Jo *et al. Appl Biol Chem* (2019) 62:60
https://doi.org/10.1186/s13765-019-0467-8

KSABC
The Korean Society for Applied Biological Chemistry

NOTE **Open Access**

Investigation of MiSeq reproducibility on biomarker identification

Hyejun Jo, Jiwan Hong and Tatsuya Unno*

Abstract

MiSeq-derived artificial sequences appeared to be of good quality, thus bioinformatics tools failed to remove MiSeq artefacts. Even after removing singleton sequences or operational taxonomic units (OTUs), it is not clear how many sequence artefacts remained. Here, 16S rRNA genes were amplified from soil, human feces, pig feces, and groundwater. These were sequenced with five separate runs of MiSeq. Subsequently, each run of MiSeq was compared through alpha and beta-diversity analyses. We found more than half the OTUs were not in consensus through the multiple MiSeq runs, resulting in varying group-specific biomarker OTUs in each MiSeq run. Thus, differential abundance test should be interpreted with caution, and we suggest that results also should be verified further with other quantification methods such as qPCR.

Keywords: Differential abundance analysis, Microbial community, MiSeq reproducibility, OTU inflation

Introduction

In recent times, MiSeq has become a major sequencing platform for microbial community analyses. The current version of MiSeq allows sequencing both sides of DNA fragments up to 50 million reads with 300 bp read length in one run. On the other hand, it has been reported that output from this platform includes some artificial sequences. Caporaso et al. [3] reported that approximately 10,000 of operational taxonomic units (OTUs) were observed when 1 million reads were obtained from a mock community containing 67 different species. This inflated number of species was considered to be machinery artefacts, thus called 'artificial OTUs' and can be minimized by removing singletons [8] or applying minimum abundance threshold [1]. While these suggestions help to reduce computational workload and increases accuracy of results, as these methods do not completely remove artificial OTUs, leaving left-over artificial OTUs that need further investigation. Linear discriminant analysis effect size (LEfSe) [7] is a method to identify metagenomic biomarkers based on effect size estimation. In many recent microbiome-based studies, this method has been widely applied to identify differentially abundant OTUs. We applied this method to identify biomarker species that are representative of either human feces, swine feces, soil, or groundwater.

Materials and methods

DNA extraction and Miseq Library preparation

Twelve DNA samples (as listed in Additional file 1: Table S1) were extracted using QIAamp PowerFecal DNA Kit or PowerWater DNA Isolation Kit (Qiagen, Hilden, Germany) and two-step PCR MiSeq library was prepared for the V4 region of 16S rRNA gene amplicons according to the manufacturer's instructions. Briefly, the first PCR was conducted to amplify V4 region of 16S rRNA gene using the primers (515F:5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGCCAGCMGCCGCGGTAA-3′ and 806R: 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGACTACHVGGGTWTCTAAT-3′) with the following condition: 95 °C for 3 min; 25 cycles of 95 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s; and 72 °C for 5 min. The second PCR was conducted to attach linkers and barcodes provided by Illumina as follows: 95 °C for 3 min; 8 cycles of 95 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s; and 72 °C for 5 min. An equimolar of the PCR amplicons was pooled and sent for sequencing at Macrogen Inc. (Seoul, Republic of Korea).

*Correspondence: tatsu@jejunu.ac.kr
Faculty of Biotechnology, College of Applied Life Sciences, SARI, Jeju National University, Jeju 63243, Republic of Korea

Miseq data processing and analysis

These 12 DNA samples were sequenced five times in five independent MiSeq runs. Sequencing data were analyzed with MOTHUR 1.42 [6]. In brief, raw sequence data were processed through paired-end assembly, aligned to SILVA database version 128 [4], and chimera removal was achieved with VSEARCH [5]. Singleton sequences were removed and the number of reads was normalized to the minimum number of reads per sample (n = 16,237) prior to downstream analysis. Resulting sequences were clustered using OptiClust [9] to assign OTUs with dissimilarity of 0.03. Alpha and beta-diversity analysis was performed using MOTHUR 'summary.single' and 'nmds' subroutines, respectively. Differentially abundant OTUs were identified using LEfSe. In addition to OTUs, exact sequence variants (ESVs) were assigned to reads using dada2 R packages [2].

