Computational and Structural Biotechnology Journal 20 (2022) 4902–4909

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

scWizard: A web-based automated tool for classifying and annotating single cells and downstream analysis of single-cell RNA-seq data in cancers

Jinfen Wei [1], Qingsong Xie [1], Yimo Qu, Guanda Huang, Zixi Chen, Hongli Du *

School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 22 March 2022
Received in revised form 27 July 2022
Accepted 12 August 2022
Available online 27 August 2022

Keywords:
Single-cell RNA-sequencing
Automated and integrated analysis shiny-based R package
Hierarchical cell annotation
Deep neural network
Tumor microenvironment

ABSTRACT

The emerging number of single-cell RNA-seq (scRNA-Seq) datasets allows the characterization of cell types across various cancer types. However, there is still lack of effective tools to integrate the various analysis of single-cells, especially for making fine annotation on subtype cells within the tumor microenvironment (TME). We developed scWizard, a point-and-click tool packaging automated process including our developed cell annotation method based on deep neural network learning and 11 downstream analyses methods. scWizard used 113,976 cells across 13 cancer types as a built-in reference dataset for training the hierarchical model enabling to automatedly classify and annotate 7 major cell types and 47 cell subtypes in the TME. scWizard provides a built-in pre-training set for user's flexible choice, and gives a higher accuracy for annotation subtypes of tumor-derived *T*-lymphocytes/natural killer cells (T/NK) and myeloid cells from different cancer types compared with the existing five methods. scWizard has good robustness in three independent cancer datasets, with an accuracy of 0.98 in annotating major cell types, 0.85 in annotating myeloid cell subtypes and 0.79 in annotating T/NK cell subtypes, indicting the wide applicability of scWizard in different cell types of cancers. Finally, the automatic analysis and visualization function of scWizard are presented by using the intrahepatic cholangiocarcinoma (ICC) scRNA-Seq dataset as a case. scWizard focuses on decoding TME and covers various analysis flows for cancer scRNA-Seq study, and provides an easy-to-use tool and a user-friendly interface for researchers widely, to further accelerate the biological discovery of cancer research.

1. Introduction

The breakthrough of the cancer research and anti-tumor therapy largely relies on the understanding of the heterogeneity of cell types and crosstalk between cells within the tumor microenvironment (TME) [1]. With the rapid development of single-cell RNA sequencing (scRNA-seq), it is now possible to provide the expressing patterns of each cell type in the TME and decipher the intercellular communication networks to explain their roles in cancer progression. The bioinformatics packages and analysis methods including Seurat [2] have been developed according to the requirements, which has accelerated cancer research.

To speed up the analysis efficiency of single cells, numerous automated software and pipelines have been developed for scRNA-seq data including scCancer [3], ASAP [4], dropClust [5], iS-CellR [6], Cerebro [7] and ascend [8]. In terms of tool applicability, these tools have limitations in personalized selection and deficiency in providing essential features, such as lacking pseudo-time trajectory analysis in most tools except Cerebro [7] and lacking cell–cell interaction analysis in most tools except scCancer [3] (Table 1). As cell-type annotation is an essential step in analyzing scRNA-seq data, which is time-consuming and subjective, therein, scCancer, dropClust and ASAP have incorporated the cell-type annotation function. However, dropClust only enables to annotate the peripheral blood mononuclear (PBMC) cells [5] and ASAP uses CellMarker and PanglaoDB databases for cell-type annotation [4], which is not always applicable for cell subtypes within the TME. scCancer uses one-class logistic regression (OCLR) model and only annotate major cell types without subtypes annotation [3]. Most importantly, there are few automation pipelines specifically for cancer scRNA-seq data except scCancer.

* Corresponding author.
E-mail address: hldu@scut.edu.cn (H. Du).
[1] These authors are joint first authors.

https://doi.org/10.1016/j.csbj.2022.08.028

Table 1
Function comparison with multiple tools.

	scWizard	scCancer	dropClust	ASAP	Cerebro	ascend	iS-CellR
Quality control	✓	✓	✓	✓	✓	✓	✓
Inference of CNV	✓	✓					
Batch processing	✓	✓	✓	✓		✓	
Cell clustering	✓	✓	✓	✓	✓	✓	✓
Cell subcluster annotation	✓	✓	✓	✓			
Cell-cell interaction	✓	✓					
Pseudo-time trajectory	✓				✓		
TF regulatory network analysis	✓						
Correlation analysis	✓	✓					
Gene set signature estimation	✓	✓	✓	✓	✓		
Visualization	✓	✓	✓	✓	✓	✓	✓

Until now, there have been several methods and tools specialized for cell type annotation developed, but few tools for classifying and annotating cell types specifically for cells within the TME. The most commonly used tools are CellAssign [9], SCSA [10], SCINA [11], scmap [12], SingleR [13], CHETAH [14], SingleCellNet [15], ACTINN [16] and scPred [17]. CellAssign, SCSA and SCINA are dependent on prior knowledge and challenged by the fact that there is not a canonical set of marker genes for all cell types and subtypes in the TME. scmap and SingleR are correlation-based tools, a general reference data set is included with the tool, but they may not perform well when a reference is not specifically matched to the query data set [18]. SingleCellNet, ACTINN and scPred are supervised learning methods, but they classify cells directly to a "terminal" cell type and overlook the hierarchical relationships between cell types. For instance, distinguishing CD8 T cells from CD4 T is relatively direct, but sometimes cannot accurately distinguish CD8+ exhausted T cells from Treg cells, directly.

