nature communications

Article

https://doi.org/10.1038/s41467-024-51891-9

scCAD: Cluster decomposition-based anomaly detection for rare cell identification in single-cell expression data

Received: 5 February 2024

Accepted: 15 August 2024

Published online: 31 August 2024

Check for updates

Yunpei Xu[1,2,3,6], Shaokai Wang[4,6], Qilong Feng[1,2,3], Jiazhi Xia[1,3], Yaohang Li[5], Hong-Dong Li[1,2,3] ✉ & Jianxin Wang[1,2,3] ✉

Single-cell RNA sequencing (scRNA-seq) technologies have become essential tools for characterizing cellular landscapes within complex tissues. Large-scale single-cell transcriptomics holds great potential for identifying rare cell types critical to the pathogenesis of diseases and biological processes. Existing methods for identifying rare cell types often rely on one-time clustering using partial or global gene expression. However, these rare cell types may be overlooked during the clustering phase, posing challenges for their accurate identification. In this paper, we propose a Cluster decomposition-based Anomaly Detection method (scCAD), which iteratively decomposes clusters based on the most differential signals in each cluster to effectively separate rare cell types and achieve accurate identification. We benchmark scCAD on 25 real-world scRNA-seq datasets, demonstrating its superior performance compared to 10 state-of-the-art methods. In-depth case studies across diverse datasets, including mouse airway, brain, intestine, human pancreas, immunology data, and clear cell renal cell carcinoma, showcase scCAD's efficiency in identifying rare cell types in complex biological scenarios. Furthermore, scCAD can correct the annotation of rare cell types and identify immune cell subtypes associated with disease, thereby offering valuable insights into disease progression.

Single-cell RNA sequencing (scRNA-seq) technologies have enabled researchers to analyze gene expression patterns at the single-cell level[1], thereby dissecting cellular heterogeneity[2], while providing new insights into understanding the composition and function of cell types within complex tissues[3]. With the advancement of sequencing technology, larger datasets become available[4], enabling not only characterizing the major cell types but also capturing low-frequency cell types[5–7]. These rare cell types exhibit low abundance and have been extensively validated for their significant roles in disease pathogenesis and biological processes such as angiogenesis and immune response mediation. For example, circulating tumor cells (CTCs) are indeed very rare in peripheral blood but their metastasis is closely associated with cancer-related death. It is estimated that CTCs account for 1 or fewer cells in every 10^5–10^6 peripheral blood mononuclear cells (PBMCs)[8]. This limited presence of CTCs poses a substantial challenge to their detection and characterization in cancer research[9,10]. Therefore, in addition to commonly used tools like Seurat[11] that comprehensively identify major cell types, developing specialized methods to accurately and effectively identify and characterize these rare cell types has become a major challenge in single-cell research.

[1]School of Computer Science and Engineering, Central South University, Changsha, China. [2]Xiangjiang Laboratory, Changsha, China. [3]Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha, China. [4]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada. [5]Department of Computer Science, Old Dominion University, Norfolk, VA, USA. [6]These authors contributed equally: Yunpei Xu, Shaokai Wang. ✉e-mail: hongdong@csu.edu.cn; jxwang@mail.csu.edu.cn

Prominent algorithms used in recent years for the identification and analysis of rare cell types include Finder of Rare Entities (FiRE)[12], CellSIUS[13], Ensemble method for simultaneous Dimensionality reduction and feature Gene Extraction (EDGE)[14], GapClust[15], GiniClust series methods[16–18], RaceID series methods[19,20], SCISSORS[21], CIARA[22], and surprisal component analysis (SCA)[23]. These methods identify rare cells from four main perspectives. The first perspective involves proposing a method for measuring cell rarity in highly variable gene space. FiRE employs an efficient Sketching process that assigns each cell a hash code multiple times, with the number of a hash bucket serving as an indicator of the rareness of its resident cells. It then assigns a consensus rareness score for each cell, identifying cells with scores above a threshold as rare. GapClust identifies rare cell types by assessing variations in Euclidean distance between cells and their *k*-nearest neighbors (KNN) within a principal component analysis (PCA) transformed subspace. The second perspective focuses on proposing a feature selection process. GiniClust introduces a novel gene selection method to identify high Gini genes specific to rare cell types and then uses a density-based clustering algorithm to cluster cells. The CIARA algorithm, utilizing KNN, identifies potentially rare cells by examining highly locally expressed genes and then applies the Louvain algorithm to cluster with the selected genes. The third perspective is based on clustering results and proposes a method tailored to identify rare subclusters. CellSIUS identifies candidate marker genes with a bimodal distribution of expression values within each cluster, then further divides cells into sub-clusters by performing one-dimensional k-means clustering based on the mean expression of each gene set with correlated expression patterns. RaceID identifies outlier cells within clusters by evaluating the transcript count variability of every gene across all cells and then reassigns each cell to the most highly correlated cluster. SCISSORS employs silhouette scoring for the estimation of heterogeneity of clusters and identifies rare cells in heterogeneous clusters by a multi-step semi-supervised reclustering process. The last perspective involves proposing dimensionality reduction methods for discriminating rare cells, such as EDGE and SCA. Furthermore, the integration of multi-omics data has emerged as a promising approach. For example, MarsGT[24] combines scRNA-seq and scATAC-seq data, using probabilistic heterogeneous graph transformers for rare cell identification.

Although certain successes have been achieved, these algorithms have limitations in terms of both accuracy and robustness. Methods relying on highly variable genes may overlook specific signals crucial for distinguishing rare cell types, thus being sensitive to the number of differentially expressed genes. When identifying rare cell subpopulations, feature selection-based methods ignore the potential dependence between different genes. Cluster-based methods may require further analysis of the genes used to distinguish rare types within each cluster. Dimensionality reduction methods may lose important information during processing or be susceptible to noise and interference, thereby complicating the accurate identification of rare cells. The method integrating multi-omics data needs to account for potential noise from batch effects and other sources of variation[25], which could complicate the identification of rare cell types.

To overcome these limitations, we propose scCAD, a Cluster decomposition-based Anomaly Detection method to effectively identify rare cell types. In contrast to the existing algorithms, scCAD offers an ensemble feature selection method aimed at maximizing the preservation of differential signals of rare cell types. During cluster decomposition, scCAD applies iterative clustering based on the most differential signals in each cluster to effectively distinguish rare types or subtypes that are initially challenging to differentiate. Finally, scCAD provides the user with several potentially rare cell clusters.

We benchmark scCAD on twenty-five real scRNA-seq datasets, showcasing its superior capability to identify rare cell types. In the majority of these datasets, scCAD exhibits higher identification accuracy compared to other state-of-the-art methods. In case studies across diverse biological scenarios, including mouse airway, brain, intestine, human peripheral blood mononuclear (PBMC), and pancreas, scCAD accurately identifies rare cell types reported in previous studies, showcasing its robustness and accuracy. In clear cell renal cell carcinoma data, scCAD corrects rare cell annotation mistakes and identifies disease-associated immune cell subtypes, providing valuable insights into disease progression. Moreover, the analysis of the results on two large-scale immunology datasets highlights the excellent scalability of scCAD.

Results

Overview of scCAD

Single-cell RNA sequencing data often consist of a diverse range of cell types, each characterized by specific functions and significant variations in cell counts. This can complicate the identification of rare cell types during initial clustering, as they may be indistinguishable from major cell types based partial or on global gene expression.

To tackle this challenge, scCAD employs an ensemble feature selection method to effectively preserve differentially expressed (DE) genes in rare cell types. Similar to GiniClust and CIARA, scCAD emphasizes the importance of the feature selection procedure, which plays a crucial role in clustering. In contrast to traditional approaches that rely solely on the most variable genes for analysis, scCAD combines the most important genes by utilizing initial clustering labels of cells based on global gene expression and a random forest model[26,27]. Then, scCAD proposes an innovative approach by decomposing the major clusters in the initial clustering through iterative clustering based on the most differential signals within each cluster. After cluster decomposition, clusters serve as the fundamental units rather than individual cells. We define the dominant cell type of a cluster as the type to which the majority of cells in the cluster belong. The rarity of specific cell types is reflected in the number of clusters they dominate. The number of clusters dominated by rare cell types is significantly lower than those dominated by major cell types. For improved computational efficiency, scCAD reduces the number of clusters by merging some of the nearest clusters. This is accomplished by merging clusters with the closest Euclidean distance between their centers. The set of clusters obtained from the initial clustering, cluster decomposition, and cluster merging are respectively defined as I-clusters (initial clusters), D-clusters (decomposed clusters), and M-clusters (merged clusters). For each cluster in M-clusters, scCAD utilizes differential expression analysis to identify a specific list of candidate DE genes. Due to limited quantity, rare cell types exhibit a higher degree of independence in the corresponding DE gene list of their respective cluster. scCAD employs an isolation forest model[28] using the candidate DE gene list to calculate the anomaly score of all cells. An independence score is computed by assessing the overlap between highly abnormal cells and those within the cluster, serving as a measure of each cluster's rarity. Figure 1 shows a schematic pipeline of scCAD, and the "Methods" section provides a comprehensive explanation of the step-by-step process in scCAD.

Benchmarking scCAD in real datasets

To comprehensively evaluate scCAD, we compare it with ten state-of-the-art methods designed for identifying rare cell types across twenty-five real scRNA-seq datasets representing diverse biological scenarios. The specifics of these datasets can be found in Supplementary Table 1. The evaluation of different methods is conducted using the F_1 score for rare cell types, which effectively captures the trade-off between precision and sensitivity (Supplementary Table 2 and Fig. 2a). As shown in Fig. 2a and Supplementary Table 2, scCAD achieves the overall highest performance (F_1 score = 0.4172) and exhibits performance improvements of 24% and 48% compared to the second and third-ranked methods (SCA: 0.3359, CellSIUS: 0.2812), respectively.

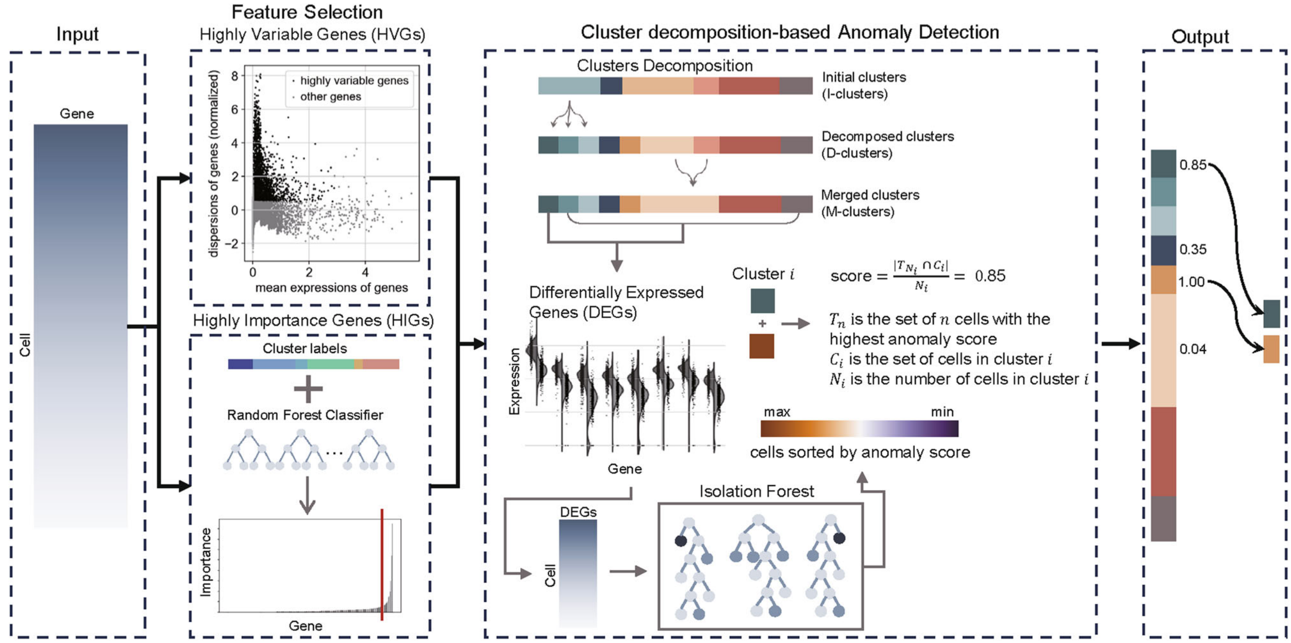
Input
Gene
Cell
Feature Selection
Highly Variable Genes (HVGs)
highly variable genes
other genes
dispersions of genes (normalized)
mean expressions of genes
Highly Importance Genes (HIGs)
Cluster labels
Random Forest Classifier
Importance
Gene
Cluster decomposition-based Anomaly Detection
Clusters Decomposition
Initial clusters (I-clusters)
Decomposed clusters (D-clusters)
Merged clusters (M-clusters)
Differentially Expressed Genes (DEGs)
Expression
Gene
Cluster i
score = $\frac{|T_{N_i} \cap C_i|}{N_i}$ = 0.85
T_n is the set of n cells with the highest anomaly score
C_i is the set of cells in cluster i
N_i is the number of cells in cluster i
max
min
cells sorted by anomaly score
Isolation Forest
DEGs
Cell
Output
0.85
0.35
1.00
0.04

Fig. 1 | Overview of scCAD. scCAD employs an ensemble feature selection approach, combining the benefits of highly variable genes (HVG) and highly important genes (HIG). It then decomposes the major clusters in I-clusters through iterative clustering. To enhance computational efficiency, certain nearest clusters are merged. For each cluster in M-clusters, scCAD conducts anomaly detection by analyzing the corresponding differentially expressed (DE) genes, assigning an independence score to each cluster. Finally, scCAD provides the user with several potential rare cell clusters according to the independence score.

In addition to the F_1 score, we employ four other measurements: the accuracy of identifying rare cell types, G-mean (geometric mean of precision and recall), Cohen's Kappa, and Matthews correlation coefficient (MCC). The accuracy of identifying rare cell types is defined as $ACC_{rare\ cell\ type} = \frac{TRC}{IC}$, where TRC represents the number of correctly identified rare cells and IC represents the total number of cells predicted as rare cell types. Since rare cell identification methods do not provide prediction probabilities, we do not use AUC as an evaluation metric. Supplementary Fig. 1 shows the distribution of the performance measured by using these four metrics across all datasets. The detailed data is provided in Supplementary Tables 3, 4, 5, and 6, respectively. As shown in Supplementary Fig. 1, scCAD demonstrates the overall highest performance (Accuracy = 0.4156, G-mean = 0.4412, Kappa = 0.3933, and MCC = 0.4162) and exhibits performance improvements of 28%, 19%, 26%, and 21% compared to the second-ranked method (SCA: Accuracy = 0.3239, G-mean = 0.3704, Kappa = 0.3128, MCC = 0.3449), respectively.

Furthermore, we showcase the distribution of rankings for different methods across five measurements, including the F_1 score, on each dataset (Supplementary Fig. 2). As shown in Supplementary Fig. 2, scCAD is one of the top three algorithms on 16 (MCC and Kappa) and 17 (Accuracy, F_1 score, and G-mean) of the 25 datasets.

During the testing process, we observed that several methods are not sufficiently adaptable to all datasets representing diverse scenarios. For example, the series methods of GiniClust may introduce errors by failing to identify high Gini genes[29]. Furthermore, certain methods (such as RaceID) may encounter challenges in generating results for datasets containing more cells due to lower computational efficiency[12]. In contrast, scCAD, EDGE, FiRE, SCA, and SCISSORS can run effectively on all datasets, showcasing their greater suitability for data analysis across a wide range of biological scenarios.

