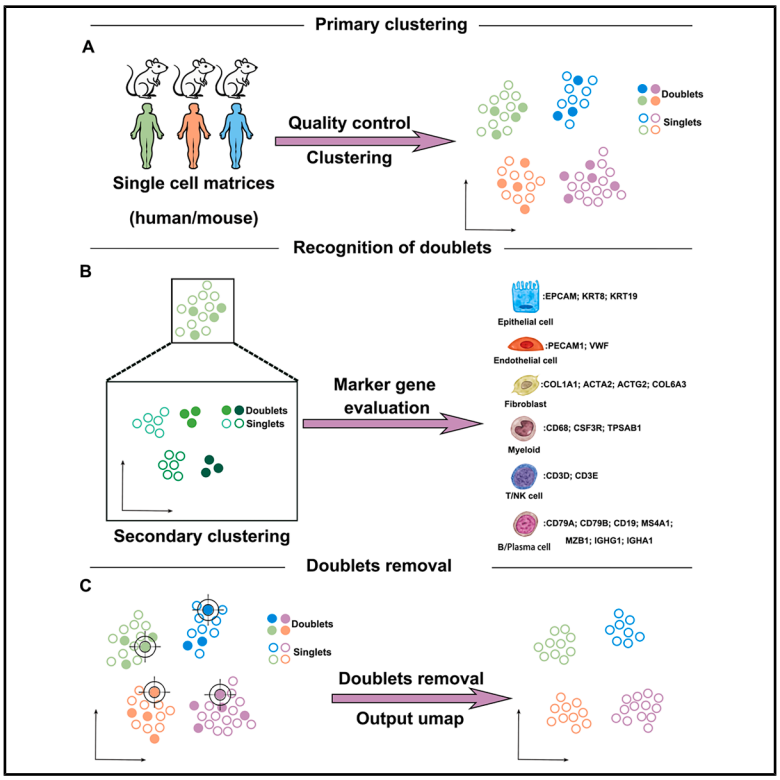
iScience

Article

scUmaper: An automated framework for doublet removal and cell-type annotation in single-cell transcriptomics

Graphical abstract



Highlights

- End-to-end R/Seurat workflow for doublet filtering and annotation of scRNA-seq
- Rule-based detection flags lineage-incompatible marker co-expression
- Global-then-within-lineage clustering exposes anomalous doublet subclusters
- Built-in human/mouse marker libraries and support for custom markers

Authors

Xushun Guo, Mudan Zhang, Zhuo Xie, Yifan Wang, Shenghong Zhang, Gaoshi Zhou

Correspondence

zhshh3@mail.sysu.edu.cn (S.Z.), zhougshi@mail3.sysu.edu.cn (G.Z.)

In brief

Biological sciences; Computer science.

Guo et al., 2026, iScience *29*, 115850
June 19, 2026
https://doi.org/10.1016/j.isci.2026.115850

CellPress

iScience

CellPress
OPEN ACCESS

Article

scUmaper: An automated framework for doublet removal and cell-type annotation in single-cell transcriptomics

Xushun Guo,[1,4] Mudan Zhang,[2,4] Zhuo Xie,[1] Yifan Wang,[3] Shenghong Zhang,[1,2,*] and Gaoshi Zhou[1,5,*]
[1]Department of Gastroenterology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China
[2]Department of Gastroenterology, Guangxi Hospital Division of The First Affiliated Hospital, Sun Yat-sen University, Nanning, China
[3]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China
[4]These authors contributed equally
[5]Lead contact
*Correspondence: zhshh3@mail.sysu.edu.cn (S.Z.), zhougshi@mail3.sysu.edu.cn (G.Z.)
https://doi.org/10.1016/j.isci.2026.115850

SUMMARY

Single-cell RNA sequencing can resolve cellular heterogeneity but is sensitive to heterotypic doublets and often requires expert-driven preprocessing and annotation. We developed single cell utility matrices processing engine (scUmaper), an R/Seurat-native workflow that integrates quality control, biologically grounded doublet filtering, and marker-library-based cell-type annotation. scUmaper codifies lineage-marker incompatibility rules and applies global clustering followed by within-lineage re-clustering to reveal anomalous subclusters with implausible cross-lineage co-expression. Across six public human organ datasets, scUmaper removed additional high-confidence heterotypic doublets that were retained by simulation-based approaches and achieved annotation agreement comparable to or higher than commonly used R-based baselines, with competitive runtime. Stress tests with simulated ambient RNA contamination and reduced sequencing depth showed stable outputs under moderate degradation. scUmaper provides an interpretable and extensible framework that lowers barriers for reproducible single-cell RNA sequencing (scRNA-seq) analysis.

INTRODUCTION

In recent years, the rapid advancement of single-cell RNA sequencing (scRNA-seq) technology has enabled high-resolution dissection of cellular functional heterogeneity within complex biological systems, offering novel insights into disease pathogenesis and the identification of potential therapeutic targets.[1,2] Compared with conventional bulk RNA sequencing, scRNA-seq allows precise characterization of gene expression profiles at the single-cell level, thereby facilitating significant breakthroughs in diverse research areas such as tumor microenvironment analysis,[3] neural development,[4] and immune cell dynamics monitoring.[5] Particularly in the context of human disease research, this technology has been effectively applied to address key scientific questions, including the identification of disease-associated cell populations, the reconstruction of cellular transformation trajectories, and the elucidation of heterogeneity in drug responses.[6]

Nevertheless, the broad adoption of scRNA-seq technology is hindered by two primary challenges. First, it is affected by technical artifacts such as doublets, which occur when multiple cells are encapsulated within a single droplet during droplet-based microfluidic separation and are assigned the same barcode. Studies indicate that the prevalence of doublets increases exponentially with higher cell throughput,[7] and their aberrant gene expression profiles can compromise the accuracy of downstream analyses, including cell clustering, differential gene expression analysis, and pseudotime trajectory reconstruction.[8] Second, the analytical process imposes dual requirements on researchers, necessitating both proficiency in bioinformatics tools for data preprocessing, quality control, dimensionality reduction, and visualization, as well as a solid biological foundation to perform cell type annotation and interpret biological implications.[9] This interdisciplinary demand presents a significant barrier for non-bioinformatics researchers, such as clinicians or molecular biologists, who may lack the necessary computational skills to independently analyze single-cell data.

To overcome these barriers, a wide range of methods have been proposed for automated cell-type annotation and doublet detection. Automated cell-type annotation methods can be broadly grouped into three paradigms. Reference-based annotation assigns labels by comparing query cells to a curated reference (e.g., SingleR, scmap, scMatch, and Azimuth)[10–12] and can be effective when a well-matched reference exists, but performance may degrade under strong domain shift, incomplete reference coverage, or batch effects. Marker-based methods (e.g., SCINA, CellAssign, scCATCH, scType, HiCat, and scSorter[13–17]) are highly interpretable but depend on marker specificity and context-dependent expression. Classifier-based approaches (e.g., CellTypist, scClassify, scSorterDL, and

statistical classifiers such as PCLDA[18–21]) can provide strong automation, but typically require labeled training data or domain-matched pretrained models and may be less transparent than marker-rule-based annotation. In addition, some practical workflows introduce cross-language or data-format burdens for users who primarily operate within R/Seurat.

Current computational approaches for doublet detection, such as the widely adopted R packages DoubletFinder,[22] scDblFinder,[23] and Scrublet,[8] predominantly utilize a "simulation-matching" strategy. This involves generating synthetic doublet datasets and applying machine learning models to identify similar patterns in real data. However, these methods exhibit notable limitations. Simulated doublet profiles often fail to capture the complexity of real experimental conditions, particularly when nonlinear interactions exist between cell types, potentially leading to false negatives or false positives in detection.[24] In contrast, manual doublet removal typically relies on the co-expression of lineage-specific marker genes. For instance, a cell co-expressing CD3E (a T cell marker) and MS4A1 (a B cell marker) is classified as a doublet.[25] Although this biologically informed criterion is scientifically valid, its implementation efficiency is heavily dependent on the analyst's dual expertise in both computational skills for single-cell analysis and in-depth knowledge of cell type-specific gene expression patterns. For researchers without a bioinformatics background, these dual requirements often result in workflow disruptions or misinterpretation of results.

Therefore, there is a pressing need to develop a comprehensive and automated analytical tool that seamlessly integrates computational algorithms with biological principles. Here we present single cell utility matrices processing engine (scUmaper in R), an R package that integrates scRNA-seq preprocessing, biologically grounded doublet removal, and cell-type annotation into a single automated Seurat-native workflow. scUmaper is designed to systematize expert marker-based heuristics (specifically, lineage-marker incompatibility rules) into a reproducible pipeline that identifies biologically implausible cross-lineage co-expression patterns. To operationalize this logic, scUmaper adopts a two-step clustering strategy commonly used in Seurat-based analyses: It first resolves major lineage-level groups across all cells and then performs within-lineage subclustering to detect anomalous subclusters whose marker expression violates lineage exclusivity. The package ships with curated, mutually exclusive marker libraries for major cell lineages in human and mouse tissues, and it supports user-defined marker sets for specialized tissues, organoids, or non-model species. Visualization of processed cell clusters is provided using uniform manifold approximation and projection (UMAP),[26] without extra methodological novelty.

Using six public human scRNA-seq datasets spanning multiple organs, and an additional public peripheral blood mononuclear cell (PBMC) benchmark with established labels, we evaluate scUmaper against commonly used R-based baselines for doublet calling and annotation. We show that codifying lineage-marker incompatibility into an automated workflow can reliably remove lineage-incompatible doublets that may be missed by simulation-based approaches, while providing practical, interpretable outputs and reducing the technical barrier to producing analysis-ready singlet objects for downstream studies.

