

Assembly and annotation pipeline for microorganism genomes from sequence to gene networks

Demenkov P.S.^{1, 2*}, Lashin S.A.^{1, 2}, Ivanisenko V.A.^{1, 2}

¹ Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

² Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

* demps@bionet.nsc.ru

Microorganisms of natural and technogenic ecosystems represent an inexhaustible pool of metabolic pathways for the utilization of some compounds and the biosynthesis of others. Modern experimental technologies of genetics and microbiology provide an effective search in microbiological collections for new strains of microorganisms that are promising for biotechnological applications, as well as complete sequencing of their genomes. The practice of the world's major biotechnology companies, accumulated over the past 30 years, indicates the critical importance of integrating experimental and information-computer approaches for constructing strains of microorganisms with target properties, including multi-parameter profiling of microorganisms based on genomic, proteomic and other data, bioinformatics and systems methods computer biology, which allows, on the basis of genomic information, to reconstruct gene networks that control the production of target substances and to build mathematical models of their functioning.

This paper proposes a software pipeline for assembly, annotation of sequenced genomes of microorganisms and reconstruction of their gene networks. The pipeline includes the following steps: 1) Quality control and sequence cleaning (FastQC, Trimmomatic [1]); 2) on the side of the genome (Spades [2]); 3) Assembly quality analysis (Quast [3]) 4) genome annotation (Prokka); 5) probable identification of the organism (ProTaxon); 6) search for enzymes (Blast, Prokka); 7) reconstruction of gene networks (ANDSystem [4, 5]). The developed pipeline was used on the data of the collection of microorganisms of the Kurchatov Genomic Center. To date, 2086 genomes have been processed. The average genome length was around 6,000,000 bp. The average number of genes in the genome is 5627. KEGG gene networks and metabolic pathways were reconstructed for all analyzed genomes.

Funding: Work was funded by the Ministry of Science and Higher Education of the Russian Federation project “Kurchatov Center for World-Class Genomic Research” number 075-15-2019-1662 from 2019-10-31.

References

1. Bolger A.M., Lohse M., Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi 10.1093/bioinformatics/btu170
2. Nurk S., Bankevich A., Antipov D. et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol*. 2013;20(10):714-737. doi 10.1089/cmb.2013.0084
3. Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-1075. doi 10.1093/bioinformatics/btt086
4. Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. 2012;11(3-4):149-161. doi 10.3233/ISB-2012-0449
5. Ivanisenko T.V., Saik O.V., Demenkov P.S. et al. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics*. 2020;21228. doi 10.1186/s12859-020-03557-8