DOI 10.18699/SBB-2022-278

Bioinformatics approaches and methods for analysis of gut microbiota

Kovtun A.S.[1, 2]*, Danilenko V.N.[1]

[1] *Vavilov Institute of General Genetics, RAS, Moscow, Russia*

[2] *Skolkovo Institute of Science and Technology, Moscow, Russia*

* *kovtunas25@gmail.com*

Key words: microbiota, metagenome, metagenomic signature

Active development of methods for the analysis of microbial communities called microbiota became possible with introduction of the next generation sequencing (NGS) platforms. Nowadays, growing number of scientific groups worldwide contribute to the knowledge about various microbial communities, such as those inhabiting soils, wastewater and natural waters, human organism, including gut microbiota, etc. Generally, the bioinformatics analysis of microbiota can be divided into taxonomic identification and functional annotation of metagenome (the overall set of genetic information contained in microbiota).

The keys goal of the taxonomic analysis is to describe, which organisms constitute the microbiota and what their abundance is. This kind of analysis can be conducted with two types of data: 16S rRNA and whole metagenome. Each of them has its advantages and disadvantages and requires specific algorithms. The 16S rRNA data can give the information about taxonomy of a microbial sample and is still widely used, thanks to its cost-effectiveness and relatively modest requirements for computational resources. Although the genes encoding the 16S rRNA are good taxonomic classifiers in theory, the short length of read sequences obtained from the NGS imposes limitations on the resolution of the method – the taxonomic abundance can be well-resolved mostly on the 'Genus' level [1]. Another issue is the generation of chimeric reads. The actively developing single-cell sequencing techniques, which allow the whole-length sequencing of the 16S rRNA gene, partly solve the resolution problem with exchange on still significantly lower quality of the data [1]. Despite these difficulties, the number of publications in metagenomics using the single-cell sequencing is growing and the corresponding algorithms are being developed and improved constantly. The whole metagenomic data gives better resolution and allows the researchers to analyze the taxonomy up to the 'Strain' level, as the taxonomy is often assigned by implementation of the whole genomes of microorganisms. However, such approach requires deeper sequencing and, thus, is way costlier, and more resource-demanding. It is also more prone for contamination with eukaryotic DNA.

Unlike the taxonomy analysis, the functional annotation of a metagenome should be conducted on the whole metagenome data. One can either conduct the assembly recreate (usually partially) the genomes, or simply align the reads against the reference database of genes and metabolic pathways. The first approach requires significantly more resources, can generate chimeric assemblies and collapses the information on relative abundances of each gene, thus, the further mapping of reads is still required. But it allows a more straightforward annotation of the genes, as they are being recreated during the assembly. The second method shows more difficulties in correct normalization, since

not only depth, but also breadth and uniformity of coverage of the reference genes should be accounted for to correctly tell, if the gene really exists in the metagenome.

The key disadvantage of the traditional metagenomic analysis is that the taxonomic and the functional results are reported separately. The unification of the information on the identified genes, their abundance and bacterial source can be implemented through metagenomic signatures. One of existing algorithms the metagenomic isoforms' recreation has been introduced by us in Kovtun et al., 2018 [2]. It has been successfully used for description of gut microbiota of healthy adults [2], its development in children of different age [3], and also for the analysis of changes in gut microbiota in children during autism spectrum disorder [4]. Currently, another research is being conducted by us, which uses the similar approach to identify the changes in gut microbiota of people with depression from Moscow region. The preliminary results demonstrate statistically significant decrease of abundance of genes involved in synthesis of butyric acid, spermidine and estradiol in *Faecalibacterium prausnitzii* and *Roseburia intestinalis*, which also correlates with the severity of the disease.

References

1. Johnson J.S. et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun*. 2019;10(1):5029.
2. Kovtun A.S. et al. *In silico* identification of metagenomic signature describing neurometabolic potential of normal human gut microbiota. *Russ J Genet*. 2018;54(9):1101-1110.
3. Averina O.V. et al. Altered neurometabolic potential of gut microbiome in healthy children of different age. *Bulletin of RSMU*. 2020;6:68-79.
4. Averina O.V. et al. The bacterial neurometabolic signature of the gut microbiota of young children with autism spectrum disorders. *J Med Microbiol*. 2020;69(4):558-571.