RESULTS

Outputs of scUmaper and two R-based baseline pipelines (DoubletFinder + SingleR; scDblFinder + scType)

We analyzed six publicly available scRNA-seq datasets derived from five distinct human organs (lung, ileum, liver, kidney, and colorectum) downloaded from the GEO database (GEO: GSE132771, GSE134809, GSE282701, GSE202052, GSE209781, GSE214695).[27–30] Each dataset was processed using scUmaper, using DoubletFinder followed by SingleR, using scDblFinder combined with scType, and manually curated as the gold standard. This yielded a total of 24 processed outcomes across the six datasets (Figures 1 and 2). We further assessed the expression patterns of marker genes across various cell types within these datasets (FiguresS1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S21, S22, S23, and S24).

Benchmarking metrics of scUmaper and R-based pipelines

Using manual doublet removal and expert annotation as the gold standard, we benchmarked scUmaper against other R-based automated pipelines, including DoubletFinder, SingleR, scDblFinder, and scType. Because the manual gold-standard doublets were conservatively defined as biologically implausible cross-lineage (heterotypic) events, the evaluated methods tend to achieve very high specificity/positive predictive value (PPV) under this definition (Table 1), and the main differences of doublet-removal are reflected in sensitivity/negative predictive value (NPV). scUmaper demonstrated sensitivity and accuracy comparable to DoubletFinder and scDblFinder (Figures 3A, 3B, 3G, and 3H; Table 1). Notably, scUmaper achieved a higher NPV than DoubletFinder (Figure 3C) and a comparable NPV to scDblFinder (Figure 3I), suggesting that its doublet-detection performance is comparable to or superior to the evaluated R-based baselines under our study definition.

Regarding automated annotation consistency, scUmaper yielded higher normalized mutual information (NMI) and comparable adjusted Rand index (ARI) relative to SingleR (Figures 3D and 3E; Table 1). It also outperformed scType in both NMI and ARI (Figures 3J and 3K; Table 1). These metrics indicate that scUmaper provides annotation consistency that is comparable to SingleR and consistently higher than scType in our benchmarks, when evaluated against the same manual labels.

In terms of computational efficiency, scUmaper was significantly faster than the widely used DoubletFinder + SingleR pipeline (Figure 3F; Table 1). Its processing speed was also comparable to the lightweight scDblFinder + scType workflow (Figure 3L; Table 1), highlighting its advantage in large-scale data processing.

Mechanistic case studies: DoubletFinder-missed but scUmaper-caught doublets

Given the observed trends of higher sensitivity and accuracy, along with significantly improved NPV, we further evaluated the presence of residual doublets in the singlet populations identified by scUmaper and DoubletFinder through marker gene

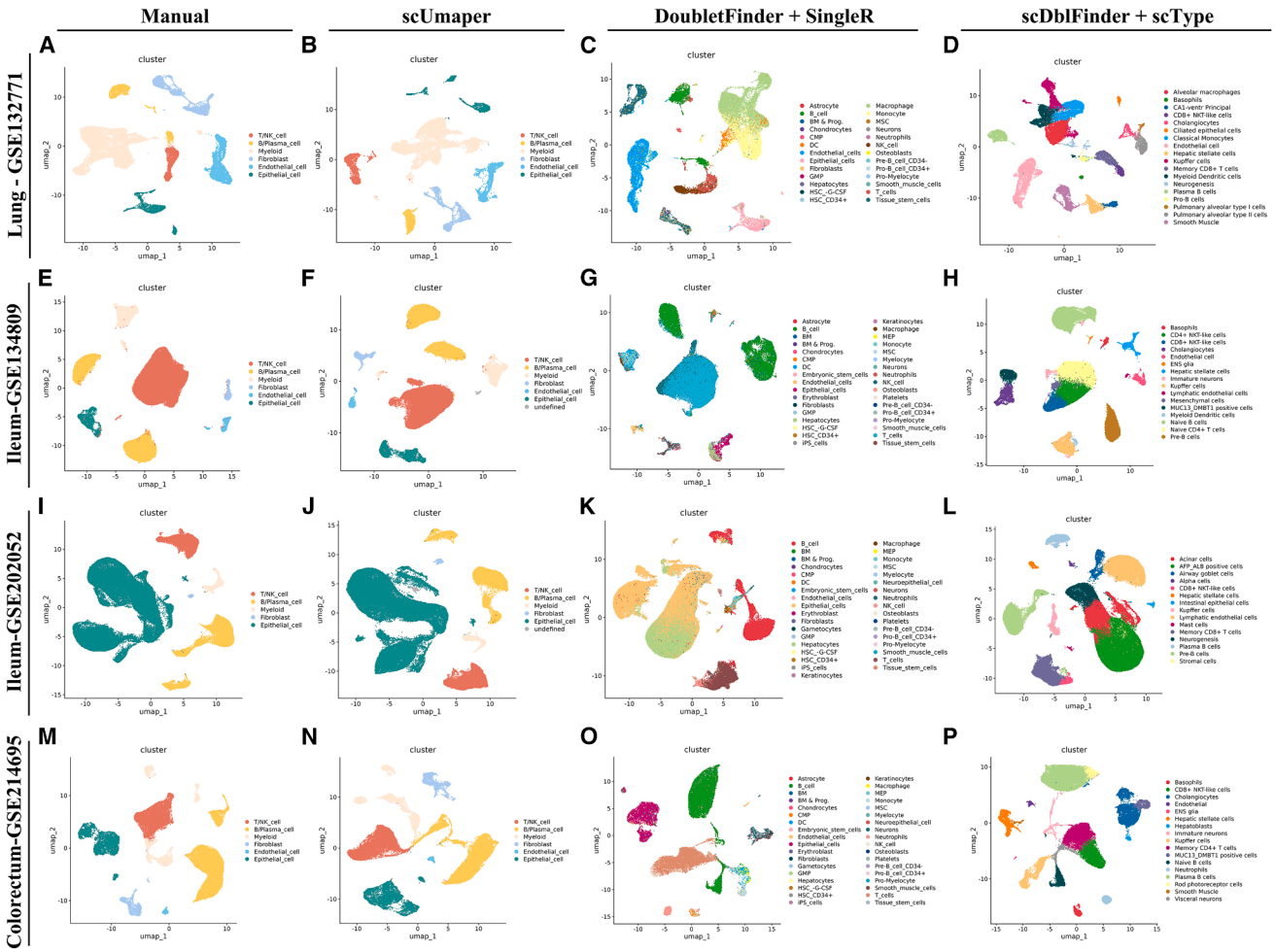


Figure 1. Comparison of outputs among manual process, scUmaper, DoubletFinder + SingleR pipeline, and scDblFinder + scType pipeline using hollow organs data

(A–D) UMAP plots of scRNA-seq data from human lung tissue (GEO: GSE132771) processed by manual, scUmaper, DoubletFinder + SingleR pipeline, and scDblFinder + scType pipeline, respectively.

(E–H) UMAP plots of scRNA-seq data from human ileum tissue (GEO: GSE134809) processed by manual, scUmaper, DoubletFinder + SingleR pipeline, and scDblFinder + scType pipeline, respectively. (I-L) UMAP plots of scRNA-seq data from human ileum tissue (GEO: GSE202052) processed by manual, scUmaper, DoubletFinder + SingleR pipeline, and scDblFinder + scType pipeline, respectively. (M-P) UMAP plots of scRNA-seq data from human colorectum tissue (GEO: GSE214695) processed by manual, scUmaper, DoubletFinder + SingleR pipeline, and scDblFinder + scType pipeline, respectively.

expression analysis. In the lung dataset (GSE132771), scUmaper effectively removed doublets, whereas DoubletFinder failed to eliminate cells co-expressing CD3E (T cell marker) and CD79A (B/plasma cell marker), which is biologically incompatible (Figures 4A and 4B). Visualizing the gene expression of these residual populations revealed co-expression of CD3D/CD3E and CD79A/MZB1 (Figure 4C), which directly triggers the T/NK versus B/plasma lineage-incompatibility rule implemented in scUmaper. Similarly, cells identified as doublets in the ileum dataset co-expressed TPSAB1/TPSB2 and CD79A/MZB1, which is characteristic of mast-cell/plasma-cell doublets (Figures 4D–4F). In the liver dataset, we observed co-expression of FCN1/S100A8 and MS4A1/BANK1, fitting the profile of monocyte/B-cell doublets (Figures 4G–4I). These biologically implausible patterns led to their correct classification as doublets by scUmaper. Notably, in these datasets, the populations missed by DoubletFinder were not always positioned as clear ‘‘intermediate’’ clusters between the two contributing lineages on the global embedding. Instead, they frequently remained adjacent to, or partially embedded within, a dominant lineage neighborhood, which can reduce the likelihood that simulation-based nearest-neighbor matching approaches flag them as doublets. Consistent results were observed in the remaining three datasets (ileum, kidney, and colorectum), where DoubletFinder failed to remove biologically implausible doublets that were effectively filtered by scUmaper (Figures S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, and S18). Collectively, these findings support that scUmaper can identify additional high-confidence