To further evaluate the performance of these algorithms, we count the total number of datasets in which each method successfully identifies at least one rare cell type (Fig. 2b). Let S_{pre} be the set of rare cells identified by one method and S_t be the set of cells for rare type t. $|S|$ denotes the size of the set S. if $\left|\frac{S_{pre} \cap S_t}{S_t}\right|$ is larger than 30%, we consider that this method can successfully identify cell type t. The setting of this approach is inspired by certain cell annotation methods[30], which suggest that accurately annotating a cell type can be achieved based on the information from 30% of cells belonging to that type. The total number of rare cell types successfully identified by different methods across all datasets is shown in Supplementary Table 7. As illustrated in Fig. 2b and Supplementary Table 7, scCAD demonstrates advantages by successfully identifying rare types in 20 datasets. Meanwhile, we further compare scCAD with four other methods (CellSIUS: 16, EDGE: 16, GapClust: 12, and SCA: 18). By combining Supplementary Tables 2 and 7, we calculate the average F_1 score of these five methods on the corresponding datasets where they successfully identified rare cell types. scCAD also demonstrates an advantage (F_1 score = 0.5208) compared to the other methods (CellSIUS: 0.3339, EDGE: 0.3954, GapClust: 0.2940, and SCA: 0.4661).

We observe significant variation in the number of rare cell types across different datasets in Supplementary Table 7, in which there are 11 datasets with two or more rare cell types and 14 datasets with only one rare cell type. As shown in Supplementary Table 7, scCAD can identify the rare cell type in, 10 of the 14 datasets and also identify 2 or more rare cell types in 8 of the 11 datasets. In summary, scCAD excels at identifying rare cell types in diverse biological scenarios.

Feature selection effectively preserves the rare cell type-specific genes

Feature selection is crucial for identifying rare cell types, as it aids in extracting and preserving key features specific to these types, thereby reducing noise and redundant information, and improving the ability to identify and distinguish these types. Most current methods for identifying rare cell types rely on a specific set of highly variable genes (HVG)[12,13,15], which exhibit significant expression changes across cells, thus potentially providing more information. Our previous study illustrated that highly important genes (HIG) based on random forests have been demonstrated to enhance clustering performance[27]. The gene selection strategy of scCAD involves merging and removing duplicates from the top 2000 HVG and the top 2000 HIG.

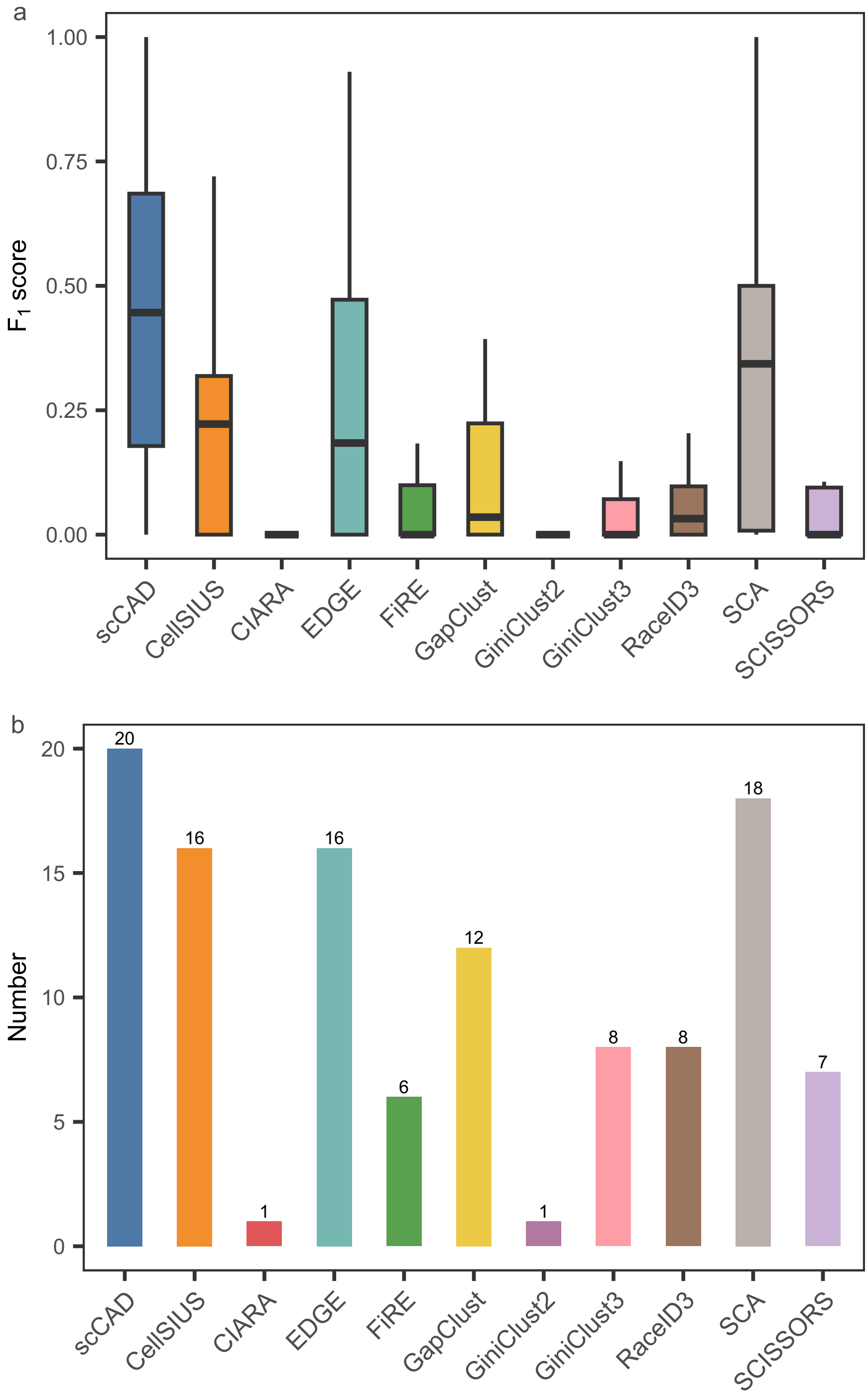

Fig. 2 | Evaluating scCAD against ten state-of-the-art methods for identifying single-cell rare types on twenty-five real datasets. a Comparing the distribution of F$_1$ scores across all datasets ($n = 25$ datasets) in identifying rare cell types. Boxes extend from the first to the third quartile (Q1–Q3) with a line in the middle that represents the median. Lines extending from both ends of the box indicate variability outside Q1 and Q3. The minimum/maximum whisker values are calculated as Q1/Q3 $-/+1.5\times$IQR. **b** Comparison of the total number of datasets in which each method successfully identifies at least one rare cell type. Source data are provided as a Source Data file.

To demonstrate the effectiveness of this strategy, we assess whether the genes selected by scCAD encompass genes specific to rare cell types. Specifically, we first apply Wilcoxon's rank sum test to identify the top 50 differentially expressed (DE) genes for each rare cell type in the dataset, which are commonly utilized to indicate the type's differential signals[31,32]. Then we collect these genes to form a reference gene set S_{ref}, which is regarded as having rare cell type-specific signals. Assume that S_{select} is the selected gene set, we define three overlap rates: *OR1*, *OR2*, and *OR3*, using the following formulas: $OR1 = \frac{|S_{ref} \cap S_{select}|}{|S_{ref}|} \times 100\%$, $OR2 = \frac{|S_{ref} \cap S_{select}|}{|S_{ref} \cup S_{select}|} \times 100\%$, $OR3 = \frac{|S_{ref} \cap S_{select}|}{|S_{select}|} \times 100\%$. A higher overlap rate indicates a stronger presence of rare cell differences in the selected features. We simultaneously compare scCAD with two individual strategies across all datasets (Supplementary Table 8). To maintain fairness, we keep the number of features for highly variable genes (HVG) and highly informative genes (HIG) the same as the number of features

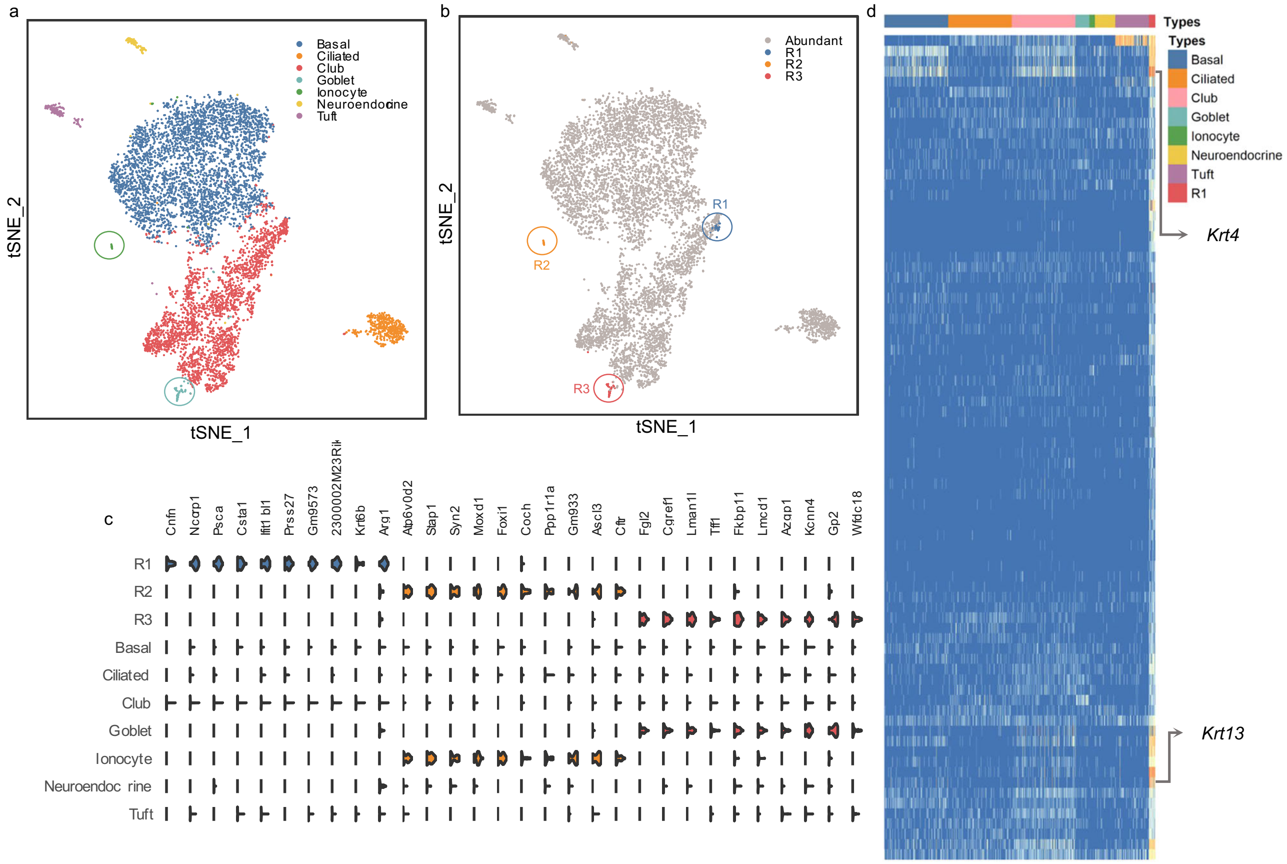

Fig. 3 | Visualization analysis of scCAD's results in airway epithelial. a The t-SNE-based 2D embedding of the cells with color-coded identities. Ionocytes and Goblet cells are specifically marked with circles. **b** The three rare cell clusters detected by scCAD are visually distinguished using different colors. **c** Violin plots showing the expression distribution of the most differentially up-regulated genes in each identified cell cluster. Additionally, seven annotated cell types reported by Montoro et al. are used for comparison. Genes within the same cell cluster are indicated with the same color. **d** The expression of all genes differentially up-regulated in cluster R1 is examined across all cell types, including cluster R1 itself. Source data are provided as a Source Data file.

ultimately selected by scCAD. Supplementary Table 8 shows the overlap rates *OR1*, *OR2*, and *OR3* between the reference gene set and the results of three gene selection strategies across all datasets. Supplementary Fig. 3a shows the distribution of the overlap rate *OR1* between genes selected by three strategies and the reference genes of rare cell types across all datasets.

As shown in Supplementary Table 8 and Supplementary Fig. 3a, the overlap rate *OR1* reveals that, on average, 86.75% of genes in the reference gene set are present in the genes selected by scCAD, while the corresponding average rates for HVG and HIG are 67.80% and 80.20%, respectively. This indicates that when selecting the same number of genes, scCAD can effectively preserve the majority of rare cell type-specific genes. Due to the significant difference in the number of genes between S_{ref} and S_{select}, the values of *OR2* and *OR3* show low relative to that of *OR1*. For the metric *OR2*, the average values for HVG, HIG, and scCAD are 2.34%, 2.74%, and 2.99%, respectively. For the metric *OR3*, the average values for HVG, HIG, and scCAD are 2.39%, 2.77%, and 3.02%, respectively. These results further illustrate that the gene set selected by scCAD contains a sufficient presence of reference genes.

To further analyze the potential impact of clustering accuracy on the reliability of genes selected by the random forest model, we first investigated the Adjusted Rand Index (ARI) of clustering results used for the model (Supplementary Fig. 4). In Supplementary Fig. 4, we find that the clustering results by Louvain indeed exhibit lower accuracy in some datasets, with ARIs around 0.2.

By combining Supplementary Fig. 4 and Supplementary Table 8, we find that the accuracy of the clustering results has a minor impact on the genes selected by scCAD. Using the Chung dataset as an example, even though the ARI is only 0.16, the genes selected by both the RF model and scCAD still encompass over 75% of the rare cell type-specific DE genes. This observation can be attributed to the inherent tendency of most rare cells of the same type to cluster together. Additionally, compared to the sole utilization of the RF model, scCAD demonstrates better robustness due to its combination of two feature selection strategies. Using the Goolam dataset as an example, while the genes selected by the RF model cover only 28% of the rare cell type-specific DE genes, scCAD's selection encompasses 61% of these DE genes.

Decomposition effectively isolates clusters dominated by rare cell types

Clusters are commonly annotated based on the primary gene expression patterns of their containing cells, which represent the characteristics of the most dominant cell type. For each cluster, we first count the number of cells of different cell types contained in the cluster based on the annotation information. Then, we identify the cell type with the highest cell count as the dominant type within the cluster. The occupy rate of the dominant cell type in cluster *i* is defined as follows: $P_i = \frac{\max(N_{i,1}, N_{i,2}, \ldots, N_{i,t})}{N_i}$, where $N_{i,j}$ is the number of cells of type *j* in cluster *i*, *t* is the total number of cell types contained in cluster *i*, and N_i is the total number of cells in the cluster *i*. A higher rate serves as an indicator of increased cluster purity, implying that more cells within the cluster belong to the same cell type. For one cell type *j* in one dataset, the proportion of cell type can be calculated as

$\text{mean}(P_{j,1}, P_{j,2}, \ldots, P_{j,l_j})$, where $P_{j,x}$ is the occupy rate of the cell type j in the dominated cluster x and l_j is the number of clusters dominated by cell type j. Subsequently, for each dataset, we separately average proportions of all cell types and rare cell types. To demonstrate the improvement, we compare the average proportions of the clusters from M-clusters with those from I-clusters across all datasets (Supplementary Table 9). For a more intuitive representation, we visually present the comparison results of rare types and all types across all datasets (Supplementary Fig. 3b and Supplementary Fig. 5), respectively.

