Bioinformatics Advances, 2024, **00**, vbae062
https://doi.org/10.1093/bioadv/vbae062
Advance Access Publication Date: 2 May 2024
Application Note

Software

scX: a user-friendly tool for scRNAseq exploration

Tomás V. Waichman [1], M.L. Vercesi[1], Ariel A. Berardino[1,2], Maximiliano S. Beckel[1,2], Damiana Giacomini[2,3], Natalí B. Rasetto[2,3], Magalí Herrero[2,3], Daniela J. Di Bella[4,5], Paola Arlotta [4,5], Alejandro F. Schinder[2,3], Ariel Chernomoretz [1,6,7,*]

[1]Integrative Systems Biology Lab, Leloir Institute, Buenos Aires, CP1405, Argentina
[2]Instituto de Investigaciones Bioquímicas de Buenos Aires, CONICET, Buenos Aires, CP1405, Argentina
[3]Laboratory of Neuronal Plasticity, Leloir Institute, Buenos Aires, CP1405, Argentina
[4]Department of Stem Cells and Regenerative Biology, Harvard University, Cambridge, MA 02138, United States
[5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02138, United States
[6]Departamento de Física, FCEN, Universidad de Buenos Aires, Buenos Aires, CP1428, Argentina
[7]INFINA, UBA-CONICET, Buenos Aires, CP 1428, Argentina

*Corresponding author. INFINA, UBA-CONICET, Buenos Aires, CP1428, Argentina & Leloir Institute, Buenos Aires, CP1405, Argentina.
E-mail: achernomoretz@leloir.org.ar or ariel@df.uba.ar

Associate Editor: Magnus Rattray

Abstract

Motivation: Single-cell RNA sequencing (scRNAseq) has transformed our ability to explore biological systems. Nevertheless, proficient expertise is essential for handling and interpreting the data.

Results: In this article, we present scX, an R package built on the Shiny framework that streamlines the analysis, exploration, and visualization of single-cell experiments. With an interactive graphic interface, implemented as a web application, scX provides easy access to key scRNAseq analyses, including marker identification, gene expression profiling, and differential gene expression analysis. Additionally, scX seamlessly integrates with commonly used single-cell Seurat and SingleCellExperiment R objects, resulting in efficient processing and visualization of varied datasets. Overall, scX serves as a valuable and user-friendly tool for effortless exploration and sharing of single-cell data, simplifying some of the complexities inherent in scRNAseq analysis.

Availability and implementation: Source code can be downloaded from https://github.com/chernolabs/scX. A docker image is available from dockerhub as chernolabs/scx.

1 Introduction

After nearly 15 years of continuous development, single-cell transcriptomics continues to have a profound impact on the biomedical research field. Over the years, various data-processing pipelines have been proposed, as well as visualization tools that aimed to ease the analysis of this type of high-throughput assays.

The existing software ecosystem is extensive, often exhibiting overlap in approaches and solutions. Table S1 presents a comprehensive comparison of several commonly used tools for single-cell RNA sequencing (scRNAseq) data exploration. This table provides insight into the extent to which each solution covers various aspects of the analysis. iSEE (Rue-Albrecht *et al.* 2018), cellxgene (Abdulla *et al.* 2023), and ShinyCell (Ouyang *et al.* 2021) are noteworthy tools that primarily concentrate on data visualization, offering a diverse range of plots and graphical data representations. ShIVA (Aussel *et al.* 2023) and CellSnake (Umu *et al.* 2023) on the other hand are solutions more focused on data processing aspects of single cell analysis. There are also more comprehensive tools, such as ASAP (David *et al.* 2020), SEQUIN (Weber *et al.* 2023), and SCHNAPPS (Jagla *et al.* 2021) that can handle multiple embeddings and visualizations, and grant interactive single-cell analysis features such as clustering, marker identification, and differential expression analysis. Each one of them addresses differently the trade-off between the extend and complexity of the offered computational calculations and design criteria in terms of usability and ease of interaction with the application. This election has profound impacts on the user side. Specifying numerous parameters may pose challenges for users who are typically more interested in extracting relevant biology from their data and may lack the expertise or criteria to define the required values for each presented option. In such cases, while comprehensiveness is desirable, it may hinder or compromise the tool's ease of use.

