represents the association between rare-cell silhouette score and F1 score at a rare-cell proportion of 0.125%, β represents the difference in the intercept between the two rare-cell proportions when $\mathrm{Sil}_i = 0$, and γ represents the difference in the slope of the silhouette score–F1 relationship between rare-cell proportions of 0.125% and 5%.

Code availability

The code is available at GitHub (https://github.com/XiuyuanJin/RareCellBenchmarking).

Competing interests

The authors declare no competing interests.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Fund 32370715, 82330108), Science Fund for Distinguished Overseas Young Scholars of Shandong Province (2023HWYQ-015), Taishan Young Scholar Program of Shandong Province (tsqn202312020), and Cheeloo Young Scholar Program of Shandong University.

Author contributions

K.L. conceived and supervised the study. X.Y.J. performed all computational analyses. All authors contributed to writing, reviewing, and editing the manuscript and approved the manuscript.

References

1. Potter, S. S. Single-cell RNA sequencing for the study of development, physiology and disease. *Nature reviews Nephrology* **14**, 479-492 https://doi.org/10.1038/s41581-018-0021-7 (2018).
2. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology* **15**, e8746 https://doi.org/10.15252/msb.20188746 (2019).
3. Grün, D.*, et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251-255 https://doi.org/10.1038/nature14966 (2015).
4. Jordan, C. T., Guzman, M. L. & Noble, M. Cancer stem cells. *The New England journal of medicine* **355**, 1253-1261 https://doi.org/10.1056/NEJMra061808 (2006).
5. Zuo, W.*, et al.* p63(+)Krt5(+) distal airway stem cells are essential for lung regeneration. *Nature* **517**, 616-620 https://doi.org/10.1038/nature13903 (2015).
6. Ouadah, Y.*, et al.* Rare Pulmonary Neuroendocrine Cells Are Stem Cells Regulated by Rb, p53, and Notch. *Cell* **179**, 403-416.e423 https://doi.org/10.1016/j.cell.2019.09.010 (2019).
7. Alix-Panabières, C. & Pantel, K. Challenges in circulating tumour cell research. *Nature reviews Cancer* **14**, 623-631 https://doi.org/10.1038/nrc3820 (2014).
8. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature reviews Genetics* **20**, 273-282 https://doi.org/10.1038/s41576-018-0088-9 (2019).
9. Jiang, L., Chen, H., Pinello, L. & Yuan, G. C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome biology* **17**, 144 https://doi.org/10.1186/s13059-016-1010-4 (2016).
10. Tsoucas, D. & Yuan, G. C. GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome biology* **19**, 58 https://doi.org/10.1186/s13059-018-1431-3 (2018).
11. Li, J., Shyr, Y. & Liu, Q. aKNNO: single-cell and spatial transcriptomics

clustering with an optimized adaptive k-nearest neighbor graph. *Genome biology* **25**, 203 https://doi.org/10.1186/s13059-024-03339-y (2024).

12. Xu, Y., *et al.* scCAD: Cluster decomposition-based anomaly detection for rare cell identification in single-cell expression data. *Nature communications* **15**, 7561 https://doi.org/10.1038/s41467-024-51891-9 (2024).
13. Fa, B., *et al.* Cell neighborhood topology directs rare cell population identification. *Nature communications* **17**, https://doi.org/10.1038/s41467-026-71180-x (2026).
14. Jindal, A., Gupta, P., Jayadeva & Sengupta, D. Discovery of rare cells from voluminous single cell expression data. *Nature communications* **9**, 4719 https://doi.org/10.1038/s41467-018-07234-6 (2018).
15. Sun, X., Liu, Y. & An, L. Ensemble dimensionality reduction and feature gene extraction for single-cell RNA-seq data. *Nature communications* **11**, 5853 https://doi.org/10.1038/s41467-020-19465-7 (2020).
16. DeMeo, B. & Berger, B. SCA: recovering single-cell heterogeneity through information-based dimensionality reduction. *Genome biology* **24**, 195 https://doi.org/10.1186/s13059-023-02998-7 (2023).
17. Quach, H., *et al.* Early human fetal lung atlas reveals the temporal dynamics of epithelial cell plasticity. *Nature communications* **15**, 5898 https://doi.org/10.1038/s41467-024-50281-5 (2024).
18. Gulati, G. S., D'Silva, J. P., Liu, Y., Wang, L. & Newman, A. M. Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nature reviews Molecular cell biology* **26**, 11-31 https://doi.org/10.1038/s41580-024-00768-2 (2025).
19. Dai, C. S., *et al.* Circulating tumor cells: Blood-based detection, molecular biology, and clinical applications. *Cancer cell* **43**, 1399-1422 https://doi.org/10.1016/j.ccell.2025.07.008 (2025).
20. Cristofanilli, M., *et al.* Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *The New England journal of medicine* **351**, 781-791 https://doi.org/10.1056/NEJMoa040766 (2004).

21. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 https://doi.org/10.1088/1742-5468/2008/10/P10008 (2008).

22. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233 https://doi.org/10.1038/s41598-019-41695-z (2019).

23. Grün, D.*, et al.* De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell stem cell* **19**, 266-277 https://doi.org/10.1016/j.stem.2016.05.010 (2016).

24. Herman, J. S., Sagar & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature methods* **15**, 379-386 https://doi.org/10.1038/nmeth.4662 (2018).

25. Pasquini, G., Rojo Arias, J. E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Computational and structural biotechnology journal* **19**, 961-969 https://doi.org/10.1016/j.csbj.2021.01.015 (2021).