Here, we developed scWizard, a shiny-based R package that aims to comprehensively process scRNA-Seq data for cancer research. Except for integrating 11 fundamental analyses (Table 1), scWizard highlights the cell annotation function, which is based on the deep neural network study to establish an accurate prediction model with hierarchical classification for classifying cell types within the TME. 113,976 cells with 7 major cell types and 47 subtypes labeled from 199 samples across 13 cancer types were built in scWizard as a reference database for pre-training the model, which is ready to automatically run on diverse scRNA-seq data sets of tumors and is able to effectively distinguish cell subtypes within the TME. Notably, scWizard is a point-and-click tool packaging automated methods into easy-to-use workflows to facilitate diverse scRNA-Seq analysis and to visualize the results for the researchers without sufficient bioinformatics experience. scWizard is available as an R package on GitHub (https://github.com/Dulab2020/scWizard).

2. Methods

2.1. Data collection and processing

All datasets used in this study were publicly available and the detailed information was presented in Table S1. We organized these datasets for building annotation model, assessing performance and using them as the case study. To create the training set of major cell types and their corresponding subtypes in TME, we downloaded 113,976 cells from 10 datasets including 199 samples across 13 cancer types. The cell type information of major cell types, endothelial cells, myeloid cells and T/NK cells were provided by the original research. For the stromal cell subtypes, we manually annotated stromal cells by canonical cell markers (Table S2) and investigated gene expression patterns of cells in breast cancer (BC), colorectal cancer (CRC), lung cancer (LC), ovarian cancer (OV), pancreatic ductal adenocarcinoma (PDAC) and squamous cell carcinoma (SCC) datasets. In the comparison of accuracy assessment with other methods, four cancer datasets including lymphoma (LYM), pancreatic adenocarcinoma of myeloid cells (Mye_PAAD), breast cancer of T/NK cells (T/NK_BRCA) and kidney renal clear cell carcinoma of T/NK cells (T/NK_KIRC) were applied. In assessment robustness of cell type annotation function, independent datasets from publicly available human Merkel cell carcinoma (MCC), CRC and LC with labeled cell type were obtained. For application description, we used a publicly available human intrahepatic cholangiocarcinoma (ICC) scRNA-seq data set. For all datasets described in Table S1, only cells that passed the quality control (QC) of the original publication and assigned cell types were included.

2.2. Clustering based on Seurat

Seurat was used to define a standardized scRNA-seq pipeline in scWizard. This step included in scWizard currently were dimension reduction, cell clustering and differential gene expression detection. For cell clustering, scWizard provided Seurat's clustering algorithm. Non-linear dimension reduction techniques were used to visualize all cells in two dimensions using Uniform Manifold Approximation and Projection (UMAP) techniques. UMAP was used to visualize the cell clusters, gene expression of interest, GSVA score, or sample information with an unannotated "single-cell map" image. The different gene analysis was based on a two-sided Wilcoxon rank-sum test with Bonferroni FDR correction.

2.3. Quality control

By default, scWizard defined cells with <200 genes detected, > 10,000 genes detected, >5 % of reads mapping to mitochondrial RNA as low-quality cells and filtered them out of the downstream analysis.

2.4. Batch processing

scWizard provided two methods to remove batch effects and performs multi-sample integration analysis by inserting the Seurat [2] and Harmony [19] algorithms. Both Harmony and Seurat are the methods to remove batch effects, in which Harmony performs well on datasets with common cell types and different techniques and the comparatively shorter runtime of Harmony also makes it suitable for initial data exploration of large datasets [20]. The methods and parameter could be chosen and adjusted according to the user's needs.

2.5. Cell type annotation model based on neural network learning

The deep neural network structures and the learning algorithms were implemented using ReLU (Rectified Linear Unit) and softmax activation function. The neural network contained one input layer, one output layer, and one hidden layer for the major cell annotation model and three hidden layers for the subtype cell annotation model, respectively. The input layer had a number of nodes equal to the number of genes in the training set after principal components analysis (PCA). The output layer had a number of nodes equal to the number of cell types in the training set. In order to allow users to have more flexible choice, the number of hidden layer nodes could be adjusted according to users' needs. How the adjustment was made and the criteria to define the number of hidden layer nodes was provided in Supplementary File-User Manual.

