scAnnoX: An R Package Integrating Multiple Public Tools for Single-Cell Annotation

by Xiaoqian Huang

Submission date: 19-Dec-2023 05:30PM (UTC+0500)
Submission ID: 2262523290
File name: peerj-93576-scAnnoX.docx (78.47K)
Word count: 6056
Character count: 36814

scAnnoX: An R Package Integrating Multiple Public Tools for Single-Cell Annotation

Xiaoqian Huang[1], Ruiqi Liu[1], Shiwei Yang[1], Xiaozhou Chen[1, *] and Huamei Li[2,*]

[1] School of Mathematics and Computer Science, Yunnan Minzu University, Kunming, Yunnan Province, P. R. China

[2] Department of Hepatobiliary Surgery, the Affiliated Drum Tower Hospital, Medical School, Nanjing University, Nanjing, Jiangsu Province, P. R. China

Corresponding Author:
Xiaozhou Chen[1, *]
Yuehua Street, Kunming, Yunnan Province, 650504, P. R. China
Email address: ch_xiaozhou@163.com
Huamei Li[2, *]
Yuehua Street, Kunming, Yunnan Province, 650504, P. R. China
Email address: ch_xiaozhou@163.com

Abstract

Background. In recent years, single-cell RNA sequencing technology has stood out and developed rapidly. Enabling researchers to gain a more comprehensive understanding of the properties and functions of individual cells. Therefore, the accurate description and classification of single cell identity has become an important and formidable challenge.

Methods. This study meticulously investigates ten widely adopted algorithms designed for the identification of cell identities within single-cell RNA sequencing data. This distinguished set of algorithms encompasses SingleR, Seurat, sciBet, scmap, CHETAH, scSorter, sc.type, cellID, scCATCH, SCINA. Leveraging these ten algorithms as a foundation, an R package, christened "scAnnoX", has been meticulously crafted. Its purpose is to harmoniously integrate these disparate algorithms for cell identity identification in single-cell RNA sequencing data, providing a cohesive framework that greatly facilitates comparative analyses among them.

Results. The overarching objective of this endeavor is to empower researchers in their pursuit of more efficient analyses of single-cell RNA sequencing data. This, in turn, equips them with the knowledge needed to make informed decisions within the intricate landscape of single-cell identity identification algorithms. The integrated environment of "scAnnoX" simplifies the processes of testing, evaluation, and comparison among a variety of algorithms. Interested parties can access the "scAnnoX" package at https://github.com/XQ-hub/scAnnoX.

Introduction

In the realm of single-cell omics research, the evolution of single-cell sequencing technology has afforded us a profound insight into the gene expression profiles and functional roles of distinct cell types within diverse biological organisms (Balzer et al. 2021; Kolodziejczyk et al. 2015; Rossin et al. 2021; Slovin et al. 2021). Single-cell identity recognition algorithms equip researchers with the means to accurately ascertain and categorize the identities of individual cells (Brendel et al. 2022; Kim et al. 2020), contributing to the identification of potential disease biomarkers or aberrant cell types (Bod et al. 2023; Hickey et al. 2023). This has a paramount bearing on the early diagnosis and treatment of a range of diseases, including cancer, immune system disorders, and neurological conditions (Chen et al. 2023; Fu et al. 2021; Wang et al. 2022). Consequently, single-cell identity recognition algorithms occupy a pivotal position in con-temporary biomedical research. Furthermore, as researchers increasingly focus on this field, a multitude of algorithms is at their disposal for selection.

Unquestionably, these algorithms autonomously annotate individual cells based on their gene expression profiles. One approach involves the annotation predicated on marker genes associated with cell types and the scoring of the presence of these marker genes within cell clusters (Pasquini et al. 2021). The second method necessitates a reference dataset containing information about cell types to compute the similarity between the expression profiles of query genes and the reference dataset. This calculation yields a similarity score between the reference and query datasets, facilitating the identification of optimal correlations between them. A recent and noteworthy approach involves the integration of machine learning techniques with single-cell identity recognition algorithms. The most frequently employed method within this category is supervised learning, which entails the training of a classifier using labeled references. Nonetheless, the selection of the appropriate algorithm, data preprocessing, model tuning, and related tasks often demand a substantial investment of time and effort. Consequently, determining the most suitable algorithm for a specific research objective is frequently a challenging undertaking. Each algorithm possesses distinctive applicability and constraints, necessitating an in-depth understanding of their intricacies to make informed choices. The process of narrowing down the selection from among numerous algorithms is labor-intensive and time-consuming. Thus, comprehending and addressing this challenge is of paramount importance.

In this context, the present study has developed an R package known as "scAnnoX" that amalgamates 10 distinct single-cell RNA sequencing data cell identity recognition algorithms into a unified framework, facilitating comparative analysis. The overarching goal is to assist researchers in efficiently analyzing scRNA-seq data, offering targeted guidance to make judicious decisions in the intricate selection of single-cell identity recognition algorithms, and simplifying the process of testing, evaluating, and comparing various algorithms within an integrated environment.

Researchers have substantiated the efficacy and stability of this R package through extensive testing on multiple authentic datasets. The development of this tool is poised to expedite the analysis of single-cell RNA sequencing data, granting researchers greater convenience and flexibility in exploring the intricacies of cell types and gene expression. This endeavor holds profound significance for the progression of the field of single-cell biology and has the potential to deepen our comprehension of cellular diversity and function.

