scientific reports

OPEN

# ShIVA: a user-friendly and interactive interface giving biologists control over their single-cell RNA-seq data

**Rudy Aussel[1], Muhammad Asif[2,3], Sabrina Chenag[1], Sébastien Jaeger[1,2,4], Pierre Milpied[1,2,4]✉ & Lionel Spinelli[1,2,4]✉**

**Single-cell technologies have revolutionised biological research and applications. As they continue to evolve with multi-omics and spatial resolution, analysing single-cell datasets is becoming increasingly complex. For biologists lacking expert data analysis resources, the problem is even more crucial, even for the simplest single-cell transcriptomics datasets. We propose ShIVA, an interface for the analysis of single-cell RNA-seq and CITE-seq data specifically dedicated to biologists. Intuitive, iterative and documented by video tutorials, ShIVA allows biologists to follow a robust and reproducible analysis process, mostly based on the Seurat v4 R package, to fully explore and quantify their dataset, to produce useful figures and tables and to export their work to allow more complex analyses performed by experts.**

In biology, single-cell RNA sequencing (scRNA-seq), and more generally single-cell genomics, have opened a new era. The scientific community has rapidly adopted those technologies as evidenced by their use in a growing number of high-impact biomedical publications[1]. The rapid growth of new applications (scRNA-seq, CITE-seq, scDNA-seq, spatial transcriptomics and scChIP-seq) is providing a large diversity of information about biological heterogeneity and dynamic processes in cell populations[2].

A major hindrance to the meaningful use of these technologies is that they require both a very good knowledge of the underlying biology and an expert level in data analysis (bioinformatics and statistics). The successful analysis of single-cell genomics datasets is therefore the result of a close collaboration between biologists and bioinformaticians. This interdependence raises two main issues:

1. In early stages of the analysis, biologists need to rapidly browse the data to grasp a first idea of the underlying biology. To do so, they rely on various methods for data visualization, made available by bioinformaticians. Upon inspection of resulting reports, scientific experts often request refinements and new analyses to push further their understanding. During this iterative process, a significant amount of time is lost in the back-and-forth exchange between collaborators, and a first level analysis often requires several weeks.
2. ScRNA-seq data analysis challenges are quite recent and many research groups are not fully equipped with required bioinformatics resources yet. In this context, in-depth analysis of experimental data is not always possible.

In recent years, the bottleneck between generation of experimental data and their analysis has been getting worse[3]. Until recently, the time required to produce datasets in the laboratory was greater than the time taken for bioinformatics analyses. Today, with the remarkable advances in biotechnologies, generating single-cell genomics datasets has become relatively easy and fast. On the other hand, time required for analysis has increased because of the large number and high complexity of analyses applied to those datasets. As a result, while research teams can produce large datasets quickly, they often have to wait long before analyzing them.

[1]CNRS, INSERM, CIML, Centre d'Immunologie de Marseille-Luminy, Aix Marseille Univ, Marseille, France. [2]Turing Centre for Living Systems (CENTURI), Aix Marseille Univ, Marseille, France. [3]Present address: Biomedical Data Science Lab, Department of Bioinformatics and Biotechnology, Government College University Faisalabad (GCUF), Faisalabad 38000, Pakistan. [4]These authors jointly supervised this work: Sébastien Jaeger, Pierre Milpied and Lionel Spinelli. ✉email: milpied@ciml.univ-mrs.fr; spinelli@ciml.univ-mrs.fr

A tool enabling biologists to execute first steps of data analysis autonomously and deeply explore the biology of their datasets would considerably speed up the process, and allow them to improve the knowledge they can infer from the exploration. In this context, data analysis experts can focus on more complex analyses, to confirm biologists' observations or bring new evidence. With that problem in mind, we have developed ShIVA, a Shiny Interface for Visualization and Analysis of single-cell datasets (Fig. 1).

