*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Applications Note

|  |
| --- |
| Sequence analysis  **rCASC: reproducible Classification Analysis of Single Cell sequencing data**  Luca Alessandrì2, Marco Beccuti1,\*, Maddalena Arigoni2, Martina Olivero3, Greta Romano1, Gennaro De Libero4, Luigia Pace5, Francesca Cordero1,# and Raffaele A Calogero2,#  1Department of Computer Sciences, University of Torino, Corso Svizzera 185, Torino, Italy, 2Department of Molecular Biotechnology and Health Sciences, University of Torino, Via Nizza 52, Torino, Italy, 3Department of Oncology, University of Torino, SP142, 95, 10060 Candiolo TO, Italy, 4Department Biomedizin, University of Basel, Hebelstrasse 20, 4031 Basel, Switzerland, 5IIGM, Via Nizza 52, Torino, Italy  #Both authors equally supervised the present work, \*Corresponding author  Associate Editor: XXXXXXX  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Abstract  **Summary:** Single-cell RNA sequencing has emerged as an essential tool to investigate cellular heterogeneity, and highlighting cell sub-population specific signatures. Nowadays, dedicated and user-friendly bioinformatics workflows are required to exploit the deconvolution of single-cells transcriptome. Furthermore, there is a growing need of bioinformatics workflows granting both functional, i.e. saving information about data and analysis parameters, and computation reproducibility, i.e. storing the real image of the computation environment. Here, we present rCASC a modular RNAseq analysis workflow allowing data analysis from counts generation to cell sub-population signatures identification, granting both functional and computation reproducibility.  **Availability and Implementation:** rCASC is part of the reproducible-bioinfomatics.org project. rCASC is a docker based application controlled by a R package available at https://github.com/kendomaniac/rCASC.  **Contact:** beccuti@di.unito.it  **Supplementary information:** Supplementary data are available at Bioinformatics online. |

# Introduction

Single cell analysis is instrumental to understand the functional differences existing between cells within a tissue. Individual cells of the same phenotype are commonly viewed as identical functional units of a tissue or organ. However, single cells sequencing results (Buettner, et al., 2015) suggest the presence of a complex organization of heterogeneous cell states producing together system-level functionalities. Single cell analysis focuses on the understanding differences characterizing any cell within a population of cells. A mandatory element of single cell RNAseq is the availability of dedicated bioinformatics workflows.

rCASC addresses the problem of functional and computational reproducibility, which is becoming a very important topic, because of the “Data Reproducibility Crisis” (Allison, et al., 2018).

# Methods

rCASC is developed under the umbrella of the Reproducible Bioinformatics Project (www.reproducible-bioinformatics.org), which is an open-source community aiming to develop reproducible bioinformatics workflows. Each module of rCASC is implemented in a docker container, and it is compliant with the rules proposed by Sandve (Sandve, et al., 2013) to guarantee reproducibility. The key elements of rCASC workflow are shown in Fig. 1A, and the main functionalities are summarized below.

**Data preprocessing:** rCASC allows processing of fastq to generate a cell count matrix (Supplementary Section 2). The most relevant preprocessing modules of rCASC (Supplementary Section 3) allow removal of low quality cells, data normalization, and detection/removal of cell cycle bias.

# 

Figure 1: A) rCASC modules, outputs for the relevant steps of the workflow are shown in capital letters in parenthesis. B) Depicting the range of clusters to be investigated. C) Cell stability score for the range of clusters in B. D) Clustering results for the most homogeneous cell stability score cluster set, i.e. five clusters. E) Heatmap of prioritized clusters-specific genes. Color bar refers to clusters colors in D. F) Cell stability score in E.

**Cell heterogeneity analysis:** The optimal number of cells partitions is detected inducing perturbations in the structure of the cell data set, i.e. removing a random subset of cells and repeating the clustering. The rational of this approach is that a robust cluster should contain a stable set of cells independently by the induced perturbations. A graph-based community detection method (https://github.com/ppapasaikas/griph), is used to identify the range of number of clusters due to dataset perturbations (Fig. 1B, Supplementary Section 4). Then, the range of number of clusters is probed via a clustering framework learning a sample-to-sample similarity measure from expression data (Wang, et al., 2017). A cell stability score (Supplementary Section 5), indicating the fraction of bootstraps in which a cell is allocated in a specific cluster, is used to identify the optimal number of clusters for the cell sub-populations representation (Fig. 1C). Cells are then plotted in each cluster with a specific symbol indicating its stability (Fig. 1D).

**Clusters specific feature selection:** The identification of clusters specific signatures is addressed with two different methods (Supplementary Section 6). The ANOVA-like method, from edgeR Bioconductor package and SIMLR gene prioritization are implemented. Gene signatures are visualized by heatmap.

**GUI:** rCASC functions are implemented within 4SeqGUI, which makes the analysis user-friendly and suitable for users lacking of scripting knowledge.

# Results

The main objective of rCASC is providing instruments for reproducible and robust analysis of single cells data. The comparison of rCASC with other four single-cell analysis workflows (Supplementary Section 8) indicate that rCASC provides unique features: i) the cell stability score, providing a better estimation of clusters stability with respect to other measurements as silhouette plot (Supplementary Section 5, Fig. 23 A,B). ii) The core clustering tool SIMLR outperforms many of the methods implemented in the other published workflows. iii) The modularity structure allows the implementation of other methods. iv) Functional and computational reproducibility is guarantee.

# Conclusion

In conclusion, rCASC is a workflow with valuable new features that could help researchers in defining cells sub-populations and detecting sub-population specific markers, under the umbrella of data reproducibility.

Funding

This work has been supported by the EPIGEN FLAG PROJECT

*Conflict of Interest:* none declared.

References

Allison, D.B., Shiffrin, R.M. and Stodden, V. Reproducibility of research: Issues and proposed remedies. *Proceedings of the National Academy of Sciences of the United States of America* 2018;115(11):2561-2562.

Buettner, F.*, et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology* 2015;33(2):155-160.

Sandve, G.K.*, et al.* Ten simple rules for reproducible computational research. *PLoS computational biology* 2013;9(10):e1003285.

Wang, B.*, et al.* Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature methods* 2017;14(4):414-416.