We designed a "two-step" annotation model for the cell-type classification, which was applied to individual cells before clustering for major cell types and after clustering for subtype annotation. First, we combined the training dataset labeled each major cell type and prediction dataset together, and performed PCA dimension reduction. Second, the training dataset and prediction dataset were separated, and training data was used to build the model. Then, the prediction data could be assigned by the model corresponding to the major cell type of each cell characteristic and scoring the possibility of each cell labeling as each cell type. The maximum likelihood was chosen to annotate this cell as a specific major cell type.

Each major cell type cluster was separated from prediction data and divided by major cell type, then each major cell type was subclustered based on a specific resolution by Seurat. The subclusters contained in each major cell type were evaluated, and the predicted subcluster corresponding to the largest number of cells within the original subcluster was defined as a specific cell subcluster. The prediction data could be assigned by the model corresponding to the cell subtype of each subcluster characteristic. The clustering before subtype cell annotation could increase the fault tolerance of the model and improve the accuracy of annotations.

To evaluate the overall performance based on the total training dataset, we performed a tenfold validation test. The dataset was divided into ten parts, nine of which were used as training data and one as test data in turn and the average accuracy of ten times was used as the estimate accuracy of the algorithm model and the reliability of the training datasets. The parameters including learning rate, hidden layer nodes and regularization rate could be adjusted in the operation interface according to users' needs.

2.6. Comparison with other cell type annotation tools

To evaluate and compare the performance of scWizard, we obtained five most commonly used and publicly available scRNA-seq classification tools span three main methodological approaches, including CellAssign, SCSA, ACTINN, SingleR and SingleCellNet. These packages were installed either through their Bioconductor or from their GitHub page. Four independent datasets, including T/NK cells and myeloid cells derived from different cancer tissues, were used to evaluate and compare the accuracy between these tools with scWizard. Each dataset was divided into training sets and testing sets, and then organized into the file format required by each tool according to the user manual. For the training sets, scWizard, ACTINN, SingleR and SingleCellNet used the same datasets for training model in annotating these cell subtypes. As CellAssign and SCSA were marker-based annotation tools, we used the classical markers of T/NK cell subtypes for annotating subtypes of T/NK cells (Table S3). As the conventional markers were not comprehensive and representative for subtypes of myeloid cells, the differentially expressed genes of each cell subcluster in the myeloid cells of original annotation information were used as markers for annotating subtypes of myeloid cells (Table S3). In the training process, the default settings were applied in each method.

The output results were the types predicted for cells in the validation datasets. To calculate the accuracy, we defined cells as "consistent" if the predicted cell types matched their annotated cell types in the original study, otherwise, they were "inconsistent" or "unknown" when the results were unmatched with original data or unable to identify, respectively. Then, we compared the number of "consistent", "inconsistent" and "unknown" cells to the total cell number of each cell cluster.

2.7. Performance evaluation of cell type and subtype annotation in independent datasets

The trained model was then applied to independent datasets with the cell information labeled to evaluate the accuracy of scWizard. The predicted results were labeled as "consistent", "unknown", and "inconsistent".

The independent MCC datasets including 11,024 cells with the labeled cell information were used to evaluate the major cell annotation capability of scWizard. 11,717 myeloid cells from CRC, and 45,555 T cells from LC were used to evaluate the robustness of scWizard in cell subtype annotation. Once the model had been evaluated for its robustness and wide applicability, it could be used to annotate single cells from independent cancer dataset for applications.

2.8. Integrative basic analysis

The cell malignancy estimation, gene set signatures estimation, cell–cell interaction, pseudo-time trajectory and transcription factor regulatory network analysis were the routine analyses for users' choice. Users could adjust parameters in the analyses process. The information of packages and methods used in scWizard was summarized in Table S4.

The methods of infer copy number alterations (CNV) were referenced from previous study [3,21] and we applied the algorithm of R package infercnv to estimate single-cell CNVs by scRNA-seq data. We used the GSVA method and gene sets from MSigDB database [22] to calculate the angiogenesis score. scWizard calculated signature scores of any hallmark gene sets and users could upload gene lists obtained from MSigDB or their own interested gene signature to the function. scWizard performed the cell–cell interaction by CellphoneDB method [23], which characterized ligand-receptor interactions extent over cell subclusters. To identify significant interactions, users could filter weak gene pairs according to the *P* value and their mean expression in cell clusters and estimate the global interaction strength between any two cell clusters by the number of the remaining gene pairs. These global interaction strengths between clusters by a bubble chart and annotated clusters with their cell-type fractions for convenience of comparison could be visualized in scWizard. scWizard integrated the Monocle R package [24] to realize its pseudo-time trajectory analysis function and sorted cells by their expression. SCENIC [25], based on co-expression and motif analysis, to calculate scRNA-seq data regulation network relationship reconstruction and cell state identification, was also introduced to scWizard.

3. Results

3.1. Comprehensive overview of the scWizard workflow

scWizard consists of four major modules including QC, cell annotation model, multi-scRNA-seq personalized analysis, and visualization, which enables the comprehensive analysis functions compared to the existing workflows (Fig. 1, Fig. S1). For utilizing the shiny webserver to analyze the user-uploaded data, users need to install scWizard and open the analysis interface. The installation guides and analysis interface guides are presented in the user manual (Supplementary File: User Manual).

