DIscBIO:a user-friendly pipeline for biomarker discovery in single-cell transcriptomics

Salim Ghannoum[1], Benjamin Ragan-Kelley[2,] Emma Jonasson[3], Anders Ståhlberg[3,4,5] and Alvaro Köhn-Luque[6]

[1] Department of Molecular Medicine, Division of Biochemistry, Faculty of Medicine, University of Oslo, Norway
[2] Simula Research Laboratory, Lysaker, Norway
[3] Sahlgrenska Cancer Center, Department of Laboratory Medicine, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden.
[4] Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Sweden.
[5] Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, Gothenburg, Sweden.
[6] Oslo Centre for Biostatistics and Epidemiology, Faculty of Medicine, University of Oslo, Norway

Keywords: Single cell sequencing, Normalization, Gene filtering, ERCC spike-ins, Clustering, DEGs, Decision trees, Protein protein interactions, Network analysis, Jupyter notebook, Binder

Abstract

- Single-cell sequencing (scRNA-seq) is expanding rapidly. Multiple scRNA-seq analysis packages and tools are becoming available. But currently, without advanced programing skills it is difficult to go all the way from raw data to biomarker discovery. Here we present DIscBIO, an open, multi-algorithmic pipeline for easy, fast and efficient analysis of cellular sub-populations and the molecular signatures that characterize them. The pipeline consists of four successive steps: data pre-processing, cellular clustering with pseudo-temporal ordering, defining differential expressed genes and biomarker identification. All the steps are integrated into an interactive notebook where comprehensive explanatory text, code, output data and figures are displayed in a coherent and sequential narrative. Advanced users can fully personalise DIscBIO with new algorithms and outputs. We also provide a user-friendly, cloud version of the notebook that allows non-programmers to efficiently go from raw scRNAseq data to biomarker discovery without the need of downloading any software or packages used in the pipeline. We showcase all pipeline features using a myxoid liposarcoma dataset.
- Availability GitHub: https://github.com/SystemsBiologist/PSCAN; Interactive demo on Binder: https://mybinder.org/v2/gh/SystemsBiologist/PSCAN/discbio-pub?filepath=DIscBIO.ipynb"
- Contact: salim.ghannoum@medisin.uio.no

Introduction

Single-cell sequencing (scRNA-seq) is a powerful technology that has already shown great potential and it is expanding rapidly [Stegle et al., 2015; Grün and van Oudenaarden, 2015]. As such, multiple computational tools for scRNA-seq analysis are being developed [Bacher and Kendziorski, 2016]. However, they are typically for programming experts and it is difficult to combine them into user-friendly and comprehensive workflows for non-experienced programmers. Due to the complex concatenation of heterogeneous programs and custom scripts via file-based inputs and outputs, as well as the program dependencies and version requirements, many computational pipelines are difficult to reproduce [Peng RD 2011; Sandve et al., 2013]. Several scRNA-seq pipelines for non-programmers have been generated using different platforms and logics [Diaz et al., 2016; DeTomaso D. and Yosef N., 2016; Gardeux et al., 2017; Mitulkumar 2018; Mah et al., 2018] but there is currently no user-friendly and integrated tool that allows biomarker discovery using decision trees and networking analysis directly from single-cell sequencing data. Hence, we developed an open Jupyter notebook called DIscBIO. It integrates in an interactive and sequential narrative, R code with explanatory text and output data and images. While R programmers can fully edit, extend and update

the pipeline, we provide a cloud version using Binder [Project Jupyter et al., 2018] that allows users without programming skills to run the pipeline even without installing the programming language R, Jupyter, or any of the software that the pipeline integrates. In order to produce the cloud-hosted Binder version, the repository includes a specification of all the software used, which enables automatic or manual reproduction or modification of the necessary computational environment needed to execute the pipeline on any computational resources [Forde et al., 2018]. Results reports can easily be generated that include high-quality figures.