Materials & Methods

This research endeavor has yielded an R package, denoted as "scAnnoX", designed to comprehensively amalgamate ten distinct algorithms for single-cell RNA sequencing data cell identity recognition. These algorithms encompass SingleR (Aran et al. 2019), Seurat (Hao et al. 2023), sciBet (Li et al. 2020), scmap (Kiselev et al. 2018), CHETAH (de Kanter et al. 2019), scSorter (Guo & Li 2021), sc.type (Ianevski et al. 2022), cellID (Cortal et al. 2021), scCATCH (Shao et al. 2020), SCINA (Zhang et al. 2019). The package further serves the purpose of facilitating comparative analyses among these algorithms. In each instance, source code packages were diligently installed, or scripts meticulously sourced from GitHub repositories. Evaluating the performance of 10 single-cell identity recognition algorithms is a multifaceted endeavor, necessitating the establishment of clearly defined methodologies and the implementation of a rigorous set of standardized experimental procedures.

Data Preprocessing

In the initial phase, the primary undertaking involves the ingestion of raw single-cell RNA sequencing data, followed by data refinement, feature extraction, and the establishment of a Seurat object to serve as a repository for this dataset. Subsequently, the application of the "NormalizeData" function, accessible through the Seurat package, normalizes the data while allowing the specification of a normalization method, typically employing logarithmic normalization. Lastly, the data undergoes Principal Component Analysis (PCA) dimensionality reduction via the utilization of the "RunPCA" function, also sourced from the Seurat package. PCA stands as a widely acknowledged dimensionality reduction technique, aimed at curtailing data dimensionality while capturing salient variations within the dataset. This process aids in facilitating a more profound comprehension and visualization of the similarities and disparities between individual cells. Data preprocessing constitutes a pivotal phase in the analysis of single-cell RNA sequencing data, with the requisite conversion of data into the Seurat format being an indispensable prerequisite, effectively executed through the implementation of the "scAnnoX" package.

Algorithm Selection

This study delves into an array of 10 prominent algorithms for identifying cell types within single-cell RNA sequencing data. These algorithms encompass diverse methodologies and features. Among these, certain tools rely on the annotation of marker genes associated with specific cell types, such as scCATCH, scSorter, SCINA and sc.type. Others leverage information derived from reference cell type datasets, exemplified by SingleR, Scmap and CHETAH. Further, certain tools are designed to train classifiers utilizing machine learning techniques, as exemplified by sciBet. The package also includes non-clustered multivariate statistical methods such as cellID, an automated tool for annotation of cellular heterogeneity based on single-cell clusters, and the Seurat method tailored for the analysis of single-cell RNA sequencing data.

Algorithm Integration

This task involves integrating and optimizing 10 different single-cell RNA sequencing data cell identity recognition algorithms. Each of these algorithms has unique strengths and applications, so cleverly combining them will provide researchers with a broader range of choices and more powerful tools. This effort aims to enhance the diversity of data processing, thereby improving the feasibility of research. To optimize algorithm integration, we need to delve into the performance and characteristics of these different algorithms and find the best way to integrate them to ensure they can work together, considering data quality and characteristics.

In the "scAnnoX" package, there is a function called "autoAnnoResult", which is used to aggregate and summarize the predictions of the 10 different algorithms. After the aggregation, the frequency (N_{pred}) of each prediction for the same sample is calculated and expressed as a ratio to the total number of methods (N_{tools}), yielding the frequency of each prediction. The result of the "autoAnnoResult" function is the prediction with the highest frequency and is determined by the formula:

$$argmax\left(p = \frac{N_{pred}}{N_{tools}}\right)$$

This result serves as the final prediction in the "scAnnoX" package, effectively integrating multiple algorithms. Through this approach, researchers will be able to analyze and interpret single-cell RNA sequencing data, providing them with more powerful tools and a wider range of choices for scientific research more effectively. In summary, by optimizing algorithm integration, we can better leverage the strengths of different algorithms, improve the efficiency and accuracy of data processing, and advance research. This will contribute to strengthening the analysis of single-cell RNA sequencing data, offering broader possibilities and more insights for various research endeavors.

Experimental Validation

Datasets originating from diverse organizational sources and various data platforms were partitioned into test and reference sets in a 6:4 ratio. The test set served the purpose of evaluating the algorithm's performance, while the reference set was employed for model training or served as a performance benchmark. Leveraging the scAnnoX package, we conducted data annotation and validation, leveraging a suite of functions for assessing the precision and consistency of single-cell RNA sequencing data. We scrutinized the alignment of their predictions on the test set against the ground truth labels within the reference set. Diverse performance metrics were employed to gauge the accuracy and reliability of the algorithm, facilitating the selection of the most appropriate algorithm to fulfill research requirements. This method assists in determining which algorithm excels in the validation of multi-source data.

Performance Assessment

Performance metrics serve as measurement standards employed to appraise the efficacy of models, algorithms, or systems within the context of specific tasks. In this context, we present a pivotal performance metric, accuracy. Accuracy stands as a ubiquitous metric utilized to assess the effectiveness of classification models or algorithms. It gauges the ratio of correctly predicted samples by the model in relation to the overall sample count. The formula for calculating accuracy is succinctly expressed as follows:

$$acc = \frac{N_{pred=ActureAnno}}{N}$$

where $N_{pred=ActureAnno}$ signifies the count of samples for which the model's or algorithm's predictions align with the authentic labels, while N denotes the aggregate sample count.