3.2. Model for classification and annotation cell major types/subtypes within the TME

By default, scWizard includes 113,976 cells with cell major type and subtype information from 10 datasets as the pretraining sets, and also optionally allows users to add a custom reference database for training. To make the classification and annotation results more accurate and easier to interpret, hierarchical classification is applied and the process is divided into two steps: major cell type and subtype annotation, including 7 cell types and 47 cell subtypes respectively (Table S5). The major cell types include epithelial cells, T/NK cells, myeloid cells, B lymphocytes, fibroblasts, endothelial cells, and mast cells. The subtypes of T/NK cells, myeloid cells, fibroblasts and endothelial cells were divided into 11, 16, 10 and 9 subtypes, respectively (Table S5). To ensure the robustness of scWizard algorithm and the reliability of the training datasets, we performed 10-fold cross-validations on the training datasets, and obtained the average 0.97 of concordances in major cell annotation and average 0.81 to 0.89 accuracy in subtype cell annotation (Table S6).

3.3. Comparison with published methods on cell type annotation

We sought to compare the cell annotation performance of scWizard with five widely used methods based on different computational algorithms, which were representatives for the marker-based, correlation-based and supervised learning-based methods, including CellAssign, SCSA, ACTINN, SingleR and SingleCellNet. We applied each method to four independent scRNA-Seq datasets for annotating cell subtypes. Due to scWizard used hierarchical model, we evaluated scWizard at two levels of the cell type hierarchy, directly annotating (scWizard) and annotating after clustering (scWizard_cluster).

As shown in Fig. 2 and Table S7, scWizard performed significant advantage in annotating cell subtypes within the TME. In the LYM dataset, scWizard and scWizard_cluster performed robust with achieving over 0.66 accuracy for annotation subtypes of Mye_LYM, ACTINN had the accuracy of 0.59. However, SingleR, SingleCellNet, SCSA and CellAssign had an accuracy lower than 0.5 (Fig. 2, Table S7, Fig. S2). In the Mye_PAAD (Fig. 3), T/NK_BRCA (Fig. S3) and T/NK_KIRC (Fig. S4), scWizard achieved the accuracy of 0.64, 0.71 and 0.66, scWizard_cluster improved the accuracy to 0.69, 0.89 and 0.76, respectively (Fig. 2, Table S7). However, the accuracy of ACTINN, SingleR, SingleCellNet, SCSA and CellAssign only had average of 0.57, 0.45, 0.5, 0.5 and 0.1 in Mye_PAAD, T/NK_BRCA and T/NK_KIRC, respectively (Fig. 2, Table S7). Besides, F1 score was also calculated and the results showed that scWizard_cluster obtained 0.740 in annotating T/NK subtype cells in T/NK_BRCA, which is higher than other methods with 0.574, 0.553, 0.572, 0.297, 0.118 in ACTINN, SingleR, SingleCellNet, SCSA and CellAssign, respectively (Table S7). Taken together, these results suggest that scWizard performed better than alternative workflows in both accuracy and F1 score for annotation cell subtypes within the TME.

3.4. Performance evaluation of cell annotation in the independent cancer scRNA-seq datasets

In order to evaluate the robustness of scWizard's performance in different cancer scRNA-seq datasets, we applied three independent datasets from human MCC, CRC, and LC. We examined whether the output cell annotation was consistent with the labels provided by the original study and found accuracy was 0.98 in major cell types from MCC (Fig. S5a, Fig. S5b, Table S8). When predicting the subtypes for myeloid cells from CRC dataset, the accuracy among 16 subtypes was 0.85. The results showed that the dendritic cells, monocytes and macrophages could be clearly and correctly distinguished between each other, while the inconsistent cells were mainly found in the internal subtypes of monocytes or macrophages, especially between Mono_CD14 and Mono_CD14CD16 subtypes (Fig. S5c, S5d, Table S9).

The accuracy of predictions on T cell subtypes was 0.79 in LC (Fig. S5e, S5f, Table S10). The results showed that the inconsistency was mainly between CD8Tn and CD8Teff subtypes. We evaluated the expression of conventional cell markers of each subtype in all clusters, aiming to inspect whether the "inconsistent" results were caused by inaccurate annotations in the original data set, cell inherent similarity or scWizard performance. Therein, 2838 CD8Tn in LC (Table S10) from original data were annotated as CD8Teff by scWizard. After evaluating the expression of canonical markers, the "inconsistent" subcluster was found to express both effector T cell marker gene *KLRG1* and naive T cell marker *TCF7* in LC (Fig. S5g), indicating these clusters may represent an intermediate cell state or gene-expression gradient. The above results show that the inconsistency might be largely due to the subclusters as an intermediate cell state.

3.5. Application of scWizard in the integrative analysis

To show how scWizard is used for the analysis of scRNA-seq data in cancer research, we performed an analysis on the ICC scRNA-seq data set to explore the particular cell subtypes and molecules involved in angiogenesis.