## Results

### Principle of the interface.

ShIVA is a software allowing any biologist to apply a correct and validated analytic process that corresponds to current state-of-the-art, relying mostly on functions from the Seurat v4 R package that was developed by single-cell data analysis experts[4]. With this tool, non-experts (biologist researchers, PhD students or engineers) can load single-cell sequencing data (scRNA-seq or CITE-seq), apply the main quality control steps, execute normalization and dimensionality reduction, perform clustering, identify marker genes and their main functions, and finally visualize the data interactively to execute supervised and non-supervised scientific analyses.

ShIVA has been designed specifically to guide biologists through the analysis workflow, offering them to visualize, quantify and explore the data at each stage before making decisions affecting downstream analysis. The interface has been designed to be as user-friendly as possible while providing maximum choices. Specifically, ShIVA has been designed to allow the biologist to easily navigate through the workflow steps to revisit a decision, modify one or more parameters and iteratively refine their understanding of the data. Moreover, once the initial analysis steps have been completed, ShIVA offers a complete and fully customizable data exploration: the biologists can create the figures of their choice (maps, boxplots, violin plots, histograms, density plots, tables, with several options for style customization), quantify precisely what interests them most and create subsets of cells (for example a cell subtype) on which they can repeat the complete analysis process to deepen their understanding. The complete analysis can be saved as a "ShiVA project" file to be loaded later for further analysis.

ShIVA does not forget that collaboration with a data analysis expert will certainly be required for more advanced analyses. To ease this collaboration, ShIVA offers to export (i) all the figures, tables, list of genes or cells created by the biologists, (ii) a complete report of the analysis performed with all the details ensuring its reproducibility, and (iii) a file (R object) that can be re-used by the bioinformatician in an R environment to take up analyses where the biologist left them.

To help biologists use ShIVA correctly, we created tutorial videos for each step of the analysis workflow, available on a YouTube channel (https://www.youtube.com/channel/UCJJ3Svi8AY6XGx4Y9r9G3Iw).

### Overview of main features.

ShIVA covers all the functions described in Fig. 2. It is able to process data from scRNA-seq and CITE-seq experiments, including samples multiplexed with "hashtag" oligonucleotides (HTO). The workflow in ShIVA is performed through a succession of interdependent modules. Activating a module is allowed only if the mandatory previous steps have been correctly performed, helping the user to grasp the rigorous analysis process to follow. In each module, the user visualizes the relevant data, makes decision on the required analysis parameters and evaluates their impact through direct observation of the result. When the result of an analysis is not optimal, the user can easily step backward in the previous modules to modify the parameters and re-run the corresponding analysis. ShIVA keeps track of the user's choice by defining a hierarchy of sub-projects, each of them containing the results of different user choices. Switching between sub-projects allows for comparison of analysis processes to optimize the deciphering of the dataset. The complete ShIVA project/sub-project structure can be saved into a file that contains the Seurat data object and the metadata linked to the analysis performed in ShiVA by the user. It allows to save then later reload a project for further investigations. Moreover, during the analysis, at the *Quality Control* and *Gene Selection* steps, the project is automatically saved on the computer disk. This saved project file can be retrieved from the *Data Import* module with the third