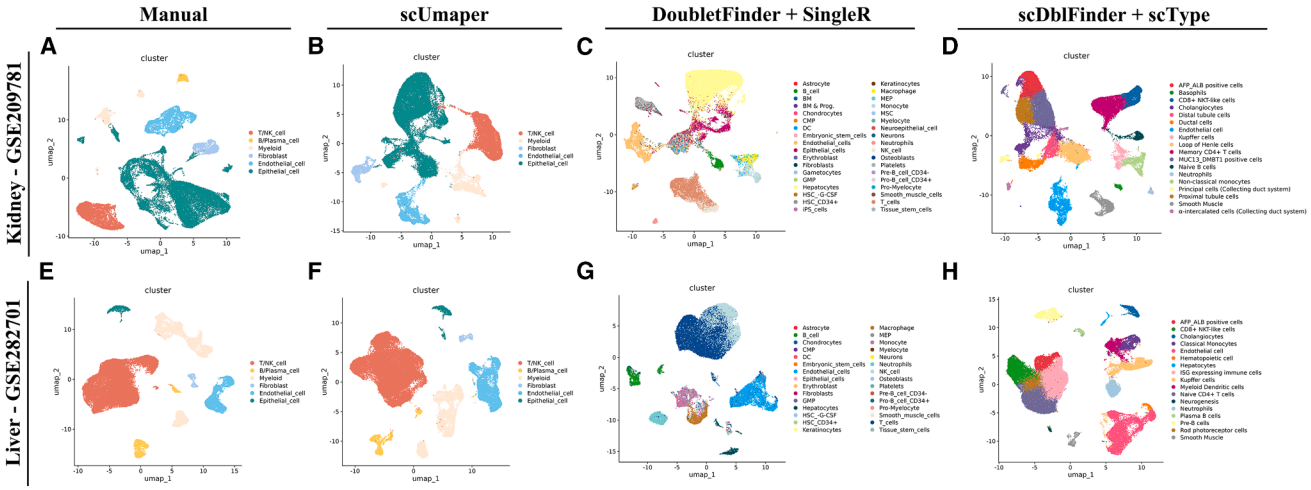


Figure 2. Comparison of outputs among manual process, scUmaper, DoubletFinder + SingleR pipeline, and scDblFinder + scType pipeline using solid organs data

(A–D) UMAP plots of scRNA-seq data from human kidney tissue (GEO: GSE209781) processed by manual, scUmaper, DoubletFinder + SingleR pipeline, and scDblFinder + scType pipeline, respectively.

(E–H) UMAP plots of scRNA-seq data from human liver tissue (GEO: GSE282701) processed by manual, scUmaper, DoubletFinder + SingleR pipeline, and scDblFinder + scType pipeline, respectively.

heterotypic doublets that may be missed by DoubletFinder in our evaluated datasets, consistent with the intended marker-incompatibility design.

Objective annotation validation on PBMC 3k

To address potential subjectivity in manual labeling, we incorporated the well-established PBMC 3k dataset from 10× Genomics for objective validation. Using the long-validated reference labels (Figure 5A), we compared the automated annotation performance of SingleR and scUmaper (Figures 5B and 5C). scUmaper achieved higher ARI and NMI compared to SingleR (Figure 5D), further substantiating the reliability of our annotation module.

Robustness of scUmaper under ambient RNA and low sequencing depth stress

To assess the impact of potential confounding factors, we performed stress tests involving ambient RNA contamination and low sequencing depth. For ambient RNA, we simulated contamination by adding a "soup" profile derived from the global gene-mean distribution. scUmaper maintained high consistency (observed agreement >0.8, Cohen's kappa >0.4) under mild-to-moderate contamination (rho $\leq$0.1), with even higher stability at lower contamination levels (rho $\leq$0.05, kappa >0.6) (Figures 6A and 6B). Similarly, in downsampling simulations, scUmaper remained stable even when sequencing depth was halved (observed agreement >0.8, kappa >0.4; Figures 6C and 6D). These results demonstrate the robustness of the framework under non-ideal experimental conditions.

DISCUSSION

The increasing adoption of scRNA-seq across diverse biological fields has highlighted a persistent gap between technological capability and analytical accessibility. While advanced computational methods for doublet removal and cell-type annotation have proliferated, they frequently necessitate specialized bioinformatic expertise and exhaustive parameter tuning, which can impede their utility for experimental researchers. Our development of scUmaper addresses this critical need by formalizing expert marker-based heuristics into a fully automated and interpretable workflow, thereby maintaining rigorous analytical standards while significantly lowering the barrier to entry for the broader scientific community.

A central design choice of scUmaper is its strategic prioritization of heterotypic doublets through a biological rule-based framework (Figure 7). Conventional simulation-based methods, such as DoubletFinder and scDblFinder, rely on the transcriptomic "neighborhood" of synthetic aggregates. Consequently, their performance is inherently sensitive to the accurate estimation of doublet rates and the optimization of nearest-neighbor parameters. These methods often exhibit reduced sensitivity when doublets do not project into clear intermediate spaces within the global manifold or when the simulated doublets fail to capture the full spectrum of technical noise. In contrast, scUmaper bypasses the stochasticity of simulation by formalizing deterministic lineage-incompatibility rules. By identifying biologically implausible co-expression patterns, such as the simultaneous detection of definitive T cell and B cell markers in a single event, scUmaper effectively captures "orphan" doublet populations that are frequently overlooked by purely algorithmic tools due to suboptimal parameter tuning. This detection capability is further augmented by our hierarchical two-step clustering strategy. While global embeddings often mask small doublet populations by clustering them within a dominant parent lineage, scUmaper's local re-clustering approach enhances the resolution of these anomalies. This heuristic high-resolution refinement allows lineage-incompatible cells to emerge as distinct subclusters, ensuring

Table 1. Benchmarking statistical metrics of scUmaper and R-based Pipelines on 6 public datasets

		Tissue	Lung	Ileum	Ileum	Kidney	Colorectum	Liver
		Dataset	GSE132771	GSE134809	GSE202052	GSE209781	GSE214695	GSE282701
Gold standard	Manual annotation	Cell count	32,802	108,065	105,196	54,097	44,964	64,051
		Doublet count	3,845	10,913	19,012	19,835	8,779	18,675
		Doublet ratio	11.7%	10.1%	18.1%	36.7%	19.5%	29.2%
Doublet removal	scUmaper	Processing time (min)	16	25	27	15	16	19
		Processing speed (cells/min)	2,050	4,323	3,896	3,606	2,810	3,371
		Sensitivity	0.609	0.601	0.338	0.038	0.177	0.135
		Specificity	1.000	1.000	1.000	1.000	1.000	1.000
		PPV	1.000	1.000	1.000	1.000	1.000	1.000
		NPV	0.930	0.950	0.867	0.636	0.812	0.731
		Accuracy	0.937	0.954	0.875	0.641	0.820	0.742
	DoubletFinder	Processing time[a] (min)	16	53	51	28	26	29
		Processing speed (cells/min)	2,050	2,039	2,063	1,932	1,729	2,209
		Sensitivity	0.313	0.290	0.211	0.104	0.188	0.131
		Specificity	1.000	1.000	1.000	1.000	1.000	1.000
		PPV	1.000	1.000	1.000	1.000	1.000	1.000
		NPV	0.911	0.920	0.842	0.633	0.824	0.718
		Accuracy	0.914	0.923	0.848	0.648	0.831	0.730
	scDblFinder	Processing time[b] (min)	9	26	30	13	15	23
		Processing speed (cells/min)	3,645	4,156	3,507	4,161	2,998	2,785
		Sensitivity	0.346	0.327	0.207	0.129	0.126	0.177
		Specificity	1.000	1.000	1.000	1.000	1.000	1.000
		PPV	1.000	1.000	1.000	1.000	1.000	1.000
		NPV	0.917	0.925	0.846	0.638	0.820	0.728
		Accuracy	0.921	0.928	0.852	0.657	0.825	0.743
Annotation	scUmaper	ARI	0.459	0.675	0.704	0.387	0.573	0.521
		NMI	0.546	0.665	0.641	0.499	0.615	0.603
	SingleR	ARI	0.495	0.703	0.427	0.236	0.621	0.359
		NMI	0.556	0.527	0.439	0.290	0.517	0.399
	scType	ARI	0.385	0.395	0.262	0.190	0.580	0.270
		NMI	0.535	0.470	0.387	0.330	0.553	0.386

PPV: positive predictive value; NPV: negative predictive value; ARI: adjusted rand index; NMI: normalized mutual information.
[a]Including the time of running DoubletFinder and SingleR.
[b]Including the time of running scDblFinder and scType.

they are robustly flagged by our marker-based filters rather than being assimilated into singlet cell populations.