3.5.1. Cell type classification and annotation

We collected cells originating from tumor and tumor-adjacent tissue and performed QC and cell annotation function. According to the results of the annotation module, a total of 8 clusters emerged (Fig. S6a) and the cells were colored based on the expression of marker genes for each major cells for verifying the accuracy of annotation (Fig. S7). Few cells were named Unknown cells and discarded in the downstream analysis. Subsequently, the gene expression of subtypes and their corresponding marker of fibroblasts (Fig. S6c), myeloid cells (Fig. S8a), T/NK cells (Fig. S8b) and endothelial cells (Fig. S8c) were obtained by scWizard.

3.5.2. Cell malignancy estimation

Using epithelial and non-epithelial cell clusters, we identified copy number alterations (CNAs) for each sample by InferCNV. An InferCNV clustered heatmap was created, which corresponded to the normalized expression values of normal cells plotted in the top panel and tumor epithelial cells in the bottom panel (Fig. S6d). We found the copy-number alterations in epithelial cancer cells exceeded non-malignant cells including endothelial cells and fibroblasts, which were consistent with ICC tumors divided from epithelial origin. In the resultant CNA heatmap, the regions of gain were depicted in red and regions of loss in blue.

Fig. 1. Overview of scWizard framework. The functionality of scWizard can be divided into three modules including Input Data, Single-Cell Analysis and Visualization. Single-Cell Analysis module include quality control, cell annotation, cell clustering, batch processing, difference analysis, gene set variation analysis, cell receptor ligand interaction analysis, pseudotime analysis, single-cell transcription factor analysis and copy number variation analysis.

3.5.3. Evaluation of angiogenesis

As angiogenesis is a hallmark of cancer, it is important to address the molecular underpinnings of angiogenesis in ICC based on the advantages of single cells. We calculated the GSVA score of angiogenesis across all cells and found score was higher in cancer samples compared with tumor-adjacent tissues, suggesting increased disordered angiogenesis in tumors (Fig. S9a). Besides, we evaluated and filtered out the cell subtypes with higher score and higher angiogenesis-related genes expression (Fig. S6b) (Fig. S9b-S9g), indicating these cell subtypes were potentially responsible for angiogenesis. Particularly, the specific subclusters harbored the highest score among each cell types. For example, Macro_NLRP3 and Mono_CD14 were exhibited the higher score in myeloid cells, Arteries-like cells were shown the highest AN score among endothelial cells, while TGF–myCAF had the lower score in fibroblasts. Besides, *VEGFA*, *MMP9* and *IL8* were the factors associated with angiogenesis and their gene expression were also evaluated and higher in Macro_NLRP3 as well as Mono_CD14, which showed the results as the same with the angiogenesis GSVA.

3.5.4. Pseudotime ordering of endothelial cells, myeloid cells and CD8 +T cells

In order to explore the evolutionary trajectories of different cell subclusters, we used scWizard to preform trajectory analysis and found arteries and activated pcv cells might be divided from capillary activated cells which was mainly in tumor adjacent tissues (Fig. S10a-S10c). For myeloid cells, Macro_NLRP3 cells were mainly in cancer samples and potentially divided from Mono_CD14 (Fig. S10d-S10f) and cDC1_CLEC9A cells were mainly in terminal (Fig. S11a-S11c). CD8+T cell analysis revealed that pseudotime began with CD8+Tn and CD8+Tcm, followed by CD8+Teff, and ended with CD8+Pro cells (Fig. S11d-S11f). Combined with above

