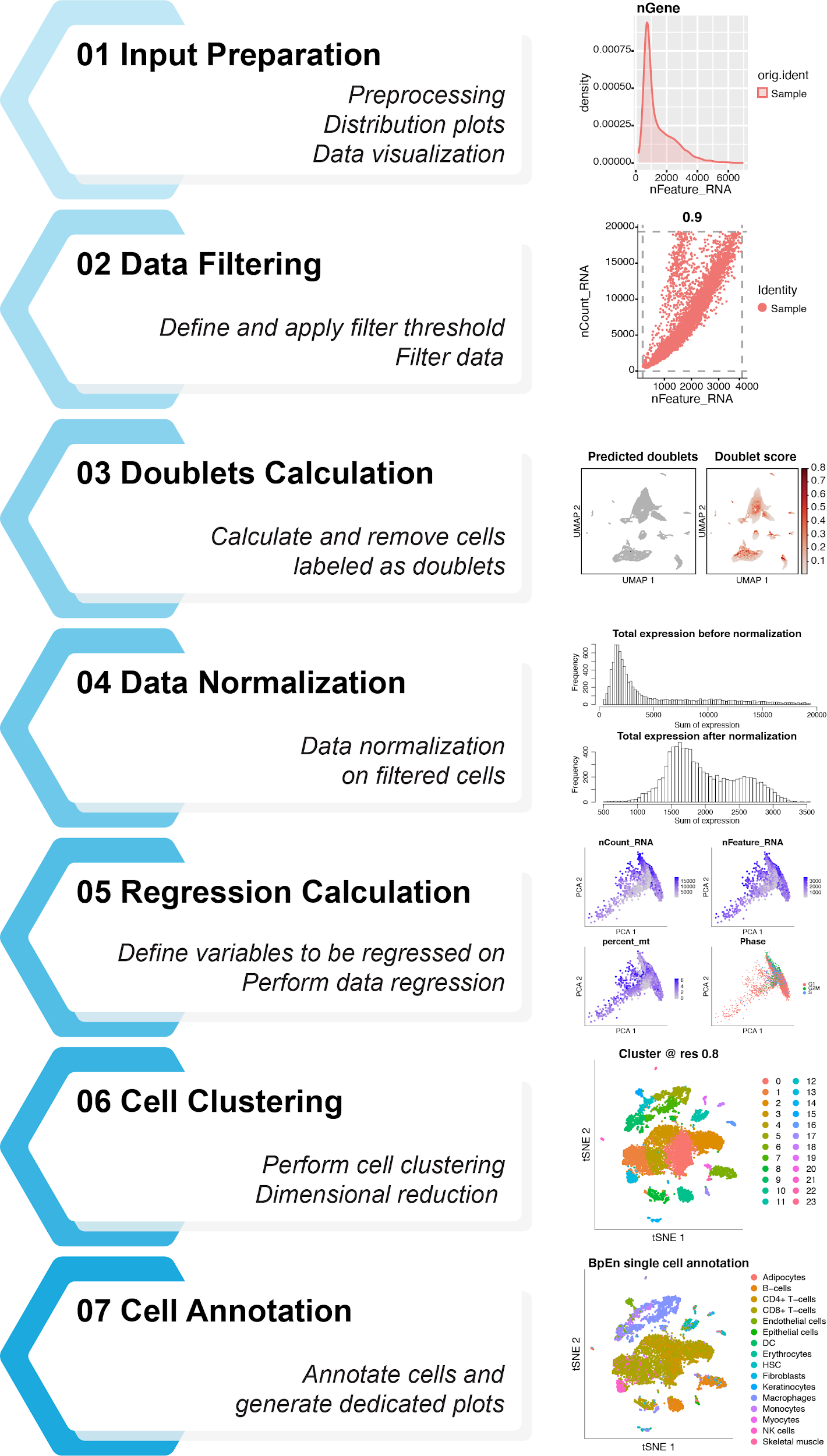
|  |
| --- |
| *Application Note*  **PoPsicle: a flexible R Package for the preprocessing and quality control analysis of single cell RNA-seq data.**  Jimmy Caroli1,2, Andrea Grilli1, Francesco Grandi1, Martina Dori1, Oriana Romano1, Mattia Forcato1, and Silvio Bicciato1\*  1 Department of Life Sciences, University of Modena and Reggio Emilia – Modena, Italy  2 Department of Drug Design and Pharmacology, University of Copenhagen - Copenhagen, Denmark  \*To whom correspondence should be addressed.  Associate Editor: XXXXXXX  Received on XXXXX; revised on XXXXX; accepted on XXXXX  **Abstract**  **Summary:** Here we present PoPsicle, a flexible R package aimed at providing most of the necessary quality controls and preliminary information for complete and detailed analysis of single cell RNA-seq data. In its workflow, PoPsicle integrates all the pivotal steps necessary to assess sample quality, apply data tailored filters and perform doublets calculation and annotation analysis. PoPsicle builds on the R package Seurat (v3.5) to perform most of its task, and integrates an R implementation of the Scrublet python module for the analysis of doublets presence. Furthermore, PoPsicle grants the possibility to investigate and analyze both human and mouse data, using organism specific integrated annotation packages. We believe that the flexible and detailed analysis pipeline of PoPsicle will provide help in both single cell data preprocessing and analysis.  **Availability and implementation:** PoPsicle is written in R language and is released under a GPL License. It can be downloaded from Github (@bicciatolab)  **Contact**: [silvio.bicciato@unimore.it](mailto:silvio.bicciato@unimore.it)  **Supplementary information:** Supplementary information is available at *Bioinformatics* online. |