Pipeline description

DIscBIO is a multi-algorithmic pipeline for an easy, fast and efficient analysis of cellular sub-populations and the molecular signatures that characterize them. The pipeline consists of four successive steps: 1) data pre-processing; 2) cellular clustering and pseudo-temporal ordering; 3) determining differentially expressed genes (DEGs); and 4) biomarker identification, including decision trees and networking analysis, see Figure 1a. Below we summarize the pipelines sections and algorithms used in the pipeline. Detailed information may also be found in each of the sections of the actual notebook:

1) **Data pre-processing.** Prior to applying data analysis methods, normalization and gene filtering are used to pre-process the raw read counts resulted from the sequencing. To account for RNA composition and sequencing depth among samples (single-cells), the normalization method "median of ratios" is used. This method makes it possible to compare the normalized counts for each gene equally between samples. Gene filtration is critical due to its dramatic impact on the downstream analysis. The key idea is to appoint the genes that manifest abundant variation across samples. In case the raw data includes ERCC spike-ins, genes will be filtered based on variability in comparison to a noise level estimated from the ERCC spike-ins according to an algorithm developed by Brennecke et al. [Brennecke et al., 2013], see Figure 1b. In case the raw data does not include ERCC spike-ins, genes will be filtered based on minimum expression in certain number of cells.

2) **Cellular clustering and pseudo-temporal ordering**. Cellular clustering is performed according to the gene expression profiles to detect cellular sub-population with unique features. This pipeline allows K-means clustering using RaceID [Grün et al., 2015] and model-based clustering [Fraley and Raftery, 2002]. It enables robustness assessment of the detected clusters in terms of stability and consistency using Jaccard's similarity statistics and silhouette coefficients. Additionally, the pipeline allows detecting outliers based on the clustering. Finally, pseudo-temporal ordering is generated to gradually order the cells based on their transcriptional profile, for example to indicate the cellular differentiation degree [Ji Z and Ji H., 2016]. Figures 1c,d,e and S1, S3 show some possible plots using tSNE, PCA and heatmaps for displaying the clustering results.

3) **Determining DEGs.** Differentially expressed genes between individual clusters are identified using significance analysis of sequencing data (SAMseq), which is a function for significance analysis of microarrays [Li J and Tibshirani R, 2013] in the samr package v2.0 [Tibshirani et al., 2015]. SAMseq is a non-parametric statistical function dependent on Wilcoxon rank statistic that equalizes the sizes of the library by a resampling method accounting for the various sequencing depths. The analysis is implemented over the pure raw dataset that has the unnormalized expression read counts after excluding the ERCCs. Furthermore, DEGs in each cluster comparing to all the remaining clusters are determined using binomial differential expression, which is based on binomial counting statistics. These DEGs can help characterizing the molecular signatures and tumorigenic capabilities of each cluster leading to developing better therapeutics. Volcano plots are used to visualise the results of differential expression analyses.

4) **Identifying biomarkers**. There are several methods to identify biomarkers, among them are decision trees and hub detection through networking analysis. To identify protein-protein interactions among DEGs as well as the decision nodes extracted from decision trees, we use STRING [Jensen et al 2009 Nucleic Acids Research]. The outcome is used as an input to check both the connectivity degree and the betweenness centrality, which reflects the communication flow in the defined PPI networks. Figures 1f,g show a decision tree and a network generated with DIscBIO.

Full information of how to use DIscBIO, all the possible options and documentation of the algorithms are included within the notebook.

Case study

To showcase DIscBIO, we analyzed a dataset consisting of single cells from a myxoid liposarcoma cell line. Myxoid liposarcoma is a rare type of tumor driven by specific fusion oncogenes, normally FUS-DDIT3 [Fletcher et al., 2013], with few other genetic changes [Ståhlberg et al., 2014; Hofvander et al., 2018]. The cells were collected based on their cell cycle phase (G1, S or G2/M), assessed by analyzing their DNA content using Fluorescence Activated Cell Sorter [Karlsson et al., 2017]. Data including ERCC spike-ins, are available in the ArrayExpress database at EMBL-EBI with accession number E-MTAB-6142). After filtering the dataset using the ERCC spike-ins (figure 1b), three clusters were obtained based on model-based clustering (Figure 2b) that match with the sorting made in the original study [Karlsson et al., 2017]. Pseudo-temporal ordering showed a gradual transition between the cells transcription profiles across the 3 clusters. Both the clustering and pseudo-time overlap with the cell cycle phases (supplementary figure S1). Cluster 2 cells, which were ordered in the foreground pseudo-temporal ordering (figure 1d), highly express genes involved in regulating proliferation, stemness and the acquisition of chemoresistance in several cancers. The networking analysis outcome (Supplementary table. S2) nominated a panel of 16 hub genes, including *CDC20, PLK1, KIF2C, BUB1, AURKB* and *PTTG1* as potential biomarkers for cluster 2 cells (supplementary figure S3). These cells may exhibit aggressive and stem-like properties and further computational investigations should validate these biomarkers candidates.

