**scdrake: a reproducible and scalable pipeline for scRNA-seq data analysis**

Novotný Jiří1,2, Kubovčiak Jan1, Kolář Michal1,2

1 Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences, Vídeňská 1083, 142 20 Prague 4, Czech Republic

2 Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology in Prague, Technická 5, 166 28 Prague 6, Czech Republic

Motivation: While the workflow for primary analysis of single-cell RNA-seq data is well established, the secondary analysis of the feature-barcode matrix is usually done by custom scripts. There is no fully automated pipeline in the R statistical environment, which would follow the current best programming practices and requirements for reproducibility.

Results: We have developed scdrake, a fully automated workflow for secondary analysis of scRNA-seq data, which is fully implemented in the R language and built within the drake framework. The pipeline includes quality control, cell and gene filtering, normalization, detection of highly variable genes, dimensionality reduction, clustering, cell type annotation, detection of marker genes, differential expression analysis, and integration of multiple samples. The pipeline is reproducible and scalable, has an efficient execution, provides easy extendability and access to intermediate results, and outputs rich HTML reports.

Availability: The source code and documentation are available under the MIT license at <https://github.com/bioinfocz/scdrake> and <https://bioinfocz.github.io/scdrake>, respectively.