Mathematical Biology and Bioinformatics
2018. V. 13. № 1. P. 159–168. doi: 10.17537/2018.13.159

========== **BIOINFORMATICS** ==========

UDC: 573, 579, 579.8

New procedure of raw Illumina MiSeq data filtering for the amplicon metagenomic libraries

Bukin Yu.S.[1,2], Buzoleva L.S.[3,4], Golozubova Y.S.[3], Galachyants Yu.P.[1,2]

[1]*Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia*
[2]*Irkutsk Scientific Center, Siberian Branch of the Russian Academy of Sciences, Russia*
[3]*Far Eastern Federal University, Vladivostok, Russia*
[4]*Somov Institute of Epidemiology and Microbiology, Vladivostok, Russia*

Abstract. In this paper we present an algorithm to filter amplicon paired-end NGS raw data which is used to capture genetic and taxonomic diversity of communities of unicellular microorganisms. The suggested approach allows one to overcome the issue of massive data loss during filtration of raw sequences and increases the static representativeness of analyzed amplicons. Furthermore, an unequal elimination of sequences belonging to different taxonomic groups was shown to occur if one applies standard trimming methods based on filtration of quality of raw reads, for instance, using sliding window approach. This bias may result in a skew of taxon counts and depletion of taxonomic diversity of analyzed communities. The suggested method does not introduce the errors of this kind. The implementation of the algorithm in R as well as a number of example files for analysis is available at https://github.com/barnsys/metagenomic_analysis.

Key words: *amplicon metagenomic, new generation sequencing, meta-barcoding, quality control.*

INTRODUCTION

Metagenomic analysis conducted on the basis of Next-Generation Sequencing of amplicons is widely used to study taxonomic diversity of microorganisms (Petrosino J.F., et al, 2009; Kim M., et al, 2013). Different targets, such as fragments of 16S and 18S rRNAs, protein-coding genes, and neutrally evolving sequences are analyzed to produce amplicon libraries using total DNA isolated from environmental samples. The length and count of reads are important parameters that should be taken into account in order to obtain statistically sound results. Illumina MiSeq instrument with v.3 500 cycles sequencing kit producing paired-end reads is a popular solution for amplicon metagenomic studies (Quail M.A. et al. 2012). Theoretically, using this approach, analysis of contiguous gene fragments of the length up to 500 bp can be performed.

Raw NGS data processing approaches can fall into one of the two major strategies (Kim M., et al, 2013):

1) Generation of sequence groups – Operational Taxonomic Units (OTUs) – which is performed on the basis of similarity of analyzed DNA fragments and their subsequent clustering; when performing comparative analysis of different communities, OTUs can be thought as groups of sequences representing a taxon of specific rank.

2) Direct classification of sequences using a reference database.

Both strategies are based on comparing sequences to an external database to assign query sequence to a specific taxon: the representative sequences of OTUs can be generated and used for taxonomic identification. However, OTU approach allows decreasing data dimensions by several orders of magnitude, making the subsequent analyses easier.

Accuracy of genetic distance measurement is an important factor, influencing the results of the OTU clustering. Biases of distance calculation may be introduced at any stage of sample preparation, sequencing, and data processing prior to clustering. For instance, "species-like" clustering distance, which is widely used for analysis of bacterial 16S rRNA amplicon sequences, is assumed to be 0.03. Applying this threshold, two 500bp sequences should differ by at least 15 positions to be assigned into different OTUs. Obviously, sequencing errors will introduce uncertainty in clustering results.

A major drawback of Illumina sequencing reagents producing long reads (more than 2×150bp) is a significant sequencing quality drop to the end of forward and reverse reads. This results in errors at the OTU generation stage which, in turn, may affect conclusions drawn from the comparison of the communities. Therefore, filtration of sequencing errors should be performed prior to clustering stage. Common methods to filter raw sequencing reads (Morgan, M., et al, 2009; Bolger A.M., et al. 2014) on the basis of quality scores generated by the NGS instrument are 1) cutting several positions from 5'- and 3'-ends of the read if these regions contain bases below the quality threshold, and 2) sliding window approach, when the sequencing quality is assessed within the region of read of defined length; the downstream part of sequence is trimmed once the average quality in current region falls below threshold. During filtration, too short reads are also removed from the resulted dataset. For paired-end data, filtration can be performed either after or before generation of consensus sequences from forward and reverse reads. In the former case, if length of amplicon exceeds read length by 50–80 %, the middle parts of consensus sequences will contain regions of low quality and trimming such sequences using sliding window approach will drop considerable portion of data. If the trimming is performed before generation of consensus sequences, a significant number of reads are truncated by 3'-end, resulting in no overlap between forward and reverse reads to be stitched. Therefore, standard trimming procedure may result in significant reduction of analyzed dataset.