Acknowledgement
We would like to thank Dominic Grün, Benjamin Ulfenborg and Hongkai Ji for their kind feedback and input during the construction of this pipeline. AS is supported by grants from Knut and Alice Wallenberg Foundation, Wallenberg Centre for molecular and translational medicine, University of Gothenburg, Sweden; Swedish Cancer Society (2016-438); Swedish Research Council (2017-01392); Swedish Childhood Cancer Foundation (2017-0043); the Swedish state under the agreement between the Swedish government and the county councils, the ALF-agreement (716321).

References

- Bacher R and Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biology 2016 17:63.
- Brennecke P, Simon A, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B, Beners V, Teichmann SA, Marioni JC and Heilser MG Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods 2013 10, 1093-1095.
- DeTomaso D., Yosef N. FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. BMC Bioinformatics 2016, 17, 315.
- Diaz A, Liu SJ, Sandoval C, Pollen A, Nowakowski TJ, Lim DA, Kriegstein A, SCell: integrated analysis of single-cell RNA-seq data. Bioinformatics 2016 32(14):2219-20.
- Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F eds. World Health Organization classification of tumors of soft tissue and bone, 4th edn. Lyon: IARC PRESS, 2013.
- Forde J, Head T. Holdgraf C, Panda Y, Nalvarte G, Ragan-Kelley B, et al. Reproducible Research Environments with Repo2Docker, Proceedings of ICML Reproducibility in Machine Learning Workshop, 2018.
- Fraley C., Raftery A. E. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 2002;97:611–631.
- Gardeux V, David FPA, Shajkofci A, Schwalie PC and Deplancke B, ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. Bioinformatics 2017 33(19): 3123-3125

- Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H and van Oudenaarden A, Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 2015 525, 251-255.
- Grün D and van Oudenaarden A, Design and analysis of single-cell sequencing experiments. Cell 2015; 163(4): 799-810.
- Hofvander J, Viklund B, Isaksson A, Brosjö O, Vult von Steyern F, Rissler P, Mandahl N and Mertens F. Different patterns of clonal evolution among different sarcoma subtypes followed for up to 25 years. Nat Commn. 2018 9(1): 3662.
- Jensen L. J., Kuhn M., Stark M., Chaffron, S., Creevey, C., Muller, J., ... and von Mering, C. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research 2009, 37(Database issue), D412–D416
- Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res*. 2016 Jul 27;44(13):e117. doi: 10.1093/nar/gkw430. Epub 2016 May 13. PubMed PMID: 27179027; PubMed Central PMCID: PMC4994863.
- Karlsson J, Kroneis T, Jonasson E, Larsson E, Ståhlberg A. Transcriptomic characterization of the human cell cycle in individual unsynchronized cells. J Mol Biol 2017 429(24):3909-3924.
- Li J and Tibshirani R, Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data. Stat Methods Med Res 2013 22(5): 519-36.
- Mah CK, Wenzel AT, Juarez EF, Tabor T, Reich MM and Mesirov JP, An accessible, interactive GenePattern notebook for analysis and exploration of single-cell trasncriptomic data [version 2; peer review; 1 approved, 1 approved with reservations]. F1000Research 2019, 7:1306
- Mitulkumar VP, iS-CellR: a user-friendly tool for analysing and visualizing single-cell RNA sequencing data. Bioinformatics 2018 34(24): 4305-4306.
- Peng RD (2011) Reproducible Research in Computational Science. Science 334, 1226- 1227.
- Project Jupyter, Bussonnier M, Forde J, Freeman J, Granger B, et al. (2018) Binder 2.0 - Reproducible, interactive, shareable environments for science at scale. Proceedings of the 17th Python in Science Conference 2018, 113-120.
- Sandve GK, Nekrutenko A, Taylor J and Hovig E, Ten simple rules for reproducible computational research. PLoS Comp Biol 2013 e1003285.
- Stegle O, Teichmann SA and Marioni JC, Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015; 16(3):133-45.
- Ståhlberg, A., C. K. Gustafsson, K. Engtröm, C. Thomsen, S. Dolatabadi, E. Jonasson, C. Y. Li, D. Ruff, S. M. Chen and P. Åman (2014). "Normal and functional TP53 in genetically stable myxoid/round cell liposarcoma." PLoS ONE 9(11).
- Tibshirani R., Chu G., Balasubramanian N. and Li J. (2015) Package ʻsamrʼ. Available at: https://cran.rproject.org/web/packages/samr/samr.pdf.