Here we present scX, an R package that deploys a user-friendly Shiny-based application developed for researchers to explore single-cell datasets. From its inception, scX was designed as a tool to efficiently bridge the computational side of the problem (e.g. data preparation, normalization, markers identification, differential expression, etc.) with various tools enabling the rapid implementation of biological analyses derived from these results (interactive 3D visualizations of low-dimensional embeddings, on-the-fly markers identification, exploratory data analysis capabilities,

Received: December 27, 2023; **Revised:** March 6, 2024; **Editorial Decision:** April 16, 2024; **Accepted:** April 30, 2024

availability of a wide array of plot types, etc.). scX becomes particularly well suited for two scenarios. For one hand, for bioinformatics laboratories looking to share scRNAseq experiment data, processed with arbitrary sophistication, with biology colleagues aiming to explore them efficiently in a user-friendly designed platform. At the same time, our package offers, if required, the ability to carry out a significant portion of typical scRNAseq data analysis (normalization, dimensionality reduction, clustering, marker identification, and calculation of differential expression) in a non-interactive preprocessing step that can complement or integrate with any existing analyses. This ensures that, even users new to computational aspects of this field have the opportunity to benefit from the streamlined analysis of their data through the tools provided by the interactive scX interface.

2 Methods

The scX app can be easily launched by executing two R functions (see Fig. 1). Starting from a provided count matrix, a SingleCellExperiment object, or a Seurat object, the function "createSCEobject" handles the offline pre-processing of the data. This function automatically executes a series of computational steps leveraging the functionality implemented in the "scran" Bioconductor package. We adopted a normalization by deconvolution scheme (Lun *et al.* 2016) to eliminate systematic differences between libraries The identification of the most variable features involves fitting a trend on the variance versus mean of log-normalized expression profiles and identifying genes with a positive biological variance component. If requested, a graph-based cell clustering procedure can be implemented. By default, it performs a Louvain clustering over a mutual-K nearest neighbor graph ($k = 20$) estimated considering Euclidean cell–cell distances in the sub-space spanned by the 20 largest PCA components. It is also possible to specify any other graph-based clustering method available from the igraph package. Advanced users can directly specify an NNGraphParam object (from the bluster Bioconductor package) to achieve maximum control in the specification of this graph-based clustering task. Differential expression analysis and marker identification are conducted on one or more user-specified partitions, relying on the functionality implemented in the "findMarkers" function. For marker identification, a Wilcoxon rank sum test is considered by default, but other options (t-test, binomial test) can also be specified. For every cluster, pairwise tests are performed against any other cluster and, by default, genes are ranked based on the maximal observed p-val (pval.type="all" in the paramFindMarkers input parameter). Optionally, the user can adopt other strategies to consolidate DE signals. Furthermore, ad-hoc pre-calculated lists of markers can also be provided.

2.1 Summary module

This module provides a summary of the primary descriptive details of the working dataset, such as the number of cells and genes, the mean number of genes detected per cell, etc. Additionally, it allows the visualization of the number of counts and detected features in connection with various metadata covariates, enabling the evaluation of potential batch-related issues.

2.2 Exploratory data analysis module

This module facilitates the exploration of relationships between the covariates included in the metadata of the SCE object (specified by the "metadataVars" and "partitionVars"

Figure 1. A schematic illustration of the scX workflow is depicted on the left-hand side. The right-hand side exhibits several instances of scX's analysis and visualization capabilities.

parameters of creatSCEobject). In the "Categories" section, bar plots can be used to analyze one- or two-dimensional distribution functions involving categorical covariates. The "Matrix" tab enables the generation of bivariate count tables. The "Field" section of this module allows the exploration of how the value of one or more continuous covariates changes concerning another variable, which can be either numerical or categorical. Various types of plots, including box plots, heatmaps, dot plots, or stacked violin plots, can be generated to assist in this analysis.

2.3 Markers module

The "Cluster markers" section facilitates the analysis of marker genes identified during the preprocessing step. When a user selects a cell displayed in the embedding window, a marker gene table is generated, including various metrics for each gene marker. By default (pval.type="all" in the paramFindMarkers input parameter of the createSCEObject function), "summary.stats" reports the weakest observed differential expression signal between the analyzed cluster and any other cluster. "log.FDR" is the logged largest observed false discovery rate (FDR)-corrected *p*-value, and "boxcor" is the Pearson correlation value between the gene expression profile and a binary indicator vector of the cluster of interest. The table can be saved in various formats (.csv, .xlsx, .pdf) or copied to the clipboard. Notably, clicking on a gene in the table allows the visualization of the corresponding expression field in the embedding window. Additional graphical characterizations are provided as violin and spike plots presented at the bottom of the page.