After cluster decomposition and merging, it becomes evident that the average proportion of cell types within their dominant clusters has significantly increased, especially for rare types, with an average increase from 0.283 to 0.704. Notably, in almost half of the datasets, the initial clustering process fails to identify any rare cell types. In addition, in Supplementary Fig. 3b and Supplementary Table 9, we observe that the average proportions of rare cell types and all cell types in M-clusters are almost higher than those in I-clusters. But we can find from Supplementary Fig. 3b that neither I-clusters nor M-clusters contain clusters dominated by the unique rare cell type present in the data in the Pollen dataset. The reason for the poor results may be due to the poor separability of this type and almost all methods can not identify the rare cell type in the dataset, which can be found in Supplementary Table 2.

To further explore the reliability of cluster decomposition, we investigate the distribution of cells from rare cell types across multiple clusters identified by scCAD. Specifically, using annotation information from the original studies of the datasets, we assess the distribution of cells from all involved rare cell types across clusters at both the initial stage (I-Clusters) and the final stage (M-Clusters).

For I-Clusters containing m-clusters, we first calculate the proportion $p_{t,i}$ of cells of the rare cell type t in cluster i relative to all cells of type t as follows: $p_{t,i} = \frac{n_t^i}{n_t} \times 100\%$, where n_t^i is the number of cells for type t in cluster i, and n_t is the total number of cells for type t in the dataset. It is clear that $\sum_{i=1}^{m} p_{t,i} = 100\%$, where m is the number of clusters in I-Clusters. Then, we calculate the proportion $q_{t,i}$ of cells of the rare cell type t in cluster i relative to all cells in cluster i as follows: $q_{t,i} = \frac{n_t^i}{n_i} \times 100\%$, where n_t^i is the number of cells for type t in cluster i, and n_i is the total number of cells in cluster i. We also calculate these two proportions, $p_{t,i}$ and $q_{t,i}$, in each cluster from M-Clusters.

We sort the proportions $\{p_{t,1}, p_{t,2}, \ldots, p_{t,m}\}$ in descending order and select the top ten clusters. Then, we conduct a joint analysis of the $p_{t,i}$ and $q_{t,i}$ of these selected clusters. Supplementary Figs. (6–10) shows the comparison of $p_{t,i}$ and $q_{t,i}$ across the selected clusters from I-Clusters and M-Clusters. As shown in Supplementary Figs. (6–10), the majority of cells (the average of $p_{t,i}$ is 88%) from rare cell types are found in the same cluster obtained by Louvain at the first stage (I-Clusters) across almost all datasets. Moreover, distinguishing rare cell types from other types during the initial clustering proves to be relatively challenging, with a lower median proportion relative to clusters (I-Clusters, the median of $q_{t,i}$ is 18%). After decomposition and merging, the majority of cells (M-Clusters, the average of $p_{t,i}$ is 79%) from rare cell types remain in the same cluster. Simultaneously, the proportion of cells for the same rare type relative to clusters significantly increases (M-Clusters, the median of $q_{t,i}$ is 81%). Using the analysis results of the Cao dataset as an example (Supplementary Fig. 6), we find that the cells of the rare cell type are distributed across six clusters. Approximately 69% of these cells of rare type are concentrated in the first cluster. The remaining 31% of rare cells are distributed across the other five clusters. After decomposition and merging, the proportion of rare cells in the first cluster is slightly reduced to about 56%, but this cluster exclusively comprises cells of this type ($q_{t,i} = 100\%$).

Overall, although rare cell types may be distributed across multiple clusters, scCAD can effectively isolate the majority of cells for almost all rare cell types in one cluster, which lays the foundation for the subsequent identification of rare cell clusters.

Evaluation of robustness and sensitivity of scCAD

To analyze the robustness and sensitivity of scCAD with respect to the number of differentially expressed (DE) genes, We conduct tests using an artificial scRNA-seq dataset and a Jurkat scRNA-seq dataset. The artificial scRNA-seq dataset comprises 2500 cells and two cell types, with the minor cell type representing approximately 1% of the total population. Further details regarding the generation of this dataset can be found in the "Methods" section. The Jurkat dataset consists of an equal-proportion in vitro mixture of 293T and Jurkat cells[33]. This dataset has been utilized in several previous studies[12–15,34] to simulate the rare cell phenomenon by adjusting the proportion of Jurkat cells. We generate a subsampled dataset by adjusting the proportion of Jurkat cells to 1%. For both datasets, we set aside the pre-identified differentially expressed (DE) genes which are selected through a stringent criterion, and retain all the non-DE genes in the dataset. Additional details about the identification of DE genes and non-DE genes in both datasets can be found in the "Methods" section.

Based on the computational efficiency of the algorithm, we compare scCAD with three rare cell detection algorithms: FiRE, GapClust, and GiniClust3. During each iteration of the experiment, an equivalent number of non-DE genes are substituted with randomly selected pre-identified DE genes. This process is repeated 10 times for each number of DE genes. The average F_1 score across iterations of different methods is compared for each count of DE genes (Supplementary Fig. 11).

As shown in Supplementary Fig. 11, all methods struggled to detect the rare cell type with only a few DE genes, consistent with previous studies[12,15]. However, scCAD progressively segregates cells of rare types from clusters through iterative clustering, thereby reducing its reliance on differential genes and potentially capturing rare cell types with low signals more effectively. As a result, with the introduction of more DE genes, scCAD's performance improved significantly, enabling more precise identification of rare cell types compared to its competitors, especially. FiRE and GapClust require more DE genes to achieve a similar stable prediction result. Among them, only GapClust can achieve the identification accuracy of scCAD when an adequate number of DE genes are utilized. GiniClust3 achieves stability with a similar number of DE genes as scCAD in the Jurkat dataset. However, compared to scCAD, its predictive performance is lower, with an F_1 score of approximately 0.6. Additionally, its performance on simulated data is poor. In summary, scCAD excels even in scenarios with weak differential expression signals among cell types, enabling precise identification of rare cell types and highlighting its robustness.

scCAD enables the identification of rare airway epithelial cell types

The airways of the lungs are a prominent site for diseases such as asthma, where rare cells play pivotal roles in maintaining airway function[35]. Montoro et al.[36] utilized scRNA-seq to examine the cellular composition and hierarchy of mouse tracheal epithelium, providing the expression profile of 7193 cells. They discovered seven cell types, including two rare ones: the *Foxi1*$^+$ lung Ionocyte and Goblet cells. The rare Ionocyte in human bronchi they detected using RNA fluorescent in situ hybridization

We apply scCAD to identify rare airway epithelial cell types. t-distributed Stochastic Neighbor Embedding (t-SNE) serves as a visualization tool for observing the distribution of the cells from various annotated cell types and the distribution of rare cells predicted by scCAD (Fig. 3a, b). The visualization result of Uniform Manifold Approximation and Projection (UMAP)[37] is shown in Supplementary Fig. 24.

scCAD identifies a total of three rare cell clusters, denoted as R1 (0.42%), R2 (0.26%), and R3 (0.57%) (Fig. 3b). To verify the true identity of these identified clusters, we first obtain the rare cell type annotation information from Montoro et al.'s original study. Then, we compare the expression of differentially up-regulated genes in the annotated rare cell types with those in the identified cell clusters. Specifically, we use Wilcoxon's rank sum test to identify differentially up-regulated genes with an FDR cutoff of 0.05 and an inter-group fold-change cutoff of 1.5 for each cluster and each annotated cell type, separately. Assume that S_i is the set consisting of differentially up-regulated genes in identified cluster *i*, and S_j is the set consisting of differentially up-regulated genes in annotated cell type *j*. The Jaccard similarity coefficient between these two gene sets can be calculated as $J_{i,j} = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$. We calculated the similarity between all identified clusters and annotated cell types (Supplementary Table 10). For better visualization, we use the top 10 differentially up-regulated genes for each cluster and compare the identified rare cell clusters with annotated cell types based on the expression distribution of these genes (Fig. 3c). As shown in Supplementary Table 10 and Fig. 3c, clusters R2 and R3 correspond to Ionocytes and Goblet cells, respectively. These two cell types, as indicated in Montoro et al.'s annotation, encompass only 0.90% and 0.36% of cells in the dataset, respectively. The top 50 differentially up-regulated genes in each cluster are detailed in Supplementary Data 1, and we discover that cells within the R2 cluster exhibit classic Ionocyte markers, such as the transgenic *Foxi1*-EGFP, the V-ATPase-subunit gene *Atp6v0d2*, the cystic fibrosis transmembrane conductance regulator (*Cftr*) gene, the transcription factor *Ascl3*, and *Smbd1* (formerly known as *Gm933*)[36,38]. Cells within the R3 cluster exhibit classic markers associated with Goblet-1, a subset of Goblet cells as given in Ref. 34. This cluster is enriched for the expression of genes encoding the key mucosal protein (*Tff1*) and secretory regulator (e.g., *Lman1l*). The visualization results and analysis of other methods are given in Supplementary Fig. 12 and Supplementary Note 1, demonstrating that only scCAD can accurately and simultaneously identify Ionocyte and Goblet cells.

In contrast to the other two clusters, cluster R1, which consists of 30 cells annotated as Club cells, does not have a corresponding annotated cell type. We visualize the expression of all genes that are specifically up-regulated in cluster R1 across both cluster R1 and all other cell types (Fig. 3d). As shown in Fig. 3d, these genes do not show significant expression in other cell types. Interestingly, we note that R1 shares striking similarities with the "hillock" cells identified by Montoro et al. in their analysis of cell differentiation trajectories. These rare transitional cells connect Basal to Club cells through the unique expression of *Krt13* and *Krt4*[39]. Deprez et al. described a population of *Krt13*$^+$ cells in the turbinates, indicating that hillock cells may also exist in other regions of the human respiratory tract[40,41].

scCAD identifies *Foxi1*$^+$ pulmonary ionocytes, hillock cells, and goblet-1 cells, all of which are confirmed by Montoro et al.[36] through immunostaining. Specifically, they confirm that ionocytes are a newly identified cell population in vivo using transgenic *Foxi1*-EGFP reporter mice and *Foxi1* immunoreactivity. They observe immunofluorescence on epithelial tissues, infer trajectories of cell differentiation, and validate the existence of hillock cells. They identify unique *Tff2*$^+$ goblet-1 cells by immunostaining.

scCAD identifies various rare cell subpopulations within the mouse brain

In general, the identification of rare cell types becomes more challenging as the dataset encompasses a larger number of cell types, particularly in datasets with multiple cell subtypes[7]. To demonstrate the effectiveness of scCAD in identifying rare cell subtypes in such datasets, we utilize an existing scRNA-seq dataset including 20,921 cells located in and around the hypothalamic arcuate-median eminence complex (Arc-ME)[42]. This dataset, as indicated in the original annotation, encompasses 36 cell subtypes, with 20 of them being considered rare cell subtypes, accounting for proportions ranging from 0.038% to 0.884%. t-SNE is applied to visualize the distribution of the cells (Fig. 4a). The visualization result of UMAP is shown in Supplementary Fig. 24. For a more intuitive comparison, the cells belonging to rare cell subtypes are color-coded to represent their respective identities in the t-SNE-based 2D embedding (Fig. 4b). scCAD identifies a total of seven rare cell clusters, denoted as R1 (0.87%), R2 (0.63%), R3 (0.40%), R4 (0.50%), R5 (0.12%), R6 (0.11%), and R7 (0.17%) (Fig. 4c). Due to the small number of significantly differentially expressed genes identified in this dataset, we utilize all differentially expressed genes rather than just the up-regulated ones. The Jaccard similarity coefficients between the sets of differentially expressed genes for each cell cluster and each cell type are shown in Supplementary Table 11. For better visualization, we select the top 3 differentially expressed genes for each cluster and compare the identified rare cell clusters with annotated cell subtypes based on the expression distribution of these genes (Fig. 4d). As shown in Supplementary Table 11 and Fig. 4d, clusters R1-R7 identified by scCAD are highly similar to the seven minor cell subtypes reported by the original study, respectively. Among them, cells within the R1 cluster exhibit gene expression patterns similar to the rare cell subtype annotated as s27.oligodendrocyte6[42]. The differentially expressed genes in each cluster are detailed in Supplementary Data 2, and we discover that the expression of several characteristic markers in R1 is associated with a subtype of oligodendrocyte known as NFO (newly formed oligodendrocytes)[43]. NFO represents a distinct stage of oligodendrocyte differentiation. Cluster R1 shows characteristic markers including *Fyn*. Additionally, it shows high expression of *Gpr17*[44], which is involved in oligodendrocyte differentiation, and epigenetic factors such as *Sirt2*, which are also highly transcribed in NFO. Clusters R2 and R7, which include markers such as *Dcn*, *Sparc*, and *Igfbp7*[45–47], show a high degree of similarity to two distinct fibroblast subtypes. Cluster R3 shows a high degree of similarity to Astrocytes, including markers *Sparcl1*, *Slc1a3*, *Slc1a2*, *Slc6a11*, *Glul*, and *Apoe*[48,49]. Cluster R4 shows a high degree of similarity to a subtype of pars tuberalis type 1C, including marker *Cyp2f2*. Cluster R5 shows a high degree of similarity to mural cells, including markers myosin light polypeptide 9 regulatory (*Myl9*), and myosin light polypeptide kinase (*Mylk*)[50]. Cluster R6 closely matches a subtype of neurons from the retrochiasmatic area that highly expresses the *Oxt* gene. Additionally, the visualization results and analysis of other methods are given in Supplementary Fig. 13 and Supplementary Note 2, showing that only scCAD can accurately identify the greatest number of rare cell subtypes without any misidentifications.

scCAD identifies various rare cell types in the crypts of the irradiated mouse intestine

The intestinal epithelium contains various rare cell types, including tuft cells and enteroendocrine cells[51]. Ayyaz et al. conducted scRNA-seq to profile the regenerating mouse intestine and discovered a distinct quiescent cell type called revival stem cell (revSC)[52], which is induced by tissue damage. They validate the rarity of this cell type by using single-molecule fluorescence in situ hybridization (smFISH) for *Clu* expression in non-irradiated small intestines. Whether it is possible to concurrently detect rare cell types, radiation-induced cell types, and revSCs in enriched crypts after irradiation (IR) is an interesting problem. To solve this problem, we utilize scCAD to analyze an existing scRNA-seq dataset containing 6644 single-cell transcriptomes of isolated crypts[52]. Ayyaz et al. reported a total of 19 cell clusters. Among them, the 9th and 10th clusters correspond to Enteroendocrine cells,