Root Mean Square Error of Prediction Performance

The root mean square error (RMSE) is a statistical metric that quantifies the disparity between predicted and actual values. It is calculated as the square root of the mean of the squared differences between predicted and actual values, divided by the total number of observations. RMSE is particularly sensitive to atypical data points, often referred to as outliers, making it a valuable tool for assessing the overall accuracy and robustness of predictive models. The formula for RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(true_i - pred_i)^2}$$

In this equation, N denotes the total number of experiments conducted, n signifies the count of predicted samples, $true_i$ represents the true value for the sample, and $pred_i$ is indicative of the predicted value for the same sample.

Results

R Package Development for Single-cell RNA Sequencing Data Annotation

Utilizing the R programming language, we have successfully engineered an R package called "scAnnoX". This meticulously crafted package integrates a comprehensive suite of 10 distinct annotation algorithms, as previously elucidated. Each of these algorithms exhibits unique applicability and inherent limitations. To facilitate users in selecting the most appropriate algorithm tailored to their specific research needs, we have thoughtfully designed comprehensive user instructions for scAnnoX. These user-friendly instructions ensure effortless accessibility and operation of all the integrated annotation algorithms. Furthermore, we have painstakingly finetuned and optimized this R package to guarantee not only its stability but also its efficiency. Our implementation adheres to the highest standards of programming practices, ensuring the long-term maintainability and extensibility of the package. This equips scAnnoX to seamlessly adapt to evolving requirements and readily accommodate the integration of new annotation algorithms. The specific architecture of the scAnnoX package is shown in Figure 1.

In summation, the scAnnoX package offers users the capability to effortlessly harness the power of 10 diverse annotation algorithms, empowering them to attain their data analysis objectives without the need for arduous investments of time and effort. The development of this R package is rooted in the objective of streamlining and enhancing data analysis processes, ultimately fostering greater convenience and efficiency.

Usage of the scAnnoX Package

The utilization of the scAnnoX package involves the organization and transformation of raw data to comply with the requirements of the Seurat data format. This ensures that the dataset's columns include gene expression values and cell identity information. The data undergo preprocessing steps such as standardization and dimension reduction. Subsequently, the testing dataset is annotated using the "autoAnnoTools" function provided within the package. The essential parameters for the "autoAnnoTools" function include the pre-processed testing dataset, the name of the single-cell annotation tool (method), and the type of single-cell annotation tool (strategy). Optional parameters encompass the reference dataset, reference cell types, and marker gene information, with their default values set to NULL. The available values for the method parameter correspond to 10 different single-cell identity recognition algorithms, namely: SingleR, Seurat, sciBet, scmap, CHETAH, scSorter, sc.type, cellID, scCATCH, SCINA. The strategy for single-cell annotation tools can be classified into two types: marker-based or reference-based. We consider scSorter, sc.type, cellID, scCATCH, SCINA as marker-based tools (see Materials and Methods).

The necessity of optional parameters depends on the value of the method. If the method is a marker-based algorithm, marker gene information needs to be provided. If the method is reference-based, both the reference dataset and reference cell types need to be provided.

The output of the "autoAnnoTools" function is the annotation results of the chosen single-cell identity recognition algorithm for the samples within the testing dataset. These annotation results assist in determining the single-cell identity of each sample. Usage examples of the scAnnoX package can be found at https://github.com/XQ-hub/scAnnoX/vignettes/example.R, with the code and output results provided therein. This exemplar serves as a valuable reference for understanding the practical implementation of the package in your research.

Annotation for Accuracy Assessment of Internal Datasets

To substantiate and compare the precision of ten annotation tools, datasets emanating from diverse tissue origins and distinct data acquisition platforms were partitioned into experimental test sets and reference sets, maintaining a 6:4 ratio. Comprehensive scrutiny was undertaken to rigorously assess the efficacy of these computational algorithms within the confines of the given datasets. In this study, we conducted a comprehensive evaluation of algorithmic performance using the scAnnoX software package on four distinct single-cell omics datasets. Specifically, our analysis centered around the human islet cells dataset by Xin Y et al. (Xin et al. 2016), where the scSorter and SCINA algorithms exhibited exceptional capabilities, achieving outstanding classification accuracy of 99.69% (Figure. 2A). Furthermore, we extended our assessment to include the human liver tissue dataset by Camp JG et al. (Camp et al. 2017) and the human brain transcriptome dataset by Darmanis S et al. (Darmanis et al. 2015), where the sciBet algorithm demonstrated remarkable performance with classification accuracies of 98.43% and 87.83%, respectively (Figures. 2B, C). In the case of the human liver tissue dataset, the sc.type algorithm also achieved a classification accuracy comparable to sciBet. Of particular significance was the performance of the SingleR algorithm, which achieved an impressive accuracy of 88.89% in classifying cell types within the human brain transcriptome dataset and an exceptional accuracy of 96.17% in the adult mouse cortical cell dataset by Tasic B et al. (Figure. 2D) (Tasic et al. 2016). In contrast, the performance of the cellID was comparatively subdued, demonstrating an accuracy of 61.78% in the human liver tissue dataset and a mere 12.91% accuracy in the human pancreatic islet cell tissue dataset. Furthermore, it is noteworthy that the performance of scmap and scCATCH, while competitive in certain contexts, exhibits considerable variability and susceptibility to the characteristics of diverse datasets.