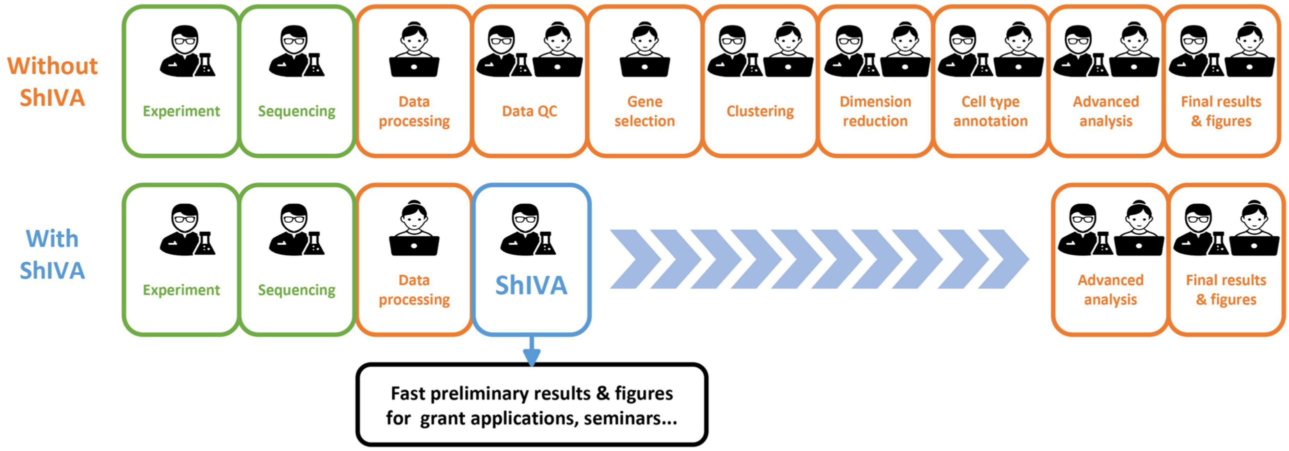


**Figure 1.** ShIVA reduces single-cell genomics data analysis time by placing the biologist project leader in power. ShIVA enables biologists to perform state-of-the-art exploratory analyses of their single-cell genomics datasets in a user-friendly interface.