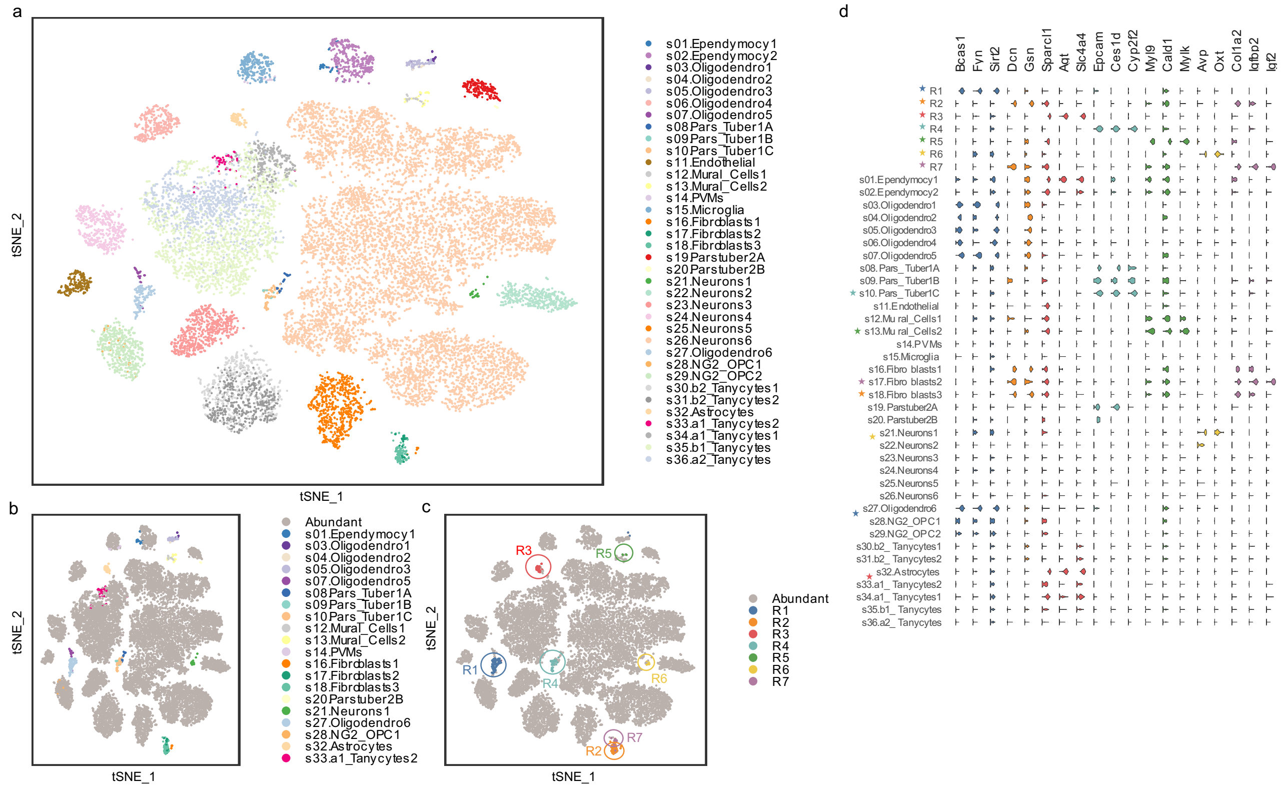

Fig. 4 | Visualization analysis of scCAD's results in mouse brain. a The t-SNE-based 2D embedding of the cells with color-coded identities. **b** The t-SNE-based 2D embedding of the cells. The cells in rare cell subtypes are color-coded to indicate their identities. **c** The seven rare cell clusters identified by scCAD are visually distinguished using different colors. **d** Violin plots showing the expression distribution of the most differentially expressed genes in each identified cell cluster. Additionally, 36 annotated cell types reported by Campbell et al. are used for comparison. Genes within the same cell cluster are indicated with the same color. Cell clusters that have been identified, along with their corresponding cell subtypes, are marked with an asterisk of the same color. Source data are provided as a Source Data file.

the 18th cluster corresponds to newly discovered revSC, and the 19th cluster corresponds to Tuft cells.

scCAD identifies a total of six rare cell clusters, denoted as R1 (0.90%), R2 (0.50%), R3 (0.63%), R4 (0.56%), R5 (0.21%), and R6 (0.69%) (Fig. 5a). Ayyaz et al. did not annotate the real cell types in this dataset and only provided the most differentially expressed genes for each cluster they reported. Therefore, we calculate the Jaccard similarity coefficient between the set of differentially expressed genes for each identified cell cluster (R1~R6) and each reported cluster (Cluster1-Cluster19) as provided by Ayyaz et al. (Supplementary Table 12). For better visualization, we use the top 10 differentially up-regulated genes for each reported cluster and compare the identified rare cell clusters with reported clusters on the expression distribution of these genes (Fig. 5b).

As shown in Supplementary Table 12 and Fig. 5b, we find that Cluster R3, R5, and R6 identified by scCAD are similar to the 9th and 10th clusters, corresponding to Enteroendocrine cells. Clusters R1 and R4 are similar to the 19th and 18th clusters, corresponding to Tuft cells and revSC, respectively. The top 50 differentially up-regulated genes in each cluster we identified are detailed in Supplementary Data 3. We discover that corresponding cell type markers, such as *Dclk1*, *Trpm5*, *Rgs13*[53], and *Chga*[54], are differentially up-regulated in cells within the R1, R3, R5, and R6 clusters. Cluster R4 exhibits gene expression characteristics similar to revSCs, indicating its potential classification as a rare subtype. Notably, we cannot find any clusters reported by Ayyaz et al. that are similar to cluster R2. Consequently, we conduct a more in-depth analysis of the expression of differentially up-regulated genes in cluster R2 across all cells (Fig. 5c).

As shown in Fig. 5c, these genes do not show significant expression in other cells. By querying the PanglaoDB[55] database for cell type markers, we get that a substantial portion (21%) of the differentially up-regulated genes in cluster R2 corresponded to macrophage markers, including CD14 and CD68[56]. Given the potential association of these rare macrophages with radiation exposure, we conduct additional analysis on other differentially expressed genes and identify *NCF2*, *NCF4*, *CYBB*, and *CYBA* among them. These genes have been observed to exhibit differential expression in the lungs of mice following exposure to IR[57]. They play a crucial role in macrophage activation and polarization towards the M2 subtype. Furthermore, the presence of these macrophages indicates alterations in the inflammatory profile of the irradiated lung tissue[58].

scCAD identifies various rare cell types in the human pancreas

The human pancreas comprises various rare cell types such as Epsilon cells[59]. To evaluate the performance of scCAD, we conduct tests on a dataset of 8569 cells from the human pancreas[60]. This dataset encompasses 14 cell types annotated in the original study, with 5 of them being considered rare cell types, accounting for proportions ranging from 0.082% to 0.642%. t-SNE is applied to visualize the distribution of the cells (Fig. 6a). The visualization result of UMAP is shown in Supplementary Fig. 24. The cells belonging to rare cell types are color-coded to represent their respective identities (Fig. 6b). scCAD identifies a total of four rare cell clusters, denoted as R1 (0.56%), R2 (0.16%), R3 (0.18%), and R4 (0.33%) (Fig. 6c). The Jaccard similarity coefficients between the sets of differentially up-regulated genes for each cell cluster and each cell type are shown in Supplementary Table 13. For better visualization, we select the top 10 significantly differentially up-regulated genes for each cluster and compare the identified rare cell clusters with annotated cell types based on the expression distribution of these genes (Fig. 6d).

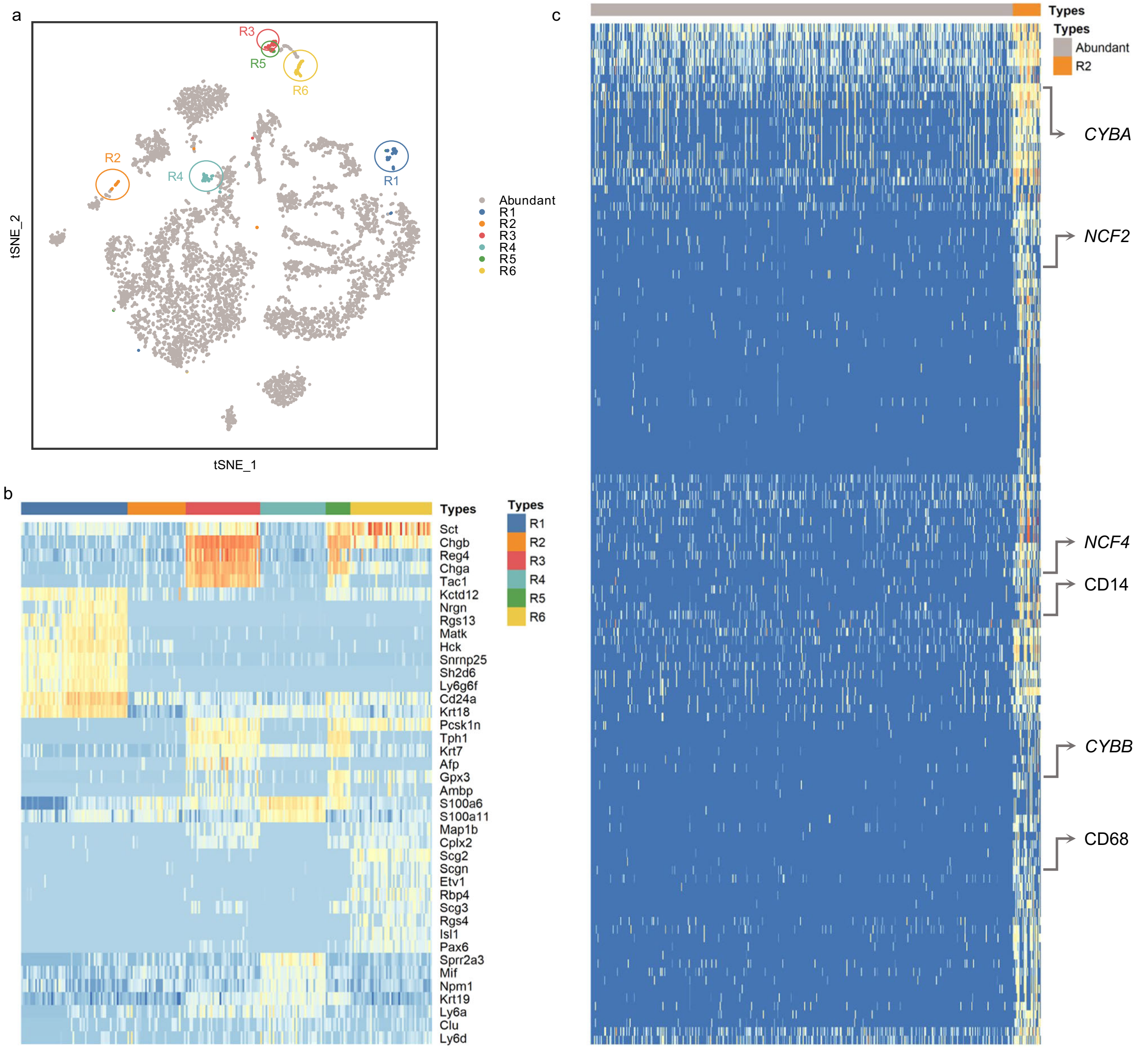

Fig. 5 | Visualization analysis of scCAD's results in mouse intestine. a The t-SNE-based 2D embedding of the cells. The rare cell clusters identified by scCAD are visually distinguished using six distinct colors. **b** Expression of the top 10 differentially up-regulated genes from four reported clusters in the cell clusters identified by scCAD. **c** The expression of differentially up-regulated genes in cluster R2 across all cells. Source data are provided as a Source Data file.

As shown in Supplementary Table 13 and Fig. 6d, we find that clusters R2, R3, and R4 correspond to Epsilon cells, Schwann cells, and Mast cells, respectively. The top 50 differentially up-regulated genes in each cluster are detailed in Supplementary Data 4. We identified distinctive markers associated with these rare cell types within the differentially up-regulated genes, including *GHRL*[61], *NGFR*, *SOX10*[62], *KIT*, and *HDC*[63]. In contrast to these clusters, cluster R1, which consists of 48 cells annotated as Beta cells, does not have a corresponding annotated cell type. By further examination, we observe differential up-regulated genes within R1 and compare these cells with other cells belonging to the Beta cell type (Fig. 6e).

As shown in Fig. 6e, these genes do not show significant expression in other Beta cells. We find that the cells in R1 represent a variant of the Beta cells described by Baron et al.[60] This variant is characterized by variable expression of genes associated with Beta cell function, such as *HERPUD1*, *HSPA5*, and *DDIT3*[64], which are involved in endoplasmic reticulum stress response. Baron et al. pointed out that further work is required to characterize this beta cell variant. The visualization results and analysis of other methods are given in Supplementary Fig. 14 and Supplementary Note 3. The visualization results clearly show that only scCAD can identify the rare Epsilon cells.

scCAD can identify known rare cell types in large-scale immunological single-cell datasets

To assess scCAD's ability to detect rare cell types and subtypes in larger single-cell datasets, we collect two immunological datasets separately. One dataset contains 73,259T cells from 8 human donors[65], and the other contains 39,563 gastrointestinal immune cells from 10 Crohn's disease patients[66]. Both of them are well-annotated by original studies and comprehensive, making the identification results of scCAD more interpretable. We use t-SNE to visualize cell distribution for both datasets (Fig. 7a, b), with color-coding cell subtypes to show their identities. The visualization result of UMAP is shown in Supplementary Fig. 24. To visualize the rare cell types in both datasets, we highlight cell types containing less than 1% of the cells in the T cell dataset (Fig. 7c) and immune cell dataset (Fig. 7d).

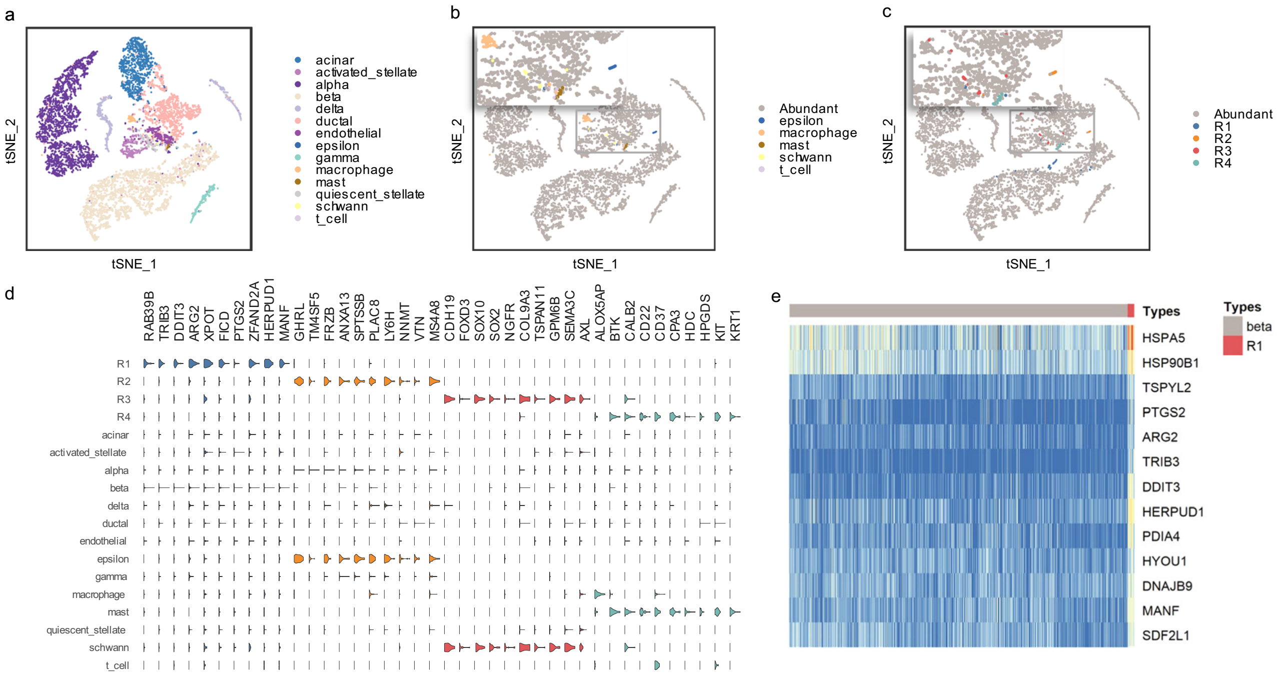

Fig. 6 | Visualization analysis of scCAD's results in human pancreas. a The t-SNE-based 2D embedding of the cells is presented, with color-coded identities indicating cell types. **b** Cells belonging to rare cell types are also color-coded. **c** The rare cell clusters identified by scCAD are visually distinguished using four distinct colors. **d** Violin plots showing the expression distribution of the most differentially expressed genes for the four identified cell clusters. Genes within the same cell cluster are indicated with the same color. **e** The expression of differentially up-regulated genes in beta cells and cells belonging to cluster R1. Source data are provided as a Source Data file.

scCAD identifies two rare cell clusters in the T cell dataset (Fig. 7e): R1 (0.21%) and R2 (0.22%). R1 primarily consists of two types of proliferating cells, CD4 and CD8, with very few annotations in the dataset (0.15% and 0.12% respectively). R1 is mainly composed of double-negative T cells (dnT), which are relatively rare in humans and mice (1~5% of all T cells)[67]. In the immune cell dataset, scCAD identifies four rare cell clusters (Fig. 7f): R1 (0.29%), R2 (0.26%), R3 (0.27%), and R4 (0.07%).