Based on the evaluations, we utilized integrated results obtained through the built-in functionalities of autoAnnoTools within the scAnnoX software package. As exemplified with human islet cells and human liver tissue datasets, we conducted two-dimensional visualizations of original cell types, scAnnoX package predicted cell types, and those predicted by one of the algorithms (Figures. 2E, F). The Uniform Manifold Approximation and Projection (UMAP) visualization demonstrated remarkable stability and robust performance within the integrated results.

Precision Assessment of Cross-Platform Datasets

The diversity of scRNA-seq techniques offers a valuable opportunity for cross-platform validation of datasets derived from the same biological tissue. To substantiate this assertion, we conducted a precision assessment experiment on cross-platform datasets using two independent and well-sequenced datasets originating from different sequencing platforms. The primary objective of this study was to evaluate the performance of the scAnnoX package comprehensively and systematically. Two distinct sets of datasets were subjected to validation in this experiment, one sourced from pancreatic tissue, as reported by Xin Y et al and Lawlor N et al (Lawlor et al. 2017), and the other from thymic tissue, as reported by Yasumizu Y et al (Yasumizu et al. 2022) and Park JE et al (Park et al. 2020). Each set of datasets, obtained from different platforms, underwent random subsampling, designating one dataset as the reference dataset and the other as the test dataset. We compared the annotation accuracy of ten annotation algorithms embedded within the scAnnoX package and subsequently derived an integrated annotation accuracy metric.

In the context of a cross-platform pancreatic tissue dataset, we utilized the dataset curated by Xin Y et al. as a reference training set and Lawlor N et al.'s dataset as the testing set, focusing on four common cell types shared between the two datasets (Figures. 3A, B). Our objective was to evaluate the performance of various computational tools in identifying and annotating the cell types in the test dataset. The results of this validation exercise clearly demonstrate the robustness of most tools in accurately characterizing and annotating the test dataset. Specifically, SingleR, sciBet, scSorter, and SCINA exhibited a remarkable predictive accuracy of 99%, while Seurat achieved an accuracy of 96.59%. It is noteworthy that SingleR displayed suboptimal performance in the identification of pancreatic polypeptide-secreting cells (PP) and delta cell types, whereas Seurat exhibited shortcomings in recognizing delta cell types (Figure. 3C). Further analysis revealed that the challenges in distinguishing these cell types can be attributed to their relatively low cell counts, especially the scarcity of pancreatic polypeptide secreting cells within the islet (Figure. 3A). Notably, scAnnoX, leveraging integrated annotations, emerged as the top performing result with an impressive accuracy of 99.69%. This exceptional performance is most striking in its perfect prediction accuracy of 100% for alpha, beta, and delta cell types, surpassing the performance of other algorithms (Figure. 3C).

In the context of a multi-platform thymic tissue dataset, we assessed the reference dataset by Park JE et al., using it as the baseline for validation against the dataset pro-vided by Yasumizu Y et al. Given the high heterogeneity in cell types and the limited sample sizes within certain cell type categories, we performed a comprehensive re-classification and aggregation of cell types within the dataset. Specifically, we have amalgamated subtypes such as mTEC(I), mTEC(II), mTEC(III), and mTEC(IV) into a unified category referred to as "mTEC" while consolidating subtypes including DC1, DC2, and aDC into a category denoted as "DC". Subsequently, we determined the predictive accuracy for each cell type (Figure. 4A). Following the data preprocessing steps, we proceeded to evaluate the annotation performance of various computational algorithms. It is imperative to note that due to the intricate nature of cell types and the potential confounding effects of batch processing, the validation results were not entirely satisfactory. The accuracy of most intrinsic methods tended to con-verge within the range of 45% to 66% (Figure. 4B). Notably, scAnnoX, after integration, achieved an accuracy of 67.2%. However, it is worth mentioning that the misclassification of cell types predominantly centered around the fine-grained subtyping of B cells and T cells (Figures. 4C, D).

This investigation underscores the complexities inherent in single-cell omics data analysis, particularly in the context of intricate cell type distinctions, and highlights the significance of algorithmic enhancements to bolster the accuracy of cell type annotations.

Stability Assessment of Annotations for scAnnoX

In the context of the experiments, we have successfully achieved a high predictive accuracy for the integrated results. Furthermore, we conducted a comprehensive analysis to assess the robustness and reliability of these integrated findings.

We conducted a comprehensive integration and synthesis of all the experiments, generating integrated predictions for each one. By leveraging the built-in functionalities of the scAnnoX software package, we obtained integration results for each experiment. In comparison to various algorithms, scAnnoX consistently exhibited outstanding predictive performance, maintaining a consistently high level of accuracy, as depicted in Figure 5A and 5B. To further assess the robustness and flexibility of scAnnoX's integrated results, we calculated the root mean square error be-tween predictive performance and actual cell types (Figure. 5C). This evaluation unequivocally demonstrates the stability and resilience of the integration results provided by the scAnnoX software package. Additionally, our study underscores the significant capability of integration results in mitigating the adverse effects of data sparsity and batch effects relative to single algorithms. This enhanced robustness and exceptional performance further underscore the stability and reliability of our approach.