The reliability of scUmaper is underscored by its high consistency with expert manual curation across diverse tissues, demonstrating doublet detection and annotation performance that is at least comparable to current mainstream R-based pipelines (Table 1; Figure 3). However, recognizing that manual labeling can be inherently subjective, particularly in the context of cell-type assignment, we incorporated the PBMC 3k standard dataset as an objective gold standard (Figure 5A). The superior ARI and NMI metrics observed in this benchmark confirm that scUmaper's annotation accuracy is a reflection of its robust underlying logic. Although some of the benchmarking results are statistically significant, the magnitude of improvement is modest. Rather than focusing on the scale of these gains, we highlight scUmaper's consistent performance across heterogeneous datasets as a marker of its reliability and its value as a stable analytical baseline. Conceptually, this robustness stems from scUmaper's position within the library-based annotation paradigm, which effectively bridges the gap between traditional manual marker-based inspection and often ''black-box'' model-based classifiers. By formalizing library-driven logic into a structured framework, scUmaper ensures that its outputs are both automated and biologically grounded. Nevertheless, it should be emphasized that, since scUmaper formalizes marker-based reasoning similar to manual annotation, ''agreement

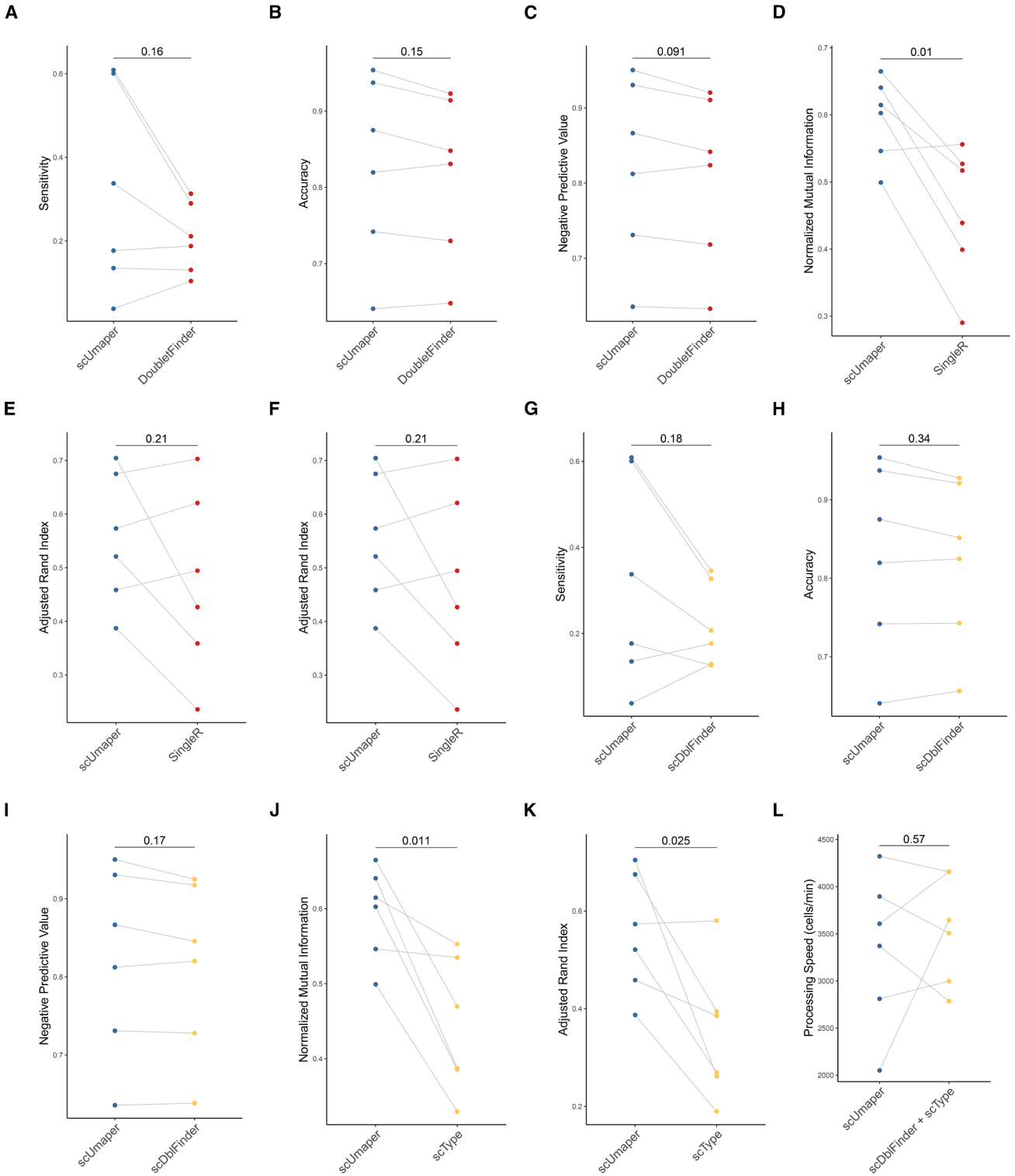


Figure 3. Benchmarking statistical metrics difference among scUmaper, DoubletFinder + SingleR pipeline, and scDblFinder + scType pipeline on 6 public datasets

(A–C) sensitivity, accuracy, and NPV differences between scUmaper and DoubletFinder, respectively.

(D and E) NMI and ARI differences between scUmaper and SingleR, respectively.

(F) processing speed differences between scUmaper and DoubletFinder + SingleR pipeline.

(legend continued on next page)

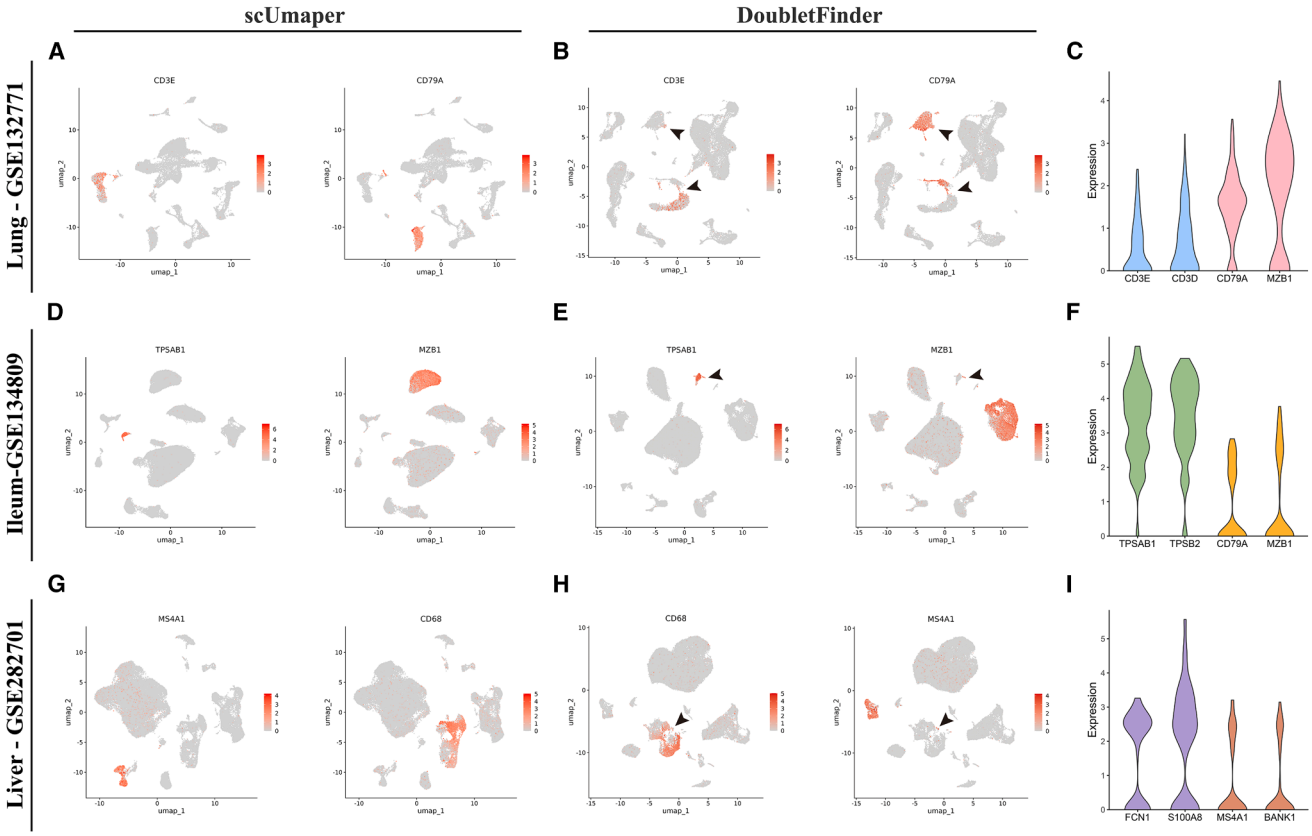


Figure 4. Comparison of doublet removal effectiveness between scUmaper and DoubletFinder by examining marker gene expression

(A and B) UMAP plots of scRNA-seq data from human lung tissue (GEO: GSE132771) processed with scUmaper (A) and DoubletFinder (B).

(C) Gene expression of representative doublets retained after DoubletFinder processing of human lung tissue.

(D and E) UMAP plots of scRNA-seq data from human ileum tissue (GEO: GSE134809) processed with scUmaper (D) and DoubletFinder (E).

(F) Gene expression of representative doublets retained after DoubletFinder processing of human ileum tissue.

(G and H) UMAP plots of scRNA-seq data from human liver tissue (GEO: GSE282701) processed with scUmaper (G) and DoubletFinder (H).

(I) Gene expression of representative doublets retained after DoubletFinder processing of human liver tissue. Arrows in UMAP plots indicate representative doublets retained after DoubletFinder processing.

with experts'' primarily measures consistency with expert biological heuristics rather than absolute truth. Importantly, this is aligned with scUmaper's goal: to systematize expert biological reasoning into a reproducible pipeline, rather than to claim an algorithmic ground truth in the absence of orthogonal validation.

Furthermore, this precision is maintained even under suboptimal experimental conditions. Our stress tests demonstrate that the framework is resilient against moderate ambient RNA contamination and low sequencing depth (Figure 6), which are challenges frequently encountered in clinical samples or large-scale cell atlases. This robustness is largely attributable to our conservative marker selection, which prevents technical noise from dominating the biological signal. For datasets with extreme ambient RNA contamination (a rare occurrence under standard sampling protocols), we recommend applying correction tools such as SoupX[31] prior to the scUmaper workflow.