26. Hu, C.*, et al.* CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* **51**, D870-D876 https://doi.org/10.1093/nar/gkac947 (2023).

27. Franzén, O., Gan, L. M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database : the journal of biological databases and curation* **2019**, https://doi.org/10.1093/database/baz046 (2019).

28. Satija Lab. SeuratData: Install and Manage Seurat Datasets. R package, https://github.com/satijalab/seurat-data (2017).

29. Wegmann, R.*, et al.* CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. *Genome biology* **20**, 142 https://doi.org/10.1186/s13059-019-1739-7 (2019).

30. Lubatti, G.*, et al.* CIARA: a cluster-independent algorithm for identifying

markers of rare cell types from single-cell sequencing data. *Development (Cambridge, England)* **150**, https://doi.org/10.1242/dev.201264 (2023).

31. Dong, R. & Yuan, G. C. GiniClust3: a fast and memory-efficient tool for rare cell type identification. *BMC bioinformatics* **21**, 158 https://doi.org/10.1186/s12859-020-3482-1 (2020).

32. Leary, J. R., *et al.* Sub-Cluster Identification through Semi-Supervised Optimization of Rare-Cell Silhouettes (SCISSORS) in single-cell RNA-sequencing. *Bioinformatics (Oxford, England)* **39**, https://doi.org/10.1093/bioinformatics/btad449 (2023).