As noted earlier, several factors may result in bias of community composition in amplicon metagenomic studies (Zhou J., et al, 2011). First, members of analyzed community may bear non-equal number of copies of the target gene. Second, at the stage of total DNA isolation from environmental sample, skew could result from non-uniform efficiency of unsealing and capturing the DNA from different microorganisms by the method used. Third, the performance of amplification of gene marker sequences could differ as a result of PCR settings, primer efficiency, differences in GC-content and secondary structure of target sequences belonging to specific taxon, etc. Finally, biases could be introduced by data analysis pipeline.

One of the possible reasons for the latter type of the artifact is as follows: marker sequences from different taxons may have different lengths resulting from accumulation of indels in the course of evolution; due to Illumina MiSeq properties, long sequences tend to be sequenced poorly and have higher chances to be dropped during filtration and trimming than the short ones. Thereby, during the pre-clustering stage, there could be non-equal elimination of sequences representing different taxons. To date, consequences of such bias were not studied.

In the present work we suggest a simple algorithm of filtration and trimming developed for raw paired-end NGS amplicon metagenomic data. This trimming procedure optimizes the number of sequences to drop off and the length of analyzed fragment. It helps to avoid significant data loss in terms of both number of fragments per amplicon and the length of analyzed fragment, assists in finding an acceptable trade-off between these two parameters, and increases the accuracy of metagenomic studies.

DESCRIPTION OF ALGORITHM

The key idea implemented by this algorithm lies in elimination of several positions from the sequence alignment under analysis, rather than deletion of whole sequences containing low quality regions. In this filtering strategy, slight alignment length reduction is accompanied by a significant increase in the number of sequences retained for further analysis. Additionally, elimination of alignment positions with poor quality reduces errors in genetic distance measurement.