From a practical perspective, scUmaper offers substantial computational advantages. By integrating preprocessing, doublet removal, and annotation into a unified single-object workflow, it eliminates the requirement for complex inter-package data conversions and mitigates the risk of error propagation. This streamlined architecture leads to a significant reduction in processing time compared to multi-step R pipelines, empowering non-bioinformaticians to process single-cell transcriptomic data rapidly, efficiently, and accurately.

Transparency and flexibility were also core design principles. While our default marker libraries (Tables 2 and 3) were curated from authoritative sources (e.g., CellMarker, PanglaoDB, and

(G–I) sensitivity, accuracy, and NPV differences between scUmaper and scDblFinder, respectively.

(J–K) NMI and ARI differences between scUmaper and scType, respectively.

(L) processing speed differences between scUmaper and scDblFinder + scType pipeline. All tests were performed using paired two-sided *t* test, the numbers indicate *p* values.

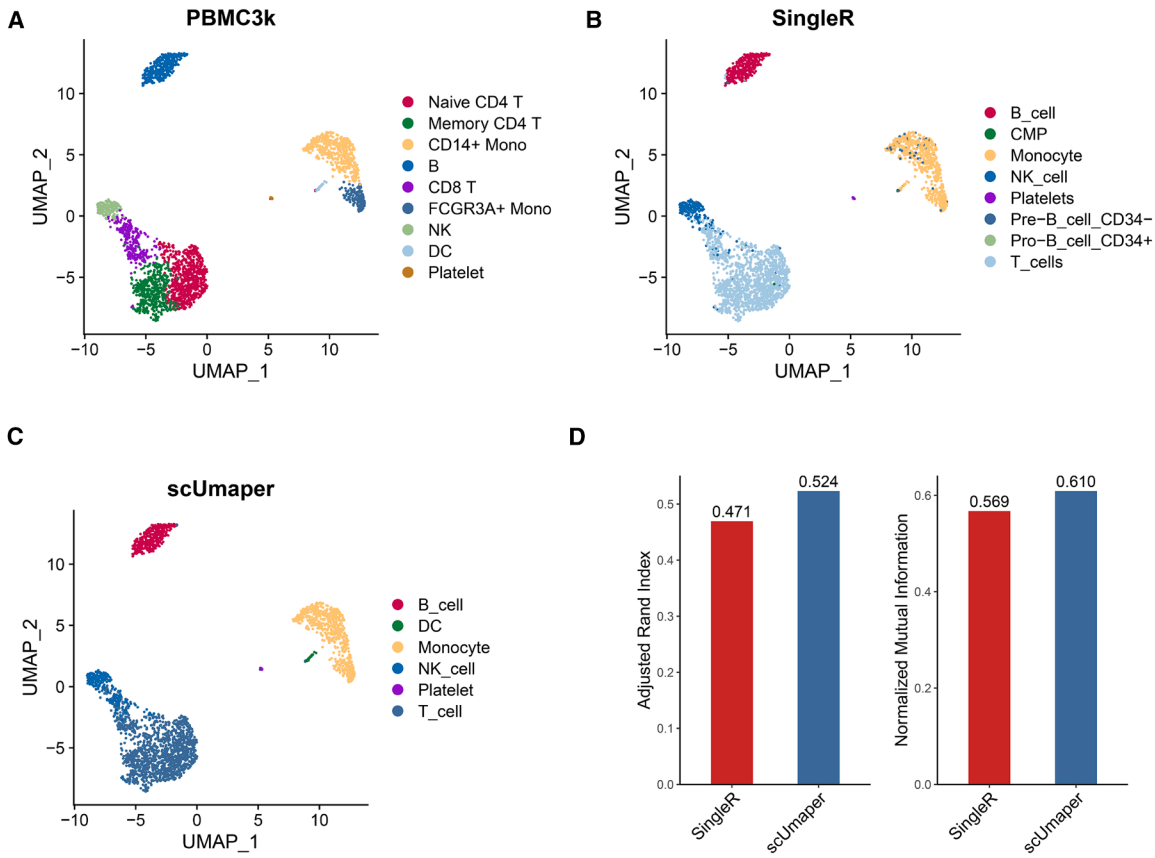


Figure 5. Annotation consistency of scUmaper and SingleR on objective PBMC 3k standard dataset

(A) UMAP plot of PBMC 3k dataset with established cell-type labels.

(B) UMAP plot of PBMC 3k dataset with cell-type annotations from SingleR.

(C) UMAP plot of PBMC 3k dataset with cell-type annotations from scUmaper.

(D) ARI and NMI difference of annotation consistency between SingleR and scUmaper on PBMC 3k dataset.

foundational literature[32–37]), we acknowledge that fixed references cannot capture every biological nuance. To address this, we implemented a custom interface (*scUmaper::run_custom_scumaper*), allowing users to supply tissue- or context-specific cell-type definitions and marker sets. However, the effectiveness of this custom mode is inherently tied to the quality of user-provided inputs. We emphasize that marker quality (specifically coverage and mutual exclusivity) is a critical parameter. Users should be aware that scUmaper's output resolution is bounded by the predefined taxonomy, and that poorly defined or overlapping markers can lead to an increase in false-positive doublet detections. This balance between "out-of-the-box" automation and expert-driven customization ensures the tool's utility across a wide range of biological systems. Looking forward, we are committed to the ongoing refinement and expansion of our marker libraries based on emerging literature and user feedback, while adhering to conservative selection criteria to ensure high specificity across diverse tissues and disease states.

Despite its strengths, scUmaper has inherent limitations. Doublets in droplet-based scRNA-seq are broadly categorized into: (1) heterotypic (cross-lineage) doublets, formed by biologically distinct lineages (e.g., T cell + B cell); and (2) homotypic (within-lineage) doublets, formed by closely related subtypes (e.g., $CD4^+$ Tfh + $CD4^+$ Treg). For heterotypic doublets, the biological criteria are direct and widely accepted: Certain markers are mutually exclusive under normal physiological conditions (e.g., CD3E for T cells vs. MS4A1 for B cells). In contrast, identifying homotypic doublets via expression alone is inherently ambiguous due to shared developmental programs and overlapping expression profiles (e.g., cytotoxic programs across $CD8^+$ T cell states), making "co-expression" an unreliable universal rule for detection. Since scUmaper is intentionally designed to systematize expert heuristics for heterotypic doublets, it does not target homotypic doublets, which remain the primary domain of simulation-based or probabilistic models. Practically, we recommend using scUmaper as a first-pass filter for clear heterotypic doublets, optionally followed by a simulation-based method (e.g., DoubletFinder) when increased coverage of potential homotypic doublets is desired. As another potential workaround, after removing cross-lineage doublets by scUmaper, users may re-run the custom scUmaper mode on a specific lineage of interest using carefully curated, mutually exclusive subtype markers where biologically justified.

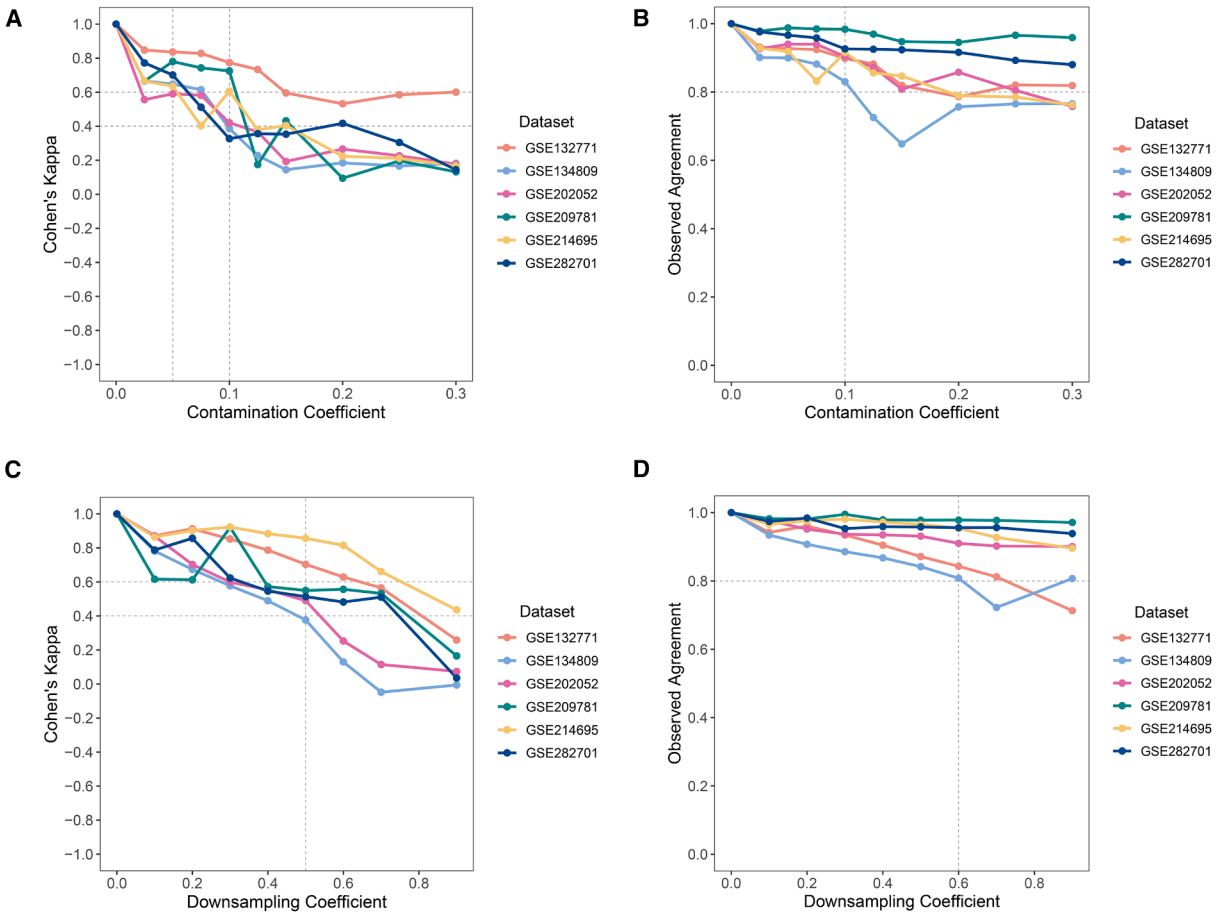


Figure 6. Robustness of scUmaper under ambient RNA and low sequencing depth stress

(A) Variations in Cohen's kappa across datasets relative to the ambient RNA contamination coefficient.