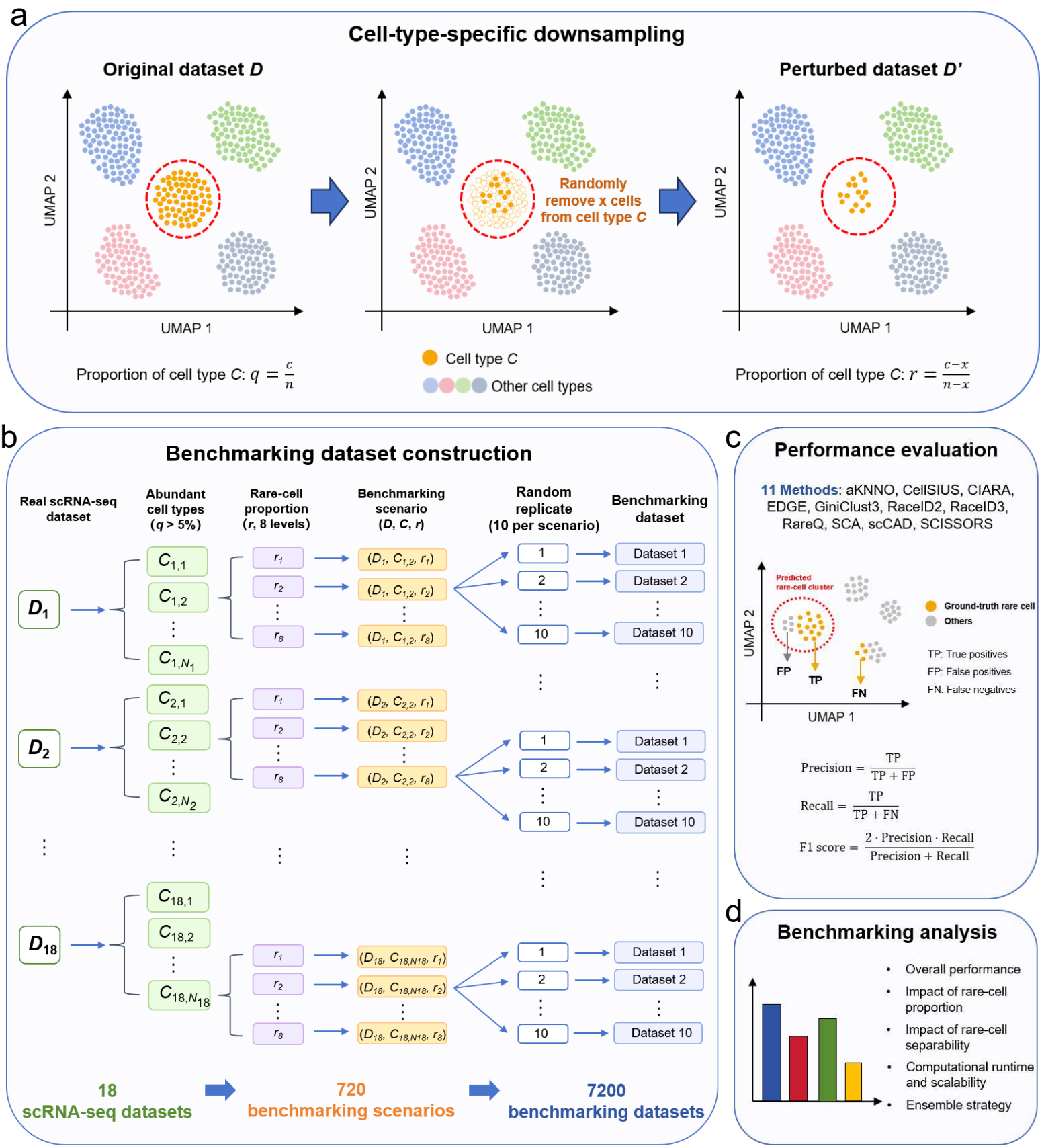

a
Cell-type-specific downsampling
Original dataset D
Perturbed dataset D'
UMAP 2
UMAP 1
Randomly remove x cells from cell type C
Cell type C
Other cell types
Proportion of cell type C: $q = \frac{c}{n}$
Proportion of cell type C: $r = \frac{c-x}{n-x}$
b
Benchmarking dataset construction
Real scRNA-seq dataset
Abundant cell types (q > 5%)
Rare-cell proportion (r, 8 levels)
Benchmarking scenario (D, C, r)
Random replicate (10 per scenario)
Benchmarking dataset
18 scRNA-seq datasets
720 benchmarking scenarios
7200 benchmarking datasets
c
Performance evaluation
11 Methods: aKNNO, CellSIUS, CIARA, EDGE, GiniClust3, RaceID2, RaceID3, RareQ, SCA, scCAD, SCISSORS
Predicted rare-cell cluster
Ground-truth rare cell
Others
TP: True positives
FP: False positives
FN: False negatives
FP
TP
FN
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1 score = 2 · Precision · Recall / (Precision + Recall)
d
Benchmarking analysis
Overall performance
Impact of rare-cell proportion
Impact of rare-cell separability
Computational runtime and scalability
Ensemble strategy

Figure 1

Figure 1 | Overall design of the study.

a, Schematic illustration of the cell-type-specific downsampling strategy. Here, n denotes the total number of cells in dataset D, and c denotes the number of cells belonging to cell type C. q and r denote the proportions of cell type C before and after downsampling, respectively.

b, Workflow for constructing benchmarking datasets.

c, Performance evaluation of the 11 rare cell detection methods. Orange points represent ground-truth rare cells, and gray points represent other cells. The red dashed outline highlights the predicted rare-cell cluster.

d, Overview of the analyses performed in this study.

