Toward big data analysis in the Baikal microbiome study

F. Malkov1,2, E. Cherkashin1,2,3,5, S. Gorsky1,2,3, A. Shigarov1,2,3, Y. Galachyants1,4, A. Morozov1,4, Y. Likhoshway1,4

1Irkutsk Scientific Center of SB RAS

2Matrosov Institute for System Dynamics and Control Theory of SB RAS

3Irkutsk State University

4Limnological Institute of SB RAS

5[National Research Irkutsk State Technical University](https://www.istu.edu/eng/)

The Baikal microbiome study is being conducted in Limnological Institute of SB RAS, the study is based on the next-generation sequencing data analysis. As a result, the study generates a big volume of metagenomic data. A seasonal monitoring of Baikal microbiome requires the study of thousands of samples.

The process of metagenomic data analysis is multi-stage, *e.g.* for amplicons up to 50 consecutive steps can be included. This requires microbiologist to be able to construct and execute a pipeline of the processing, which consists of command-line utilities like Mothur or QIIME, scripts of the Python programming language, statistical processing with R modules. Additionally, the execution of the pipeline requires interaction with a storage and management software of metagenomic data. At the current state of the automation of the analysis, microbiologist is not always able to cope with the complex task independently.

In this work, we present an environment that supports visual planning and parallel execution of pipelines of the metagenomic data analyses. Visual planning is implemented as a plug-in module of Rapid miner system, and the parallel execution is provided by instrumental system Orlando. The big data storage facility is based on Hadoop infrastructure. The environment under construction simplifies the research activity of microbiologist, making the metagenomic analysis to be more accessible for researcher without in-depth programming and data science skills.