(B) Variations in observed agreement across datasets relative to the ambient RNA contamination coefficient.

(C) Variations in Cohen's kappa across datasets relative to the downsampling coefficient.

(D) Variations in observed agreement across datasets relative to the downsampling coefficient.

Furthermore, we observed that scUmaper's doublet discovery rate was less than ideal in highly specialized tissues, such as the kidney, a limitation that is also observed in the evaluated simulation-based baselines on this tissue. The kidney represents one of the most complex human organs, containing highly specialized epithelial subtypes (e.g., proximal tubule, collecting duct). At the "major-group:" level, default settings for highly variable genes (3,000) may capture strong epithelial subtype programs, suggesting that treating "epithelial" as a single major group may be too coarse-grained. In such instances, users would benefit from defining more granular, tissue-specific major groups and marker sets within the custom mode to increase resolution and improve doublet discovery.

In conclusion, scUmaper represents a practical step toward biologically informed and reproducible scRNA-seq preprocessing by codifying expert marker-based heuristics into an automated Seurat-native workflow. While maintaining high computational efficiency and robustness, it provides a rigorous, accessible, and interpretable framework for the single-cell community. We anticipate that by democratizing rigorous preprocessing, scUmaper will not only accelerate individual discovery but also foster greater reproducibility and collaboration in the field of single-cell transcriptomics.

Limitations of the study

scUmaper is designed to prioritize biologically implausible heterotypic doublets using lineage-marker incompatibility rules; it is not intended to sensitively detect homotypic doublets where expression programs are shared across closely related subtypes. Because the doublet filter and annotation rely on predefined marker libraries and mutual exclusivity assumptions, performance may be reduced in tissues with highly specialized or poorly covered lineages, or when user-supplied markers overlap across categories. Although stress tests indicated robustness to moderate ambient RNA contamination and reduced sequencing depth, extreme contamination or strong technical artifacts may still introduce false-positive lineage mixing; applying dedicated correction tools (e.g., ambient RNA removal) upstream may be beneficial in such

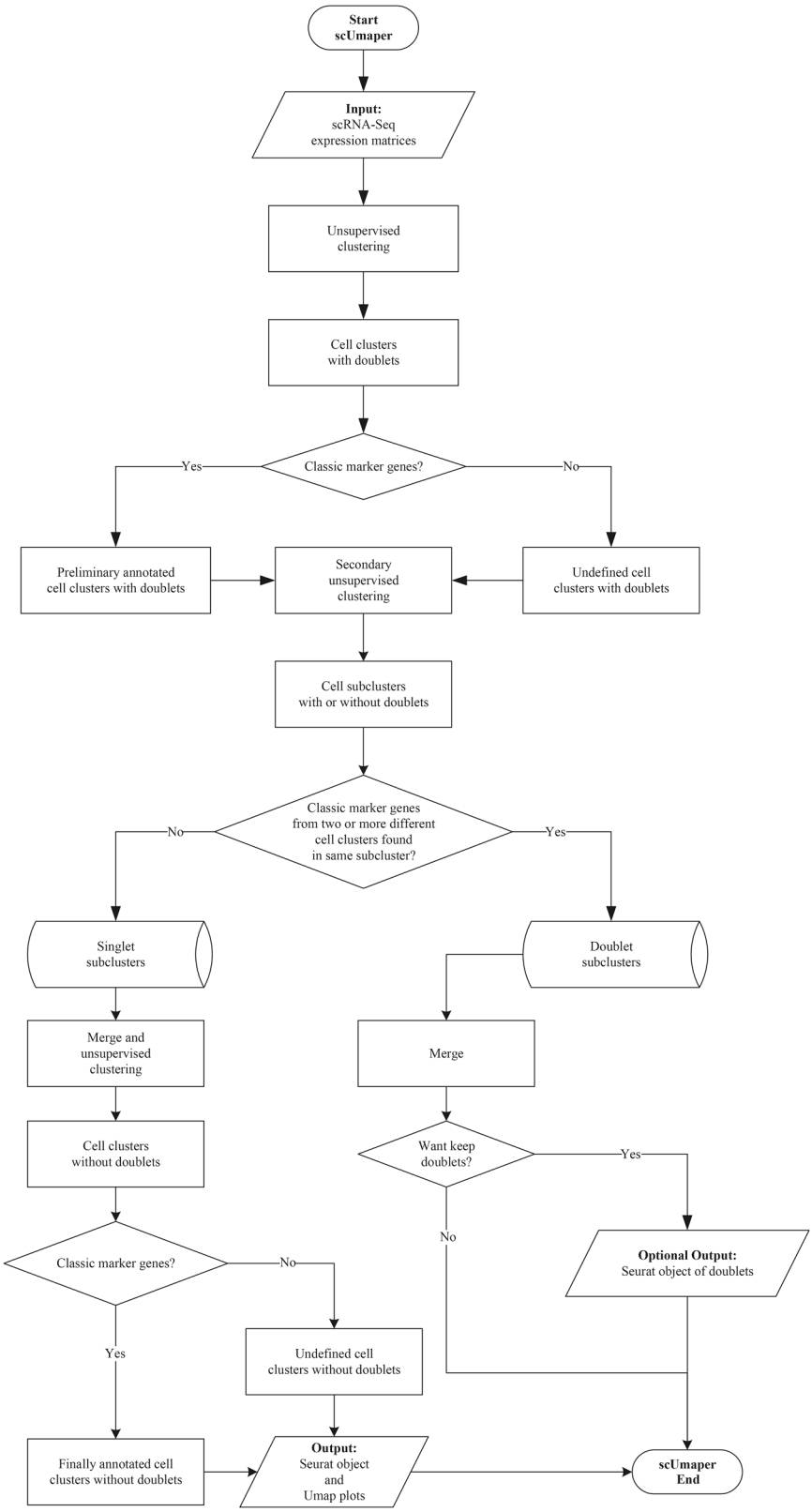


Figure 7. Workflow of the scUmaper pipeline for automated scRNA-seq data processing

Table 2. Reference marker genes for major human cell types used in the scUmaper pipeline

Cell type	Marker genes
T/NK cell	CD3D; CD3E
B/Plasma cell	CD79A; CD79B; CD19; MS4A1; MZB1; IGHG1; IGHA1
Myeloid	CD68; CSF3R; TPSAB1
Fibroblast	COL1A1; ACTA2; ACTG2; COL6A3
Endothelial cell	PECAM1; VWF
Epithelial cell	EPCAM; KRT8; KRT19

cases. Finally, agreement with expert curation reflects consistency with marker-based biological heuristics rather than an orthogonal ground truth, and downstream validation remains important for specific biological claims.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Gaoshi Zhou (zhougshi@mail3.sysu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This article does not report original data.
- scRNA-seq data were downloaded from publicly available datasets, accessible at GEO (GEO: GSE132771, GSE134809, GSE202052, GSE209781, GSE214695, GSE282701) and http://cf.10xgenomics.com/samples/cell-exp/1.1.0/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz.
- All original code has been deposited at GitHub repositories (https://github.com/guoxsh3/scUmaper and https://github.com/guoxsh3/scUmaper-validation) and is publicly available at Zenodo: https://doi.org/10.5281/zenodo.19096359 and Zenodo: https://doi.org/10.5281/zenodo.19096428 as of the date of publication.
- Any additional information required to reanalyze the data reported in this article is available from the lead contact upon request.

ACKNOWLEDGMENTS

We extend our sincere gratitude to Mr. Jinzhe Huang from the scUmaper Project Community for his dedicated contributions to server configuration, performance optimization, and code refinement during the package's public release. We also wish to thank Mr. Zhuohui Wu from The First Affiliated Hospital, Sun Yat-sen University, for his invaluable assistance in sourcing and curating the public datasets used for validation. Their support was crucial to the practical implementation and benchmarking of this work. This study was supported by the Guangdong Basic and Applied Basic Research Foundation, grant no. #2024A1515013032.