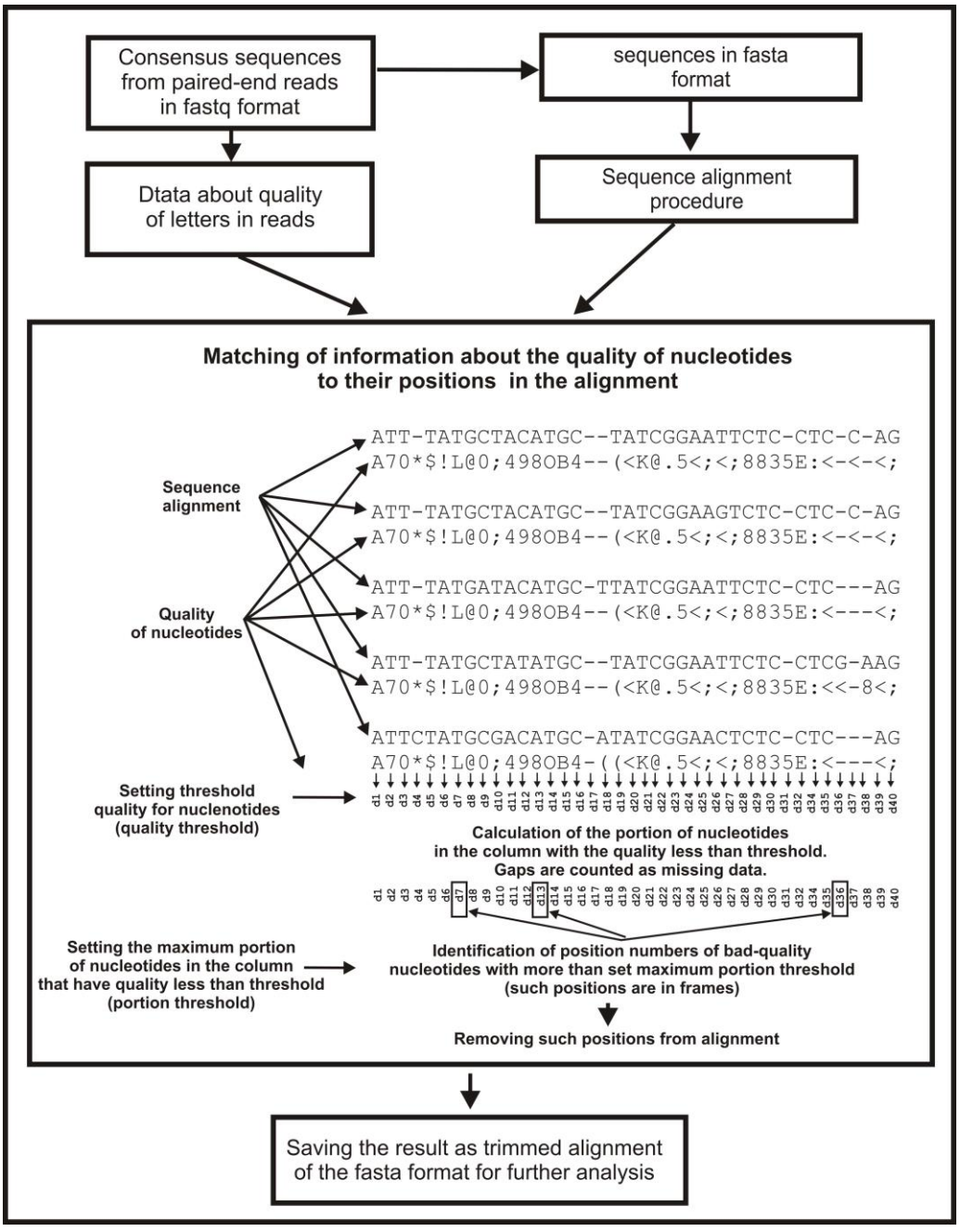

Fig. 1. Scheme of the trimming algorithm.

The proposed algorithm includes the following stages:

1. Generation of the consensus sequences from paired-end reads. Optionally, filtering paired-end reads can be performed to remove too short or too long sequences, not suitable for analysis. We generated consensus sequences using FLASH program (Magoč T., Salzberg S.L., 2011), which is widely used in amplicon metagenomic studies (Fosso B., et al, 2015;

Tennant R.K., et al, 2017). Prior the next stage, fastq-file with consensus sequences should be converted into fasta format.

2. Sequence alignment. We generated 16S rRNA sequence alignment using Mothur package (Schloss P.D., et al, 2009) and SILVA reference database v.123 (Quast C., et al, 2013). To generate alignment of neutrally-evolved sequences such as ITS of rRNA and sequences of protein-coding genes, multiple sequence alignment programs, for instance MAFT (Katoh K., Toh H., 2010), can be used.

3. Matching of information about the quality of nucleotides to their positions in the alignment.

2. Setting threshold quality for nuclenotides (quality threshold – defined by user).

4. Setting the maximum portion of nucleotides in the column that have quality less than threshold (portion threshold – defined by user).

5. Collection of quality statistics for consensus sequences for alignment positions.

6. Identification of position numbers of bad-quality nucleotides with more than set maximum portion threshold.

7. Defining and target positions and their removal from alignment; generating the final alignment.

The general scheme of the trimming algorithm is shown in Figure 1.

Stages 3–6 were implemented as an R script using ShortRead package (Morgan, M., et al 2009). This script includes three functions: 1) to visualize portion of alignment positions having quality below threshold; 2) to remove positions below quality threshold; 3) to remove sequences having low-quality positions in start- and end-regions of consensus sequence similarly to standard trimming procedure.

The default quality threshold for the procedure is set to 20 (1 % sequencing error probability in consensus sequence) and maximum portion of nucleotides per position below this quality is 0.1 or 10 % (portion threshold = 0.1).

Algorithm in R is available at: https://github.com/barnsys/metagenomic_analysis.

TESTING THE ALGORITHM

For testing we used data obtained by sequencing of 16S rRNA fragments with Illumina MiSeq instrument and v.3 sequencing reagents to generate paired-end reads of 250 bp. Amplicons were generated with PCR primers 9F and 541R (Chun, J., et al, 2010) targeting V1-V3 region of 16S rRNA. Total DNA isolated from bacterial community of water sample of Round Bay (Japan Sea) in August 2015.Sampling took place from the surface water (10–15 cm depth). It was used as a template (data is available at: https://github.com/barnsys/metagenomic_analysis). Generation of consensus sequences was performed with FLASH. Trimmomatic program (Bolger A.M., et al. 2014) was used for standard trimming of consensus sequences. Trimmed and non-trimmed datasets were aligned to SILVA database v.123 using Mothur package. Mothur was then used to remove chimeric sequences, non-bacterial sequences and sequences aligned outside of V-V3 region of 16S rRNA from both datasets. The developed algorithm was finally applied for a set of non-trimmed sequences. Hereinafter, dataset trimmed by Trimmomatic is referred to as “STD” and the one obtained with the suggested algorithm – as “ALIOPT”.

Next, for both datasets, clusters of OTUs at distance 0.03 and their representative sequences were generated in Mothur package. Taxonomic classification was performed using SILVA database v.123. OTUs were sorted in the order of decreased sequence count. We have examined both all OTUs of dataset and major OTUs contributing 95 % of total dataset fragments. Using Mothur package we also computed pairwise distances for representative sequences of STD and ALIOPT datasets to find identical and different OTUs.

Convergence of sequencing data was assessed with rarefaction analysis and bootstrap index (Smith E.P., van Belle G, 1984), which evaluates the number of underestimated taxons

in analyzed dataset. Computations were performed in R using vegan package (Dixon P., 2003).

RESULTS AND CONCLUSIONS

NGS dataset contained 26987 paired-end 2 × 250 bp reads. Typically for MiSeq data, the quality of sequencing gradually falls to the 3'-end of reads (Fig. 2,A), with the portion of reads that have base call quality below 20 after 200 sequencing cycles reaching 5 % and 15 % in the forward and reverse reads, respectively.