Cluster R1 predominantly consists of mast cells, R2 predominantly consists of pericytes and smooth muscle cells, R3 predominantly consists of lymphocytes, and R4 predominantly consists of glial cells. It's worth noting that these cell types are the top five rarest annotated in this data.

scCAD identifies various unannotated rare cell subtypes in the clear cell renal cell carcinoma dataset

Renal cell carcinomas (RCCs) are a diverse group of malignancies believed to originate from kidney tubular epithelial cells. Various RCC subtypes exhibit a broad range of histomorphology, proteogenomic alterations, immune cell infiltration patterns, and clinical behaviors. The most prevalent subtype is clear cell renal cell carcinoma (ccRCC). We collected a total of 6046 cells annotated into 26 cell clusters from benign adjacent kidney tissues (6 samples from 5 patients) and a total of 20,748 cells annotated into 13 cell types from 7 ccRCC samples[68]. Both of them are utilized to assess the effectiveness of scCAD in the complex tumor microenvironment. The cell type annotation information originates from their original studies. As the visualization results of t-SNE are less discriminative for cell types in these two datasets, we visualize the datasets and their respective annotated rare cell types using the 2D UMAP embedding results (Fig. 8a, b, d, e). The visualization results of t-SNE are shown in Supplementary 20.

In the benign kidney data, scCAD identifies a total of 12 rare cell clusters (0.26%~0.86%) (Fig. 8c). Upon comparing the detailed annotation information, we discover that the dominant cell types of these clusters encompass multiple rare cell types. For instance, cluster R5 primarily consists of B cells, while R9 is mainly composed of mesangial cells. Notably, scCAD identifies two rare proximal tubule (PT) cell subtypes reported by previous studies[68,69], namely PT-B (R12) and PT-C (R1). Zhang et al. confirmed the presence of these two subtypes of cells using RNA in situ hybridization (RNA-ISH) on independent benign kidney tissue samples with select markers.

In the ccRCC data, scCAD identifies a total of 7 cell clusters (0.10% ~0.56%) (Fig. 8f). In addition to CD8$^+$ T cells (R1, R6), mast cells (R2), and plasma cells (R3) annotated as rare cell types, scCAD also identifies three rare cell clusters (R4, R5, and R7). By further examination, we observed differential up-regulated genes within these clusters and compared these cells to other cells of the same annotated type (Fig. 8g–i).

Cluster R4 is annotated as T cells. The top 50 differentially up-regulated genes in these three cell clusters are detailed in Supplementary Data 5. We find that cells in R4 should belong to a rare subtype of effector CD4$^+$ T, named CD4$^+$ effector-GNLY, characterized by high expression of genes associated with cytotoxicity, including *NKG7*, *GZMB*, *GZMH*, and *GNLY*, as given in a previous study[70].

Cluster R5 cells are initially annotated as macrophages. However, we identify multiple markers for dendritic cells, such as CD1C, CD207, and *FCER1A*[71]. Interestingly, Kaplan-Meier analyses of the top 10 differentially expressed genes in R5 reveal an association between high expression levels and increased overall survival in ccRCC (TCGA-KIRC). As shown in Supplementary Fig. 15, high expression of these genes is a positive survival indicator, suggesting that the rare cluster identified by scCAD may provide valuable prognostic information for ccRCC patients.

Cluster R7 comprises 81 cells from the 239 cells annotated as "ua" (Unanalyzed). However, we observe that its differentially expressed genes are all related to hemoglobin, including *AHSP*, *HBD*, and *HEMGN*, indicating that this rare cell cluster may be related to hemoglobin synthesis or related biological processes. From the list of highly

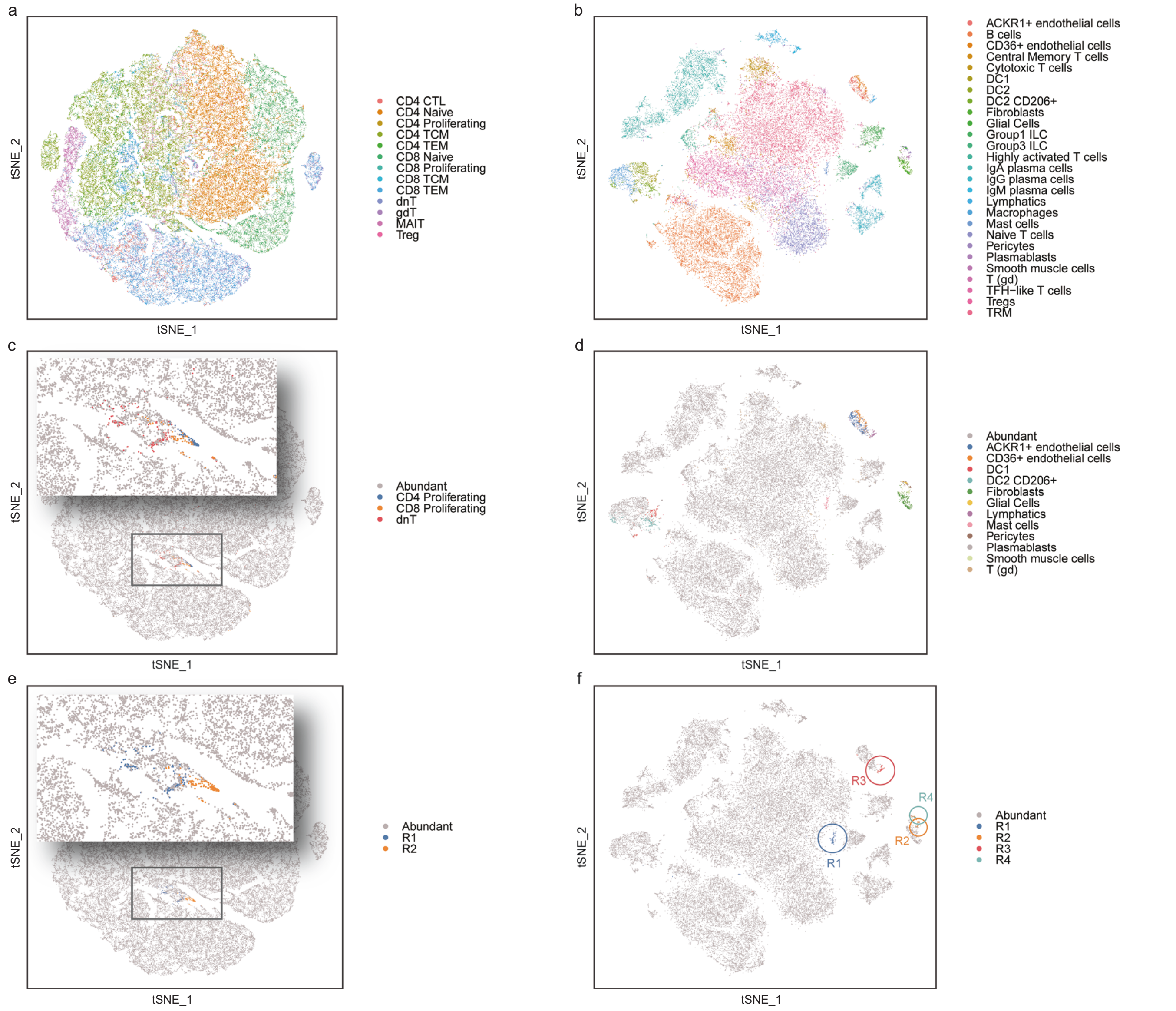

Fig. 7 | Visualization analysis of scCAD's results in two large-scale immunological single-cell datasets. a The t-SNE-based 2D cell embedding with color-coded identities for cell types in the T cell dataset. CTL cytotoxic T cells, TCM T Central Memory, TEM T Effector Memory, dnT double-negative T, gdT gamma-delta T, Treg regulatory T. **b** The t-SNE-based 2D cell embedding with color-coded identities for cell types in the immune cell dataset. DC Dendritic Cell, ILC innate lymphoid cells, T(gd) gamma-delta T, TFH T follicular helper, Tregs regulatory T cells, TRM tissue-resident memory T cells. **c** Cell types comprising less than 1% of cells in the T cell dataset are color-coded. **d** Cell types comprising less than 1% of cells in the immune cell dataset are color-coded. **e** The rare cell clusters identified by scCAD are visually distinguished using two and four distinct colors on the T cell dataset. **f** The rare cell clusters identified by scCAD are visually distinguished using two and four distinct colors on the immune cell dataset. Source data are provided as a Source Data file.

expressed genes (in reads per kilobase per million transcripts) for each stage of erythroid differentiation[72], we conclude that cells in cluster R7 are likely in the polychromatic erythroblast stage. Supplementary Fig. 16 illustrates individual cells color-coded on a 2D embedding plot derived from UMAP, reflecting the RNA expression levels of different marker genes. Overall, scCAD not only accurately identifies rare cell subtypes but also proves useful in correcting rare cell type annotation mistakes. Furthermore, it has the great potential to identify disease-related immune cell subtypes, providing insights into disease progression.

Comparative performance of scCAD against multi-omics approach

The advancement in sequencing technology facilitates the integrative analyses of different types of single-cell omics data, providing insights that are more comprehensive than those from a single type of single-cell omics data[73]. This has the potential to enhance downstream analysis performance. However, this progress also presents challenges, including the introduction of noise due to batch effects among different omics data[74]. We conduct a comparison between scCAD solely based on scRNA-seq data, and MarsGT[24], which integrates both scRNA-seq data and single-cell ATAC sequencing (scATAC-seq) data.

Specifically, we first conduct a comparison between scCAD and MarsGT on four real datasets (PBMC-bench-1, 2, 3, and PBMC-test) obtained from human peripheral blood mononuclear cells, which coincide with the datasets used by MarsGT. scCAD solely utilizes the scRNA-seq data in each dataset, and the specific details of these datasets can be found in Supplementary Table 1. We present the performance of scCAD and MarsGT in identifying rare cell types on these datasets, as measured by F_1 score, precision, and recall

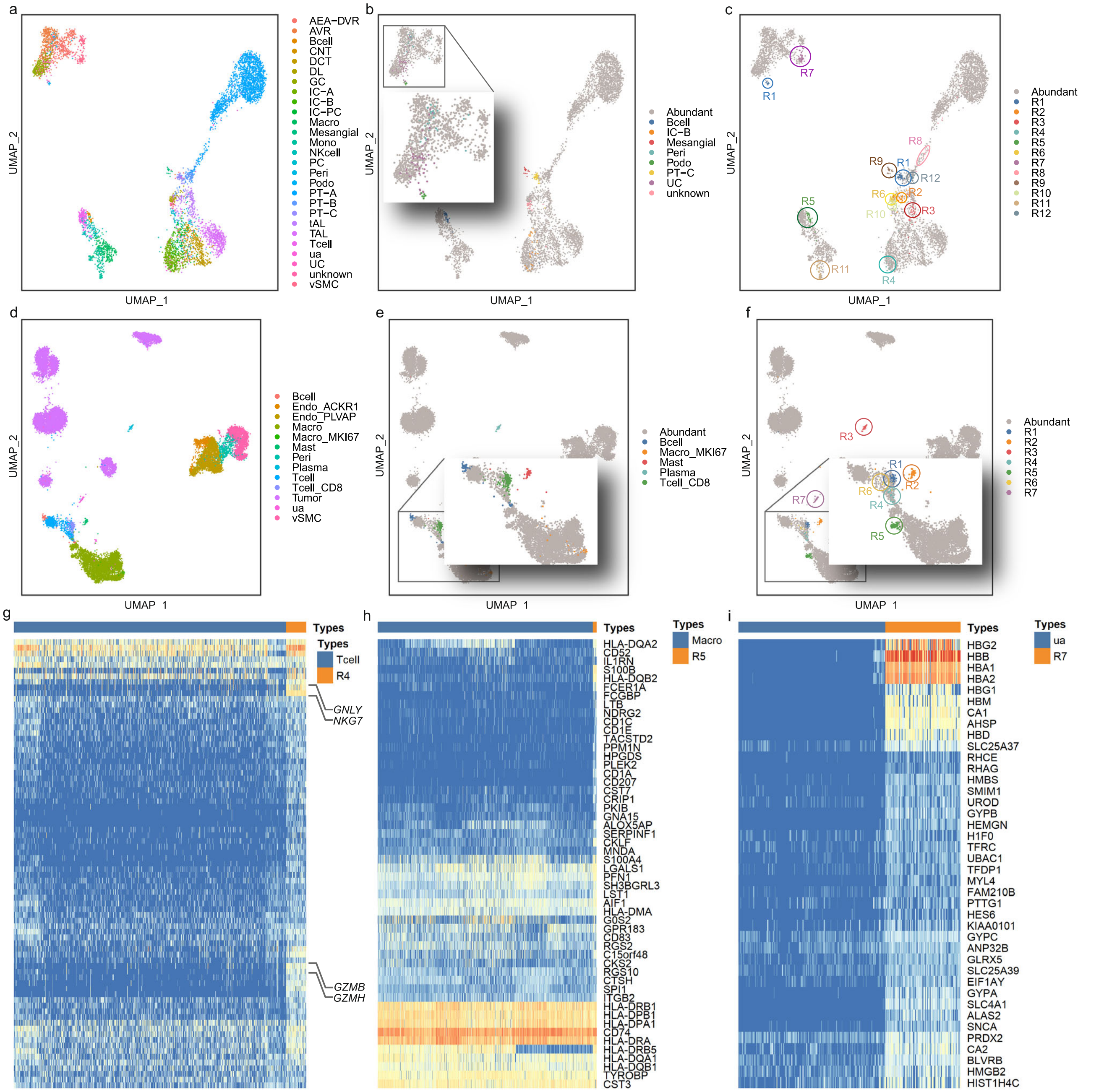

Fig. 8 | Visualization analysis of scCAD's results in clear cell renal cell carcinoma dataset. a UMAP-based 2D visualization depicts cells from the benign kidney, with distinct cell types represented by different color codes. **b** Cell types comprising less than 1% of cells are color-coded. **c** The rare cell clusters identified by scCAD are visually distinguished using twelve distinct colors. **d** UMAP-based 2D visualization depicts cells from the ccRCC, with distinct cell types represented by different color codes. **e** Cell types comprising less than 1% of cells are color-coded. **f** The rare cell clusters identified by scCAD are visually distinguished using seven distinct colors. **g**–**i** Comparing the expression of differentially expressed genes in the identified rare cell cluster and other cells annotated as the same type, from left to right: R4 (**g**), R5 (**h**), R7 (**i**). AEA-DVR afferent/efferent arterioles/descending vasa recta, AVR ascending vasa recta, CNT connecting duct, DCT distal convoluted tubule, DL descending limb, GC glomerular capillaries, IC intercalated cells, PC principal cells, Macro macrophages, Mono monocytes, NK natural killer cells, Peri pericytes, Podo podocytes, PT Proximal tubule, tAL thin ascending limb, TAL thick ascending limb, ua unanalyzed, UC uncharacterized, vSMC vascular smooth muscle cells, Endo endothelial. Source data are provided as a Source Data file.