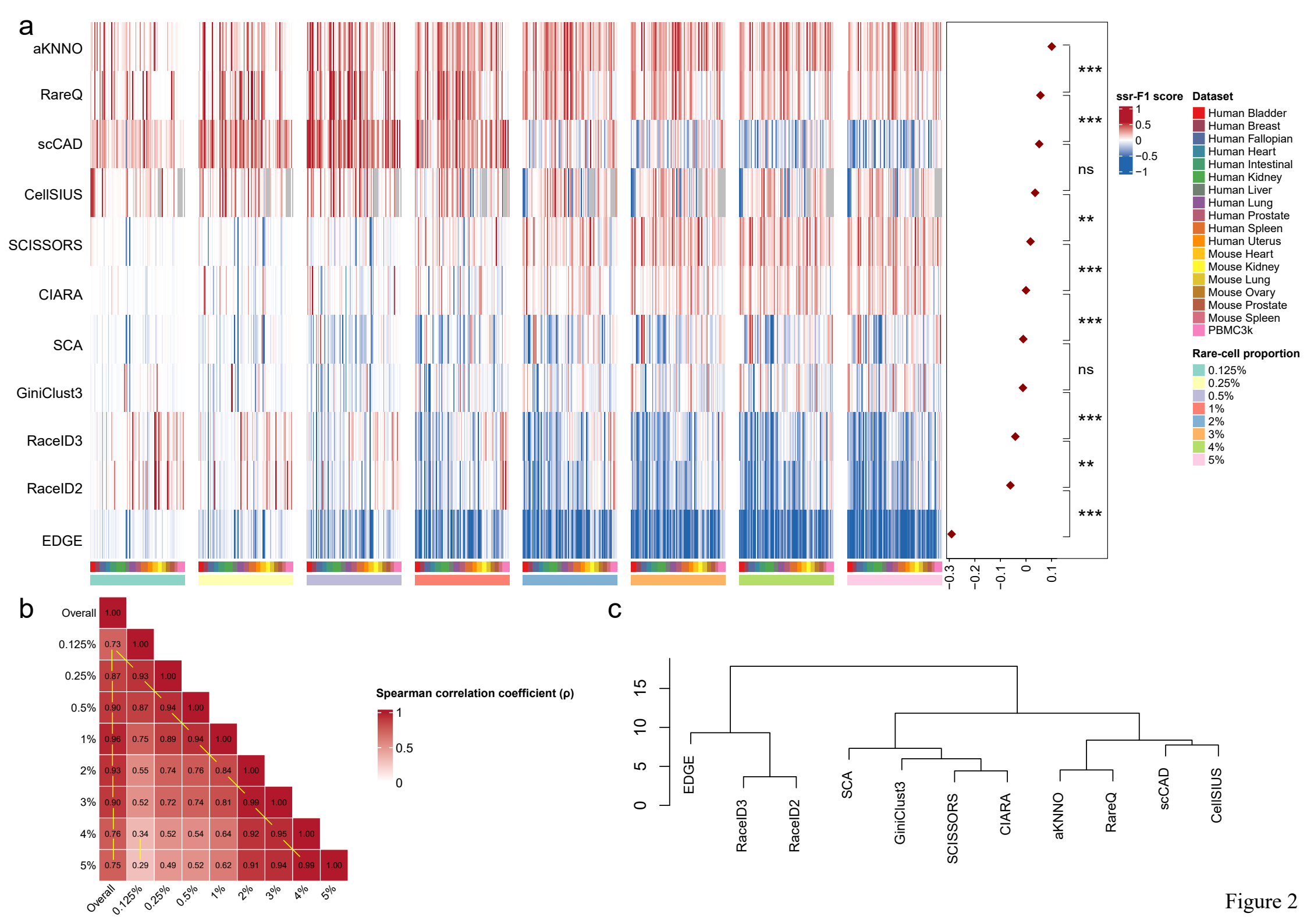

a
aKNNO
RareQ
scCAD
CellSIUS
SCISSORS
CIARA
SCA
GiniClust3
RaceID3
RaceID2
EDGE
***
***
ns
**
***
***
ns
***
**
***
−0.3
−0.2
−0.1
0
0.1
ssr-F1 score
1
0.5
0
−0.5
−1
Dataset
Human Bladder
Human Breast
Human Fallopian
Human Heart
Human Intestinal
Human Kidney
Human Liver
Human Lung
Human Prostate
Human Spleen
Human Uterus
Mouse Heart
Mouse Kidney
Mouse Lung
Mouse Ovary
Mouse Prostate
Mouse Spleen
PBMC3k
Rare-cell proportion
0.125%
0.25%
0.5%
1%
2%
3%
4%
5%
b
Overall 1.00
0.125% 0.73 1.00
0.25% 0.87 0.93 1.00
0.5% 0.90 0.87 0.94 1.00
1% 0.96 0.75 0.89 0.94 1.00
2% 0.93 0.55 0.74 0.76 0.84 1.00
3% 0.90 0.52 0.72 0.74 0.81 0.99 1.00
4% 0.76 0.34 0.52 0.54 0.64 0.92 0.95 1.00
5% 0.75 0.29 0.49 0.52 0.62 0.91 0.94 0.99 1.00
Overall
0.125%
0.25%
0.5%
1%
2%
3%
4%
5%
Spearman correlation coefficient (ρ)
1
0.5
0
c
0
5
10
15
EDGE
RaceID3
RaceID2
SCA
GiniClust3
SCISSORS
CIARA
aKNNO
RareQ
scCAD
CellSIUS

Figure 2

Figure 2 | Comparison of performance based on benchmarking scenarios.

a, Heatmap of the scenario-specific relative F1 (ssr-F1) score matrix, with rows representing methods and columns representing benchmarking scenarios. Red and blue indicate positive and negative ssr-F1 scores, respectively, whereas gray indicates missing values. Columns are grouped by rare-cell proportion, with bottom annotations indicating the original dataset and rare-cell proportion. The right panel summarizes the median ssr-F1 score for each method across all benchmarking scenarios. The *P* values were calculated using paired Wilcoxon signed-rank tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ns, not significant.

b, Spearman correlation matrix comparing method rankings derived from the overall benchmark and from each rare-cell proportion. The left yellow line highlights the correlations between the overall ranking and the rankings at each rare-cell proportion. The middle yellow line highlights the correlations between the ranking at 0.125% and those at 4% and 5%. The right yellow line highlights the correlations between rankings at adjacent rare-cell proportions.

c, Hierarchical clustering of methods based on their ssr-F1 score profiles across benchmarking scenarios.