FLASH program was used to generate consensus sequences with the following settings: minimal overlap – 50 bp, maximal portion of unmatched positions within overlap – 25 %. Consensus sequences outside the length range 440–475 bp were removed. This resulted in dataset of 18870 consensus sequences in ALIOPT dataset. The quality of consensus sequences is decreased in the middle part, corresponding to overlap of the forward and reverse reads (Fig. 2,B). In several positions, the share of bases, which have the consensus quality value below 20, reaches 28 %.

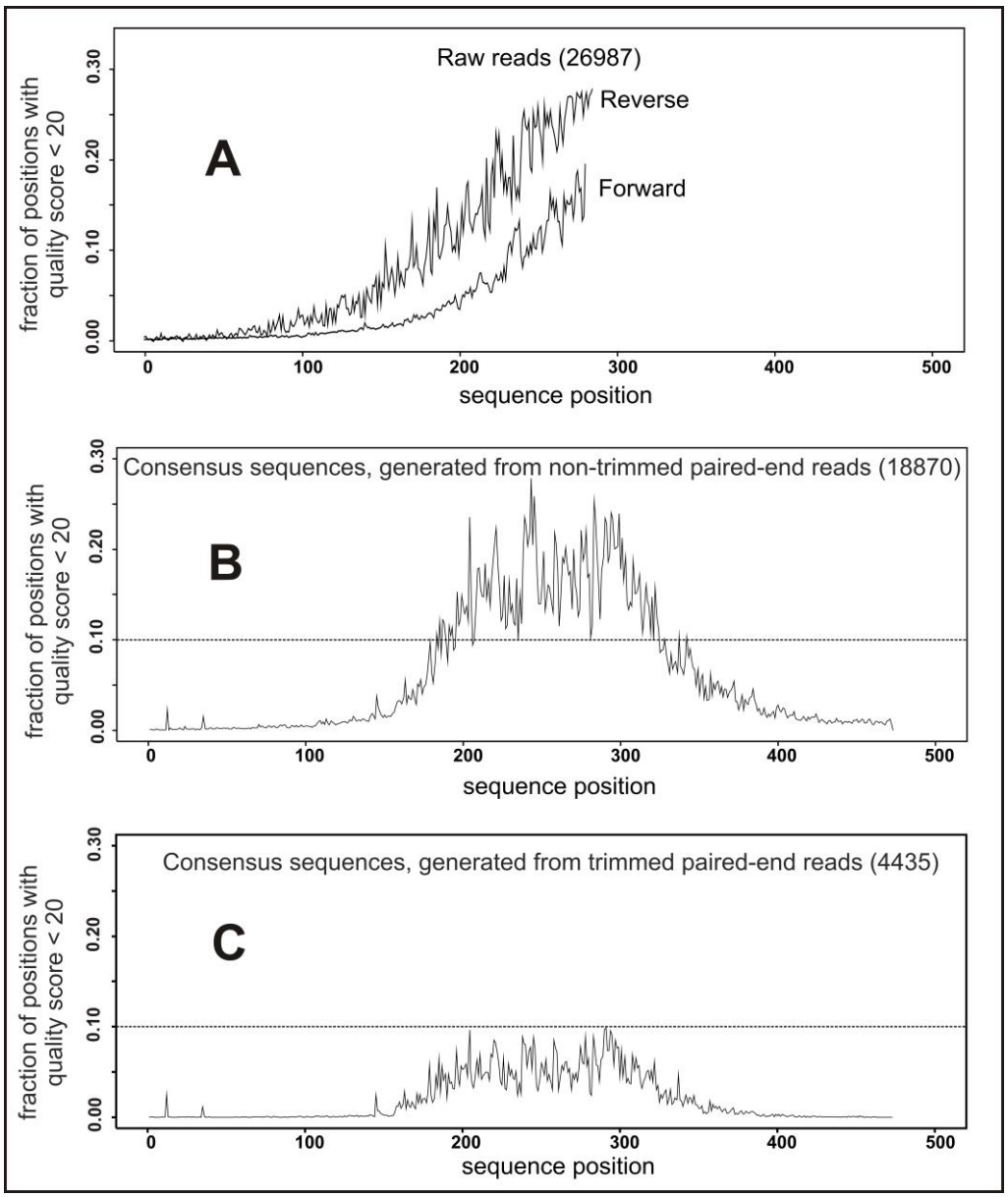

Fig. 2. Quality of raw paired-end reads and consensus sequences.

For STD dataset, Trimmomatic settings were as follows: size of sliding window – 10 bp, average base call quality – 23. After generation of consensus sequences and length filtering, this resulted in 4435 sequences. The portion of sequences having consensus quality above 20 for certain position does not exceeded 10 % (Fig. 2,C).

