DOI 10.18699/BGRSSB-2018-230

Workflows for classification of NGS metagenomic data

D. Sukhomlinov[1, 2]*, Y. Algaer[2], A. Tiunov[2], E. Pushkova[2], O. Golosova[2]
[1] *Novosibirsk State University, Novosibirsk, Russia*
[2] *Unipro Center of Information Technologies, Novosibirsk, Russia*
** e-mail: dsukhomlinov@unipro.ru*

Key words: bioinformatics, NGS, data analysis, metagenomics, classification, workflow, graphical user interface

Motivation and Aim: Next Generation Sequencing (NGS) technologies, invented a few decades ago, have opened new opportunities for scientists. In particular, the technologies are used for identification of microorganisms in metagenomic clinical and environmental samples. Although, there is an extensive set of computational tools for classification of the sequencing data, the tools are commonly command-line and require additional configuration. Also, usage of multiple tools, that may be useful for improving quality of classification, requires a lot of routines from a scientist in case of stand-alone tools. We are motivated to provide convenient and user-friendly graphical user interface for some of these tools to make a work process more effective and simple. We believe that a well-organized process of the classification will improve viral and bacterial pathogen detection and discovery as well as disease control and prevention.
Methods and Algorithms: Unipro UGENE [1] is a desktop multiplatform software package that integrates dozens of widely used bioinformatics tools. The Workflow Designer component of the software allows one to use graphical interface to create and run workflows, composed of different tools. In addition, it is intended to store and investigate the results and re-run workflows on different datasets. The described infrastructure has served as the basis for a new framework for whole-genome NGS data classification, which includes the following popular tools: CLARK [2] (CLAssifier based on Reduced K-mers), supplied with NCBI RefSeq viral and bacterial database; Kraken [3], supplied with MiniKraken database; DIAMOND [4], a sequence aligner, similar to NCBI BLAST, supplied with UniRef50 and UniRef90 databases and WEVOTE [5] (WEighted VOting Taxonomic idEntification), used to ensemble classification data, produced by other tools.
Results: The NGS data classification framework was integrated into the open source Unipro UGENE software and contains sample workflows. The first serial workflow sequentially runs Kraken, CLARK and DIAMOND tools, filtering out NGS reads after each step, and reports classification information, produced by the tools. The second parallel workflow runs these tools and ensembles the output data using WEVOTE. The third workflow classifies NGS scaffolds, assembled de novo by SPAdes [6]. Advanced users can also use individual blocks of these workflows to create a new one, suitable for their purposes. The results are available on Linux and macOS platforms.
Conclusion: The new framework is intended for optimization of a scientist's work in the area of whole-genome NGS data classification, easy to set up for a biologist and require no additional configuration.
Availability: http://ugene.net/download.html
Acknowledgements: The project was supported by the VIROGENESIS consortium [7].

References:
1. Okonechnikov K., Golosova O., Fursov M. (2012) the UGENE team, Unipro. Bioinformatics. 28:1166-1167. DOI 10.1093/bioinformatics/bts091
2. Ounit R. et al. BMC Genomics (2015) 16:236. DOI 10.1186/s12864-015-1419-2
3. Wood D.E., Salzberg S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Gen Biol. 15:R46.
4. Buchfink B., Xie C., Huson D. (2015) Nature Methods. 12:59-60.
5. Metwally A., Dai Y., Finn P.W., Perkins D.L. (2016) WEVOTE: Weighted Voting Taxonomic Identification Method of Microbial Sequences. PloS ONE.
6. Bankevich A. et al. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal Computational Biol.
7. VIROGENESIS project web site: http://www.virogenesis.eu/