Comparative Analysis of Computational Runtime

Building upon experiments validating our in-house dataset, our study undertook a comprehensive analysis that unveiled profound disparities among ten distinct single-cell identity recognition algorithms concerning their computational execution times. This investigation underscores the significance of our work in shedding light on the temporal dynamics of these algorithms, a crucial dimension in the ever-evolving landscape of single-cell omics research.

In the pancreatic islet cell dataset, we grappled with an extensive volume of data, encompassing 38,008 genes and 1,809 samples. Notably, sc.type and the SCINA method exhibited exceptional efficiency in this regard, completing the analysis within 0.30 and 0.55 seconds, respectively (Figure. 6A). In fact, they boasted the shortest processing times among the ten algorithms we evaluated, an achievement that merits strong emphasis. Conversely, scSorter and cellID necessitated relatively longer durations to fulfill the task. In the liver and brain tissue datasets, featuring 465 and 466 samples respectively and approximately twenty thousand genes, sc.type and SCINA continued to deliver outstanding performance, with execution times remaining under 0.6 seconds, and even dipping to 0.3 seconds in the case of the hepatic tissue dataset (Figures. 6B, C). In the mouse cortical cell dataset, encompassing 1,600 and 1,809 samples, and harboring complex cell types, sc.type still managed to provide predictions within 0.5 seconds, while scCATCH reached 221.75 seconds (Figure. 6D). In summation of the time assessments from these experiments, it is evident that sc.type and SCINA consistently exhibit highly favorable performance, whereas scSorter and cellID require relatively longer durations to complete their tasks. scCATCH demonstrates an increase in runtime when faced with datasets featuring complex cell types.

The experimental analysis results regarding the running times of various algorithms across different datasets reveal a noteworthy trend: a substantial increase in sample size or data complexity corresponds to an increment in time consumption. To be more specific about the runtime implications, certain algorithms exhibit substantial variations, while others remain relatively stable.

Initially, our observations demonstrate a significant augmentation in the running times of the CellID, sciBet, scmap, Seurat, and SingleR algorithms in response to enlarged sample sizes. This phenomenon is attributed to their necessity to process an in-creased number of data points and undertake more computationally demanding tasks. On the contrary, the scCATCH algorithm displays an atypical behavior as sample sizes expand. In the comparative analysis between the human islet cell tissue dataset and the human liver tissue dataset, the running time of scCATCH decreases with the enlargement of sample size. In contrast, certain algorithms, such as sc.type and SCINA, appear to be less influenced by variations in sample size. This observation underscores the superior stability and efficiency of these algorithms in handling extensive datasets.

In summation, these findings illuminate the varying performances of distinct algorithms when confronted with different sample sizes. Researchers should be mindful of the sample size's influence when choosing an algorithm, ensuring it aligns with the research requirements and can execute the analysis within a reasonable timeframe.

Discussion

With the advancements in single-cell RNA sequencing technologies, a plethora of single-cell annotation algorithms has emerged. Given the distinct data formats, applicability, and limitations associated with each algorithm, researchers face the intricate task of selecting an appropriate algorithm. This necessitates a profound understanding of the algorithmic structure embedded within the source code. Subsequently, data preprocessing, model fine-tuning, and other operations tailored to the input-output formats of each algorithm become imperative, demanding substantial investments of time and effort.

Against this backdrop, this study establishes a comprehensive framework tailored to accommodate ten prominent single-cell annotation algorithms: SingleR, Seurat, sciBet, scmap, CHETAH, scSorter, sc.type, cellID, scCATCH, and SCINA. Within this framework, these algorithms share a standardized data input and output schema. Consequently, researchers can streamline their efforts by conducting a singular round of data preprocessing in adherence to the framework's specified input format. This unified approach facilitates the validation of diverse algorithmic methodologies, significantly alleviating the preparatory workload and time investment for researchers.

This innovative framework not only enhances the efficiency of algorithm selection but also provides a unified platform for the scientific community to benchmark and compare the performance of various single-cell annotation tools. The integration of these algorithms within a standardized framework contributes to a more streamlined and reproducible approach in the realm of single-cell omics research.

In this study, leveraging the devised framework, we have developed an R package termed "scAnnoX". This package seamlessly integrates ten distinct single-cell RNA sequencing data cell identity recognition algorithms into the established framework, facilitating comparative analyses. Additionally, within scAnnoX, a function named "autoAnnoResult" has been implemented. This function serves the purpose of generating the integrated predictions of scAnnoX, which, following validation across diverse datasets, attests to the commendable robust performance of the integrated predictions achieved by scAnnoX.

Researchers, utilizing the scAnnoX package, have the flexibility to select and validate one or more algorithms embedded within the package, enabling comparative analyses across diverse algorithms. Tailoring their investigations to align with specific research objectives, the researchers conducted extensive downstream analyses in this study. Specifically, we validated and compared the runtime performance of ten algorithms, elucidating variations in their execution times. Furthermore, the investigation unveiled temporal fluctuations in algorithmic performance across distinct datasets, facilitating an understanding of the differential impacts of various datasets on algorithmic behavior. This experimental evidence contributes to the assessment of algorithmic robustness and resilience.

The overarching objective of this study is to facilitate the effective analysis of single-cell RNA sequencing data, providing targeted guidance to researchers for making informed decisions within the intricate landscape of single-cell identity recognition algorithms. The research aims to streamline the processes of testing, evaluation, and comparison, offering valuable insights to the scientific community. This work endeavors to empower researchers with the tools needed to navigate the complexities of algorithm selection, ultimately contributing to the simplification of the testing, assessment, and comparative analysis processes in the realm of single-cell RNA sequencing data analysis.