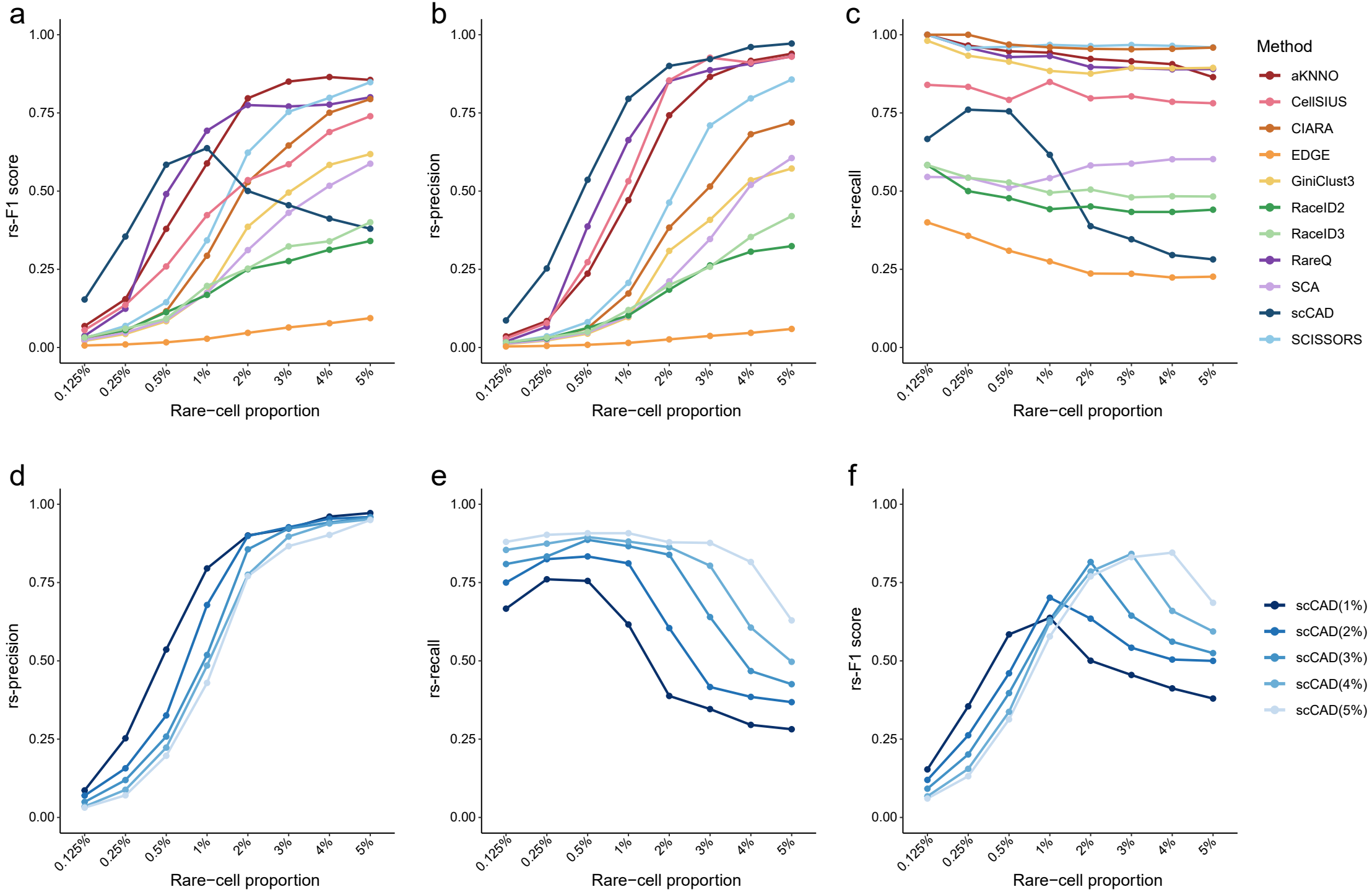

a
b
c
d
e
f
rs-F1 score
rs-precision
rs-recall
1.00
0.75
0.50
0.25
0.00
0.125%
0.25%
0.5%
1%
2%
3%
4%
5%
Rare−cell proportion
Method
aKNNO
CellSIUS
CIARA
EDGE
GiniClust3
RaceID2
RaceID3
RareQ
SCA
scCAD
SCISSORS
scCAD(1%)
scCAD(2%)
scCAD(3%)
scCAD(4%)
scCAD(5%)

Figure 3

Figure 3 | Effect of rare-cell proportion on performance.

a–c, Line plots showing the relationships between rare-cell proportion and (**a**) rs-F1 score, (**b**) rs-precision, and (**c**) rs-recall for the 11 evaluated methods.

d–f, Line plots showing the relationships between rare-cell proportion and (**d**) rs-precision, (**e**) rs-recall, and (**f**) rs-F1 score for scCAD with different values of the *rare_h* parameter.

In all six panels, the x-axis represents rare-cell proportion, and the y-axis represents the median value calculated across the corresponding benchmarking scenarios.