Distribution of lengths of consensus sequences in STD dataset is changed after trimming procedure (Fig. 3,A) mainly by decrease of the portion of long sequences. This can be explained by the non-uniform elimination of consensus sequences during filtration: longer fragments have shorted overlap with lower overall consensus quality (Fig. 2).

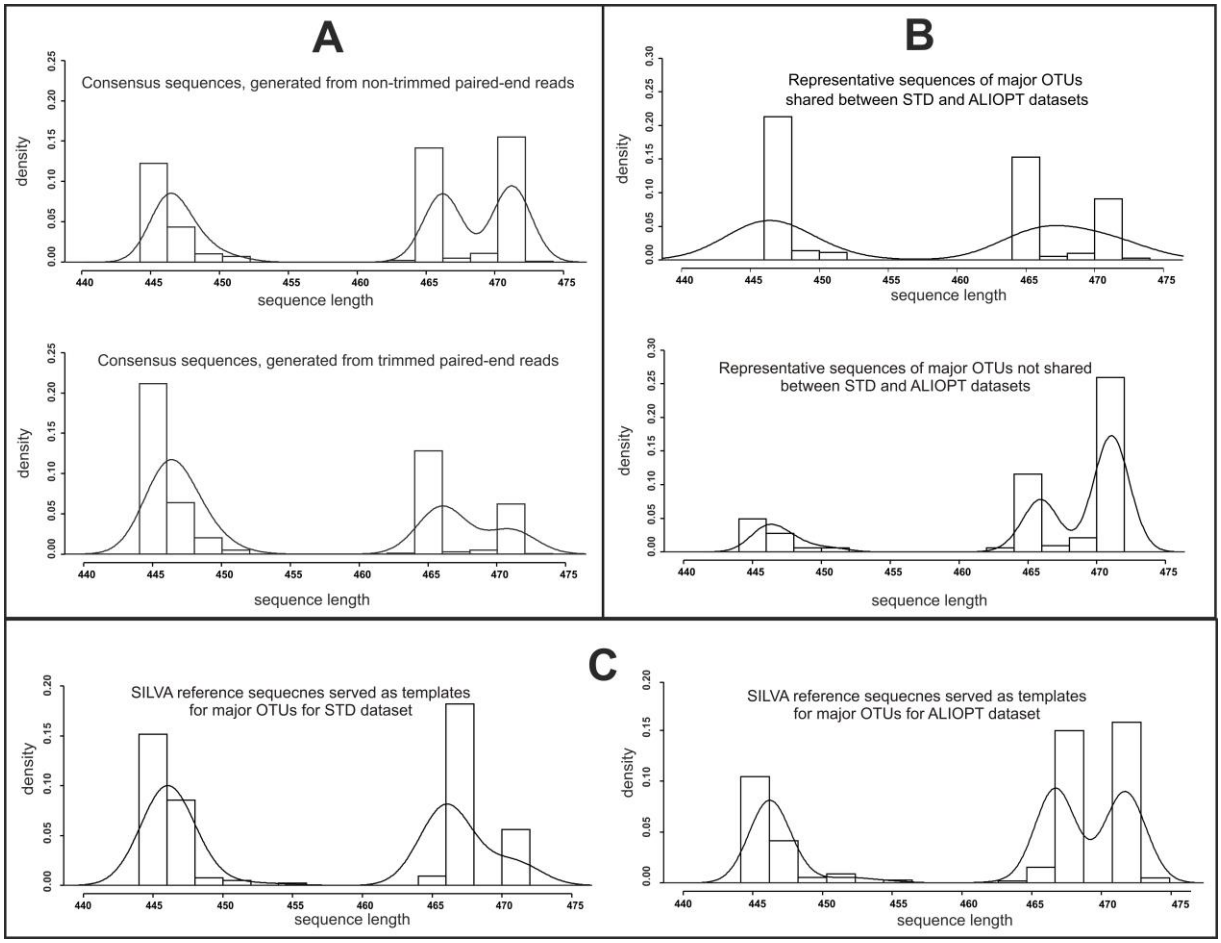

Fig. 3. Sequence length distribution on different stages of filtering pipelines.

Consensus sequences of trimmed and non-trimmed datasets were aligned with sequences of 16S rRNAs from SILVA database. Sequences aligned outside V1-V3 region of 16S rRNA and classified as non-bacterial were removed. After this stage, trimmed and non-trimmed datasets contained 3840 and 17160 sequences, respectively. For non-trimmed data, the developed procedure of filtration was applied, generating ALIOPT dataset.