Conclusions

Our study, grounded in the field of single-cell omics, has resulted in the development of an R package named "scAnnoX" by integrating ten distinct single-cell RNA sequencing data identification algorithms, including SingleR, Seurat, sciBet, scmap, CHETAH, scSorter, sc.type, cellID, scCATCH, and SCINA. The primary objective of this software package is to provide a unified framework that alleviates the dilemma faced by researchers when selecting the most suitable single-cell RNA sequencing data identification algorithm for their specific research methods, objectives, and datasets. scAnnoX reduces the time and effort required for data preprocessing and model optimization.

The development of the "scAnnoX" package was driven by a need to gain a deeper understanding of the intricacies of each algorithm and to synthesize a common input schema applicable to all single-cell RNA sequencing data identification algorithms. This allows researchers to obtain experimental results from these ten algorithms by using only the "scAnnoX" package and further improve predictive accuracy through the "autoAnnoResult" function. As a result, researchers can significantly reduce the preparatory workload.

Employing the "scAnnoX" software package, our study conducted a comparative assessment of the accuracy and runtime performance of ten algorithms across diverse datasets. This assessment encompassed both internal validation experiments and cross-platform validation experiments.

The results underscore the pivotal role of "scAnnoX" in providing researchers with a vital decision-making tool, enabling them to make informed selections based on their research objectives. These experimental findings not only validate the significance of algorithm integration and comparison but also offer robust support for researchers to make prudent algorithm choices in specific research scenarios.

Moreover, this research underscores the critical importance of performance evaluations, encompassing accuracy and runtime, in the realm of single-cell RNA sequencing data analysis. This provides a potent tool for analyzing single-cell RNA sequencing data and holds the potential to drive substantial advancements in biomedical research.

Acknowledgements

We thank all the authors involved in this study for data collection, preparation, quality control and manuscript writing. This work was carried out on the High-performance Computing platform of Yunnan Minzu University.

References

Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, and Bhattacharya M. 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 20:163-172. 10.1038/s41590-018-0276-y

Balzer MS, Ma Z, Zhou J, Abedini A, and Susztak K. 2021. How to Get Started with Single Cell RNA Sequencing Data Analysis. *J Am Soc Nephrol* 32:1279-1292. 10.1681/ASN.2020121742

Bod L, Kye YC, Shi J, Torlai Triglia E, Schnell A, Fessler J, Ostrowski SM, Von-Franque MY, Kuchroo JR, Barilla RM, Zaghouani S, Christian E, Delorey TM, Mohib K, Xiao S, Slingerland N, Giuliano CJ, Ashenberg O, Li Z, Rothstein DM, Fisher DE, Rozenblatt-Rosen O, Sharpe AH, Quintana FJ, Apetoh L, Regev A, and Kuchroo VK. 2023. B-cell-specific checkpoint molecules that regulate anti-tumour immunity. *Nature* 619:348-356. 10.1038/s41586-023-06231-0

Brendel M, Su C, Bai Z, Zhang H, Elemento O, and Wang F. 2022. Application of Deep Learning on Single-cell RNA Sequencing Data Analysis: A Review. *Genomics Proteomics Bioinformatics* 20:814-835. 10.1016/j.gpb.2022.11.011

Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, Kanton S, Kageyama J, Damm G, Seehofer D, Belicova L, Bickle M, Barsacchi R, Okuda R, Yoshizawa E, Kimura M, Ayabe H, Taniguchi H, Takebe T, and Treutlein B. 2017. Multilineage communication regulates human liver bud development from pluripotency. *Nature* 546:533-538. 10.1038/nature22796

Chen WJ, Dong KQ, Pan XW, Gan SS, Xu D, Chen JX, Chen WJ, Li WY, Wang YQ, Zhou W, Rini B, and Cui XG. 2023. Single-cell RNA-seq integrated with multi-omics reveals SERPINE2 as a target for metastasis in advanced renal cell carcinoma. *Cell Death Dis* 14:30. 10.1038/s41419-023-05566-w

Cortal A, Martignetti L, Six E, and Rausell A. 2021. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nat Biotechnol* 39:1095-1102. 10.1038/s41587-021-00896-6

Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Hayden Gephart MG, Barres BA, and Quake SR. 2015. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* 112:7285-7290. 10.1073/pnas.1507125112

de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, and Holstege FCP. 2019. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res* 47:e95. 10.1093/nar/gkz543

Fu H, Sun H, Kong H, Lou B, Chen H, Zhou Y, Huang C, Qin L, Shan Y, and Dai S. 2021. Discoveries in Pancreatic Physiology and Disease Biology Using Single-Cell RNA Sequencing. *Front Cell Dev Biol* 9:732776. 10.3389/fcell.2021.732776

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

Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, and Satija R. 2023. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol*. 10.1038/s41587-023-01767-y

Hickey JW, Becker WR, Nevins SA, Horning A, Perez AE, Zhu C, Zhu B, Wei B, Chiu R, Chen DC, Cotter DL, Esplin ED, Weimer AK, Caraccio C, Venkataraaman V, Schurch CM, Black S, Brbic M, Cao K, Chen S, Zhang W, Monte E, Zhang NR, Ma Z, Leskovec J, Zhang Z, Lin S, Longacre T, Plevritis SK, Lin Y, Nolan GP, Greenleaf WJ, and Snyder M. 2023. Organization of the human intestine at single-cell resolution. *Nature* 619:572-584. 10.1038/s41586-023-05915-x