Table 3. Reference marker genes for major mouse cell types used in the scUmaper pipeline

Cell type	Marker genes
T/NK cell	CD3d; Cd3e
B/Plasma cell	Cd79a; Cd19; Ms4a1; Mzb1; Sdc1; Ighg1; Igha; Jchain
Myeloid	Cd68; Csf3r; S100a8; Mcpt4
Fibroblast	Col1a1; Col1a2; Acta2; Actg2
Endothelial cell	Pecam1; Eng; Cdh5
Epithelial cell	Epcam; Krt8; Krt18

AUTHOR CONTRIBUTIONS

Conceptualization, G.Z. and S.Z.; methodology, X.G.; investigation, X.G., M.Z., Z.X., and Y.W.; writing – original draft, X.G. and M.Z.; writing – review and editing, G.Z. and S.Z.; funding acquisition, G.Z.; resources, X.G. and G.Z.; supervision, G.Z. and S.Z.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Algorithm workflow
  - Validation datasets and procedures
  - Robustness testing through simulated data degradation
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2026.115850.

Received: December 5, 2025
Revised: February 6, 2026
Accepted: April 20, 2026
Published: April 22, 2026

REFERENCES

1. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.
2. Gulati, G.S., D'Silva, J.P., Liu, Y., Wang, L., and Newman, A.M. (2025). Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. Nat. Rev. Mol. Cell Biol. *26*, 11–31. https://doi.org/10.1038/s41580-024-00768-2.
3. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (New York, N.Y.) *352*, 189–196. https://doi.org/10.1126/science.aad0501.
4. Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (New York, N.Y.) *347*, 1138–1142. https://doi.org/10.1126/science.aaa1934.
5. Villani, A.C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science (New York, N.Y.) *356*, eaah4573. https://doi.org/10.1126/science.aah4573.

6. Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol. *18*, 35–45. https://doi.org/10.1038/nri.2017.76.

7. McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., Hu, J.L., Murrow, L.M., Weissman, J.S., Werb, Z., et al. (2019). MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods *16*, 619–626. https://doi.org/10.1038/s41592-019-0433-8.

8. Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. *8*, 281–291.e9. https://doi.org/10.1016/j.cels.2018.11.005.

9. Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746. https://doi.org/10.15252/msb.20188746.

10. Kiselev, V.Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. Nat. Methods *15*, 359–362. https://doi.org/10.1038/nmeth.4644.

11. Hou, R., Denisenko, E., and Forrest, A.R.R. (2019). scMatch: a single-cell gene expression profile annotation tool using reference datasets. Bioinformatics *35*, 4688–4695. https://doi.org/10.1093/bioinformatics/btz292.

12. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

13. Zhang, Z., Luo, D., Zhong, X., Choi, J.H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski, E.W., Modrusan, Z., et al. (2019). SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. Genes *10*, 531. https://doi.org/10.3390/genes10070531.

14. Shao, X., Liao, J., Lu, X., Xue, R., Ai, N., and Fan, X. (2020). scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. iScience *23*, 100882. https://doi.org/10.1016/j.isci.2020.100882.

15. Ianevski, A., Giri, A.K., and Aittokallio, T. (2022). Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat. Commun. *13*, 1246. https://doi.org/10.1038/s41467-022-28803-w.

16. Bi, C., Bai, K., and Zhang, X. (2025). HiCat: a semi-supervised approach for cell type annotation. Brief. Bioinform. *26*, bbaf428. https://doi.org/10.1093/bib/bbaf428.

17. Guo, H., and Li, J. (2021). scSorter: assigning cells to known cell types according to marker genes. Genome Biol. *22*, 69. https://doi.org/10.1186/s13059-021-02281-7.

18. Domínguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T., Howlett, S.K., Suchanek, O., Polanski, K., King, H.W., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science (New York, N.Y.) *376*, eabl5197. https://doi.org/10.1126/science.abl5197.

19. Lin, Y., Cao, Y., Kim, H.J., Salim, A., Speed, T.P., Lin, D.M., Yang, P., and Yang, J.Y.H. (2020). scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. Mol. Syst. Biol. *16*, e9389. https://doi.org/10.15252/msb.20199389.

20. Bai, K., Moa, B., Shao, X., and Zhang, X. (2025). scSorterDL: a deep neural network-enhanced ensemble LDAs for single cell classifications. Brief. Bioinform. *26*, bbaf446. https://doi.org/10.1093/bib/bbaf446.

21. Bai, K., Moa, B., Shao, X., and Zhang, X. (2025). PCLDA: An interpretable cell annotation tool for single-cell RNA-sequencing data based on simple statistical methods. Comput. Struct. Biotechnol. J. *27*, 3264–3274. https://doi.org/10.1016/j.csbj.2025.07.019.

22. McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. *8*, 329–337.e4. https://doi.org/10.1016/j.cels.2019.03.003.

23. Germain, P.L., Lun, A., Garcia Meixide, C., Macnair, W., and Robinson, M.D. (2021). Doublet identification in single-cell sequencing data using scDblFinder. F1000Res. *10*, 979. https://doi.org/10.12688/f1000research.73600.2.

24. Xi, N.M., and Li, J.J. (2021). Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. Cell Syst. *12*, 176–194.e6. https://doi.org/10.1016/j.cels.2020.11.008.

25. Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. Nat. Rev. Genet. *20*, 257–272. https://doi.org/10.1038/s41576-019-0093-7.

26. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44. https://doi.org/10.1038/nbt.4314.

27. Tsukui, T., Sun, K.H., Wetter, J.B., Wilson-Kanamori, J.R., Hazelwood, L.A., Henderson, N.C., Adams, T.S., Schupp, J.C., Poli, S.D., Rosas, I.O., et al. (2020). Collagen-producing lung cell atlas identifies multiple subsets with distinct localization and relevance to fibrosis. Nat. Commun. *11*, 1920. https://doi.org/10.1038/s41467-020-15647-5.

28. Martin, J.C., Chang, C., Boschetti, G., Ungaro, R., Giri, M., Grout, J.A., Gettler, K., Chuang, L.S., Nayar, S., Greenstein, A.J., et al. (2019). 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. https://doi.org/10.1016/j.cell.2019.08.008.

29. Maddipatla, S.C., Kolachala, V.L., Venkateswaran, S., Dodd, A.F., Pelia, R.S., Geem, D., Yin, H., Sun, Y., Xu, C., Mo, A., et al. (2023). Assessing Cellular and Transcriptional Diversity of Ileal Mucosa Among Treatment-Naïve and Treated Crohn's Disease. Inflamm. Bowel Dis. *29*, 274–285. https://doi.org/10.1093/ibd/izac201.

30. Garrido-Trigo, A., Corraliza, A.M., Veny, M., Dotti, I., Melón-Ardanaz, E., Rill, A., Crowell, H.L., Corbí, Á., Gudiño, V., Esteller, M., et al. (2023). Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. Nat. Commun. *14*, 4506. https://doi.org/10.1038/s41467-023-40156-6.

31. Young, M.D., and Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. GigaScience *9*, giaa151. https://doi.org/10.1093/gigascience/giaa151.

32. Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., et al. (2023). CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. *51*, D870–D876. https://doi.org/10.1093/nar/gkac947.

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

34. Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303–309. https://doi.org/10.1038/s41586-020-2157-4.

35. Bandyopadhyay, S., Duffy, M.P., Ahn, K.J., Sussman, J.H., Pang, M., Smith, D., Duncan, G., Zhang, I., Huang, J., Lin, Y., et al. (2024). Mapping the cellular biogeography of human bone marrow niches using single-cell transcriptomics and proteomic imaging. Cell *187*, 3120–3140.e29. https://doi.org/10.1016/j.cell.2024.04.013.

36. Shi, Q., Chen, Y., Li, Y., Qin, S., Yang, Y., Gao, Y., Zhu, L., Wang, D., and Zhang, Z. (2025). Cross-tissue multicellular coordination and its rewiring in cancer. Nature *643*, 529–538. https://doi.org/10.1038/s41586-025-09053-4.

37. Chen, Y., Wang, D., Li, Y., Qi, L., Si, W., Bo, Y., Chen, X., Ye, Z., Fan, H., Liu, B., et al. (2024). Spatiotemporal single-cell analysis decodes cellular

dynamics underlying different responses to immunotherapy in colorectal cancer. Cancer Cell *42*, 1268–1285.e7. https://doi.org/10.1016/j.ccell.2024.06.009.

38. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172. https://doi.org/10.1038/s41590-018-0276-y.

39. Korsunsky, I., Nathan, A., Millard, N., and Raychaudhuri, S. (2019). Presto scales Wilcoxon and auROC analyses to millions of observations. bioRxiv. https://doi.org/10.1101/653253. https://www.biorxiv.org/content/10.1101/653253v1.

40. Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., and Ruzzo, W.L. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics *17*, 977–987. https://doi.org/10.1093/bioinformatics/17.10.977.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Lung scRNA-seq dataset	Tsukui et al.[27] 2020	GEO: GSE132771
Ileum scRNA-seq dataset 1	Martin et al.[28] 2019	GEO: GSE134809
Ileum scRNA-seq dataset 2	Maddipatla et al.[29] 2023	GEO: GSE202052
Kidney scRNA-seq dataset	Lv et al., 2025	GEO: GSE209781
Colorectum scRNA-seq dataset	Garrido-Trigo et al.[30] 2023	GEO: GSE214695
Liver scRNA-seq dataset	Ximing S, 2025	GEO: GSE282701
PBMC 3k dataset	10x Genomics, 2017	http://cf.10xgenomics.com/samples/cell-exp/1.1.0/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz
Software and algorithms		
R (4.3.3)	R Core Team	https://www.r-project.org/
Seurat (5.2.1)	Satija et al.	https://doi.org/10.1038/s41587-023-01767-y
DoubletFinder (2.0.4)	Chris McGinnis	https://github.com/chris-mcginnis-ucsf/DoubletFinder
scDblFinder (1.16.0)	Germain et al.[23] 2021	https://doi.org/10.12688/f1000research.73600.2
SingleR (2.4.1)	Aran et al.[38] 2019	https://doi.org/10.1038/s41590-018-0276-y
scType	Aleksandr et al.	https://doi.org/10.1038/s41467-022-28803-w
presto (1.0.0)	Korsunsky et al.[39] 2019	https://github.com/immunogenomics/presto
mclust (6.1.1)	Scrucca et al.	https://doi.org/10.1201/9781003277965
aricode (1.0.3)	Chiquet et al.	https://cran.r-project.org/package=aricode
scUmaper (0.2.1)	Guo et al.	Zenodo: https://doi.org/10.5281/zenodo.19096359

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Omitted as our study does not involve biological models.