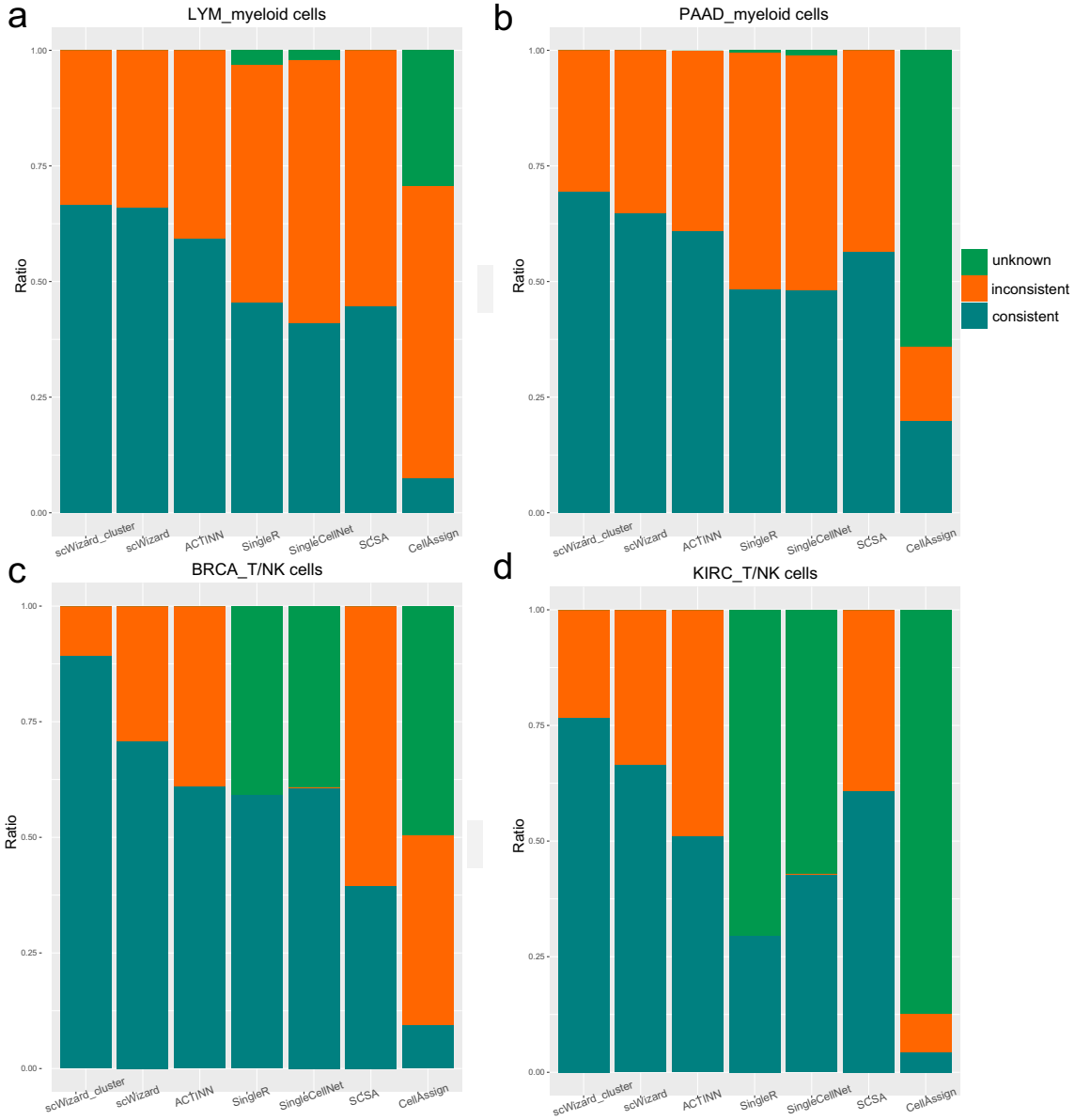

Fig. 2. Comparison scWizard against five methods on cell subtype classification and annotation of the myeloid cells and T/NK cell datasets. The proportion of cells assigned "consistent", "inconsistent", and "unknown" by scWizard, ACTINN, SingleCellNet, SingleR, SCSA, and CellAssign on myeloid cells from lymphoma (LYM) (a), myeloid cells from pancreatic cancer (PAAD) (b), T/NK cells from breast cancer (BRCA) (c) and T/NK cells from kidney renal clear cell carcinoma (KIRC) (d).

results, cells involved in disordered angiogenesis were tumor-specific cell subclusters.

3.5.5. Ligand-receptor interaction between cell subclusters within different cell types

As angiogenesis is the result of the interaction of multiple cells including endothelial cells with macrophages [26] as well as fibroblasts [27], we conducted the cell–cell interaction between these cell subclusters. The results showed that tumor endothelial cells would receive angiogenic stimulatory signals from macrophages and fibroblasts through *VEGFA/VEGFB* and its receptor *FLT1* and *KDR*, as the key mediators to activate the angiogenesis program (Fig. S10g, S10h). These results revealed that the communication with other cells was crucial for endothelial cells to promote tumor angiogenesis. Besides, transcription factor activity for endothelial cells was also analyzed (Fig. S11g).

4. Discussion

Here, we developed scWizard, a user-friendly tool for automated analysis of cancer single cell data, establishing cell classification and annotation model for cells within the TME, integrating existing methods to construct an automatic analysis process, aiming to shorten the time of single cell analysis process and improve the efficiency of researchers. In addition to highlighting the easy-operating on routine analysis of single cell study, scWizard also focuses on providing analytical thinking that reveling the specific cells and cell states contributing to hallmarks of cancer, which is expected to greatly improve our understanding of the diversity and complexity of tumor-derived cells. Following the user manual, users can easily upload data to the scWizard and start the analysis in the user-friendly interface, and visualize each step of the analysis results according to their own needs. We believe that scWizard is a useful software package for cancer subtype annotation and sin-