Clustering results for STD and ALIOPT datasets are presented in Table 1. The portion of orphan OTUs is relatively small in both datasets. This suggests reasonable error rate in raw NGS data, as well as acceptable quality of sequencing and preliminary data filtration. Slight decrease of average length of consensus sequences in ALIOPT dataset seems to have no influence to the number of orphan OTUs. On the other hand, clustering results for STD and ALIOPT datasets differ significantly in terms of major (184 vs 319) and underestimated (291 vs 499) OTUs. Due to the fact that both datasets characterize the same sampled bacterial community, the difference of these estimates is unexpected. Indeed, the portion of underestimated OTUs should be higher for STD dataset since it contains only 3840 consensus sequences, whereas ALIOPT dataset has 17160. Yet, the proportion of underestimated OTUs is similar in both datasets (63 %). In summary, these estimates of α-diversity suggest that the

diversity of STD dataset is lower than that of ALIOPT dataset. In other words, non-uniform elimination of sequences representing specific taxons (OTUs) occurs during standard trimming procedure.

Table 1. Diversity of bacterial community calculated for STD and ALIOPT datasets

Index	STD dataset	ALIOPT dataset
Observed number of OTUs	405	1199
Number of non-orphan OTUs	268	567
Number of orphan OTUs	137	632
Portion of consensus sequences in orphan OTUs	0.036	0.037
Number of major OTUs (top 95 % of consensus sequences)	184	319
Bootstrap index (expected number of major OTUs)	291	499
Ratio of number of observed major OTUs to number of expected major OTUs	0.63	0.63
Shannon index calculated for major OTUs	4.23	4.93

Rarefaction analysis performed for both datasets confirms α-diversity estimates (Fig. 4). Asymptotes of rarefaction curves for STD dataset (major OTUs – 184, all OTUs – 268) have lower values than those for ALIOPT dataset (major OTUs – 319, all OTUs – 567). Again, because of the lower number of sequences in STD dataset, one could expect STD curve to have no asymptote and converge with ALIOPT curve, as this dataset is, roughly speaking, a superset of STD dataset. The opposite behavior of STD curve argues for the depletion of taxonomic spectrum in STD dataset during filtration of raw data.

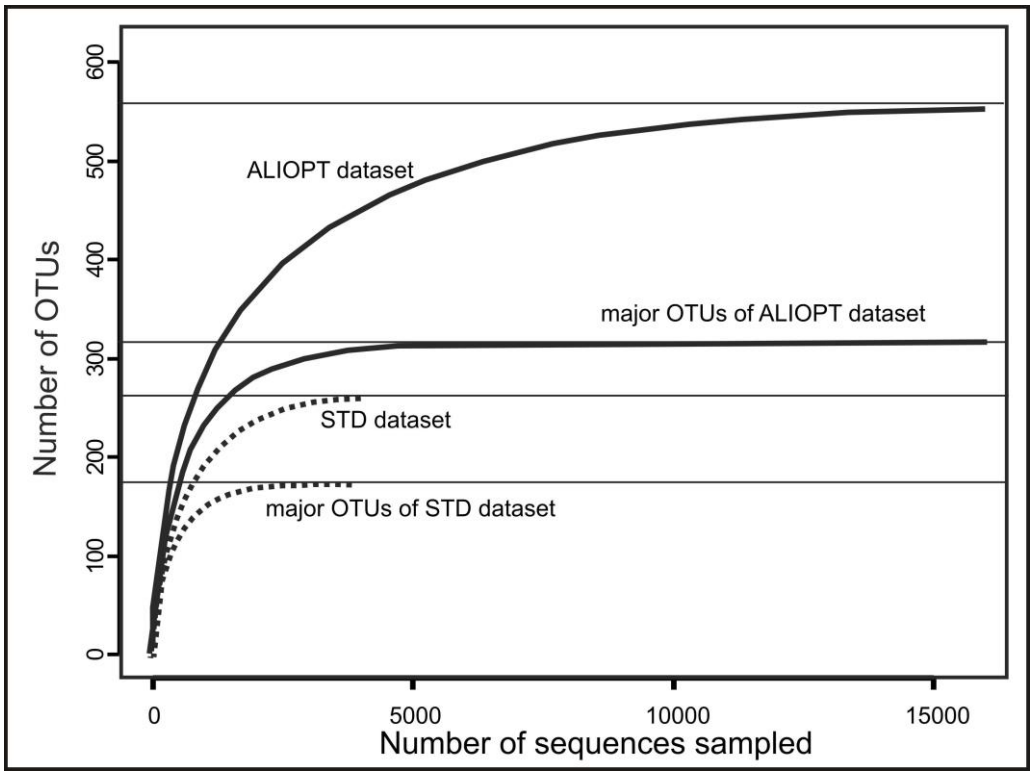

Fig. 4. Rarefaction curves.

To test whether taxonomic composition of bacterial community differs between STD and ALIOPT datasets, we performed taxonomic classification of sequences representing major

OTUs up to family rank (Fig. 5). Fractions of top 5 major OTUs are significantly different in two datasets. For instance, there is almost fourfold difference of the fraction of the Alteromonadaceae family in STD and ALIOPT datasets (1.8 % vs. 6.7 %.). This suggests that the observed taxonomic composition of communities differs with respect to filtration procedure applied to raw data.