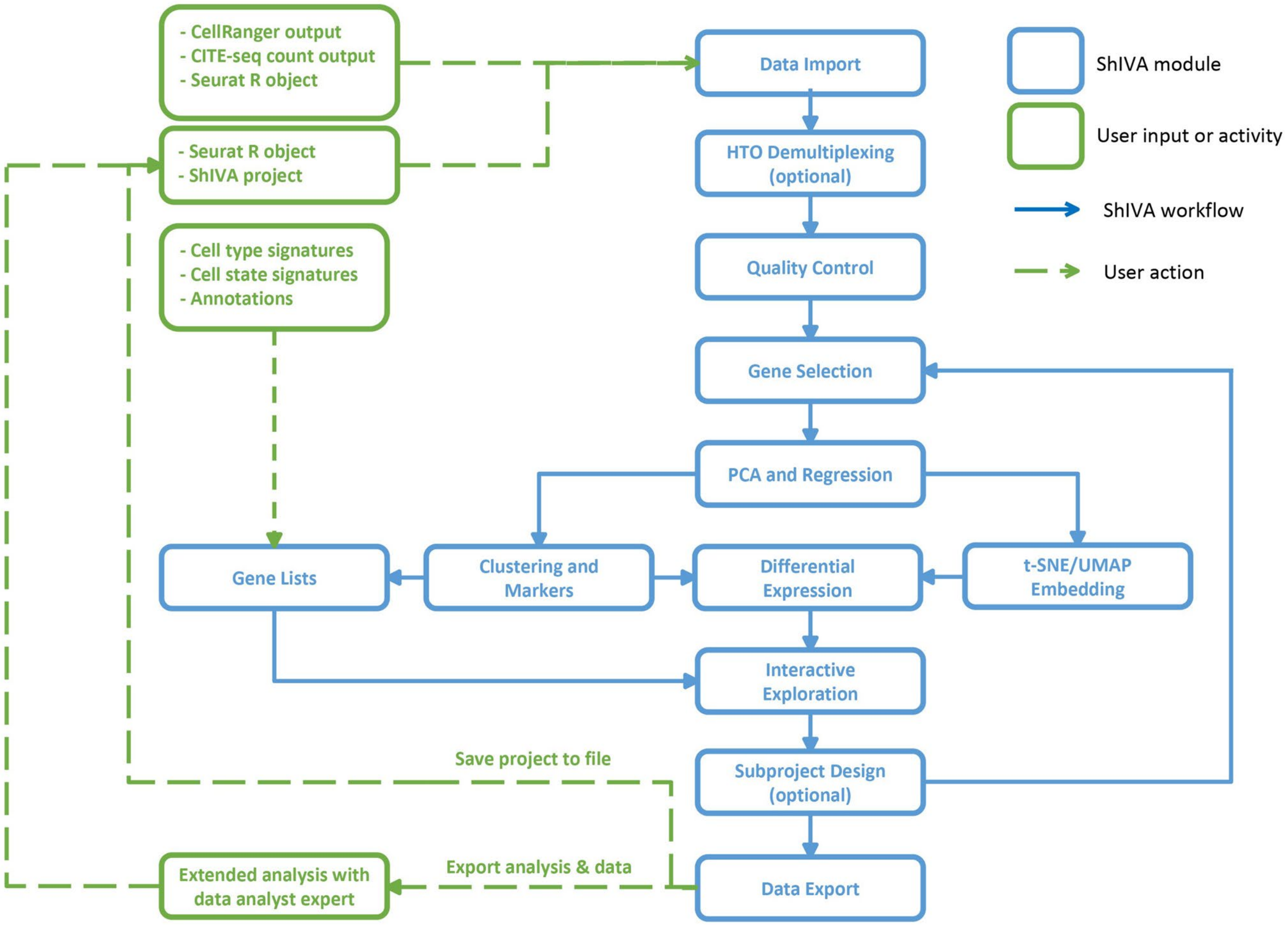
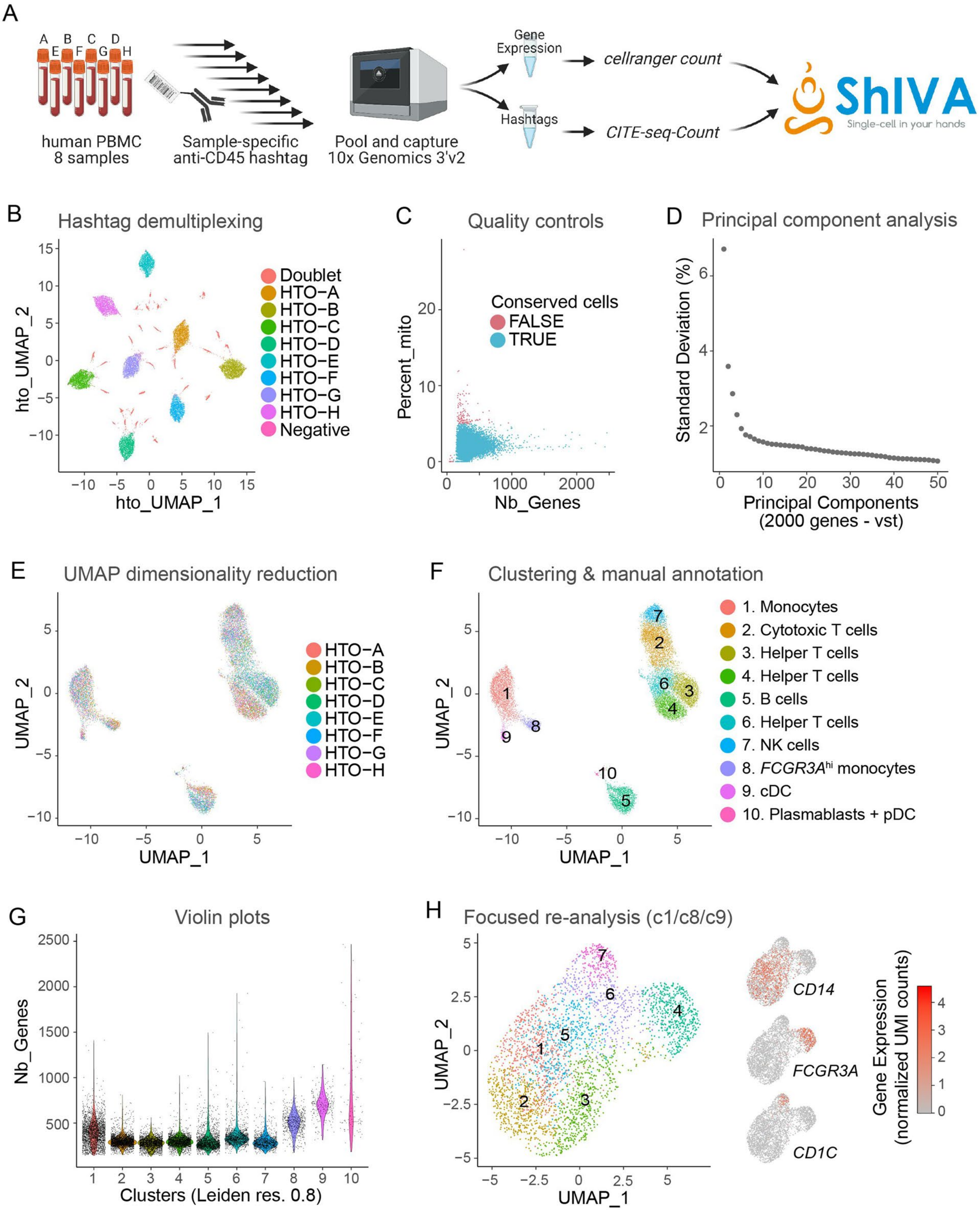