(Supplementary Table 14). As shown in Supplementary Table 14, scCAD demonstrates slightly superior performance compared to MarsGT in terms of F_1 score and recall, particularly noticeable in the independent test dataset (PBMC-test), which is the dataset primarily used by MarsGT to illustrate its performance. Upon re-examination of these four datasets, we ascertain that they originate from a common dataset totaling 69,249 cells, with each dataset representing a distinct batch and displaying remarkably similar cell type distributions. In Supplementary Table 14, scCAD exhibits greater stability (F_1 score standard deviation of 0.1101) compared to MarsGT (0.1747). This difference may result from the effects of technical variations and noise often encountered in the integrative analyses of diverse single-cell omics data types. Further analysis of the identification results of scCAD on these four PBMC datasets can be found in Supplementary Note 4. The visualization analyses of scCAD's results in these PBMC datasets are shown in Supplementary Figs. 17–20. Jaccard similarity coefficients between the sets of differentially up-regulated genes for each identified cell cluster and each real cell type in these datasets are detailed in

Supplementary Tables 17–20. The genes that are differentially up-regulated in the identified cell clusters across these datasets are detailed in Supplementary Data 7–10. According to the cell type annotation information from their original studies, we find that scCAD not only identifies diverse minor cell types but also uncovers unannotated subtypes. Furthermore, scCAD consistently identifies the same minor cell types across datasets, showcasing its potential for analyzing multiple batches of datasets.

Then, we test whether scCAD, using solely scRNA-seq data, could identify rare cell types in the two single-cell Multi-omics datasets employed in MarsGT's case studies. The two datasets consist of 9383 cells from the mouse retina[75] (Retina dataset) and 14,148 cells obtained from a flash-frozen intra-abdominal lymph node tumor (B_lymphoma dataset) (Supplementary Table 1).

In the Retina dataset, MarsGT reported 12 rare cell clusters, comprising one amacrine cell (AC) cluster, seven bipolar cells (BC) clusters, one horizontal cell (HC) cluster, two Müller glia cell (MG) clusters, and one Rod cell cluster. In contrast, scCAD identifies more rare cell clusters (R1~R19, totaling 19 clusters). According to the annotations provided in [75], these clusters correspond to two AC clusters (R10, R13), one HC cluster (R8), four Rod cell clusters (R7, R9, R17, R18), six BC clusters (R1, R3, R11, R12, R14, R16), and six retinal ganglion cell (RGC) clusters (R2, R4, R5, R6, R15, R19). Given that BC populations are known to encompass numerous rare populations, we further investigated six clusters associated with BC. We visualize the expression of marker genes specific to the BC subpopulation across the six BC clusters (R1, R3, R11, R12, R14, R16) and the 10 BC subtypes annotated in [75] (Fig. 9a).

For better visualization, we compute the Pearson correlation coefficients between the six BC clusters and the 10 BC subtypes based on the average expression values of these marker genes and present them in a heatmap (Fig. 9b). As shown in Fig. 9a, b, we observe that these clusters correspond to distinct BC subtypes, particularly R3, which represents the rarest BC subtype (BC10), accounting for only 3% of all BCs[75], and was not identified by MarsGT. Additionally, RGCs also exhibit multiple subtypes[76], prompting us to further analyze the six RGC clusters identified by scCAD.

Rheaume et al.[77] classify RGCs into 40 subtypes and validate the markers of these subtypes in purified RGCs by fluorescent in situ hybridization (FISH) and immunostaining. We compile a total of 115 uniquely enriched marker genes from 40 RGC subtypes reported by Rheaume et al. (Supplementary Table 15). We visualize the expression of these enriched marker genes across all RGC-related clusters (R2, R4, R5, R6, R15, R19) in Fig. 9c, and we find that these clusters represent various RGC subtypes. Notably, MarsGT only identified a major RGC cluster.

In the B_lymphoma dataset, MarsGT reported a rare state named B lymphoma-state-1. t-SNE is utilized to visualize the cell distribution in the lymphoma dataset (Fig. 9d). scCAD identifies a total of five rare cell clusters (R1~R5) (Fig. 9e). The Jaccard similarity coefficients between the sets of differentially up-regulated genes for each cell cluster identified by scCAD and each cell type annotated by 10X Genomics are presented in Supplementary Table 16. According to Supplementary Table 16, these clusters include one Mono/T mix cluster (R2, 0.27%), one plasmacytoid dendritic cells (pDC) cluster (R3, 0.29%), and one Stromal cell cluster (R5, 0.74%). The top 50 differentially up-regulated genes in each cluster are detailed in Supplementary Data 6. We identified distinctive markers associated with these rare cell types within the differentially up-regulated genes, including CD163[78], *IL3RA*[79], and *CALD1*[80]. Unlike the other clusters, neither cluster R1 (0.35%) nor cluster R4 (0.13%) has a corresponding annotated cell type. Through the analysis of their differentially expressed genes in Supplementary Data 6, we conclude that they likely correspond to gamma-delta T cells and mucosal-associated invariant T (MAIT) cells, as indicated by the up-regulated expression of marker genes *CENPF*[81] and *KLRB1*[82], respectively. In contrast to scCAD, MarsGT did not identify these rare cell types.

Moreover, scRNA-seq data is more readily accessible, thereby streamlining the data acquisition and processing workflow and reducing experimental costs. In summary, scCAD holds advantages in performance, stability, and cost-effectiveness.

scCAD effectively identifies well-validated dendritic cell subtypes

Dendritic cells (DCs) play a central role in pathogen sensing, phagocytosis, and antigen presentation. DCs are one of the rarest types of immune cells, constituting only 1–2% of peripheral blood mononuclear cells (PBMCs)[83]. Villani et al.[79] identified six distinct subtypes of dendritic cells (DCs) by analyzing their expression profiles using fluorescence-activated cell sorting (FACS). They validated the existence of these subtypes by flow cytometry.

To further test the reliability of the rare clusters identified by scCAD, we apply scCAD to the widely used ~68k PBMC dataset and investigate whether any dendritic cell subtypes not captured in the original annotation could be identified. This dataset encompasses 11 cell types annotated in the original study, accounting for proportions ranging from 0.14% to 30.29%. t-SNE is applied to visualize the distribution of the cells (Fig. 10a).

scCAD identifies a total of four rare cell clusters, denoted as R1 (0.50%), R2 (0.24%), R3 (0.13%), and R4 (0.12%) (Fig. 10b). Upon comparing the original annotation, we find that R2 consists of megakaryocytes, a type that makes up only 0.4% of the entire dataset. Additionally, R1 and R4 mainly consist of DCs annotated in the original study, while R3 mainly consists of $CD19^+$ B cells. The top 50 differentially up-regulated genes in these three cell clusters are detailed in Supplementary Data 11.

To explore the true identities of these two DC clusters (R1 and R4), we calculate the correlation between their average expression and that of all well-validated DC subtypes on the same marker gene set. First, we construct a gene set using the top 50 markers for each DC subtype reported by Villani et al.[79] Due to differences between datasets, the filtered gene set consists of 245 markers. Next, we calculate the average expression of this gene set for each DC subtype in the Villani et al. dataset. Simultaneously, we calculate the average expression of this gene set for R1, R4, and all DCs in the ~68k PBMC dataset. Finally, we compute the Pearson correlation coefficient between them (Fig. 10c).

As shown in Fig. 10c, we observe that between the two datasets, clusters R1 and R4 show the highest similarities to DC subtypes DC1 and DC6 (pDC), with similarities of 0.8 and 0.74, respectively, significantly higher than those of other subtypes. Violin plots illustrate the expression distribution of the top markers for each DC subtype across R1, R4, and all DCs (Fig. 10d). Clusters R1 and R4 exhibit significant expression of top markers belonging to DC subtypes DC1 and DC6. By combining Fig. 10c, d, we can confidently determine cluster R1 mapping to $CLEC9A^+$ DCs and cluster R4 to pDCs. By further examination, we observed differential up-regulated genes within R1 and R4 and compared these cells with other DCs (Supplementary Fig. 21a). This highlights their rarity in the dataset.

Cluster R3 cells are initially annotated as $CD19^+$ B cells. Supplementary Fig. 21b compares the expression of differentially up-regulated genes in R3 with other B cells. However, we identify multiple markers for plasma cells, such as CD27 (*TNFRSF17*), *MZB1*, *DERL3*, *ITM2C*, and *IGLL5*[84]. Furthermore, other studies[85,86] have also reported the rarity of plasma cells in this dataset, thus validating our findings.

Discussion

scCAD offers an ensemble feature selection method to maximize the preservation of differential signals of rare cell types, thereby enabling the accurate identification of rare cells. During cluster decomposition, scCAD applies iterative clustering based on the most differential

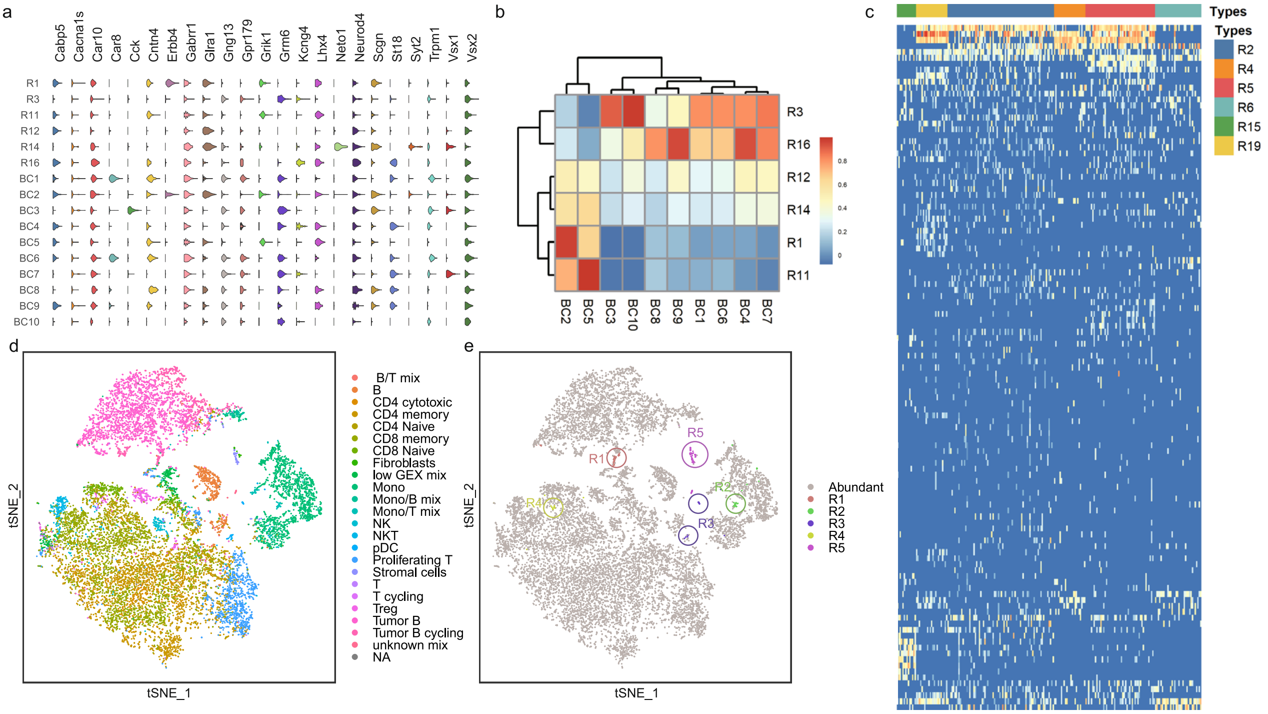

Fig. 9 | Visualization analysis of scCAD's results in mouse retina dataset and human lymphoma dataset. a Violin plots showing the expression distribution of the known marker genes related to BC subtypes across the six identified BC clusters and annotated 10 BC subtypes. **b** The Pearson correlation heatmap between the 6 identified BC clusters and the 10 BC subtypes, is calculated based on the average expression values of BC marker genes. **c** The expression of enriched marker genes from 40 RGC subtypes is examined across all RGC-related clusters (R2, R4, R5, R6, R15, R19). **d** The t-SNE-based 2D embedding of the cells with color-coded identities in the lymphoma dataset. **e** The five rare cell clusters detected by scCAD are visually distinguished using different colors. BC bipolar cells, GEX gene expression, Mono monocytes, NK natural killer Cells, NKT natural killer T Cells, pDC plasmacytoid dendritic cells, Treg regulatory T. Source data are provided as a Source Data file.

signals within clusters to effectively distinguish rare types or subtypes that are initially challenging to differentiate. With the application of the anomaly detection algorithm, scCAD can identify clusters dominated by rare cell types within the cluster decomposition results. Extensive experiment results show that scCAD demonstrates performance advantages across diverse biological scenarios.

Several computational methods have been developed specifically for identifying rare cell types, broadly categorized into four groups based on their methodological characteristics: feature selection, clustering, dimensionality reduction, and rarity measurement. However, single-cell data often arise from diverse and complex biological scenarios, and the performance of many methods is limited due to imperfect assumptions. In contrast, scCAD adeptly addresses these challenges and offers corresponding solutions.

Indeed, several studies have underscored the pivotal role of feature selection in downstream single-cell data analysis[87–89]. It is commonly acknowledged that the efficient selection of marker genes capable of distinguishing rare cells is crucial for their accurate identification. Extensive testing across various scRNA-seq datasets reveals that highly variable genes often do not fulfill this objective well. However, as shown in Supplementary Table 8, rare cell type-specific genes identified by scCAD accounted for over 86% on average among the selected genes.

Given that single-cell data often contain multiple cell types that differ significantly in number and function, it is difficult for initial clustering to distinguish rare types. To overcome this challenge, scCAD conducts cluster decomposition by iteratively clustering based on the most differential signals in each cluster for the first time. The substantial average proportion of dominant rare types within the clusters after decomposition underscores the effectiveness of this approach in isolating rare cell types.

Clusters in M-clusters dominated by rare cell types exhibit stronger rare signals than individual cells. scCAD utilizes an anomaly detection algorithm in the space of cluster-specifically expressed genes. It calculates an independence score based on the overlap of cells in the cluster with highly abnormal cells to determine the rarity. Essentially, the features of rare clusters exhibit higher independence, leading to a clear distinction of cells within them from cells belonging to the major cluster in this feature space.

In the benchmark on 25 real-world datasets, scCAD outperforms 10 other state-of-the-art methods in accuracy, successfully identifying rare cell types in 20 of them. In-depth case studies across various complex biological scenarios, including mouse airway epithelium, hypothalamic arcuate-median eminence complex (Arc-ME), irradiated mouse intestinal crypts, human pancreas, and large-scale human immunology cells, demonstrate scCAD's capability in accurately identifying rare cell types and subtypes. This holds even in cases involving radiation and multiple subtypes. Furthermore, in the clear cell renal cell carcinoma (ccRCC) dataset, scCAD not only rectifies the annotation of rare cell types but also identifies rare subtypes not discovered in the original article. Importantly, in diverse datasets containing immune cells, scCAD identifies multiple immune cell subtypes associated with disease, potentially providing valuable insights into disease progression.

In summary, scCAD has demonstrated its effectiveness as a tool for identifying rare cells, showcasing its high accuracy, sensitivity, and robust generalization capabilities across various biological scenarios.