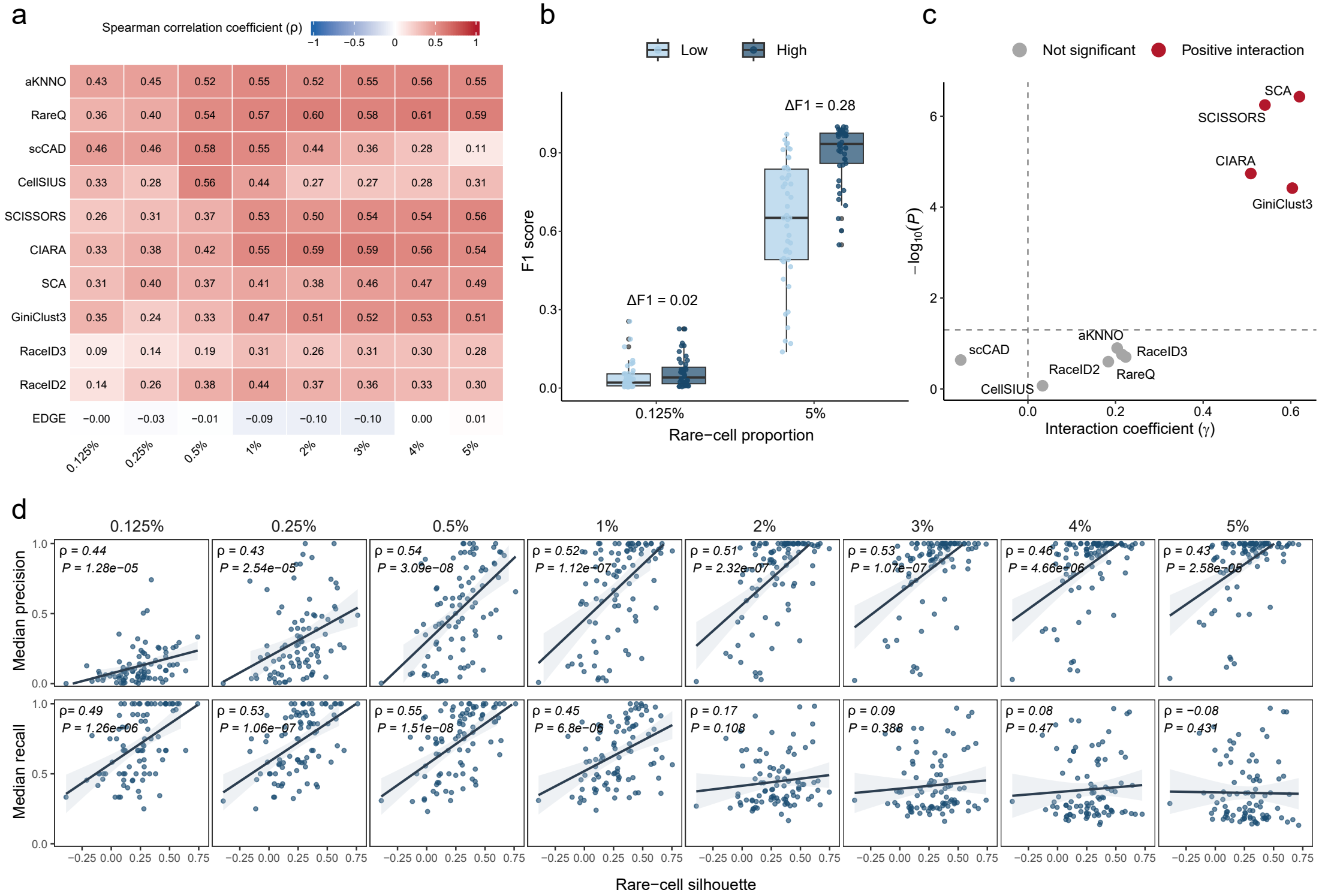

a
Spearman correlation coefficient (ρ)
−1
−0.5
0
0.5
1
aKNNO 0.43 0.45 0.52 0.55 0.52 0.55 0.56 0.55
RareQ 0.36 0.40 0.54 0.57 0.60 0.58 0.61 0.59
scCAD 0.46 0.46 0.58 0.55 0.44 0.36 0.28 0.11
CellSIUS 0.33 0.28 0.56 0.44 0.27 0.27 0.28 0.31
SCISSORS 0.26 0.31 0.37 0.53 0.50 0.54 0.54 0.56
CIARA 0.33 0.38 0.42 0.55 0.59 0.59 0.56 0.54
SCA 0.31 0.40 0.37 0.41 0.38 0.46 0.47 0.49
GiniClust3 0.35 0.24 0.33 0.47 0.51 0.52 0.53 0.51
RaceID3 0.09 0.14 0.19 0.31 0.26 0.31 0.30 0.28
RaceID2 0.14 0.26 0.38 0.44 0.37 0.36 0.33 0.30
EDGE −0.00 −0.03 −0.01 −0.09 −0.10 −0.10 0.00 0.01
0.125% 0.25% 0.5% 1% 2% 3% 4% 5%
b
Low
High
ΔF1 = 0.28
ΔF1 = 0.02
F1 score
0.0
0.3
0.6
0.9
0.125%
5%
Rare−cell proportion
c
Not significant
Positive interaction
SCA
SCISSORS
CIARA
GiniClust3
aKNNO
scCAD
RaceID3
RaceID2
RareQ
CellSIUS
$-\log_{10}(P)$
0
2
4
6
0.0
0.2
0.4
0.6
Interaction coefficient (γ)
d
0.125%
0.25%
0.5%
1%
2%
3%
4%
5%
Median precision
ρ = 0.44 P = 1.28e−05
ρ = 0.43 P = 2.54e−05
ρ = 0.54 P = 3.09e−08
ρ = 0.52 P = 1.12e−07
ρ = 0.51 P = 2.32e−07
ρ = 0.53 P = 1.07e−07
ρ = 0.46 P = 4.66e−06
ρ = 0.43 P = 2.58e−05
Median recall
ρ= 0.49 P = 1.26e−06
ρ = 0.53 P = 1.06e−07
ρ = 0.55 P = 1.51e−08
ρ = 0.45 P = 6.8e−06
ρ = 0.17 P = 0.108
ρ = 0.09 P = 0.388
ρ = 0.08 P = 0.47
ρ = −0.08 P = 0.431
0.0
0.5
1.0
−0.25 0.00 0.25 0.50 0.75
Rare−cell silhouette

Figure 4

Figure 4 | Association between rare-cell separability and performance.

a, Heatmap of Spearman correlations between rare-cell silhouette score and F1 score for each method across eight rare-cell proportions. Colors and cell values indicate Spearman correlation coefficients (ρ).

b, Boxplots comparing F1 scores between low- and high-silhouette groups for SCISSORS at rare-cell proportions of 0.125% and 5%. In each box, the central line represents the median value and the bounds represent the 25th and 75th percentiles (interquartile range). The whiskers encompass 1.5 times the interquartile range. ΔF1 indicates the difference in median F1 score between the high- and low-silhouette groups.

c, Volcano plot of the interaction analysis between rare-cell silhouette score and rare-cell proportion. The x-axis represents the fitted interaction coefficient (γ), and the y-axis represents −log10-transformed *P* values. The two dashed lines indicate $\gamma = 0$ and $P = 0.05$, respectively.

d, Scatterplots showing the relationships between rare-cell silhouette score and precision (top) or recall (bottom) for scCAD at eight rare-cell proportions. Each point represents a benchmarking scenario, and lines indicate fitted linear trends. Spearman correlation coefficients (ρ) and corresponding *P* values are shown in each panel. The *P* values were calculated using Spearman's rank correlation test with an asymptotic *t* approximation.