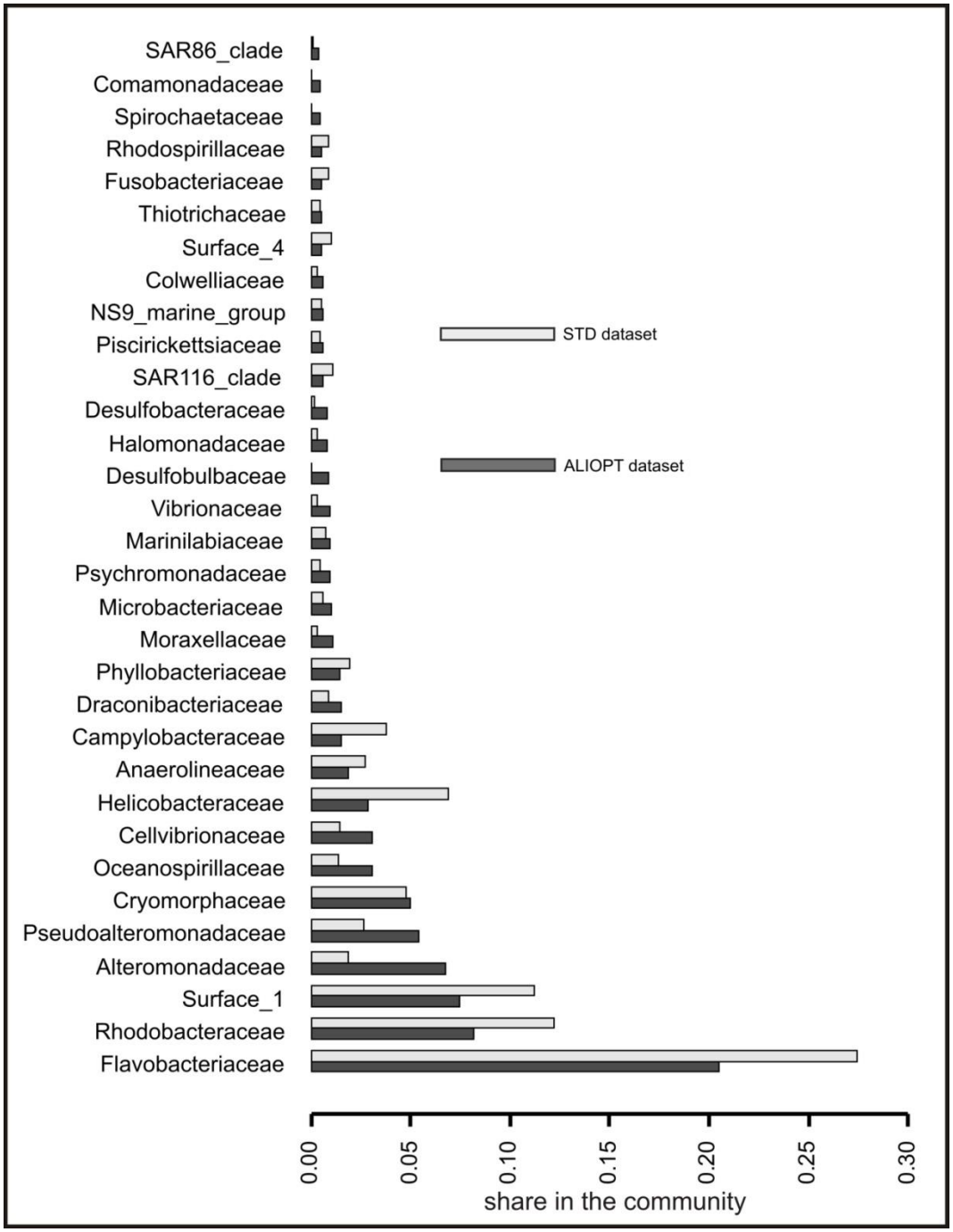

Fig. 5. Taxonomic composition of bacterial community by family rank for major OTUs.

We tested in two ways whether non-uniform elimination of consensus sequences of different length occurs during standard filtration procedure. First, we compared distributions of length of representative sequences of major OTUs from STD and ALIOPT datasets. Representative sequences of major OTUs from both datasets were merged and aligned. OTUs shared between datasets were found using the pairwise distance matrix. OTU was marked as shared if the genetic distance between the representative sequences was less than 0.03. 174 out of 184 major OTUs of STD dataset were shared with OTUs of ALIOPT dataset. Thus,

only 5 % of major OTUs obtained by standard trimming procedure were absent in ALIOPT dataset. On the other hand, 145 out of 319 major OTUs of ALIOPT dataset were not shared with STD dataset. Importantly, a portion of representative sequences falling into length range 445–455 bp is higher within a set of OTUs shared between STD and ALIOPT (Fig. 3,B), whereas a set of representative sequences unique to one of the datasets is enriched by longer sequences (465–475 bp). Finally, we analyzed 16S rRNA sequences from reference SILVA database to test, whether the template sequences belonging to different taxons have lengths variability. We extracted the SILVA database sequences that served as templates for reference alignment of OTU representative sequences. The region of SILVA-templates that corresponds to analyzed 16S rRNA fragment was extracted by applying the “hard-mask” in Mothur. Overall, spectrum of template lengths for STD dataset is shifted to the left, whereas templates for ALIOPT dataset are enriched by longer sequences (Fig. 3,C). These two results complement each other and argue for non-uniform elimination of sequences of different length during standard filtration procedure.