In the "Find new markers" section, users can investigate markers for specific sets of cells, selected on the fly from 2D embeddings using the box or lasso tools (see the online manual for an animated GIF tutorial). Putative markers are identified by ranking genes in decreasing order based on their estimated boxcor values (above a minimum value of 0.3). The marker table, along with the corresponding cell list, can be downloaded. Similarly, to the previous section, clicking on a marker row generates a visualization of the marker's expression pattern in the embedding dataset, and additional graphical characterizations in the form of violin and spike plots are also produced.

2.4 Gene expression module

The "Gene Expression" module facilitates the exploration of expression patterns for one or more genes of interest. Expression changes in response to different categorical and/or numerical covariates can be assessed, and coexpression patterns between pairs of genes can be analyzed. In the "Categories" section, one or more genes of interest can be selected (or uploaded from a file). The module displays the average expression of these genes across the embedded dataset in the "Scatter" window. Visualization options, including heatmaps, dot plots, and stacked violin plots, are also available for analyzing the expression of these genes concerning different categorical covariates found in the metadata. The "Field" section allows for the analysis of gene expression in conjunction with numerical covariates that may be present in the metadata within the SCE object. This can include variables like the number of counts or pseudotime values. Finally, the "Co-expression" section allows for the examination of coexpression patterns between selected gene pairs in the embedding space window. The percentage of co-detection events within categorized groups of cells can also be assessed and visualized.

2.5 Differential expression module

In this section, the results of the differential expression analysis can be assessed. An interactive selection of threshold values for both logFC (logarithm of fold change) and the FDR significance level is available. The list of differentially expressed genes is downloadable in various formats (csv, pdf, and xlsx). This section also generates a Volcano plot graphical representation, along with visualizations (violin plots, spike plots, heatmaps, and dot plots) that facilitate a more comprehensive understanding of expression patterns for up- and down-regulated genes.

2.6 Visual tools module

This module provides many tools to produce pdf plots with more complex or specific layouts involving gene expression patterns and covariate variables.

2.7 Case study

We focused on the study conducted by Tusi and collaborators on hematopoietic lineages in mouse basal bone marrow cells (Tusi *et al.* 2018). The authors performed a population balance analysis (PBA) to predict cell fate probabilities and identified seven putative commitment probabilities for each hematopoietic progenitor. We employed scX to preprocess and re-visit their scRNAseq data (metadata and raw counts were downloaded from the paper's supporting webpage, and the used R script was included as Supplementary Material File). For our analysis, we retained the original 2D data projection (generated using a force-directed graph layout algorithm on a knn graph), and a Louvain partition of the filtered 4763 cells. Upon initial exploration a pronounced batch effect associated with "basal_bm1" cells, originating from a library processed in a specific sequencing run (seq_run_1), was identified (see Fig. 2A). Consequently, we filtered out this run, retained 4016 cells, and re-created the SCEObject for further analysis.

We then considered the 2D force-directed layout (FDL) representation shown in Fig. 2B, where different colors were used for the 12 clusters of the original Louvain partition. We focused on characterizing the terminal group identified as basophilic or mast cells (Ba in Fig. 2B). This group exhibited high commitment probabilities to the Ba attracting state, as visualized in the $P(Ba)$ field over the graph (inset of Fig. 2B). Notably, these cells were included in a broader Louvain cluster (cluster 6, gray dots in 2b). The "find-new-marker" functionality was then employed to identify specific gene markers for this set of cells. The top five ranked marker genes found by our tool were: Cpa3, Ms4a2, Gzmb, Cyp11a1, and Fcer1a, with boxcor values of 0.553, 0.509, 0.506, 0.496, and 0.486, respectively. The expression field of Cpa3 is shown in Fig. 2C. To validate these putative markers we referred to the work of Miao and collaborators, who developed the single-cell clustering assessment framework (SCCAF) strategy for cell type discovery from single-cell expression data (Miao *et al.* 2020). We found that the first three marker genes identified by scX were also top-ranked features in SCCAF's logistic regression model for recognizing the Ba cell group [cluster 10 in Fig. 5d of (Miao *et al.* 2020)]. Additionally, we found further support for the complete set of scX-identified markers in several bibliographic references

