rCASC: reproducible Classification Analysis of Single Cell Sequencing Data

[2]Luca Alessandrì, [1]Marco Beccuti, [2]Maddalena Arigoni, [3]Martina Olivero, [1]Greta Romano, [4]Gennaro De Libero, [5]Luigia Pace, [1]Francesca Cordero and [2]Raffaele A Calogero

[1]Department of Computer Sciences, University of Torino, Corso Svizzera 185, Torino, Italy, [2]Department of Molecular Biotechnology and Health Sciences, University of Torino, Via Nizza 52, Torino, Italy, [3]Department of Oncology, University of Torino, SP142, 95, 10060 Candiolo TO, Italy, [4]Department Biomedizin, University of Basel, Hebelstrasse 20, 4031 Basel, Switzerland, [5]IIGM, Via Nizza 52, Torino, Italy""

Abstract

Summary: Recent technological advances made possible to obtain genome-wide transcriptome data from single cells using high-throughput sequencing. Single-cell RNA sequencing has emerged as an essential tool to investigate cellular heterogeneity, identifying cell subsets linked to a particular phenotype and highlighting cell sub-population specific signatures. Dedicated bioin-formatics workflows are required to exploit the deconvolution of single-cells transcriptome and for the organization of tissue cells in clusters. A critical issue of these bioinformatics workflows is to guarantee both functional, i.e. the information about data and the utilized tools are saved in terms of meta-data, and computation reproducibility, i.e. the real image of the computation envi-ronment used to generate the data is stored. Here we present RCASC a RNAseq analysis work-flow dedicated to the dissection of tissues cell organization granting both functional and compu-tation reproducibility. RCASC workflow embeds different modules from counts table generation to cell sub-population specific signatures identification. **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.

Introduction

Single cell analysis is instrumental to understand the functional dif-ferences 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 suggests the presence of a complex organization of heterogeneous cell states pro-ducing together system-level functionalities (Buettner, et al., 2015). Single cell analysis focuses on the understanding the differences char-acterizing any cell within a population of cells. A mandatory element of single cell RNAseq is the availability of dedicated bioinformatics work-flows. In the

last few years a lot of tools have been developed for the identification of tissue cell sub-populations (Hwang, et al., 2018). How-ever, sub-population identification might require some preprocessing steps, depending on the sequencing technology in use and on some spe-cific cell characteristics, e.g. cell state (resting, cycling, dead cells). Furthermore, after cell partitioning, extra steps are required to identify cell sub-population specific signatures. RCASC, Cluster Analysis of Single Cells, allows direct processing of 10XGenomics, inDrop and whole transcripts single-cell sequences from fastq till the definition of cell sub-population specific signatures. Furthermore, 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).

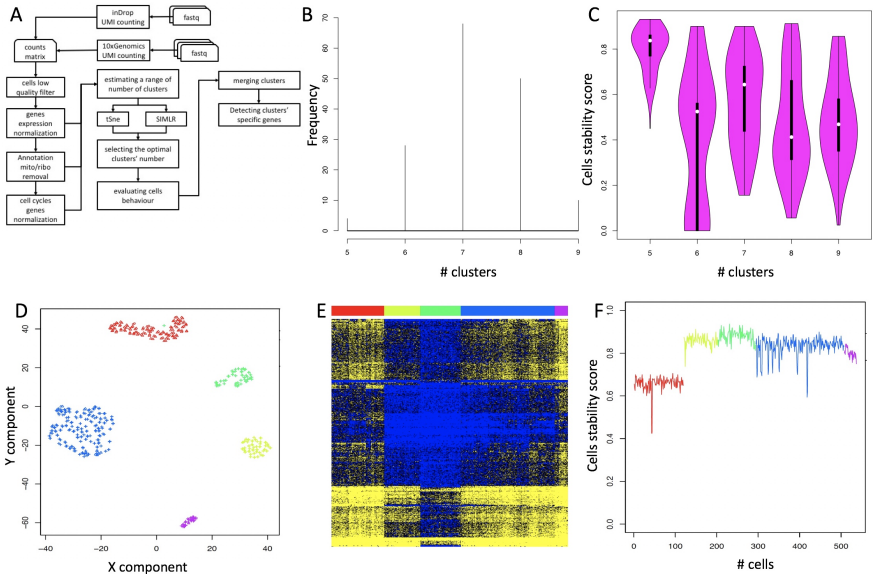

Figure 1: rCASC. A) rCADC workflow. B) Definining the range of clusters to be analysed with SIMLR/tSne. C) Cell stability score within the selected range of clusters. D) SIMLR analysis with bootstaps of the optimal number of clusters, i.e. 5. E) Heatmap of prioritized clusters-specific genes. F) Cell stability score within clusters-specific genes.

Methods

rCASC is developed within the umbrella of the Reproducible Bioinfor-matics Project (www.reproducible-bioinformatics.org), which is an open-source community aiming to develop reproducible bioinformatics workflows. Each module of RCASC workflow 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 the RCASC workflow are shown in Fig. 1A and the main functionalities of RCASC are summarized below.