METHOD DETAILS

Algorithm workflow

scUmaper accepts as input raw or otherwise preprocessed single-cell expression matrices arranged per sample (Cell Ranger output structure or comparable matrix/feature/barcode triplets). The expression matrices are converted into a Seurat object, followed by quality control and dimensionality reduction.[12] Subsequently, scUmaper performs unsupervised clustering in a two-step hierarchical design: a global clustering step to obtain major lineage-level groups; followed by re-clustering within each major group to identify anomalous subclusters for rule-based doublet filtering. In the first step, scUmaper performs global clustering to obtain a sufficient number of major lineage-level groups. This procedure follows the standard Seurat pipeline: after matrix normalization, the framework identifies highly variable genes (HVGs; default = 3,000), performs data scaling, and conducts principal component analysis (PCA). Subsequently, k-nearest neighbor (KNN)/shared nearest neighbor (SNN) graph construction and Louvain (graph-based) clustering are executed. During this process, users can customize the number of HVGs (typically ranging from 2,000 to 3,000 depending on tissue complexity, higher complexity generally warrants a larger value), while all other parameters are set to Seurat's default values. Next, marker genes of each cell group are identified using the Presto algorithm[39] and compared

with the reference marker gene libraries of major cell types (Tables 2 and 3) to initially classify each cell group as T/NK cells, B/plasma cells, myeloid cells, fibroblasts, endothelial cells, or epithelial cells by default for human or mouse tissues. The marker libraries were curated by integrating CellMarker and PanglaoDB resources and cross-checking with multiple authoritative scRNA-seq studies,[32–37] using markers selected for high lineage specificity and mutual exclusivity. Cell groups without marker gene matches to any major category are labeled as undefined. After the initial annotation, cells undergo another round of quality control by calculating the mitochondrial gene percentage. At this stage, immune cells are subject to a stricter quality threshold, with a default upper limit of 5%, while non-immune cells are allowed up to 25%.Subsequently, in the second step, each cell group undergoes additional unsupervised clustering to identify distinct subgroups (following the same dimensionality reduction and clustering procedures as applied at the global population level). For each subgroup, marker genes are identified and cross-compared with those of other major cell types. If a subgroup within any preliminarily annotated major cell type expresses marker genes that are exclusive to a different major cell type, it is labeled as a doublet. For example, if a subgroup within a preliminarily annotated T/NK cell group expresses MS4A1 (CD20), it is labeled as a doublet, as CD3E and MS4A1 are not co-expressed in the same cell based on biological principles. Similarly, if subgroups originating from stromal or epithelial cell groups express PTPRC (CD45), they are also labeled as doublets. For undefined cell groups, subgroups are classified as doublets if they co-express marker genes from two or more distinct major cell types.

Furthermore, cells identified as doublets are separated from singlet cells and can optionally be retained as a Seurat object if the user wishes to review and validate the removed cells during the automatic analysis. Singlet cells, after doublet removal, are merged and reprocessed for dimensionality reduction, clustering, and final annotation, following the same procedures as the preliminary annotation step. Uniform manifold approximation and projection (UMAP) plots and feature plots of marker genes for the final Seurat object of singlet cells can also be generated upon request. The complete workflow of scUmaper is illustrated in Figure 7. The scUmaper package is implemented in R and available for download via Github. Installation instructions, tutorials, source code, and detailed vignettes are available on its GitHub repository: https://github.com/guoxsh3/scUmaper.

Validation datasets and procedures

To evaluate the robustness and generalizability of scUmaper, we applied it to six publicly available scRNA-seq datasets derived from five different human organs (lung, ileum, liver, kidney, and colorectum) obtained from the GEO database (GEO: GSE132771, GSE134809, GSE282701, GSE202052, GSE209781, GSE214695).[27–30] scUmaper was used to automatically perform clustering, remove doublets, and annotate cell types, and its performance was systematically compared to that of the widely used doublet-removal algorithm DoubletFinder[22] concatenated to automatic annotation algorithm SingleR,[38] and compared to another widely used doublet-removal algorithm scDblFinder[23] concatenated to automatic annotation algorithm scType.[15] For all methods, computations were performed using their respective default parameters. Specifically, SingleR utilized the Human Primary Cell Atlas (HPCA) as the reference dataset, while scType employed its comprehensive reference library of cell types and marker genes across all available tissues. An external validation dataset, PBMC 3k standard dataset with established labels from 10× Genomics, was also added to provide an objective validation of annotation. scUmaper and SingleR were run respectively on PBMC 3k dataset to compare annotation performance. To ensure comparability, all the algorithms were executed on the same Linux-based server with the same running resources.

Meanwhile, manual doublet-removal and annotation were performed by two experienced bioinformatic experts and used as the gold standard, the procedures are shown as follow. After initial clustering into major groups, within-group unsupervised subclustering were performed. For each subcluster, Expert 1 inspected marker expression and flagged clusters co-expressing markers from two or more major lineages as doublets. Expert 2 independently reproduced the analysis scripts and re-evaluated marker expression for those flagged subclusters. Discrepancies were resolved through joint review and consensus adjudication. During these procedures, disagreements were infrequent (<5% of reviewed cells) and were resolved by consensus review. In addition, two experts performed curation blinded to outputs of scUmaper and other tested algorithms during the procedures above. The full curation scripts are available in our GitHub repository (https://github.com/guoxsh3/scUmaper-validation) to facilitate transparency and reproducibility.

Robustness testing through simulated data degradation

To evaluate the algorithmic resilience against background noise, we simulated ambient RNA contamination, a common artifact in droplet-based single-cell sequencing. We defined the Soup Profile as the aggregated transcriptomic distribution across all captured cells. For each gene g, its relative abundance in the background S_g was calculated as:

$$S_g = \frac{\sum_{i=1}^{n} C_{g,i}}{\sum_{g \in G} \sum_{i=1}^{n} C_{g,i}}$$

where $C_{g,i}$ represents the raw unique molecular identifier (UMI) count of gene g in cell i, n is the total number of cells, and G is the set of all detected genes.

The "polluted" count for each cell was then generated by adding a noise component proportional to the cell's original library size and a coefficient factor ρ. The polluted count $C'_{g,i}$ is expressed as:

$$C'_{g,j} = C_{g,j} + \left(S_g \times \rho \times \sum_{g \in G} C_{g,i} \right)$$

This model simulates the worst-case scenario where the ambient RNA profile perfectly reflects the pooled transcriptome of the sample, thereby challenging the specificity of cell-type-marking transcripts. Across the six test datasets, we randomly selected 20,000 cells per dataset and contaminated the expression matrices according to the aforementioned procedure. These contaminated matrices were then utilized as input to execute the scUmaper pipeline.

On the other hand, to assess the performance of our method under limited sequencing depth, we performed stochastic downsampling on the UMI count matrix. We define ρ' as the downsampling coefficient, representing the fraction of signal removed during the simulation. Consequently, the probability of any single UMI molecule being retained is 1 - ρ'. The downsampled count $C''_{g,i}$ was determined using a binomial sampling process:

$$C''_{g,i} \sim \text{Binomial}\left(n = C_{g,i}, p = 1 - \rho'\right)$$

This model preserves the discrete nature of UMI counts and accurately captures the increased sparsity and technical "dropout" events observed when sequencing depth is curtailed. Furthermore, we performed downsampling on the expression matrices of the six test datasets following the aforementioned procedure and used the resulting matrices as input for scUmaper.

Subsequently, we evaluated the consistency between the scUmaper results obtained from the contaminated and downsampled matrices and those from the original, unprocessed matrices. To quantify this consistency, we calculated both the observed agreement and Cohen's kappa coefficients.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis was performed using R (version 4.3.3). The performance of scUmaper was statistically analyzed in two parts: doublet removal and annotation. In the doublet removal section, the doublets identified by the scUmaper DoubletFinder, and scDblFinder algorithms were respectively compared with the gold standard results of manual doublet removal, and accuracy metrics such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated. In the annotation section, the annotation results of the scUmaper, SingleR, and scType algorithms were compared with the gold standard manual annotation results using the mclust[40] and aricode (https://github.com/jchiquet/aricode) packages to calculate the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) values to assess annotation consistency. For the PBMC 3k dataset, the consistency between the annotation results of scUmaper and SingleR and the ground-truth labels was also evaluated using NMI and ARI. We also recorded the processing time for running scUmaper alone and for the concatenated execution of DoubletFinder and SingleR. Robustness under ambient RNA contamination and downsampling was assessed using observed agreement and Cohen's kappa. Furthermore, we compared the differences in statistical metrics between scUmaper and the other algorithms at the sample level using paired t-tests to evaluate the robustness and generalizability of scUmaper. Exact *p* values and statistics are reported in the corresponding figures and tables.