**Figure 2.** ShIVA modules and workflow. Schematic view of ShIVA analysis modules (blue boxes) enabling the analysis of single-cell genomics datasets from various standard user inputs (green boxes).

| Module name | Parameters | Available visualizations/results |
|---|---|---|
| Data import | Project name, species, normalization method, cells and genes thresholds | Creation of a ShIVA project |
| HTO demultiplexing | Normalization method, quantile threshold (expert) | Number of cells by HTO, negative cells and multiplets; HTO UMAP plot; HTO Ridge plots |
| Quality control | Range of: % of mitochondrial genes, % of ribosomal genes, number of features per cell, number of UMI per cell | Covariable violin plots; covariable scatterplots; Number of cells filtered out |
| Gene selection | Gene selection method; parameters of the selected method | Table of selected genes; scatterplot of selected genes |
| PCA and regression | Number of PC to compute; covariable to regress (optional) | Heatmap of loading genes per PC; heatmap of PC and covariables correlations; Scatterplots of PC against covariables |
| T-SNE/UMAP embedding | PCA analysis to use; number of PC to use; perplexity; number of neighbors to consider | t-SNE plots; UMAP plots |
| Clustering | Number of PC to use; clustering method; parameters of the selected method | Plots of the clusters in PCA or t-SNE or UMAP plots |
| Markers genes | Minimal expression percentage; minimal log fold change; analysis method | Table of all makers; table of markers by cluster; feature plots of top markers by cluster |
| Differential expression | Parameters to define the two group of cells; minimal expression percentage; minimal log fold change; analysis method | Table of all DEG; feature plots of top DEG; functional enrichment of DEG in GO BP, MF, CC and KEGG pathways |
| Interactive exploration—2D plot | Type of plot (PCA/t-SNE/UMAP); variable to use for colors; variables for layout (facet); theme for axis and background; size of dots; selection of cells by manual gating (optional) | 2D plot with chosen colors, layout and theme; List of selected cells |
| Interactive exploration—1D plot | Type of plots (violin/histogram/density); variables to plot; variable to use for colors; theme for axis and background | 1D plot of the chosen variables, layout and theme |
| Interactive exploration—table | Type of table (counts/expression); variables to compute in the table; layout of table | Table of counts/expression of the chosen variables |

**Table 1.** List of ShIVA modules. Each row describes a module with its name (left column), the parameters available to user customization (middle column), and the different visualizations and results available (right column).