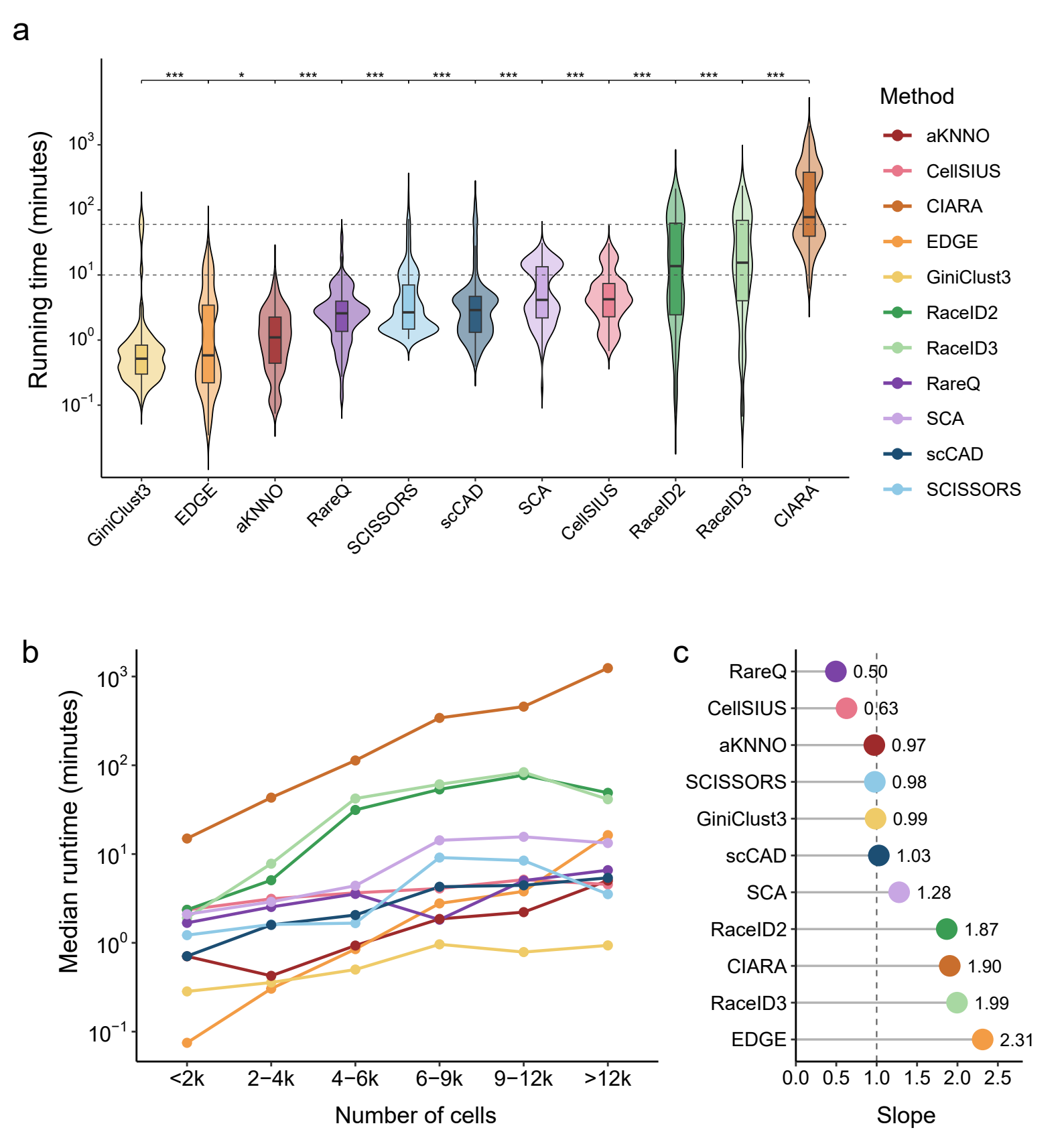

a
Running time (minutes)
10^3
10^2
10^1
10^0
10^{-1}
***
*
***
***
***
***
***
***
***
***
GiniClust3
EDGE
aKNNO
RareQ
SCISSORS
scCAD
SCA
CellSIUS
RaceID2
RaceID3
CIARA
Method
aKNNO
CellSIUS
CIARA
EDGE
GiniClust3
RaceID2
RaceID3
RareQ
SCA
scCAD
SCISSORS
b
Median runtime (minutes)
10^3
10^2
10^1
10^0
10^{-1}
<2k
2−4k
4−6k
6−9k
9−12k
>12k
Number of cells
c
RareQ 0.50
CellSIUS 0.63
aKNNO 0.97
SCISSORS 0.98
GiniClust3 0.99
scCAD 1.03
SCA 1.28
RaceID2 1.87
CIARA 1.90
RaceID3 1.99
EDGE 2.31
0.0 0.5 1.0 1.5 2.0 2.5
Slope

Figure 5

Figure 5 | Runtime efficiency of evaluated methods.

a, Violin plots showing the distribution of runtime for the evaluated methods. In each box, the central line represents the median value and the bounds represent the 25th and 75th percentiles (interquartile range). The y-axis is shown on a log10 scale, and the dashed horizontal lines indicate runtimes of 10 and 60 minutes. The *P* values were calculated using paired Wilcoxon signed-rank tests. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

b, Line plots showing the relationship between runtime and the total number of cells. The y-axis is shown on a log10 scale.

c, Ranking of methods according to their estimated slope values. The dashed vertical line indicates a slope of 1.