Methods

Data preprocessing

We preprocess the gene expression matrix as follows. First, we filter out genes with low expression rates, which may not provide effective

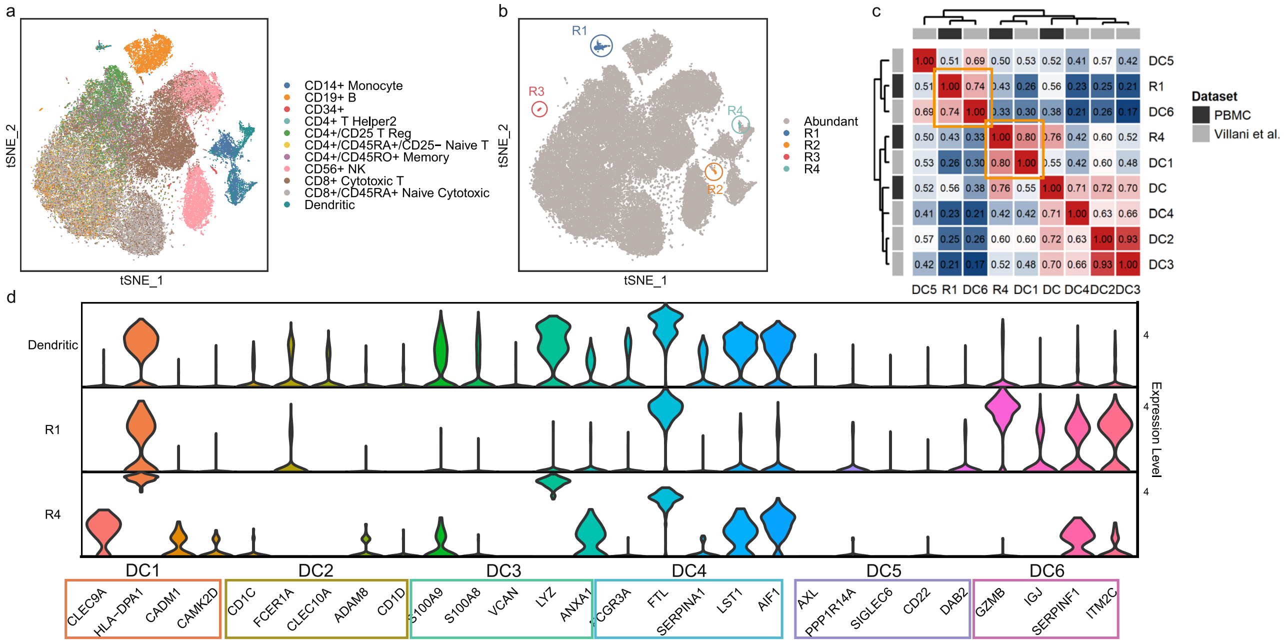

Fig. 10 | Visualization analysis of scCAD's results in ~68k PBMC dataset. a The t-SNE-based 2D embedding of the cells is presented, with color-coded identities indicating cell types. Reg regulatory, NK natural killer Cells. **b** The rare cell clusters identified by scCAD are visually distinguished using four distinct colors. **c** The Pearson correlation heatmap compares the two identified dendritic cell (DC) clusters, all annotated DCs, and six validated DC subtypes, based on the average expression values of common DC marker genes across the two datasets. **d** The expression distribution of the top marker genes related to six DC subtypes across the two DC clusters and all annotated DCs. Source data are provided as a Source Data file.

information. Specifically, genes that are expressed in at least three cells are retained for downstream analysis. Each scRNA-seq dataset is normalized by using the log-normalization procedure including the calculation of cell-specific size factor based on the sequencing depths, and normalization. The normalized matrix is then log2-transformed after adding 1 as a pseudo-count.

Rapid clustering module for single-cell analysis

scCAD utilizes this rapid single-cell clustering module multiple times to assign cluster labels to all or a subset of cells. Similar to previous works[90,91], scCAD first applies PCA to obtain the top 40 principal components (PCs) that represent the most differential signals in the data. Then, the cell undirected graph is constructed using Euclidean distance and the KNN algorithm with the nearest neighbor parameter k set to 15. Finally, the graph-based community detection algorithm, such as Louvain[92], is used to assign cluster labels to cells.

The procedure of scCAD method

scCAD integrates an ensemble feature selection method and a cluster decomposition-based anomaly detection score step. Specifically, scCAD involves the following detailed procedures after data preprocessing.

1. The ensemble feature selection process involves selecting highly variable genes[11] and highly discriminative genes[27]. Specifically, scCAD calculates the mean and a dispersion measure (variance/mean) for each gene across all single cells, selecting the top 2000 most variable genes that exhibit high variability compared to genes with similar average expression. At the same time, a random forest model is trained using the preprocessed gene expression matrix and cluster labels, and the importance of each gene is calculated based on the Gini impurity obtained from a set of decision trees[26]. Next, scCAD selects the top 2,000 genes with the highest importance. Finally, the combined set of genes from both selections is retained for subsequent analysis.
2. Using the preprocessed expression matrix with selected genes, scCAD performs cell clustering and initially partitions n cells into several clusters. The set of clusters obtained from the initial clustering is defined as I-clusters (initial clusters). Then, scCAD iteratively decomposes each cluster containing more than $R = 1\%$ of the total number of cells ($R * n$) through clustering until no new clusters are generated by using the Louvain method, or until all clusters become smaller than $R * n$. The set of clusters obtained from cluster decomposition is defined as D-clusters (decomposed clusters). Subsequently, clusters in D-clusters will be merged based on a threshold to form the final clusters set (M-clusters). Specifically, scCAD first determines the centers for each cluster. The center of cluster i is calculated as: $\mathbf{V_i} = (\frac{1}{N_i}\sum_{j=1}^{N_i} x_{j,1}^i, \frac{1}{N_i}\sum_{j=1}^{N_i} x_{j,2}^i, \ldots, \frac{1}{N_i}\sum_{j=1}^{N_i} x_{j,g}^i$, where $\mathbf{V_i}$ is a vector with a magnitude of g, g is the number of selected genes, N_i represents the number of cells in cluster i, and $x_{j,k}^i$ represents the expression value of gene k in cell j belonging to cluster i. Then, scCAD calculates the Euclidean distance between all cluster centers: $D\left(\mathbf{V_i}, \mathbf{V_j}\right) = \sqrt{(\mathbf{V_i} - \mathbf{V_j})^2}$, where $\mathbf{V_i}$ and $\mathbf{V_j}$ represent the centers of cluster i and j, respectively. Finally, scCAD determines the threshold of merging, $THM = \text{median}(d_1, d_2, \ldots, d_m)$, where d_i is the Euclidean distance between cluster i and its nearest neighboring cluster, m is the number of clusters in D-clusters. If $D\left(\mathbf{V_i}, \mathbf{V_j}\right) < THM$, clusters i and j are merged. The resulting merged clusters form M-clusters (Merged clusters): $\{C_1, C_2, \ldots, C_{m'}\}$. I-clusters, D-clusters, and M-clusters consist of clusters obtained from initial clustering, clusters derived after decomposition, and clusters formed after merging, respectively.
3. For each gene in cluster i in M-clusters, scCAD calculates the difference between the median of the gene expressions of all cells within cluster i and the median of those outside of cluster i[93]. Assume that $X_{C_i}^k$ represents the vector composed of gene k

expression values for all cells in cluster i, $X^k_{C_i'}$ represents the vector composed of gene k expression values for all cells outside of cluster i, the median difference of gene k is calculated as: $d_k = |\text{median}\left(X^k_{C_i}\right) - \text{median}\left(X^k_{C_i'}\right)|$. Finally, scCAD selects the top 20 genes with the largest differences to generate the candidate gene set S_i for cluster i.

4. The gene expression matrix, which contains the genes in S_i, is then fed into an isolation forest model[28] to calculate an anomaly score for each cell. The isolation forest model builds a collection of isolation trees where each tree is constructed by randomly selecting a subset of cells and recursively partitioning them into smaller subsets based on their expression in randomly selected candidate genes. This process continues until each cell is isolated in its leaf node. The anomaly score is computed by normalizing the average path length for each cell, achieved by comparing it to the average path length of a randomly generated cell from the same dataset. The resulting score represents the degree of abnormality exhibited by each cell with the candidate genes. As described in[28], the ensemble anomaly score of cell j based on the candidate genes in S_i is calculated with:

$$AS_j^{S_i} = 2^{-\frac{E(h(j))}{c(n)}} \tag{1}$$

   where $h(j)$ is the path length of cell j in an isolation tree, which is the number of edges traversed in an isolation tree from the root node to the node containing cell i. $E(h(j))$ is the average of $h(j)$ across all the isolation trees in the isolation forest model. $c(n)$ represents the average path length when the total number of cells is n, and its formula is as follows[28]:

$$c(n) = \begin{cases} 2H(n-1) - \frac{2(n-1)}{n} & n > 2 \\ 1 & n = 2 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

   where $H(n-1)$ is the harmonic number that can be estimated by $\ln(n-1) + 0.5772156649$ (Euler's constant)[28].

5. scCAD assigns an independence score to cluster i in M-clusters based on the list composed of the corresponding anomaly scores of all cells: $\{AS_1^{S_i}, AS_2^{S_i}, \ldots, AS_n^{S_i}\}$. The independence score (IS) of cluster i is defined as follows:

$$\text{IS}_i = \frac{|T_{N_i} \cap C_i|}{N_i} \tag{3}$$

   where N_i is the number of cells in cluster i, T_{N_i} is the set of the top N_i cells with the highest anomaly scores, and C_i is cluster i in M-clusters. A higher independence score indicates that the differentially expressed genes of the corresponding cluster effectively distinguish and characterize its encompassing cells.

6. scCAD executes steps 3-5 for each cluster in M-clusters until obtaining the independence score for all clusters: $\{\text{IS}_1, \text{IS}_2, \ldots, \text{IS}_{m'}\}$. Finally, clusters with an independence score exceeding the threshold I (default is 0.7) ($\{C_i \in \text{M} - \text{clusters}|\text{score}_i > I\}$) are predicted as rare cell types and are outputted, along with the corresponding candidate genes.

Parameter value selection in scCAD

Clusters containing more than $R * n$ cells are considered for decomposition through iterative clustering. Based on a comprehensive review of previous studies[18,94,95] defining the size of rare cell types, we set this parameter to 1% on all larger datasets. For smaller datasets, especially when the total number of cells is below 3000, we set the threshold to $\frac{30}{n}$ to prevent the generation of excessively small clusters, thereby enhancing the interpretability and reliability of clustering outcomes.

After cluster decomposition, scCAD merges clusters if their distance is smaller than the threshold THM, which is denoted as $THM = \text{median}(d_1, d_2, \ldots, d_m)$, where d_i is the Euclidean distance between cluster i and its nearest neighboring cluster. We test the number of clusters generated after merging and the average proportions of all cell types and rare cell types in their dominant clusters by using different THM values in the Arc-ME dataset (Supplementary Fig. 22). As shown in Supplementary Fig. 22, lower THM values (such as zero and the lower quartile) may incur higher computational overhead due to a larger number of analyzed clusters, while higher THM values (such as the upper quartile and the 90th percentile) may significantly increase the likelihood of merging clusters dominated by rare cell types, potentially diminishing the effectiveness of the decomposition step. To enhance efficiency and reduce the number of analyzed clusters, we use the median as the default parameter across all datasets.

scCAD identifies a cluster as rare when its independence score exceeds a threshold value, I. We display the distribution of independence scores calculated by scCAD for each cluster on four datasets in Supplementary Fig. 23. As shown in Supplementary Fig. 23, clusters dominated by rare cell types exhibit significantly higher independence scores compared to other clusters, and using the default threshold can effectively distinguish them. It is important to note that reducing this threshold may result in the identification of multiple clusters dominated by larger cell types. We default to applying $I = 0.7$ across all datasets.

In Supplementary Note 5 and Supplementary Table 21, we provide an estimation of the runtime for each step in the scCAD workflow.

Usage of comparative methods

To evaluate the performance of scCAD for identifying rare cells, we conduct a benchmark analysis comparing scCAD with other methods. The CellSIUS package is obtained from GitHub (Novartis/CellSIUS). The initial major cell types are determined using a single-cell clustering workflow in the Seurat package[11]. The CellSIUS algorithm provides the results of sub-clusters assigned to each cell. The CIARA package is obtained from GitHub (ScialdoneLab/CIARA). The CIARA algorithm merges the clustering results obtained by the standard algorithm (Louvain) based on the HVG and identified genes, especially labeling rare cell types. The EDGE package is obtained from GitHub (shawnstat/EDGE). For each dataset, we utilize the data matrix preprocessed by the Seurat package[11] as the input for this method. Based on the 2-dimensional embedding results generated by EDGE, we construct a k-nearest neighbor graph and apply the Louvain algorithm to obtain the global clustering results. The FiRE package is obtained from GitHub (princethewinner/FiRE, R version). FiRE assigns a score to each cell and outputs predicted rare cells that meet the thresholding criteria based on the interquartile range (IQR). The GapClust package is obtained from GitHub (fabotao/GapClust). Similar to scCAD, GapClust generates multiple sets of predicted rare cells as output. The GiniClust2 and GiniClust3 packages are obtained from GitHub (dtsoucas/GiniClust2, rdong08/GiniClust3), respectively. Both of them return global consensus clustering results. The RaceID3 package is obtained from GitHub (dgrun/RaceID3_StemID2_package). RaceID3 returns a list containing predicted rare cells. SCA is implemented in Python, and the latest version is obtained from GitHub (bendemeo/shannonca). Following their previous recommendations, we construct a 15-nearest Euclidean neighbor graph in the 50-dimensional space of SCA and use Leiden clustering with the default resolution of 1.0 to obtain the final clustering results. The SCISSORS package is obtained from GitHub (jrleary7/SCISSORS). We reclustering the clusters with an average silhouette coefficient calculated by SCISSORS smaller than the overall average, resulting in the final clustering results.

For each algorithm, all parameters are set to their default values. For algorithms that directly output sets of predicted rare cells, such as scCAD, GapClust, FiRE, and RaceID3, we combine all the predicted cells from the result to obtain the final binary prediction outcome. For algorithms that return global clustering results, such as CellSIUS, CIARA, EDGE, GiniClust2, GiniClust3, SCA, and SCISSORS, clusters with a cell population smaller than the corresponding threshold (1% or 5%) are identified as rare clusters. We combine all the cells from the predicted rare clusters to obtain the final binary prediction outcome.

Generation and description of simulation data

To analyze the sensitivity of scCAD to rare cell type identity, we generate artificial scRNA-seq data using the splatter R package[96]. The following command is used to generate these data:

splatSimulate(group.prob = c(0.99, 0.01), method = 'groups', verbose = F, batchCells = 2500, de.prob = c(0.4, 0.4), out.prob = 0, de.facLoc = 0.4, de.facScale = 0.8, nGenes = 5000, seed = 2023)

The dataset consists of 2500 cells, each containing 5000 genes. Of these cells, 2476 represent the major cell type, while 24 define the minor type.

After data preprocessing, we use Wilcoxon's rank sum test to identify DE genes with an FDR cutoff of 0.05 and an inter-group absolute fold-change cutoff of 1.5. Assume that the set of these DE genes is S_{DEGs}. We remove the DE genes obtained from the randomly permuted labels from the set S_{DEGs}. This step is repeated ten times. The final 220 DE genes are removed from the data and preserved as a separate set. Additionally, 3226 genes with a p-value exceeding 0.05 are retained as a distinct set of non-differential genes.

The subsampled Jurkat dataset consists of 1556 cells, each containing 32,738 genes. Of these cells, 1540 represent the major 293T cell type, while 16 define the minor Jurkat cell type. To increase the number of DE genes for analysis, the inter-group absolute fold-change cutoff is adjusted to 1. The final 108 DE genes are removed from the data and preserved as a separate set. Additionally, 30,479 genes with a p-value exceeding 0.05 are retained as a distinct set of non-differential genes.

