scRNAseq_KNIME workflow: A Customizable, Locally Executable, Interactive and Automated KNIME workflow for single-cell RNA seq

Kausar Samina, Muhammad Asif, Anaïs Baudot

scRNAseq_KNIME workflow: A Customizable, Locally Executable, Interactive and Automated KNIME workflow for single-cell RNA seq

Samina Kausar[1,†], Muhammad Asif [2,†,] *, and Anaïs Baudot[1,3,4] *

[1]Aix Marseille Université, INSERM, MMG, Marseille, France,

[2]Biomedical Data Science Lab, Department of Bioinformatics and Biotechnology, Government College University Faisalabad, Faisalabad 38000, Pakistan,

[3]CNRS, Marseille, France

[4] Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain.

*Correspondence:

Anaïs Baudot: anais.baudot@univ-amu.fr

Muhammad Asif: muhasif123@gmail.com

[†]Equal contributing authors

Abstract

Summary: Single-cell RNA sequencing (scRNA-seq) is nowadays widely used to measure gene expression in individual cells, but meaningful biological interpretation of the generated scRNA-seq data remains a complicated task. Indeed, expertise in both the biological domain under study, statistics, and computer programming are prerequisite for thorough analysis of scRNA-seq data. However, biological experts may lack data science expertise, and bioinformatician's limited understanding of the biology may lead to time-consuming iterations.

A user-friendly and automated workflow with possibility for customization is hence of a wide interest for both the biological and bioinformatics communities, and for their fruitful collaborations. Here, we propose a locally installable, user-friendly, interactive, and automated workflow that allows the users to perform the main steps of scRNA-seq data analysis. The interface is composed of graphical entities dedicated to specific and modifiable tasks. It can easily be used by biologists and can also serve as a customizable basis for bioinformaticians.

Availability and implementation: The workflow is developed in KNIME; its tasks were defined by R scripts using KNIME R nodes. The workflow is publicly available at https://github.com/Saminakausar/scRNAseq_KNIME.

Contact: anais.baudot@univ-amu.fr; muhasif123@gmail.com

Introduction and motivation

Single-cell RNA sequencing (scRNA-seq) aims to quantify the gene expression of individual cells. Each run of scRNA-seq generates large amounts of noisy and sparse data, making rigorous data analysis and biological interpretation challenging.

The main steps of a scRNA-seq analysis are quality check, normalization, denoising and dimensionality reduction, clustering of cells, and identification of differentially expressed genes. All these steps are usually coupled with extensive data visualization. In bioinformatics, R and Python languages are frequently used to develop pipelines for the analysis of biological datasets. The Seurat [1] and Scanpy [2] packages, developed in R and Python, respectively, have been widely adopted for scRNA-seq data analysis. These packages perform the different tasks necessary for scRNA-seq data analysis. For example, they allow the users to normalize raw counts (gene/feature counts matrix), apply Principal Component Analysis (PCA) for dimension reduction and different clustering methods to identify clusters of cells. However, using such

packages requires prior knowledge of R or Python. This prerequisite expertise in computer programming hampers the engagement of biological experts in scRNA-seq data analysis, which limits reliable and robust biological interpretations.

In this context, several web tools have been developed [3–7]. These tools are accessible but present some limitations. First, some tools are restricted to few steps of the scRNA-seq analysis pipeline. For example, visnormsc performs only the normalization [8]. Second, web tools follow a strict and previously-defined architecture, thus preventing the users with expertise in computer programming customizing or extending the tool. Moreover, web tools process and analyze data on their hosted servers and often lack well-defined policy for data protection and integrity, which raises concerns related to data security. In addition, scRNA-seq dataset are large and can crash web applications, forcing the users to re-run the analyses from the beginning.

Frameworks such as snakemake [9], nextflow (https://www.nextflow.io/index.html), Galaxy (https://usegalaxy.org/), and Konstanz Information Miner (KNIME) [10] can also be employed to implement workflows composed of several data analysis steps. For example, WASP is a snakemake workflow dedicated to scRNA-seq data analysis [11]. Importantly, the installation, execution, and modification of such frameworks require programming expertise. KNIME is a free and open-source data analysis framework with graphical interface. KNIME allows the creation of locally installable and reusable workflows for big data analysis and visualization. For instance, Kausar et al. proposed a KNIME workflow for machine learning in drug design [12].

Here, we propose scRNAseq_KNIME workflow, a KNIME-based workflow encompassing the main steps of scRNAseq analysis. scRNAseq_KNIME workflow is interactive and automated. It is also executable locally, thus eliminating the concern of data security and integrity. Overall, scRNAseq_KNIME workflow can be used by researchers lacking programming skills to run a complete scRNAseq analysis. It can simultaneously be exploited by bioinformaticians as a basis to customize and extend each step of scRNA-seq data analysis.