Ianevski A, Giri AK, 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. 10.1038/s41467-022-28803-w

Kim D, Chung KB, and Kim TG. 2020. Application of single-cell RNA sequencing on human skin: Technical evolution and challenges. *J Dermatol Sci* 99:74-81. 10.1016/j.jdermsci.2020.06.002

Kiselev VY, Yiu A, and Hemberg M. 2018. scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* 15:359-362. 10.1038/nmeth.4644

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, and Teichmann SA. 2015. The technology and biology of single-cell RNA sequencing. *Mol Cell* 58:610-620. 10.1016/j.molcel.2015.04.005

Lawlor N, George J, Bolisetty M, Kursawe R, Sun L, Sivakamasundari V, Kycia I, Robson P, and Stitzel ML. 2017. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res* 27:208-222. 10.1101/gr.212720.116

Li C, Liu B, Kang B, Liu Z, Liu Y, Chen C, Ren X, and Zhang Z. 2020. SciBet as a portable and fast single cell type identifier. *Nat Commun* 11:1818. 10.1038/s41467-020-15523-2

Park JE, Botting RA, Dominguez Conde C, Popescu DM, Lavaert M, Kunz DJ, Goh I, Stephenson E, Ragazzini R, Tuck E, Wilbrey-Clark A, Roberts K, Kedlian VR, Ferdinand JR, He X, Webb S, Maunder D, Vandamme N, Mahbubani KT, Polanski K, Mamanova L, Bolt L, Crossland D, de Rita F, Fuller A, Filby A, Reynolds G, Dixon D, Saeb-Parsy K, Lisgo S, Henderson D, Vento-Tormo R, Bayraktar OA, Barker RA, Meyer KB, Saeys Y, Bonfanti P, Behjati S, Clatworthy MR, Taghon T, Haniffa M, and Teichmann SA. 2020. A cell atlas of human thymic development defines T cell repertoire formation. *Science* 367. 10.1126/science.aay3224

Pasquini G, Rojo Arias JE, Schafer P, and Busskamp V. 2021. Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J* 19:961-969. 10.1016/j.csbj.2021.01.015

Rossin EJ, Sobrin L, and Kim LA. 2021. Single-cell RNA sequencing: An overview for the ophthalmologist. *Semin Ophthalmol* 36:191-197. 10.1080/08820538.2021.1889615

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. 10.1016/j.isci.2020.100882

Slovin S, Carissimo A, Panariello F, Grimaldi A, Bouche V, Gambardella G, and Cacchiarelli D. 2021. Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. *Methods Mol Biol* 2284:343-365. 10.1007/978-1-0716-1307-8_19

Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, Bertagnolli D, Goldy J, Shapovalova N, Parry S, Lee C, Smith K, Bernard A, Madisen L, Sunkin SM, Hawrylycz M, Koch C, and Zeng H. 2016. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* 19:335-346. 10.1038/nn.4216

Wang Y, Wang Q, Xu Q, Li J, and Zhao F. 2022. Single-cell RNA sequencing analysis dissected the osteo-immunology microenvironment and revealed key regulators in osteoporosis. *Int Immunopharmacol* 113:109302. 10.1016/j.intimp.2022.109302

Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, Murphy AJ, Yancopoulos GD, Lin C, and Gromada J. 2016. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metab* 24:608-615. 10.1016/j.cmet.2016.08.018

Yasumizu Y, Ohkura N, Murata H, Kinoshita M, Funaki S, Nojima S, Kido K, Kohara M, Motooka D, Okuzaki D, Suganami S, Takeuchi E, Nakamura Y, Takeshima Y, Arai M, Tada S, Okumura M, Morii E, Shintani Y, Sakaguchi S, Okuno T, and Mochizuki H. 2022. Myasthenia gravis-specific aberrant neuromuscular gene expression by medullary thymic epithelial cells in thymoma. *Nat Commun* 13:4230. 10.1038/s41467-022-31951-8

Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, Mahrt E, Guo W, Stawiski EW, Modrusan Z, Seshagiri S, Kapur P, Hon GC, Brugarolas J, and Wang T. 2019. SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. *Genes (Basel)* 10. 10.3390/genes10070531

scAnnoX: An R Package Integrating Multiple Public Tools for Single-Cell Annotation

ORIGINALITY REPORT

11%	8%	8%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	github.com Internet Source	1%
2	www.wjgnet.com Internet Source	1%
3	openscholarship.wustl.edu Internet Source	1%
4	watermark.silverchair.com Internet Source	<1%
5	Shaoqi Chen, Bin Duan, Chenyu Zhu, Chen Tang, Shuguang Wang, Yicheng Gao, Shaliu Fu, Lixin Fan, Qiang Yang, Qi Liu. "Privacy-preserving integration of multiple institutional data for single-cell type identification with scPrivacy", Science China Life Sciences, 2022 Publication	<1%
6	repositorio.ucm.cl Internet Source	<1%
7	thno.org Internet Source	<1%