Data preprocessing: RCASC embeds the direct processing of fastq derived by 10XGenomics and inDrop platforms to generate a cell count matrix annotated using ENSEMBL gene model (Supplementary Materi-al Section 2). Furthermore, any cell count matrix, using ENSEMBL gene model, can be processed within RCASC. The most relevant pre-processing modules of RCASC allow the visualization of the numbers of genes detected in each cell with respect to the cells total reads (Sup-plementary Material Section 3.2), the removal

of low quality cells using the Lorenz statistic (Diaz, et al., 2016) (Supplementary Material Sec-tion 3.3), the removal of ribosomal and mitochondrial genes and the association of gene symbol to the ENSEMBL gene identifier (Supple-mentary Material Section 3.4), data normalization (Bacher, et al., 2017) (Supplementary Material Section 3.6), detection of possible cell cycle bias (Liu, et al., 2017) and removal of such effects from the data (Barron and Li, 2016) (Supplementary Material Section 3.8).

Cell heterogeneity analysis: The optimal cells partitioning 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 of cells should contains the same set of cells independently by the perturbation of the overall dataset. The bootstrapped dataset is analyzed with a graph-based community detec-tion method (https://github.com/ppapasaikas/griph), allowing the identi-fication of the range of number of clusters observable perturbing the cells dataset structure (Fig. 1B, Supplementary Material Section 4). Then, the range of number of clusters is probed using SIMLR (Wang, et al., 2017), a k-mean algorithm based on a framework learning a sample-to-sample similarity measure from expression data observed for heterogenous samples. A cell stability score (Supplementary Material Section 5), indicating the fraction of bootstraps in which a cell is allo-cated in a specific cluster, is used to identify the ideal number of clus-ters for the cell sub-populations representation (Fig. 1C) and then cells are plotted in each cluster with a specific symbol indicating its stability (Fig. 1D). Furthermore, the shuffling of unstable cells between nearby clusters can be visualized in a video in which each bootstrap is a frame of a video (Supplementary Material Section 5.1).

Clusters specific feature selection: The identification of clusters spe-cific signatures is addressed with two different options. In case it is available a reference cluster, e.g. in a cells activation experiment it could be the cluster of resting cells undergoing to activation by an ex-ternal stimulus ((Pace, et al., 2018)), the ANOVA-like method from edge-R (Robinson, et al., 2010) is implemented (Supplementary Materi-al Section 6.1). In case a reference cluster is not available SIMLR (Wang, et al., 2017) provides a gene prioritization, measuring how gene expression values across cells correlate with the learned cell-to-cell similarity. This information combined with dataset bootstraps allows the identification of genes that are the main players in clusters organi-zation (Supplementary Material Section 6.2). These genes can be then visualized with a supervised heatmap ordering cells according to the belonging cluster (Supplementary Material Section 6.3) 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 the identification of the most robust partitioning of cell sub-populations, i.e. clusters. The cluster's robust-ness is evaluated measuring the persistence of cells in a cluster as con-sequence of jackknife resampling from the full set of cells data (Cells stability score, Supplementary Material Section 5, Fig. 22), which pro-vides a better estimation of clusters stability with respect to other meas-urements as the silhouette plot (Supplementary Material Section 5, Fig. 23A, B). As clustering tool we

selected SIMLR because outperforms other tools (Wang, et al., 2017). However, SIMLR requires in input the number of clusters in which the dataset has to be split, since ith is a k-mean clustering approach. We have used griph (https://github.com/ppapasaikas/griph), which uses louvain modularity to identify the optimal number of clusters, combined with Jackknife resampling to identify a range of number of clusters to be used with SIMLR. was applied We observed that the perturbations affect the number of clusters detectable by griph (Fig. 1A). With this approach we can de-fine a most probable range or number of cluster to be queried using using a k-mean clustering tool such as SIMLR (Fig. 1B). To identify the optimal clustering parameter in RCASC we have implemented a cell stability score (Supplementary Material Section 5, Fig. 22), which pro-vides a better estimation of clusters stability with respect to other meas-urements as the silhouette plot (Supplementary Material Section 5, Fig. 23A, B)

Conclusion

Funding

This work has been supported by the EPIGEN FLAG PROJECT

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.

Bacher, R., et al. SCnorm: robust normalization of single-cell RNA-seq data. Nature methods 2017;14(6):584-586.

Barron, M. and Li, J. Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. Sci Rep 2016;6:33892.

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.

Diaz, A., et al. SCell: integrated analysis of single-cell RNA-seq data. Bioinformatics 2016;32(14):2219-2220.
Hwang, B., Lee, J.H. and Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018;50(8):96.

Liu, Z.H., et al. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nature Communications 2017;8.

Pace, L., et al. The epigenetic control of stemness in CD8+ T cell fate commitment. Science (New York, N Y) 2018;359(6372):177-186.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26(1):139-140.

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.

Wong, S et al. ARC 2017 Proceedings, Wong S, Beck AC, Bertels K, Carro L Eds Springer 2017