Implementation and scRNAseq_KNIME workflow architecture

scRNAseq_KNIME workflow is developed in KNIME. In KNIME, a specific task, for example reading/writing files, is defined by introducing a node. A node is both a graphical entity and a basic processing unit. Some nodes are associated with predefined tasks, for example a node

dedicated to reading CSV files. Other nodes are associated with tasks that can be defined by writing scripts in different languages, such as R, Groovy, Matlab and Python.

The scRNAseq_KNIME workflow is implemented with a modular structure (Supplementary figure 1), in which each module is composed of KNIME predefined nodes and nodes defined by custom R scripts (Figure 1A). The nodes are connected to each other through edges defining their relationships. The input module takes 10x genomics files (barcode, genes/features and matrix files) as input. The Quality Control (QC) module then assesses the quality of input data and applies user-defined cutoffs to exclude low quality data. The data normalization and denoising module prepares data for further analysis and identifies highly variable features. The dimension reduction and clustering module allow users to apply linear and non-linear dimension reduction approaches and identify clusters of cells. In the next step, various statistical tests can be applied on the identified clusters to highlight important genes. The next module allows marker-based cell type annotation of the clusters. Interactive visualizations are provided by the final visualization module.

scRNAseq_KNIME workflow provides extensive and interactive visualization. For example, clusters of cells can be visualized in 2D and 3D interactive plots. Many other data visualization options including VlnPlot, FeaturePlot, DotPlot, Heatmap, and interactive FeaturePlot are provided. Results of downstream analyses such as identification of differentially expressed genes are also provided through interactive data tables. To showcase the scRNAseq_KNIME workflow and its usage, we applied it to the analysis of the single-cell RNAseq dataset of Peripheral Blood Mononuclear Cells (PBMC), downloaded from 10x Genomics (Figure 1 B).

Overall, the scRNAseq_KNIME workflow implements the main steps required for scRNA-seq data analysis. In addition, each node of scRNAseq_KNIME workflow is interactive, and the users can test multiple parameters (Figure 1A and supplementary figure 2). Importantly, the R nodes can further be customized by modifying the R scripts.

