

# Automatic identification of contaminating sequences in amplicon sequencing data

Fedorov D.<sup>1\*</sup>, Galeeva J.<sup>1</sup>, Manolov A.<sup>1</sup>, Klimina K.<sup>2</sup>, Veselovsky V.<sup>2</sup>, Pavlenko A.<sup>1</sup>, Ilina E.<sup>1</sup>

<sup>1</sup> *Institute of Systems Biology and Medicine, Moscow, Russia*

<sup>2</sup> *Research Institute of Physical-Chemical Medicine, Moscow, Russia*

\**fedorov.de@gmail.com*

**Key words:** microbiome, contamination, amplicon sequencing, metagenomics

**Motivation and Aim:** Amplicon sequencing allows for efficient study of microbiological communities without the need for microorganism cultivation. However, the PCR amplification stage introduces bias into the composition and quantity of the obtained amplicons. This effect can lead to sequencing of sequences from reagents, laboratory plastics, the environment and have a serious impact on the results and their interpretation [1]. This is especially critical when sequencing media with a low bacterial load. In this work, swabs from the nasopharynx of patients with a new coronavirus infection, as well as healthy volunteers, were analyzed. We observed the presence of contamination phenomena of unknown origin and proposed a method for automated search for anomalous structures in data and their subsequent cleaning.

**Methods:** 16s rRNA sequencing data were obtained from the Illumina MiSeq instrument, amplicon sequence variants (ASV) and their abundance in samples were obtained using the DADA2 [2] algorithm, the decontam [3] algorithm was used at the first stage of decontamination, the correlation network was obtained using the SPIEC-EASI [4] method, hierarchical clustering of sequences on the basis of the correlation matrix was performed by the ward.D2 algorithm using the R language.

**Results:** Using a combination of the decontam algorithm and hierarchical clustering based on a correlation matrix, it was possible to achieve automatic search and removal of contaminating sequences. Data cleaning allowed us to reduce the effect of the batch on the estimated composition of the community and to achieve a uniform distribution of samples in a multidimensional space. We observed that the representation of sequences identified as contaminants was specific to individual sequencer runs.

**Conclusion:** A method for the automatic identification and removal of contaminating sequences is proposed, which makes it possible to reduce the bias introduced by individual runs of the sequencer and provides an opportunity for accurate interpretation of the observed compositions of microbiological communities.

**Acknowledgements:** This study was supported by Russian Science Foundation No. 21-15-00431.

## References

1. Salter S.J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 2014;12:87.
2. Callahan B.J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016;13(7):581-583.
3. Davis N.M. et al. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome.* 2018;6:226.
4. Kurtz Z.D. et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biol.* 2015;11:e1004226.