A
A B C D
E F G H
human PBMC
8 samples
Sample-specific
anti-CD45 hashtag
Pool and capture
10x Genomics 3'v2
Gene
Expression
Hashtags
cellranger count
CITE-seq-Count
ShIVA
Single-cell in your hands
B
Hashtag demultiplexing
hto_UMAP_2
hto_UMAP_1
Doublet
HTO−A
HTO−B
HTO−C
HTO−D
HTO−E
HTO−F
HTO−G
HTO−H
Negative
C
Quality controls
Percent_mito
Nb_Genes
Conserved cells
FALSE
TRUE
D
Principal component analysis
Standard Deviation (%)
Principal Components
(2000 genes - vst)
E
UMAP dimensionality reduction
UMAP_2
UMAP_1
HTO−A
HTO−B
HTO−C
HTO−D
HTO−E
HTO−F
HTO−G
HTO−H
F
Clustering & manual annotation
UMAP_2
UMAP_1
1. Monocytes
2. Cytotoxic T cells
3. Helper T cells
4. Helper T cells
5. B cells
6. Helper T cells
7. NK cells
8. FCGR3A^hi monocytes
9. cDC
10. Plasmablasts + pDC
G
Violin plots
Nb_Genes
Clusters (Leiden res. 0.8)
H
Focused re-analysis (c1/c8/c9)
UMAP_2
UMAP_1
CD14
FCGR3A
CD1C
Gene Expression
(normalized UMI counts)

◂**Figure 3.** Example of analysis with ShIVA. (**A**) Description of test dataset. Peripheral blood mononuclear cell (PBMC) samples from 8 human donors were stained with anti-CD45 hashtag antibodies, pooled and captured for 10× Genomics 3′v2 scRNA-seq and HTO library preparation, as described in (Stoeckius et al.[8]). Gene expression reads were processed with cellranger count to generate the cell × gene UMI count matrix, HTO reads were processed with CITE-seq-Count to generate the cell × HTO UMI count matrix. Both count matrices were downloaded from the Seurat tutorial website (https://satijalab.org/seurat/articles/hashing_vignette.html), combined into a Seurat object in R, and analyzed in ShIVA. (**B**) UMAP computed on normalized HTO counts showing the results of HTO-based sample demultiplexing in ShIVA. Each dot is a cell, colored according to HTO-based sample assignation. Doublets and Negative cells are excluded from further analysis. (**C**) Scatter plot of number of genes detected (x-axis) and percentage of transcripts from mitochondrial genes (y-axis) used for quality controls. Each dot is a cell, colored according to quality control results; only TRUE cells are conserved for further analysis. (**D**) Scree plot of Standard Deviation percentage explained by Principal Components. Only the top 50 PC are shown, after computing PCA on the top 2000 highly variable genes defined by vst method. (**E**) Gene expression based UMAP embedding computed on 30 PCs. Each dot is a cell, colored according to sample origin. (**F**) Results of Leiden clustering presented on the gene expression based UMAP embedding. Each dot is a cell, colored according to cluster identity. Clusters were annotated manually based on cluster-specific marker genes. (**G**) Violin plot of number of genes detected in cells according to cluster identity. (**H**) Gene expression based UMAP embedding and clustering of myeloid cells after focused re-analysis of clusters 1, 8 and 9 from the analysis in (**F**). The expression of discriminating marker genes *CD14*, *FCGR3A* and *CD1C* is shown as feature plots. Each dot is a cell, colored according to cluster identity (left) or gene expression levels (right).

option "Shiva project". In this option, users can load a saved ShiVA project file or choose one of the automatically saved project.

The workflow of ShIVA is the following (Fig. 2). Data is imported and demultiplexed if HTOs are used. A quality control step allows to eliminate low quality cells. The most variable genes are then selected for further analysis. Principal component analysis (PCA) is applied to select the most variable components. This step also allows to analyse the risks of experimental or technological dependence of the observed variability. A dimensionality reduction step (UMAP and/or t-SNE) is then used to allow a two-dimensional visualization of the data. In order to group the cells, a clustering step is performed. This step allows to detect the marker genes of each cluster, which can be refined by a detection step of differentially expressed genes and their functional enrichment. Once the data is ready, the user can proceed to the data exploration step, in order to visualize and quantify the data and yield observations of biological interest. The user can also import signatures of cell types or states to identify them in the data. If needed, the user can focus on part of the dataset and resume the complete analysis process on this subset of cells of interest. The complete analyses can finally be exported as binary files and HTML reports that can be used by an analysis expert to extend/complement the analyses already performed. These enriched analyses can be reloaded into ShIVA to directly use the exploration module and allow users to continue their analysis.