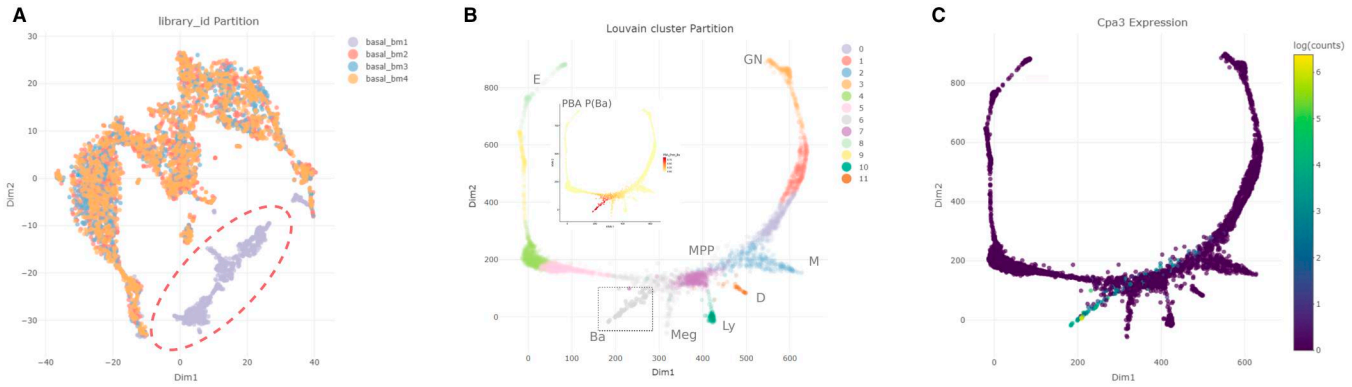

Figure 2. (A) 2D tSNE embedding used to highlight the large batch effect affecting "basal_bm1" cells. (B) FDL visualization of Tusi dataset. Louvain clusters and labels for the seven PBA inferred terminal states can be appreciated (Ba, basophilic or mast cell; D, dendritic; E, erythroid; GN, granulocytic neutrophil; Ly, lymphocytic; M, monocytic; Meg, megakaryocytic; MPP, multipotential progenitors). The dotted square schematizes the interactive cell selection process. The field of commitment probability values to the Ba state is displayed in the inset. (C) Expression field of the top ranked Cpa3 putative marker gene.

(Metcalfe *et al.* 2016, Siddhuraj *et al.* 2017, Hiroyasu *et al.* 2021, Silva-Gomes *et al.* 2021, Miyake *et al.* 2023).

2.8 Computational demands

To test the computational demands of a typical scX pipeline we considered the mouse nervous system's scRNAseq data from Zeisel *et al.* (2018). Running-time and peak memory consumption tests were conducted on an Intel(R) Xeon(R) Silver 4116 CPU @ 2.10 GHz, 514G RAM server, considering incrementally subsampled datasets. Details and results are summarized in Table S2. The most demanding step was data preprocessing (createSCEobject function), requiring a peak of 24 Gb of RAM and 180 minutes for $N = 160\,000$ cells. We found that peak memory usage and running time scaled linearly at a rate of 1.3 Gb/10k-cells and 10.4 minutes/10k-cells, respectively (see Supplementary Material Fig. SF1). It is important to note that the number of partitions and clusters can influence the processing time of a given dataset. For visualization purposes a subsampling strategy can be specified at pre-processing time (default setting of 50 000 cells) to ensure a smooth interactive experience.

3 Conclusions

We developed scX, a Shiny-based application that enhances collaboration between bioinformaticians and experimental biologists in joint projects. The platform is user-friendly and highly interactive, fostering a collaborative environment that could significantly advance the development of joint projects.

Acknowledgements

D.G., A.C., and A.F.S. are investigators in the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). N.B.R., A.A.B., M.H., and M.B. were supported by CONICET fellowships.