## **1 Introduction**

Single cell RNA-seq experiments are currently becoming one of the most common biological assay, securing increasing importance in the modern medical research field **(Navin *et al.*, 2011; Pollen *et al.*, 2014; Treutlein *et al.*, 2017; Tanay and Regev, 2017)**. These experiments grant the ability to deeply investigate the genomic structure of every single cell sampled, uncovering connections and intricacies previously unknown or superseded due to less sensible methods used, leading to improved understanding of molecular regulations, such as in cancer environment, for instance **(Giustacchini *et al.*, 2017)**. While these experiments are able to shed light on previously overlooked information, dedicated analysis remains extremely complex to be performed, requiring both high computational skills and thorough biological knowledge. Recently, several methods **(Hwang *et al.*, 2018)** and pipelines **(Franzén and Björkegren, 2020; Mohanraj *et al.*, 2020)** have been designed to help researchers solve the conundrums scattered throughout the single cell data analyses. Specifically, some dataset may need to refine analysis parameters during the analysis itself, making the definition of an *a priori* pipeline of investigation a challenging task. Furthermore, different filtering steps and approaches are currently upon debate in the literature, meaning that no gold standard analysis pipeline has been set yet. Thus, to solve some of these issues we present PoPsicle, a flexible and interactive R package containing an analysis pipeline for the preprocess and investigation of single cell RNA sequencing data. PoPsicle builds on the Seurat v3 R package **(Stuart *et al.*, 2019)** to perform most of its analysis, and provides several built in functions for plotting the data, applying filters upon visual data investigation, performing clustering analysis and annotating results obtained.

## **2 Package workflow description**

PoPsicle is designed to provide a simple yet efficient analysis pipeline throughout **seven** sequential steps, each one accessible via a dedicated function call of the package.

**Figure 1:** complete PoPsicle workflow. PoPsicle is structured in seven sequential steps, handling multiple quality control checks, providing filtering options and performing the most common data analyses on single cell RNA-seq samples. Figure reports each single step along a brief description of the procedure involved and an example of the several plots generated throughout the analysis.

In the first (i) step, data are preprocessed and number of cells, number of genes per cell and gene expression values are plotted in dedicated graphs to be inspected by the user. Preprocessing is performed using the Seurat package, generating a Seurat object that will be maintained throughout the analysis pipeline. Each single function returns a modified Seurat object to be further investigated. The second step (ii) consists in the filtering of the Seurat object, allowing users to provide personal thresholds based on the plots generated previously. After filter application, PoPsicle performs doublet calculations (iii), granting the ability to retain or eliminate putative doublets in the analyzed dataset. Calculations in this step are performed using an R transposition of the Scrublet python module **(Wolock *et al.*, 2019).** Step four (iv) performs data normalization, which is directly followed by the regression analysis (v): in this step the user may select on which variable to perform the regression (if any), by inspecting both UMAP and tSNE plots provided after normalization. After the regression analysis, PoPsicle performs cell clustering (vi), plotting the cluster distribution in dedicated graphs. Finally, the last step (vii) of analysis handles cell annotation. This analysis is carried out according to the organism investigated: in case of human data, SingleR is employed using as reference the cell populations determined by the Blueprint and Encode consortia and from the Human Primary Cell Atlas **(Aran *et al.*, 2019)**; conversely, when analyzing mouse data, the scMCA package is invoked, using as reference the cell populations from Mouse Cell Atlas **(Schaum *et al.*, 2018)**. Dedicated folders containing plots and reports for each step are automatically generated during the analysis, and on screen messages will display their location to the user along statistics of the processed data.