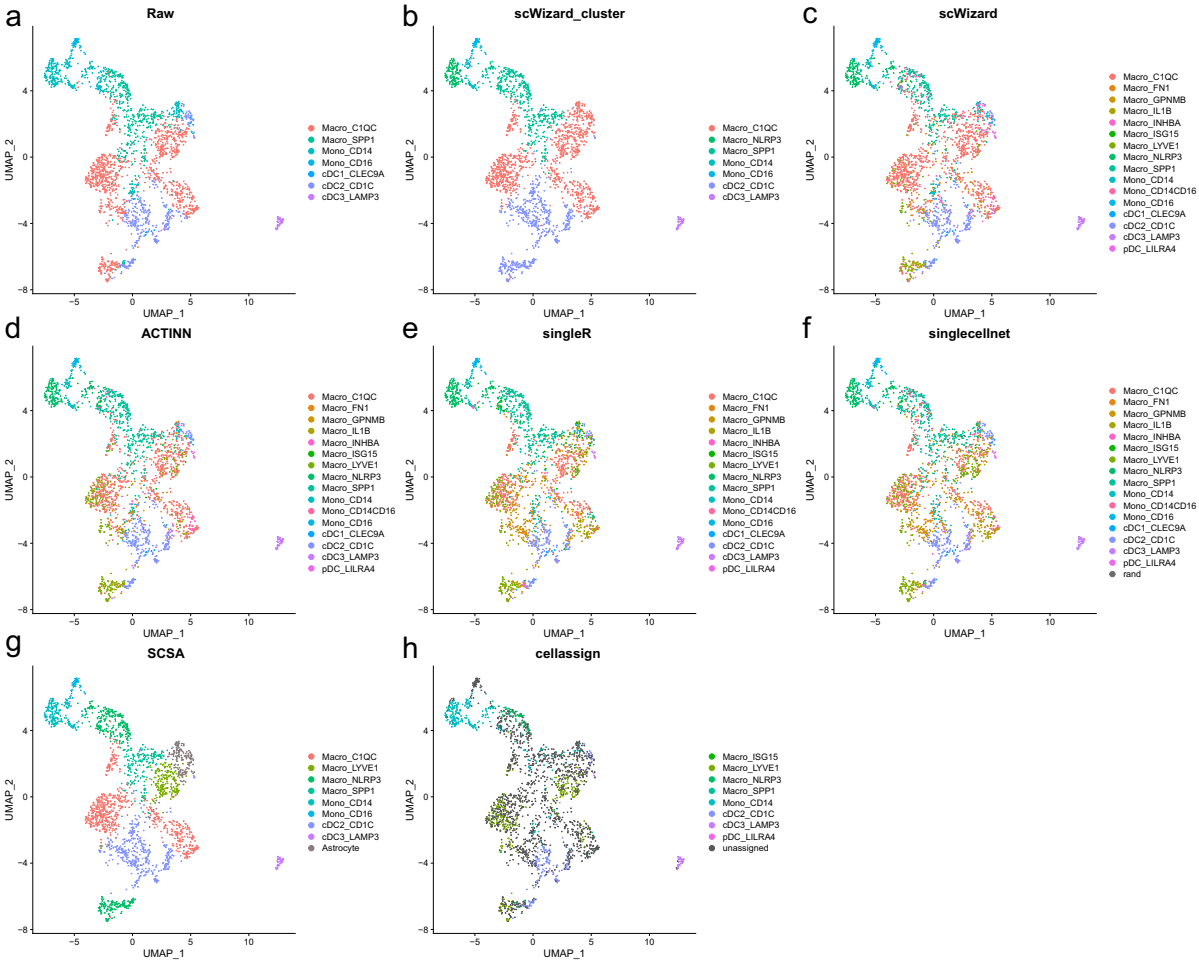

Fig. 3. Myeloid cell subtypes from pancreatic cancer (PAAD) predicted by six tools. (a) Original annotation. (b) Annotation by scWizard_cluster. (c) Annotation by scWizard. (d) Annotation by ACTINN. (e) Annotation by SingleR. (f) Annotation by SingleCellNet. (g) Annotation by SCSA. (h) Annotation by CellAssign. Conflict of interest.

gle cell downstream analysis, which will facilitate us to reveal the biological discovery and understand the complexity of cancer.

Collectively, scWizard outperforms the five existing tools in annotating T/NK cell and myeloid cell subtypes from different tumor tissues. Due to the biology complexity of multicellular systems and the heterogeneity of tumors, clustering subtypes and states is relatively difficult [28]. The strength of scWizard is that it uses hierarchical classification model for accurate cell type identification and avoiding misclassification. The major breakthrough of scWizard, as compared with most other methods, is that it provides avenues for annotating comprehensive cell types and subtypes within the TME, which is important because stromal cells [29] and immune cells [30] in the TME also play a key role in tumor development, but the biology of these cell subtypes is still incompletely revealed at the single cell resolution due to the tough annotation task in cancer studies [31]. Specifically, there are 47 subtypes spanning subtypes of T cells, myeloid cells, endothelial cells and fibroblasts within the TME in our repository and scWizard is prepared to generalize to other solid cancer scRNA-seq dataset for cell annotating and downstream analysis.

Currently, the package mainly focuses on universal cell types within the TME. For tissue specific cell types, scWizard may not be very applicable, such as the alveolar cells in lung cancers may not be identified and named in Unknown cluster. As the quick accumulation of more single-cell data of different cell types, scWizard will be updated and improved to identify more cell types including tissue specific subtypes and cellular states in the future to solve cell heterogeneity and more cancer-specific problems.