8	www.mdpi.com Internet Source	<1 %
9	Liuting Zeng, Kailin Yang, Tianqing Zhang, Xiaofei Zhu, Wensa Hao, Hua Chen, Jinwen Ge. "Research progress of single-cell transcriptome sequencing in autoimmune diseases and autoinflammatory disease: A review", Journal of Autoimmunity, 2022 Publication	<1 %
10	assets.researchsquare.com Internet Source	<1 %
11	www.frontiersin.org Internet Source	<1 %
12	Bassel Ghaddar, Subhajyoti De. "Hierarchical and automated cell-type annotation and inference of cancer cell of origin with Census", Bioinformatics, 2023 Publication	<1 %
13	Xiuhui Yang, Koren K. Mann, Hao Wu, Jun Ding. "scCross: A Deep Generative Model for Unifying Single-cell Multi-omics with Seamless Integration, Cross-modal Generation, and In-silico Exploration", Cold Spring Harbor Laboratory, 2023 Publication	<1 %
14	Nan Wang, Zhenhua Liu, Jing-Yuan Ma, Jingju Liu et al. "Sustainability Perspective Oriented	<1 %

Synthetic Strategy for Zinc Single-atom Catalysts Boosting Electrocatalytic Reduction of Carbon Dioxide and Oxygen", ACS Sustainable Chemistry & Engineering, 2020
Publication

15	elifesciences.org Internet Source	<1 %
16	Submitted to Bournemouth University Student Paper	<1 %
17	Shangru Jia, Artem Lysenko, Keith A Boroevich, Alok Sharma, Tatsuhiko Tsunoda. "scDeepInsight: a supervised cell-type identification method for scRNA-seq data with deep learning", Briefings in Bioinformatics, 2023 Publication	<1 %
18	Chloe X. Wang, Lin Zhang, Bo Wang. "One Cell At a Time (OCAT): a unified framework to integrate and analyze single-cell RNA-seq data", Genome Biology, 2022 Publication	<1 %
19	minerva-access.unimelb.edu.au Internet Source	<1 %
20	Giovanni Pasquini, Jesus Eduardo Rojo Arias, Patrick Schäfer, Volker Busskamp. "Automated methods for cell type annotation on scRNA-	<1 %

seq data", Computational and Structural Biotechnology Journal, 2021
Publication

21	dokumen.pub Internet Source	<1 %
22	pubmed.ncbi.nlm.nih.gov Internet Source	<1 %
23	pure.hw.ac.uk Internet Source	<1 %
24	scholarworks.iupui.edu Internet Source	<1 %
25	www.cancerbiomed.org Internet Source	<1 %
26	www.ncbi.nlm.nih.gov Internet Source	<1 %
27	"Poster Exhibition", Hepatology International, 2009 Publication	<1 %
28	Alex M. Ascensión, Sandra Fuertes-Álvarez, Olga Ibañez-Solé, Ander Izeta, Marcos J. Araúzo-Bravo. "Human Dermal Fibroblast Subpopulations Are Conserved across Single-Cell RNA Sequencing Studies", Journal of Investigative Dermatology, 2020 Publication	<1 %

29	Peng Tian, Jie Zheng, Yue Xu, Tao Wu et al. "scPharm: identifying pharmacological subpopulations of single cells for precision medicine in cancers", Cold Spring Harbor Laboratory, 2023 Publication	<1 %
30	Sibo Zhu, Tao Qing, Yuanting Zheng, Leming Shi. "Advances in single-cell RNA sequencing and its applications in cancer research", Oncotarget, 2017 Publication	<1 %
31	genome.cshlp.org Internet Source	<1 %
32	spiral.imperial.ac.uk Internet Source	<1 %
33	www.deepdyve.com Internet Source	<1 %
34	www.nature.com Internet Source	<1 %
35	www.researchgate.net Internet Source	<1 %
36	Changde Cheng, Wenan Chen, Hongjian Jin, Xiang Chen. "A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell–Cell Communication", Cells, 2023 Publication	<1 %

37	Christos Tzaferis, Evangelos Karatzas, Fotis A. Baltoumas, Georgios A. Pavlopoulos, George Kollias, Dimitris Konstantopoulos. "SCALA: A web application for multimodal analysis of single cell next generation sequencing data", Cold Spring Harbor Laboratory, 2022 Publication	<1 %
38	Dorothy Ellis, Dongyuan Wu, Susmita Datta. " : A review on statistical analytics of single-cell RNA sequencing data ", WIREs Computational Statistics, 2021 Publication	<1 %
39	Liyen Loh, Salomé Carcy, Harsha S. Krovi, Joanne Domenico et al. "Unraveling the Phenotypic States of Human innate-like T Cells: Comparative Insights with Conventional T Cells and Mouse Models", Cold Spring Harbor Laboratory, 2023 Publication	<1 %
40	Yu Zhang, Feng Zhang, Zekun Wang, Siyi Wu, Weidong Tian. "scMAGIC: accurately annotating single cells using two rounds of reference-based classification", Nucleic Acids Research, 2022 Publication	<1 %
41	www.biorxiv.org Internet Source	<1 %

42	Oscar González-Velasco, Dulce Papy-García, Gael Le Douaron, José M. Sánchez-Santos, Javier De Las Rivas. "Transcriptomic landscape, gene signatures and regulatory profile of aging in the human brain", Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 2020 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

scAnnoX: An R Package Integrating Multiple Public Tools for Single-Cell Annotation

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14