CONCLUSIONS

Standard procedures of NGS metagenomic amplicon data filtration and trimming were revealed to be prone to introduce bias in genetic diversity of analyzed community and therefore to alter its taxonomic composition. We explored the reasons of this bias and suggested a new method of data filtration which avoids such type of errors. This method can be efficiently applied to Illumina MiSeq paired-end amplicon data generated for arbitrary gene marker to improve results of metagenomic analysis.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Irkutsk Supercomputer Center of SB RAS for providing the access to HPC-cluster «Akademik V.M. Matrosov». We also thank to Ivan Sidorov, system administrator of HPC-cluster, for help in performing computations. The work was carried out with the financial support of the integration project 4.1.2 "Application of the NGS-BD (Next Generation Sequencing – Big Data) methods for solving environmental issues".

REFERENCES

1. Bolger A.M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014. V. 30. № 15. P. 2114–2120. doi: 10.1093/bioinformatics/btu170
2. Chun J., Kim K.Y., Lee J.H., Choi Y. The analysis of oral microbial communities of wild-type and toll-like receptor 2-deficient mice using a 454 GS FLX Titanium pyrosequencer. *BMC Microbiology*. 2010. V. 10. № 1. P. 101. doi: 10.1186/1471-2180-10-101
3. Dixon P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*. 2003. V. 14. № 6. P. 927–930. doi: 10.1658/1100-9233(2003)014[0927:VAPORF]2.0.CO;2
4. Fosso B., Santamaria M., Marzano M., Alonso-Alemany D., Valiente G., Donvito G., Monaco A., Notarangelo P., Pesole G. BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS. *BMC Bioinformatics*. 2015. V. 16. № 1. P. 203. doi: 10.1186/s12859-015-0595-z
5. Katoh K., Toh H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*. 2010. V. 26. № 15. P. 1899–1900. doi: 10.1093/bioinformatics/btq224
6. Kim M., Lee, K.H., Yoon S.W., Kim B.S., Chun J., Yi H. Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics & Informatics*. 2013. V. 11. V. 3. P. 102–113. doi: 10.5808/GI.2013.11.3.102

7. Magoč T., Salzberg S.L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011. V. 27. № 21. P. 2957–2963. doi: 10.1093/bioinformatics/btr507
8. Morgan M., Anders S., Lawrence M., Aboyoun P., Pages H., Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*. 2009. V. 25. № 19. P. 2607–2608. doi: 10.1093/bioinformatics/btp450
9. Petrosino J.F., Highlander S., Luna R.A., Gibbs R.A., Versalovic J. Metagenomic pyrosequencing and microbial identification. *Clinical Chemistry*. 2009. V. 55. № 5. P. 856–866. doi: 10.1373/clinchem.2008
10. Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., Connor T.R., Bertoni A., Swerdlow H.P., Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*. 2012. V. 13. № 1. P. 341. doi: 10.1186/1471-2164-13-341
11. Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J., Glöckner F.O. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*. 2013. V. 41. № D1. P. D590–D596. doi: 10.1093/nar/gks1219
12. Schloss P.D., Westcott S.L., Ryabin T., Hall J.R., Hartmann M., Hollister E.B., Lesniewski R.A., Oakley B.B., Parks D.H., Robinson C.J., Sahl J.W., Stres B., Thallinger G.G., Van Horn D.J., Weber C.F. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*. 2009. V. 75. № 23. P. 7537–7541. doi: 10.1128/AEM.01541-09
13. Smith E.P., van Belle G. Nonparametric estimation of species richness. *Biometrics*. 1984. V. 40. № 1. P. 119–129. doi: 10.2307/2530750
14. Tennant R.K., Sambles C.M., Diffey G.E., Moore K.A., Love J. Metagenomic Analysis of Silage. *Journal of Visualized Experiments: JoVE*. 2017. V. 119. doi: 10.3791/54936
15. Zhou J., Wu L., Deng Y., Zhi X., Jiang Y.H., Tu Q., Yang Y. Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME Journal*. 2011. V. 5. № 8. P. 1303–1313. doi: 10.1038/ismej.2011.11

Received 02 October 2017.
Published 15 May 2018.