Author contributions

Tomás V. Waichman (Conceptualization [lead], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—review and editing [supporting]), M.L. Vercesi (Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—review and editing [equal]), Ariel A. Berardino (Methodology [supporting], Software [equal], Validation [Supporting], Visualization [supporting], Writing—review and editing [equal]), Maximiliano S. Beckel (Data curation [equal], Methodology [Supporting], Validation [supporting], Writing—review and editing [Supporting]), Damiana Giacomini (Conceptualization [supporting], Validation [equal]), Natalí B. Rasetto (Methodology [supporting], Validation [supporting], Visualization [supporting]), Magalí Herrero (Methodology [supporting], Validation [supporting], Visualization [supporting]), Daniela J. Di Bella (Validation [supporting], Writing—review and editing [supporting]), Paola Arlotta (Conceptualization [equal], Funding acquisition [equal], Validation [supporting]), Alejandro F. Schinder (Conceptualization [equal], Funding acquisition [Equal], Methodology [supporting], Supervision [equal], Validation [equal], Writing—review and editing [equal]), and Ariel Chernomoretz (Conceptualization [lead], Investigation [lead], Methodology [equal], Project administration [equal], software [equal], Supervision [lead], Validation [equal], Visualization [equal], Writing—original draft [equal]).

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported by grants the National Institute of Neurological Disorders and Stroke (NINDS) and Fogarty International Center (FIC) (R01NS103758) to P.A. and A.F.S., and the Argentine Agency for the Promotion of Science and Technology (PICT-2020–0046 and PICT-2021–0077) to A.F.S., (PICT 2018–03713) to A.C. and M.S.B. (postdoctoral fellowship) and (PICT 2017–0389) to D.G.

Data availability

Source code can be downloaded from https://github.com/chernolabs/scX. User manual available at https://chernolabs.github.io/scX/. A docker image is available from dockerhub as chernolabs/scx.

References

Abdulla S, Aevermann B, Assis P *et al.* Cz cell×gene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. bioRxiv. https://doi.org/10.1101/2023.10.30.563174, 2023, preprint: not peer reviewed.

Aussel R, Asif M, Chenag S *et al.* ShIVA: a user-friendly and interactive interface giving biologists control over their single-cell RNA-seq data. *Sci Rep* 2023;**13**:14377.

David FPA, Litovchenko M, Deplancke B *et al.* ASAP 2020 update: an open, scalable and interactive web-based portal for (single-cell) omics analyses. *Nucleic Acids Res* 2020;**48**:W403–14. 05.

Hiroyasu S, Zeglinski MR, Zhao H *et al.* Granzyme b inhibition reduces disease severity in autoimmune blistering diseases. *Nat Commun* 2021;**12**:302.

Jagla B, Libri V, Chica C *et al.* SCHNAPPs - single cell sHiNy APPlication(s). *J Immunol Methods* 2021;**499**:113176.

Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Res* 2016;**5**:2122.

Metcalfe DD, Pawankar R, Ackerman SJ *et al.* Biomarkers of the involvement of mast cells, basophils and eosinophils in asthma and allergic diseases. *World Allergy Organ J* 2016;**9**:7.

Miao Z, Moreno P, Huang N *et al.* Putative cell type discovery from single-cell gene expression data. *Nat Methods* 2020;**17**:621–8.

Miyake K, Ito J, Nakabayashi J *et al.* Single cell transcriptomics clarifies the basophil differentiation trajectory and identifies pre-basophils upstream of mature basophils. *Nat Commun* 2023;**14**:2694.

Ouyang JF, Kamaraj US, Cao EY *et al.* ShinyCell: simple and sharable visualization of single-cell gene expression data. *Bioinformatics* 2021;**37**:3374–6.

Rue-Albrecht K, Marini F, Soneson C *et al.* ISEE: interactive SummarizedExperiment explorer. *F1000Res* 2018;**7**:741.

Siddhuraj P, Mori M, Bjermer L *et al.* Distinct tryptase and cpa3-positive basophil phenotypes in healthy individuals, asthma, and copd. *Eur Respiratory J* 2017;**50**(**suppl 61**):OA4848.

Silva-Gomes R, Mapelli SN, Boutet M-A *et al.* Differential expression and regulation of MS4A family members in myeloid cells in physiological and pathological conditions. *J Leukoc Biol* 2021;**111**:817–36.

Tusi BK, Wolock SL, Weinreb C *et al.* Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* 2018;**555**:54–60.

Umu SU, Rapp Vander-Elst K, Karlsen VT *et al.* Cellsnake: a user-friendly tool for single-cell RNA sequencing analysis. *Gigascience* 2023;**12**:giad091.

Weber C, Hirst MB, Ernest B *et al.* SEQUIN is an R/shiny framework for rapid and reproducible analysis of RNA-seq data. *Cell Rep Methods* 2023;**3**:100420.

Bioinformatics Advances, 2024, 00, 1–5
https://doi.org/10.1093/bioadv/vbae062
Application Note