Data availability

All the datasets used in this paper are publicly available and applied in the supplementary file. The downloaded accessible number, web link and corresponding references are presented in Table S1. scWizard is available as an R package on GitHub (https://github.com/Dulab2020/scWizard). The detailed documentation on how to acquire, install, and run the software are provided in the Supplementary File: User Manual. To facilitate the first usage of scWizard, the example_data.rds file of testing set including 10,000 single cells of various cell types is randomly chosen and uploaded in Figshare (https://figshare.com/s/8f568e156d943754915e), users can selectively download and quickly get started with scWizard.

Funding

This work has been supported by the National Key R&D Program of China (2018YFC0910201), the Key R&D Program of Guangdong Province (2019B020226001) and China Postdoctoral Science Foundation (2021M701253).

Author contributions

H.D. and J.W. designed and concepted the study. J.W. and Q.X. collected the data and led the data analysis. Q.X. developed the method, analyzed and implemented the Shiny app. J.W. interpreted the results and wrote the manuscript. Y.Q. contributed to preliminary background research survey and prepared supplemental files. G.H. implemented the R package. Z.C. evaluated the method. H.D. conceived of the project, supervised it and revised the manuscript. All authors read, revised and approved the final version of the manuscript. J.W. Resource, Methodology, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.08.028.

References

[1] Hinshaw DC, Shevde LA. The tumor microenvironment innately modulates cancer progression. Cancer Res 2019;79(18):4557–66.
[2] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36(5):411–20.
[3] Guo W, Wang D, Wang S, Shan Y, Liu C, Gu J. scCancer: a package for automated processing of single-cell RNA-seq data in cancer. Brief Bioinform 2021;22(3).
[4] David F, Litovchenko M, Deplancke B, Gardeux V. ASAP 2020 update: an open, scalable and interactive web-based portal for (single-cell) omics analyses. Nucleic Acids Res 2020;48(W1):W403–14.
[5] Sinha D, Kumar A, Kumar H, Bandyopadhyay S, Sengupta D. dropClust: efficient clustering of ultra-large scRNA-seq data. Nucleic Acids Res 2018;46(6):e36.
[6] Patel MV. iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. Bioinformatics 2018;34(24):4305–6.
[7] Hillje R, Pelicci PG, Luzi L. Cerebro: interactive visualization of scRNA-seq data. Bioinformatics 2020;36(7):2311–3.
[8] Senabouth A, Lukowski SW, Hernandez JA, et al. ascend: R package for analysis of single-cell RNA-seq data. GigaScience 2019;8(8).
[9] Zhang AW, O, Flanagan C, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat Methods. 2019. 16(10): 1007-1015.
[10] Cao Y, Wang X, Peng G. SCSA: A cell type annotation tool for single-cell RNA-seq data. Front Genet 2020;11:490.
[11] Zhang Z, Luo D, Zhong X, et al. SCINA: A semi-supervised subtyping algorithm of single cells and bulk samples. Genes (Basel) 2019;10(7).
[12] Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods 2018;15(5):359–62.
[13] Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 2019;20(2):163–72.
[14] de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege F. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res 2019;47(16):e95.
[15] Tan Y, Cahan P. SingleCellNet: A computational tool to classify single cell RNA-seq data across platforms and across species. Cell Syst 2019;9(2):207–213.e2.
[16] Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics 2020;36(2):533–8.
[17] Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol 2019;20(1):264.
[18] Abdelaal T, Michielsen L, Cats D, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 2019;20(1):194.
[19] Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 2019;16(12):1289–96.
[20] Tran H, Ang KS, Chevrier M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol 2020;21(1):12.
[21] Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014;344 (6190):1396–401.
[22] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27 (12):1739–40.
[23] Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. Cell PhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc 2020;15(4):1484–506.
[24] Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32(4):381–6.
[25] Aibar S, González-Blas CB, Moerman T, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods 2017;14(11):1083–6.
[26] Fu LQ, Du WL, Cai MH, Yao JY, Zhao YY, Mou XZ. The roles of tumor-associated macrophages in tumor angiogenesis and metastasis. Cell Immunol 2020;353:104119.
[27] Unterleuthner D, Neuhold P, Schwarz K, et al. Cancer-associated fibroblast-derived WNT2 increases tumor angiogenesis in colon cancer. Angiogenesis 2020;23(2):159–77.
[28] Wahl GM, Spike BT. Cell state plasticity, stem cells, EMT, and the generation of intra-tumoral heterogeneity. NPJ Breast Cancer 2017;3:14.
[29] Chen Y, McAndrews KM, Kalluri R. Clinical and therapeutic relevance of cancer-associated fibroblasts. Nat Rev Clin Oncol 2021.
[30] Philip M, Schietinger A. CD8(+) T cell differentiation and dysfunction in cancer. Nat Rev Immunol 2021.
[31] Clarke ZA, Andrews TS, Atif J, et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat Protoc 2021;16(6):2749–64.