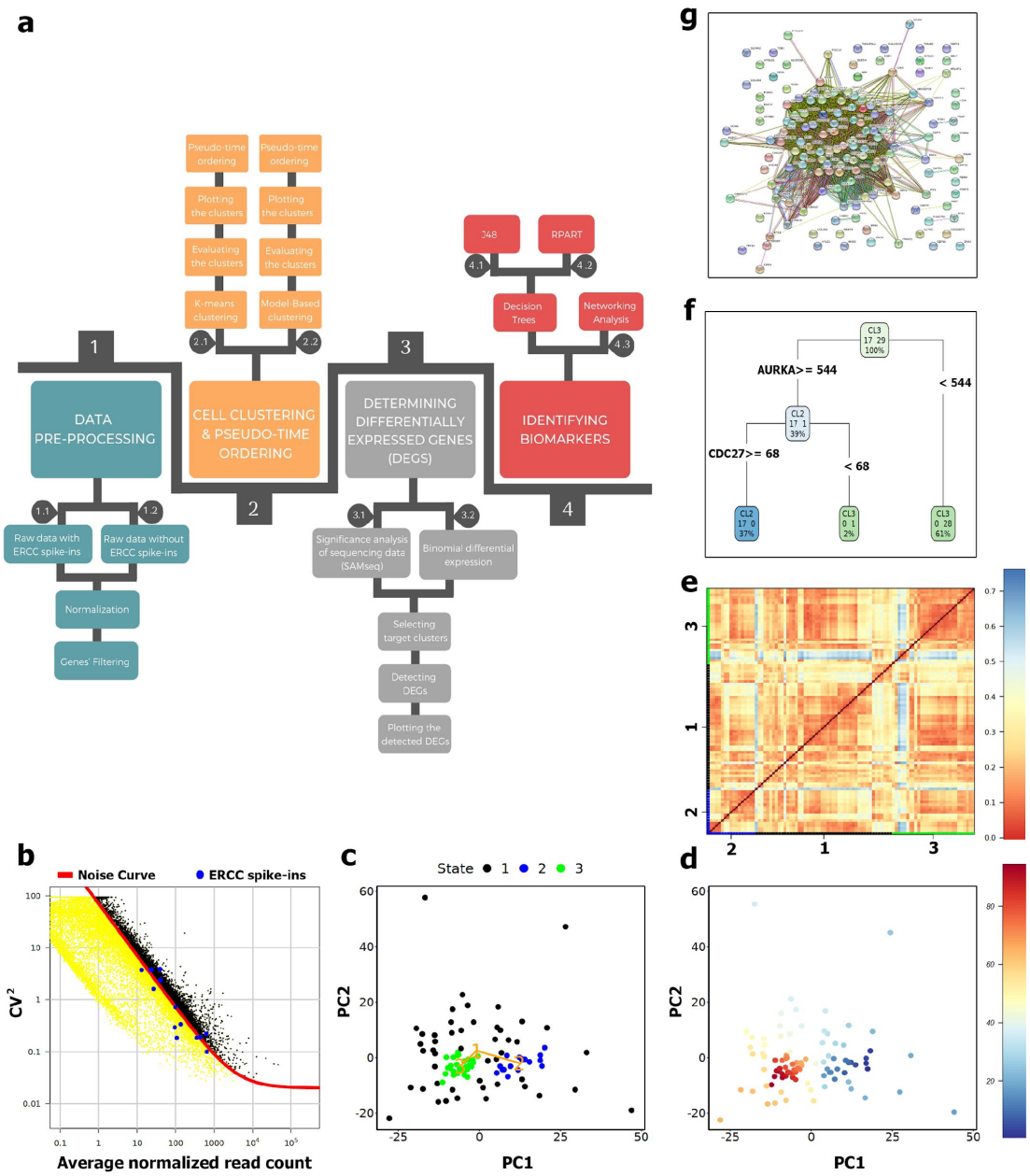

Figure 1: DIscBIO pipeline. Panel a shows and overview of the pipeline sections and algorithms. Panels b, c, d, e, f, g shows some of the graphical outputs of the pipeline when analysing and dataset of 94 myxoid liposarcoma cells.