Statistics and reproducibility

Differentially expressed genes were identified using the *FindMarkers* function from the Seurat R package (version 4.5.0). This analysis employed a two-sided Wilcoxon rank sum test, with a false discovery rate (FDR) cutoff of 0.05 and an inter-group absolute fold-change cutoff of 1.5. Fold-change values were calculated based on the mean expression levels of each gene between groups. P-values were adjusted using Bonferroni correction, accounting for the total number of genes in the dataset. In box plot representations, the horizontal line denotes the median value, the lower and upper quartiles represent the 25th (Q1) and 75th percentiles (Q3), respectively. The interquartile range (IQR) is defined as the range between Q1 and Q3, and the whisker values calculated as $\mathrm{Q1}/\mathrm{Q3} - / + 1.5 \times \mathrm{IQR}$ as indicated in figure legends.

No statistical method was used to pre-determine sample size. No data were excluded from the analysis, all genes in datasets were used throughout all analyses. The selection of pre-identified differentially expressed genes was randomized, all other experiments were not randomized. The investigators were blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The details of the datasets used in this study are reported in Supplementary Table 1. All described datasets are obtained from various public websites under accession codes provided in Supplementary Table 1, including NCBI Gene Expression Omnibus (GEO) [https://www.ncbi.nlm.nih.gov/geo/], ArrayExpress [https://www.ebi.ac.uk/arrayexpress/], Sequence Read Archive (SRA) [https://www.ncbi.nlm.nih.gov/sra]. 10X PBMC is obtained at Github [https://github.com/ttgump/scDeepCluster/blob/master/scRNA-seq%20data/10X_PBMC.h5]. 68k PBMC and Jurkat datasets are obtained from the website of 10X genomics ([https://www.10xgenomics.com/datasets/fresh-68-k-pbm-cs-donor-a-1-standard-1-1-0], [https://www.10xgenomics.com/datasets/50-percent-50-percent-jurkat-293-t-cell-mixture-1-standard-1-1-0]). The worm neuron cells dataset Cao is sampled from a dataset obtained from the sci-RNA-seq platform (single-cell combinatorial indexing RNA sequencing) [http://atlas.gs.washington.edu/worm-rna/docs/]. The preprocessed human tonsil data, named Tonsil, and Crohn data are available from Broad Institute Single Cell Portal ([https://singlecell.broadinstitute.org/single_cell/study/SCP2169/slide-tags-snrna-seq-on-human-tonsil], [https://singlecell.broadinstitute.org/single_cell/study/SCP359/ica-ileum-lamina-propria-immunocytes-sinai]). The mouse retina data and B_ lymphoma data are available at Github [https://github.com/OSU-BMBL/marsgt/tree/main/Data]. Source data are provided with this paper.

Code availability

scCAD is publicly available at GitHub [https://github.com/xuyp-csu/scCAD] and Zenodo[97].

References

1. Potter, S. S. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.* **14**, 479–492 (2018).
2. Choi, Y. H. & Kim, J. K. Dissecting cellular heterogeneity using single-cell RNA sequencing. *Mol. Cells* **42**, 189–199 (2019).
3. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
4. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).
5. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
6. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol. JASN* **30**, 23–32 (2019).
7. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).
8. Ross, A. et al. Detection and viability of tumor cells in peripheral blood stem cell collections from breast cancer patients using immunocytochemical and clonogenic assay techniques [see comments]. *Blood* **82**, 2605–2610 (1993).
9. Paterlini-Brechot, P. & Benali, N. L. Circulating tumor cells (CTC) detection: clinical impact and future directions. *Cancer Lett.* **253**, 180–204 (2007).
10. Joosse, S. A., Gorges, T. M. & Pantel, K. Biology, detection, and clinical implications of circulating tumor cells. *EMBO Mol. Med.* **7**, 1–11 (2015).
11. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
12. Jindal, A., Gupta, P., Jayadeva & Sengupta, D. Discovery of rare cells from voluminous single cell expression data. *Nat. Commun.* **9**, 4719 (2018).
13. Wegmann, R. et al. CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. *Genome Biol.* **20**, 142 (2019).

14. Sun, X., Liu, Y. & An, L. Ensemble dimensionality reduction and feature gene extraction for single-cell RNA-seq data. *Nat. Commun.* **11**, 5853 (2020).
15. Fa, B. et al. GapClust is a light-weight approach distinguishing rare cells from voluminous single cell expression profiles. *Nat. Commun.* **12**, 4197 (2021).
16. Jiang, L., Chen, H., Pinello, L. & Yuan, G.-C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.* **17**, 144 (2016).
17. Tsoucas, D. & Yuan, G.-C. GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome Biol.* **19**, 58 (2018).
18. Dong, R. & Yuan, G.-C. GiniClust3: a fast and memory-efficient tool for rare cell type identification. *BMC Bioinform.* **21**, 158 (2020).
19. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
20. Herman, J. S., Sagar & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018).
21. Leary, J. R. et al. Sub-cluster identification through semi-supervised optimization of rare-cell silhouettes (SCISSORS) in single-cell RNA-sequencing. *Bioinformatics* **39**, btad449 (2023).
22. Lubatti, G. et al. CIARA: a cluster-independent algorithm for identifying markers of rare cell types from single-cell sequencing data. *Development* **150**, dev201264 (2023).
23. DeMeo, B. & Berger, B. SCA: recovering single-cell heterogeneity through information-based dimensionality reduction. *Genome Biol.* **24**, 195 (2023).
24. Wang, X. et al. MarsGT: multi-omics analysis for rare population inference using single-cell graph transformer. *Nat. Commun.* **15**, 338 (2024).
25. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
26. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
27. Xu, Y. et al. CellBRF: a feature selection method for single-cell clustering using cell balance and random forest. *Bioinformatics* **39**, i368–i376 (2023).
28. Liu, F. T., Ting, K. M. & Zhou, Z.-H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* 413–422 (IEEE, Pisa, Italy, 2008).
29. Gerniers, A., Bricard, O. & Dupont, P. MicroCellClust: mining rare and highly specific subpopulations from single-cell expression data. *Bioinformatics* **37**, 3220–3227 (2021).
30. Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
31. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
32. Peng, J. et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* **29**, 725–738 (2019).
33. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
34. Xie, K., Huang, Y., Zeng, F., Liu, Z. & Chen, T. scAIDE: clustering of large-scale single-cell RNA-seq data reveals putative and rare cell types. *NAR Genom. Bioinform.* **2**, lqaa082 (2020).
35. Davis, J. D. & Wypych, T. P. Cellular and functional heterogeneity of the airway epithelium. *Mucosal Immunol.* **14**, 978–990 (2021).
36. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
37. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
38. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
39. Hewitt, R. J. & Lloyd, C. M. Regulation of immune responses by the airway epithelial cell landscape. *Nat. Rev. Immunol.* **21**, 347–362 (2021).
40. Deprez, M. et al. A single-cell atlas of the human healthy airways. *Am. J. Respir. Crit. Care Med.* **202**, 1636–1645 (2020).
41. Song, H., Seddighzadeh, B., Cooperberg, M. R. & Huang, F. W. Expression of ACE2, the SARS-CoV-2 receptor, and TMPRSS2 in prostate epithelial cells. *Eur. Urol.* **78**, 296–298 (2020).
42. Campbell, J. N. et al. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* **20**, 484–496 (2017).
43. Chen, R., Wu, X., Jiang, L. & Zhang, Y. Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Rep.* **18**, 3227–3241 (2017).
44. Chen, Y. et al. The oligodendrocyte-specific G protein–coupled receptor GPR17 is a cell-intrinsic timer of myelination. *Nat. Neurosci.* **12**, 1398–1406 (2009).
45. Lendahl, U., Muhl, L. & Betsholtz, C. Identification, discrimination and heterogeneity of fibroblasts. *Nat. Commun.* **13**, 3409 (2022).
46. Joost, S. et al. The molecular anatomy of mouse skin during hair growth and rest. *Cell Stem Cell* **26**, 441–457.e7 (2020).
47. Ascensión, A. M., Fuertes-Álvarez, S., Ibañez-Solé, O., Izeta, A. & Araúzo-Bravo, M. J. Human dermal fibroblast subpopulations are conserved across single-cell RNA sequencing studies. *J. Invest. Dermatol.* **141**, 1735–1744.e35 (2021).
48. Morel, L. et al. Molecular and functional properties of regional astrocytes in the adult brain. *J. Neurosci.* **37**, 8706–8717 (2017).
49. Jurga, A. M., Paleczna, M., Kadluczka, J. & Kuter, K. Z. Beyond the GFAP-astrocyte protein markers in the brain. *Biomolecules* **11**, 1361 (2021).
50. He, L. et al. Analysis of the brain mural cell transcriptome. *Sci. Rep.* **6**, 35108 (2016).
51. Gerbe, F., Legraverend, C. & Jay, P. The intestinal epithelium tuft cells: specification and function. *Cell. Mol. Life Sci.* **69**, 2907–2917 (2012).
52. Ayyaz, A. et al. Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature* **569**, 121–125 (2019).
53. Middelhoff, M. et al. Dclk1-expressing tuft cells: critical modulators of the intestinal niche? *Am. J. Physiol. Gastrointest. Liver Physiol.* **313**, G285–G299 (2017).
54. Engelstoft, M. S. et al. Research resource: a chromogranin a reporter for serotonin and histamine secreting enteroendocrine cells. *Mol. Endocrinol.* **29**, 1658–1671 (2015).
55. 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* **2019**, baz046 (2019).
56. Hunyadi, J., Simon, M., Kenderessy, A., Sz & Dobozy, A. Expression of monocyte/macrophage markers (CD13, CD14, CD68) on human keratinocytes in healthy and diseased skin. *J. Dermatol.* **20**, 341–345 (1993).
57. Xu, Q. et al. NADPH oxidases are essential for macrophage differentiation. *J. Biol. Chem.* **291**, 20030–20041 (2016).
58. Chung, E. J. et al. Natural variation in macrophage polarization and function impact pneumocyte senescence and susceptibility to fibrosis. *Aging* **14**, 7692–7717 (2022).
59. Dominguez Gutierrez, G. et al. Gene signature of the human pancreatic ε cell. *Endocrinology* **159**, 4023–4032 (2018).
60. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360.e4 (2016).

61. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394.e3 (2016).
62. Xue, M. et al. Schwann cells regulate tumor cells and cancer-associated fibroblasts in the pancreatic ductal adenocarcinoma microenvironment. *Nat. Commun.* **14**, 4600 (2023).
63. Eissmann, M. F. et al. IL-33-mediated mast cell activation promotes gastric cancer through macrophage mobilization. *Nat. Commun.* **10**, 2735 (2019).
64. Sharma, R. B. et al. Insulin demand regulates β cell number via the unfolded protein response. *J. Clin. Invest.* **125**, 3831–3846 (2015).
65. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
66. Martin, J. C. et al. Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. *Cell* **178**, 1493–1508.e20 (2019).
67. D'Acquisto, F. & Crompton, T. CD3 + CD4 − CD8− (double negative) T cells: saviours or villains of the immune response? *Biochem. Pharmacol.* **82**, 333–340 (2011).
68. Zhang, Y. et al. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. *Proc. Natl Acad. Sci. USA* **118**, e2103240118 (2021).
69. Stewart, B. J. et al. Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–1466 (2019).
70. Zhang, J.-Y. et al. Single-cell landscape of immunological responses in patients with COVID-19. *Nat. Immunol.* **21**, 1107–1118 (2020).
71. Maier, B. et al. A conserved dendritic-cell regulatory program limits antitumour immunity. *Nature* **580**, 257–262 (2020).
72. An, X. et al. Global transcriptome analyses of human and murine terminal erythroid differentiation. *Blood* **123**, 3466–3477 (2014).
73. Lee, J., Hyeon, D. Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.* **52**, 1428–1442 (2020).
74. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. *Trends Biotechnol.* **38**, 1007–1022 (2020).
75. Dou, J. et al. Bi-order multimodal integration of single-cell data. *Genome Biol.* **23**, 112 (2022).
76. Langer, K. B. et al. Retinal Ganglion Cell Diversity and Subtype Specification from Human Pluripotent Stem Cells. *Stem Cell Rep.* **10**, 1282–1293 (2018).
77. Rheaume, B. A. et al. Single cell transcriptome profiling of retinal ganglion cells identifies cellular subtypes. *Nat. Commun.* **9**, 2759 (2018).
78. Møller, H. J. et al. Soluble CD163: a marker molecule for monocyte/macrophage activity in disease. *Scand. J. Clin. Lab. Invest.* **62**, 29–33 (2002).
79. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
80. Calon, A. et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **47**, 320–329 (2015).
81. MacParland, S. A. et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018).
82. Koay, H.-F. et al. A divergent transcriptional landscape underpins the development and functional branching of MAIT cells. *Sci. Immunol.* **4**, eaay6039 (2019).
83. Kleiveland, C. R. Peripheral blood mononuclear cells. In *The Impact of Food Bioactives on Health: in vitro and ex vivo models* (Springer, Cham, 2015).
84. da Silva, F. A. R. et al. Whole transcriptional analysis identifies markers of B, T and plasma cell signaling pathways in the mesenteric adipose tissue associated with Crohn's disease. *J. Transl. Med.* **18**, 44 (2020).
85. Wang, Z. et al. Celda: a Bayesian model to perform co-clustering of genes into modules and cells into subpopulations using single-cell RNA-seq data. *NAR Genom. Bioinform.* **4**, lqac066 (2022).
86. Stassen, S. V. et al. PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells. *Bioinformatics* **36**, 2778–2786 (2020).
87. Yang, P., Huang, H. & Liu, C. Feature selection revisited in the single-cell era. *Genome Biol.* **22**, 1–17 (2021).
88. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).
89. Ranjan, B. et al. DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nat. Commun.* **12**, 5849 (2021).
90. Wang, J. et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat. Commun.* **12**, 1882 (2021).
91. Yu, Z. et al. ZINB-based graph embedding autoencoder for single-cell RNA-seq interpretations. *Proc. AAAI Conf. Artif. Intell.* **36**, 4671–4679 (2022).
92. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
93. Scherf, U. et al. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* **24**, 236–244 (2000).
94. Märtens, K. et al. Rarity: discovering rare cell populations from single-cell imaging data. https://doi.org/10.1101/2022.07.15.500256 (2022).
95. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
96. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
97. Xu, Y. et al. scCAD: Cluster decomposition-based anomaly detection for rare cell identification in single-cell expression data. *scCAD* https://doi.org/10.5281/zenodo.13121480 (2024).

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (No.2021YFF1201200), the National Natural Science Foundation of China under Grants (Nos. 62350004, 62332020), the Project of Xiangjiang Laboratory (No. 23XJ01011) to J.X.W., the National Natural Science Foundation of China under Grants (No. U22A2041) to H.D.L., and the Science Foundation for Distinguished Young Scholars of Hunan Province (No. 2023JJ10080) to J.Z.X. This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

Author contributions

J.X.W. and H.D.L. conceived and designed this project. Y.P.X., S.K.W., J.X.W., and H.D.L. conceived, designed, and implemented the scCAD. Y.P.X. and S.K.W. collected datasets and conducted experiments. Y.P.X., S.K.W., J.Z.X., and H.D.L. performed the analysis. Y.P.X., Y.H.L., Q.L.F., and J.X.W. wrote the paper. All authors have read and approved the final version of this paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-51891-9.

Correspondence and requests for materials should be addressed to Hong-Dong Li or Jianxin Wang.

Peer review information *Nature Communications* thanks Anil Giri and Shibiao Wan for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.