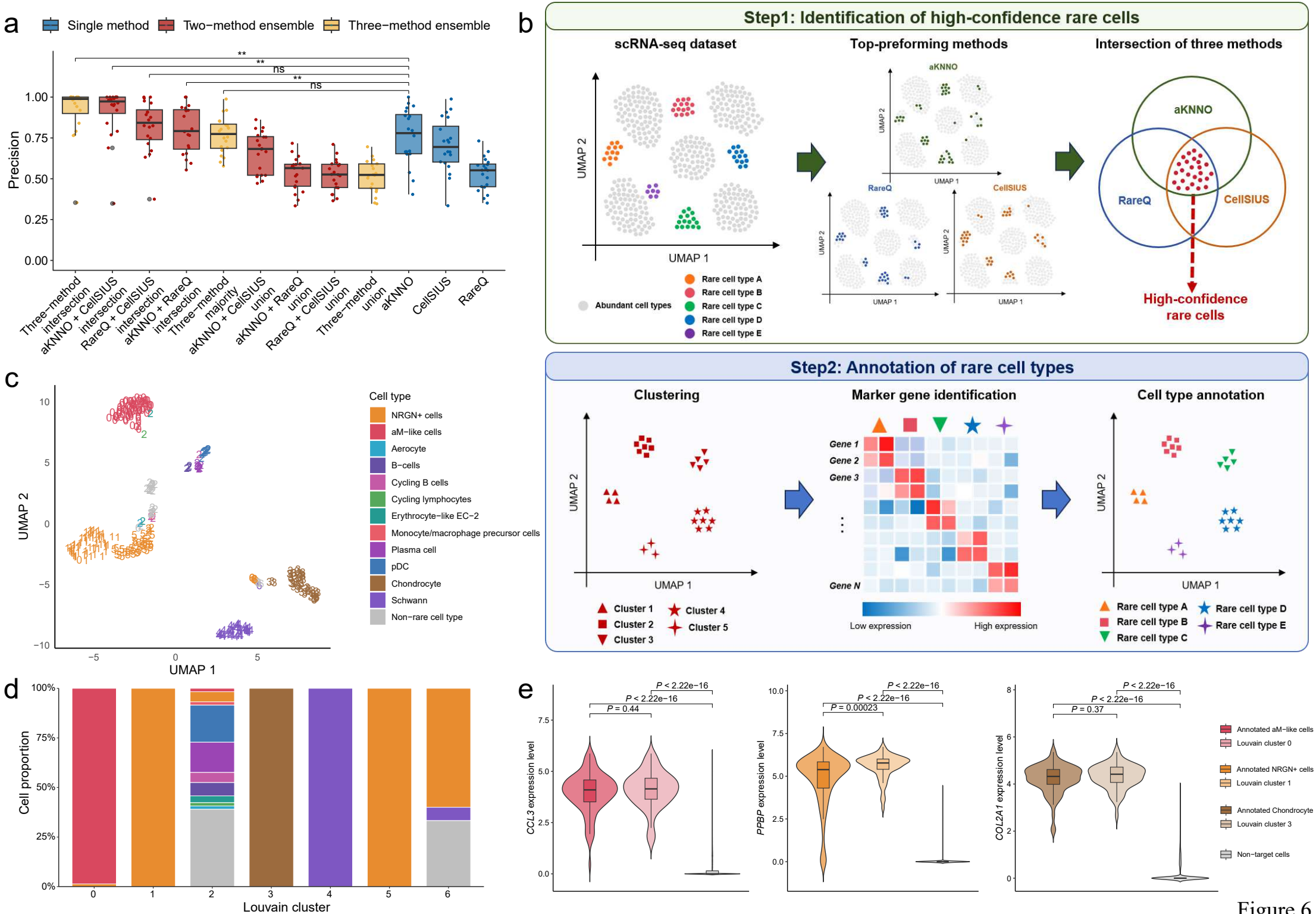

a
Single method
Two−method ensemble
Three−method ensemble
Precision
Three−method intersection
aKNNO + CellSIUS intersection
RareQ + CellSIUS intersection
aKNNO + RareQ intersection
Three−method majority
aKNNO + CellSIUS union
aKNNO + RareQ union
RareQ + CellSIUS union
Three−method union
aKNNO
CellSIUS
RareQ
b
Step1: Identification of high-confidence rare cells
scRNA-seq dataset
Top-preforming methods
Intersection of three methods
Abundant cell types
Rare cell type A
Rare cell type B
Rare cell type C
Rare cell type D
Rare cell type E
High-confidence rare cells
Step2: Annotation of rare cell types
Clustering
Marker gene identification
Cell type annotation
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Low expression
High expression
c
Cell type
NRGN+ cells
aM−like cells
Aerocyte
B−cells
Cycling B cells
Cycling lymphocytes
Erythrocyte−like EC−2
Monocyte/macrophage precursor cells
Plasma cell
pDC
Chondrocyte
Schwann
Non−rare cell type
d
Cell proportion
Louvain cluster
e
Annotated aM−like cells
Louvain cluster 0
Annotated NRGN+ cells
Louvain cluster 1
Annotated Chondrocyte
Louvain cluster 3
Non−target cells

Figure 6

Figure 6 | Practical evaluation of ensemble strategies for rare cell detection.

a, Boxplots of the precision of multiple ensemble strategies. For reference purposes, the precision of each method is also shown. In each box, the central line represents the median value and the bounds represent the 25th and 75th percentiles (interquartile range). The whiskers encompass 1.5 times the interquartile range. The *P* values were calculated using paired Wilcoxon signed-rank tests. **, $P < 0.01$; ns, not significant.

b, Schematic overview of the proposed two-step workflow for rare cell identification and annotation.

c, UMAP visualization of the high-confidence rare cells identified by the three-method intersection in sample GSM8217713. Numbers indicate cluster indices, and colors indicate the original cell type annotations.

d, Cellular composition of the seven clusters shown in panel (**c**). Each stacked bar shows the proportion of cells with different original cell type annotations within each cluster.

e, Violin plots showing the distribution of marker gene expression levels in originally annotated rare cell types, corresponding clusters, and non-target cells in sample GSM8217713. In each embedded box, the central line represents the median value and the bounds represent the 25th and 75th percentiles (interquartile range). The *P* values were calculated using Wilcoxon rank-sum tests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.materials.pdf