DOI 10.18699/BioGenEvo-2018-19

Single-cell bioinformatics: practical implementations

O. Kuleshova
A.O. Kovalevsky Institute of Marine Biological Research of RAS, Sevastopol, Russia
e-mail: v_olgo4ka@inbox.ru

Key words: single-cell analysis, bioinformatics, single-cell software

Motivation and Aim: Rapid development of the next-generation sequencing (NGS) technologies observed the last 15 years. High potential of the NGS methods for a single cell RNA-seq (scRNA-seq) has been showed [1]. This revolutionary analysis gives additional information relative to RNA-seq to bulk sample by analyzing expression profiles at the cellular level. These technologies have opened up opportunities for understanding biological processes at a fundamentally new level of the cellular heterogeneity detections, cell lines tracering, subpopulations identification and clarification of the cell-specific characteristics [2]. Modern scRNA-seq technologies require a development of the special approaches for the big-data sets analysis. In this regard, rapid improving of analytical techniques of the processing of increasing data stream, as well as the implementing software packages, has observed. So, current task is to study and analyze software methods and tools for scRNA-seq bioinformatic analysis.
Methods and Algorithms: There are two main stages in the bioinformatics analysis of the scRNA-seq data: primary and secondary analysis. Primary analysis includes (1) data preprocessing (adapter trimming, paired-end data processing, nucleotide quality filter), (2) constraction of expression matrix including reads quality control, alignment, mapping, quantification. Secondary analysis includes data data preprocessing (filtering and normalization), estimation of the differential expression, clustering, dimensionality reduction (visualization), subpopulation detection, pseudo-time construction. The primary stage is implemented using a set of specialized program packets (e. g., Trimmomatic, FASTQC, STAR, HiSeq or functional analogues), or automated pipelines (e. g., CellRanger, zUMIs, scPipe, Dr.seq2). For normalization, it could be used one of the packages: MAST, SCDE, Monocle2, BCSeq et al. Evaluation of differential expression could be performed using MAST, SCDE, Monocle 2, D3E, DESeq, edgeR, Seurat, Scanpy et al. Most of the variant clustering and visualization methods are implemented in the software packages mentioned above. Pseudo-time construction is available, for example, in the Monocle2 and Scanpy packages. Subpopulation detection could be performed using packages SCUBA, SCENT, Scanpy et al.
Conclusion: Bioinformatic tools for scRNA-seq data analysis are very rapidly evolving to meet the needs for the analysis of the huge increasing data flows. Currently, secondary analysis packages Seurat and Scanpy could be view as the most optimal. However, bioinformatic tools improve rapidly and constantly, that is why continuous monitoring, evaluation and selection of software packages for specific tasks are required.
Acknowledgements: Supported by the Ministry of Education and Science of the Russian Federation grant No. 14.W03.31.001

References

1. D'Argenio V. (2018) The High-Throughput Analyses Era: Are We Ready for the Data Struggle? High-Throughput. 7:8.
2. Poirion O.B., Zhu X., Ching T., Garmire L. (2016) Single-Cell Transcriptomics Bioinformatics and Computational Challenges. Front. Genet. 7:163.