Table 1 summarizes for each analysis module the parameters available to user customization, the different visualizations available, and the results.

**Advantages of ShIVA.** Several tools already exist to manage and visualize single-cell datasets (Refs.[5–7], BBrowser, CellxGene, SeqGeq …). To our knowledge, no freely available solution offers the combination of high level of simplicity and deep data exploration ShIVA provides. Existing user-friendly solutions are limited in the ability for user to control, compare and explore the analysis results. Compared to others tools, the main advantages of ShIVA are (i) user experience orientation, allowing a non-expert to easily understand and control each step of the analysis, (ii) modularity, allowing to easily follow the rigorous workflow of analysis, (iii) interactivity and iterability, allowing to acquire a better understanding of the underlying biology of the dataset, (iv) exportability and reproducibility, allowing to easily share and push forward the analysis with expert collaborators.

**Example of analysis with ShIVA.** We tested ShIVA on a public dataset of human peripheral blood mononuclear cells (PBMC) from 8 donors (Fig. 3A, as described in Ref.[8]. Samples were stained with barcoded anti-CD45 hashtag antibodies prior to capture for 10× Genomics 3'v2 scRNA-seq. We downloaded gene expression and hashtag count matrices from the Seurat tutorial website (https://satijalab.org/seurat/articles/hashing_vignette.html) and combined both matrices into a Seurat object in R. The object was then loaded into ShIVA for analysis, with log normalization of the gene expression count matrix. We performed hashtag count matrix normalization and automatic demultiplexing to identify samples from the eight donors, and excluded doublets and negative events from further analysis (Fig. 3B). We then performed gene expression quality controls, excluding cells expressing more than 5% mitochondrial gene transcripts and fewer than 100 genes (Fig. 3C). We then selected 2000 highly variable genes with the vst method, and performed principal component analysis (PCA) (Fig. 3D). We selected the first 30 components for low dimension embedding and clustering. Cells from all donors were mixed in different clusters in the UMAP embedding (Fig. 3E). After clustering with Leiden method at resolution 0.8, we discriminated 10 clusters corresponding to common peripheral blood cells that we annotated manually based on cluster-specific marker genes (Fig. 3F). We visualized the number of genes detected per cell for all clusters as violin plots, highlighting that myeloid cells (clusters 1, 8 and 9) expressed more genes than lymphoid cells (Fig. 3G). For higher resolution analysis of those myeloid cell clusters, we created a new object containing only cells from clusters 1, 8 and 9, and repeated the computation of 2,000 highly variable genes, PCA, UMAP embedding and clustering. That re-analysis revealed several sub-clusters of *CD14+* monocytes, but did not reveal additional sub-clusters of *FCGR3A+* monocytes or *CD1C+* conventional dendritic cells (cDC) (Fig. 3H). The exploratory analysis of that dataset in ShIVA took approximately 2 h from loading the dataset to exporting all figure plots in vectorized pdf format.

## Discussion

In conclusion, ShIVA offers a user-friendly and user-oriented interface dedicated to the analysis of single-cell RNA-seq and single-cell CITE-seq data with an emphasis on ease of use, application of state-of-the-art methods, and reproducibility. ShIVA supports cell hashing analysis and provides great flexibility in visualization, whether by dimensionality reduction maps, boxplots, violin plots, histograms, density plots, or count tables. ShIVA is designed to give non-experts the power of state-of-the-art algorithms, allowing them to gain a deep understanding of their dataset while preparing for an optimized collaboration with data analyst experts in subsequent extensive analyses. A complete series of tutorial videos are available on YouTube (https://www.youtube.com/channel/UCJJ3Svi8AY6XGx4Y9r9G3Iw).