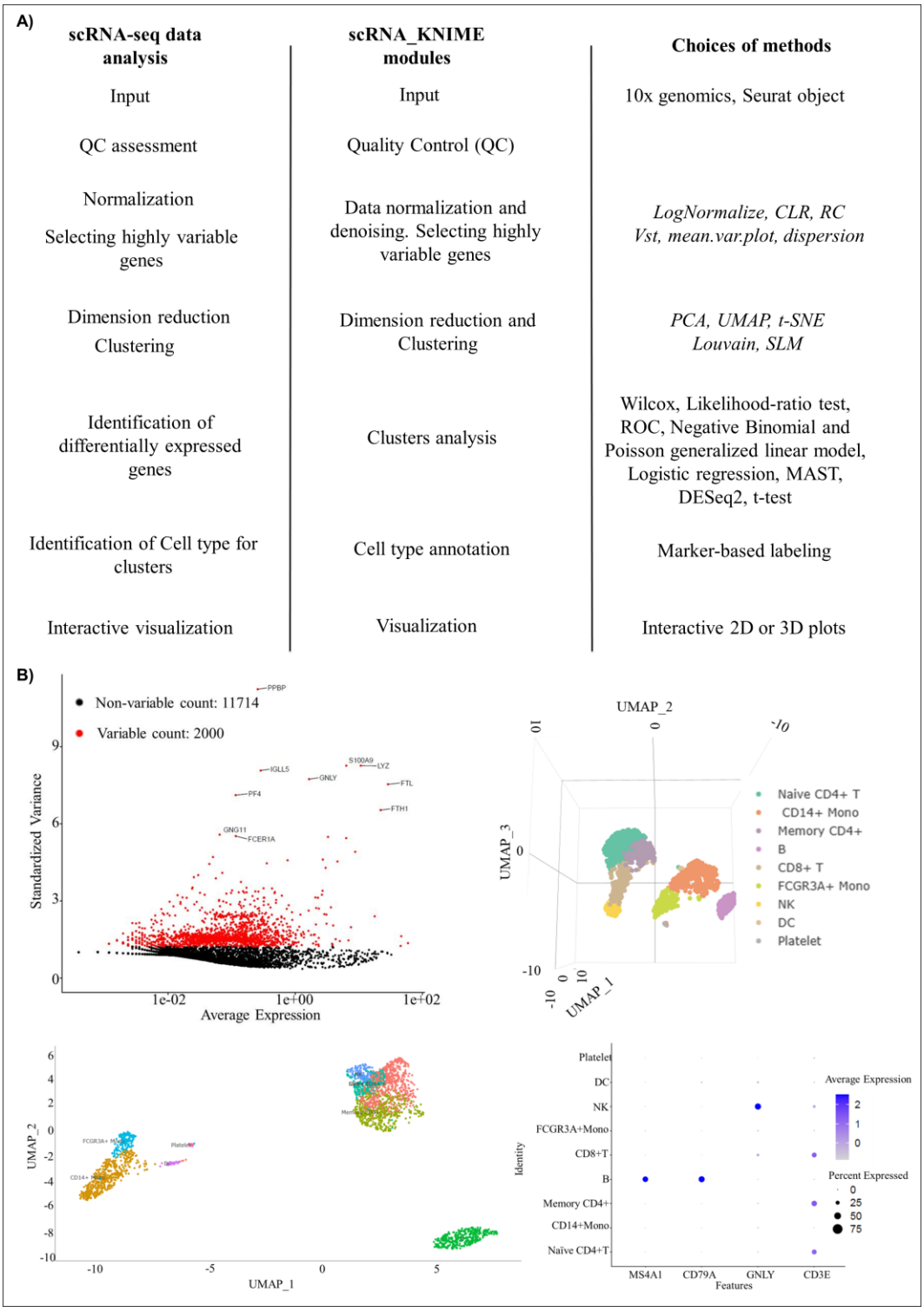

Figure 1: A) scRNAseq_KNIME workflow vs main steps of scRNA-seq data analysis workflow. MDS: Multidimensional scaling. B) Examples of interactive visualizations generated by the scRNAseq_KNIME workflow

The main features of scRNAseq_KNIME workflow are:

- Easy installation and execution: The workflow is developed with the KNIME platform, which comes with extensive documentation for its installation and execution. We further provide detailed documentation for running scRNAseq_KNIME workflow on Mac, Linux and Windows operating systems.
- Completeness: scRNAseq_KNIME workflow covers the main steps of scRNA-seq data analysis, allowing users (biologists and bioinformaticians) to run complete scRNA-seq data analysis and generate publication-ready plots.
- Possibility of customization: The scRNAseq_KNIME workflow modules are composed of R nodes containing R scripts. Users with expertise in R can customize R nodes by modifying the existing code.
- Modularity: Like snakemake, scRNAseq_KNIME workflow modules are independent from each other, meaning a change in one module of workflow does not require processing of previous modules.

scRNAseq_KNIME workflow is a locally installable, executable, and user-friendly framework that allows biologists to perform scRNA data analysis without knowing computer programming and statistical methods. The main steps of scRNA data analysis in scRNAseq_KNIME workflow can be customized by bioinformaticians by altering the code in R nodes of workflow.

Funding

This work was supported by the AFM-Téléthon and by the French National Research Agency (ANR)

References

1. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021. https://doi.org/10.1016/j.cell.2021.04.048.

2. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. Genome Biol. 2018. https://doi.org/10.1186/s13059-017-1382-0.

3. Patel M V. iS-CellR: A user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty517.

4. Feng D, Whitehurst CE, Shan D, Hill JD, Yue YG. Single Cell Explorer, collaboration-driven

tools to leverage large-scale single cell RNA-seq data. BMC Genomics. 2019. https://doi.org/10.1186/s12864-019-6053-y.

5. Serra A, Serra A, Saarimäki LA, Saarimäki LA, Fratello M, Fratello M, et al. BMDx: A graphical Shiny application to perform Benchmark Dose analysis for transcriptomics data. Bioinformatics. 2020. https://doi.org/10.1093/bioinformatics/btaa030.

6. Lawlor N, Marquez EJ, Lee D, Ucar D. V-SVA: An R Shiny application for detecting and annotating hidden sources of variation in single-cell RNA-seq data. Bioinformatics. 2020. https://doi.org/10.1093/bioinformatics/btaa128.

7. Aussel R, Asif M, Chenag S, Jaeger S, Milpied P, Spinelli L. ShIVA – A user-friendly and interactive interface giving biologists control over their single-cell RNA-seq data. bioRxiv. 2022;:2022.09.20.508636.

8. Tang L, Zhou N. visnormsc: A Graphical User Interface to Normalize Single-cell RNA Sequencing Data. Interdiscip Sci – Comput Life Sci. 2018. https://doi.org/10.1007/s12539-017-0277-9.

9. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research. 2021. https://doi.org/10.12688/f1000research.29032.1.

10. Fillbrunn A, Dietz C, Pfeuffer J, Rahn R, Landrum GA, Berthold MR. KNIME for reproducible cross-domain analysis of life science data. Journal of Biotechnology. 2017.

11. Hoek A, Maibach K, Özmen E, Vazquez-Armendariz AI, Mengel JP, Hain T, et al. WASP: a versatile, web-accessible single cell RNA-Seq processing platform. BMC Genomics. 2021. https://doi.org/10.1186/s12864-021-07469-6.

12. Kausar S, Falcao AO. An automated framework for QSAR model building. J Cheminform. 2018. https://doi.org/10.1186/s13321-017-0256-5.