## **2.1 Comparison with other tools**

To investigate and assess PoPsicle’s quality controls and analysis pipeline, we compared our tool with scCancer **(Guo *et al.*, 2020)**, an R package which focuses on processing and analyzing scRNA-seq data for cancer research. The analysis pipeline provided by scCancer features only two sequential steps: data preprocessing (*scStatistics*) and the complete calculation process (*scAnnotation*). Even though this structure greatly simplifies the overall analysis process, it lacks the possibility to tailor data parameters for each single analysis performed, limiting its usability by requiring several runs to investigate more accurate setups. For sake of comparison, when testing PoPsicle performances we opted to analyze the same data that were investigated by scCancer authors in their original manuscript. PoPsicle and scCancer provide very similar quality controls checks and report consistent plots and data when using the same filters for both tools. While PoPsicle generates a sample directory with subdirectories named after the analysis step performed, scCancer deploys all of his outputs into a single folder, along with an HTML report file describing plots and results. Given these insights, the main upside of PoPsicle resides in its flexibility: several different single functions tailored for specific purposes offer researchers a more curated analysis pipeline, along with the possibility to modify parameters and rerun specific analysis rather than the whole workflow. Furthermore, computational comparisons in terms of time and memory usage were very similar between the two tools (xxx for PoPsicle, xxx for scCancer).

## **2.2 Installation and usage**

PoPsicle is written in R language, it has been developed in version 3.6.3 and requires several packages to be run, listed in the Supplementary Materials. PoPsicle is freely available at **Github/bicciatolab**, where it can be directly installed using devtools in R. Moreover, we provide an Anaconda yaml file, granting the possibility to create a simple yet complete working environment to perform analysis with PoPsicle via the Anaconda platform.

## **3 Conclusion**

Here we present PoPsicle, a flexible R package designed for addressing quality control checks and perform analysis on single cell RNA sequencing data. PoPsicle provides a simple and efficient workflow for a complete and exhaustive pipeline of analysis. Its structure comprises seven different functions designed to perform dedicated tasks for the investigation of single cell data. This feature, coupled to its lightweight computational requirements and fast calculation time, characterizes PoPsicle as an useful tool for analyzing both 10X and smart-seq single cell RNA sequencing data. We believe that its flexibility and ease of use would greatly benefit researchers in addressing new and unknown biological connections from single cell experiments.

*Conflict of interest*: none declared.

## **References**

Aran,D. *et al.* (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.*, **20**, 163–172.

Franzén,O. and Björkegren,J.L.M. (2020) alona: a web server for single-cell RNA-seq analysis. *Bioinformatics*, **36**, 3910–3912.

Giustacchini,A. *et al.* (2017) Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat Med*, **23**, 692–702.

Guo,W. *et al.* (2020) scCancer: a package for automated processing of single-cell RNA-seq data in cancer. *Briefings in Bioinformatics*, bbaa127.

Hwang,B. *et al.* (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, **50**, 1–14.

Mohanraj,S. *et al.* (2020) CReSCENT: CanceR Single Cell ExpressioN Toolkit. *Nucleic Acids Res*, **48**, W372–W379.

Navin,N. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**, 90–94.

Pollen,A.A. *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, **32**, 1053–1058.

Schaum,N. *et al.* (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**, 367–372.

Stuart,T. *et al.* (2019) Comprehensive Integration of Single-Cell Data. *Cell*, **177**, 1888-1902.e21.

Tanay,A. and Regev,A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature*, **541**, 331–338.

Treutlein,B. *et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.

Wolock,S.L. *et al.* (2019) Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst*, **8**, 281-291.e9.