## Methods

ShIVA is a Shiny Application built in an R framework (v. 4.0.3). ShIVA is deployed using Docker. ShIVA uses the Seurat package (v. 4.0.0, Ref.[4]) for the main analysis workflow. HTO demultiplexing is performed with MULTI-seq[9]. Dimensionality reduction is performed using PCA (ade4 v.1.7.16, Ref.[10]), t-SNE[11] and UMAP (v.0.2.7, Ref.[12]). Clustering is performed through the methods proposed by Seurat among which the Leiden algorithm (v. 0.3.10, Ref.[13]). Visualization is performed using packages ggplot2 (v.3.3.3), plotly (v.4.9.3), pheatmap (v1.0.12) and heatmaply (v.1.2.1). Functional enrichment is performed with clusterProfiler (v.3.18.1, Ref.[14]). Complete details on the other R packages used are available in the dockerfile provided on GitHub.

## Data availability

Code and documentation are available from the GitHub repository: https://github.com/CIML-bioinformatic/ShIVA. Complete video tutorials are available on YouTube: https://www.youtube.com/channel/UCJJ3Svi8AY6XGx4Y9r9G3Iw. Docker image is available on DockerHub: https://hub.docker.com/r/cb2m/shiva.

Received: 27 February 2023; Accepted: 19 August 2023
Published online: 01 September 2023

## References

1. Svensson, V., Da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, 073 (2020).
2. Perkel, J. M. Single-cell analysis enters the multiomics age. *Nature* **595**, 614–616 (2021).
3. Eisenstein, M. Single-cell RNA-seq analysis software providers scramble to offer solutions. *Nat. Biotechnol.* **38**, 254–257 (2020).
4. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
5. Jiang, A., Lehnert, K., You, L. & Snell, R. G. ICARUS, an interactive web server for single cell RNA-seq analysis. *Nucleic Acids Res.* **50**, W427–W433 (2022).
6. Hasanaj, E., Wang, J., Sarathi, A., Ding, J. & Bar-Joseph, Z. Interactive single-cell data analysis using cellar. *Nat. Commun.* **13**, 1998 (2022).
7. Moreno, P. *et al.* User-friendly, scalable tools and workflows for single-cell RNA-seq analysis. *Nat. Methods* **18**, 327–328 (2021).
8. Stoeckius, M. *et al.* Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
9. McGinnis, C. S. *et al.* MULTI-seq: Sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
10. Dray, S. & Dufour, A.-B. The ade4 package: Implementing the duality diagram for ecologists. *J. Stat. Soft.* **22**, 4 (2007).
11. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
12. McInnes, L., Healy, J., Saul, N. & Großberger, L. U. M. A. P. Uniform manifold approximation and projection. *JOSS* **3**, 861 (2018).
13. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
14. Yu, G., Wang, L., Han, Y. & He, Q. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**(5), 284–287 (2012).

## Acknowledgements

The authors thank Guillaume Voisinne for his help in designing the basic architecture of the software. They thank Bernard Malissen for supporting the project through human resources. They would like to thank all our beta-testers from Centre d'Immunologie de Marseille-Luminy, Institut de Neurobiologie de la Méditérranée, Institut de Biologie du Développement de Marseille, Centre de Recherche en Cancérologie de Marseille, Centre de Recherche en CardioVasculaire et Nutrition, Theories and Approaches of Genomic Complexity, Unité de biologie Moléculaire, Cellulaire et du Développement, Institut de Recherche Saint Louis, Institut de Biologie Paris-Seine, RESTORE, and Institut de la Vision.

## Author contributions

S.J., P.M. and L.S. designed the project, produced the feature specifications and supervised the development. R.A., M.A. and S.C. developed the software. P.M. and L.S. wrote the manuscript with input from all authors. All authors read and approved the manuscript.

## Funding

The ShIVA project has been funded by the European Union's Horizon 2020 Framework Programme (Grant Agreement 787300 [BASILIC] to B. Malissen), the CENTURI Tech Transfer program and the Avesian ITMO GGB Amorçage program.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to P.M. or L.S.

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

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023