Results and discussion

Results from alpha-diversity analysis showed that variation of species richness (Chao) within the five multiple runs was small except for one groundwater sample (Gwater#3) and three soil samples [Additional file 1: Fig. S1(A)]. Gwater#3 was found to have higher concentrations of nitrate and *E. coli* compared to the other groundwater samples (Additional file 1: Table S1), indicative of fecal matter or fertilizer contamination. As a result of fecal matter contamination, Gwater#3 showed higher species richness likely due to the presence of fecal bacteria. Including soil samples, these samples with higher richness showed variations in Chao as seen in the five MiSeq runs. On the other hand, species evenness (Shannon) showed a little variation among runs [Additional 1: Fig. S1(B)], suggesting that species evenness was not affected by multiple MiSeq runs, unlike species richness. Non-metric multidimensional scaling (NMDS) analysis showed that samples were clustered based on sample types (i.e., feces, soil, and groundwater) as well as IDs (i.e., #1, #2, and #3) [Additional file 1: Fig. S1(C)], except for Human#1, Soil#2, and Soil#3 which differed once (Additional file 1: Fig. S2).

The number of shared-OTUs among MiSeq runs are summarized in Fig. 1. Environmental samples (i.e., groundwater and soil) showed a high number of run-specific OTUs that were not shared between MiSeq runs. On the other hand, fecal samples (i.e., human and pig feces) showed that the number of consensus OTUs (found through five MiSeq runs) was almost the same as that of run-specific OTUs. These results suggested that environmental samples were more prone to suffer from OTU-inflation. Regardless of the sample types, our results suggest that more than half the OTUs are possibly a consequence of sequence artefacts. The maximum read abundance of these OTUs is summarized in Additional file 1: Table S2. Interestingly, non-consensus OTUs (OTUs that were not identified through all the MiSeq

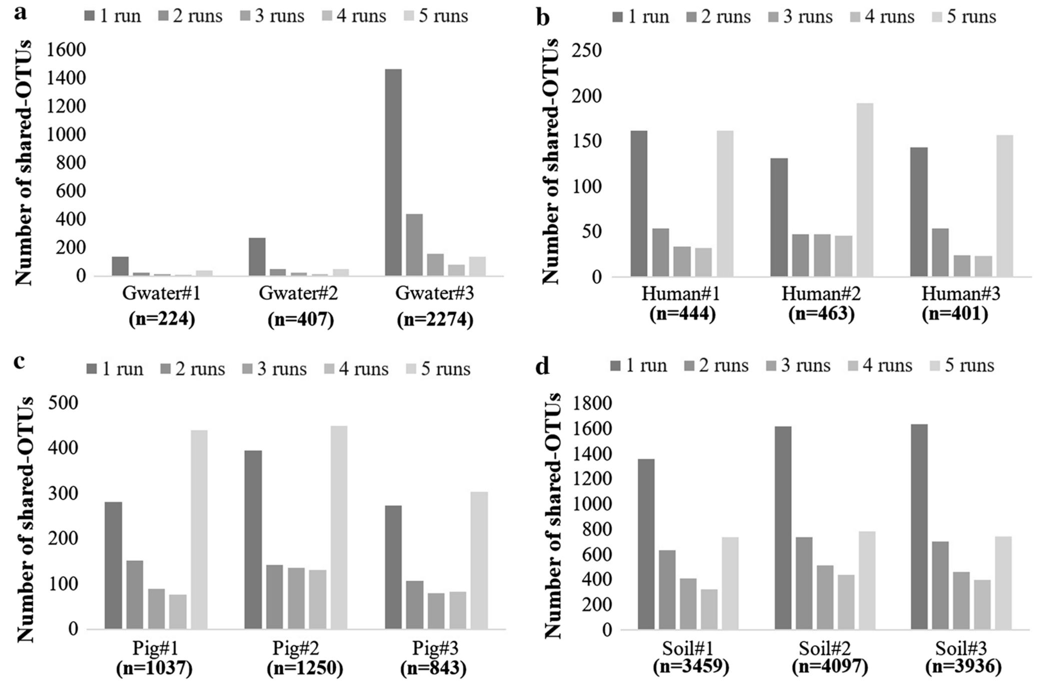

Fig. 1 Number of shared operational taxonomic units (OTUs) among multiple MiSeq runs for DNA obtained from groundwater (**a**), human feces (**b**), pig feces (**c**), and soils (**d**)

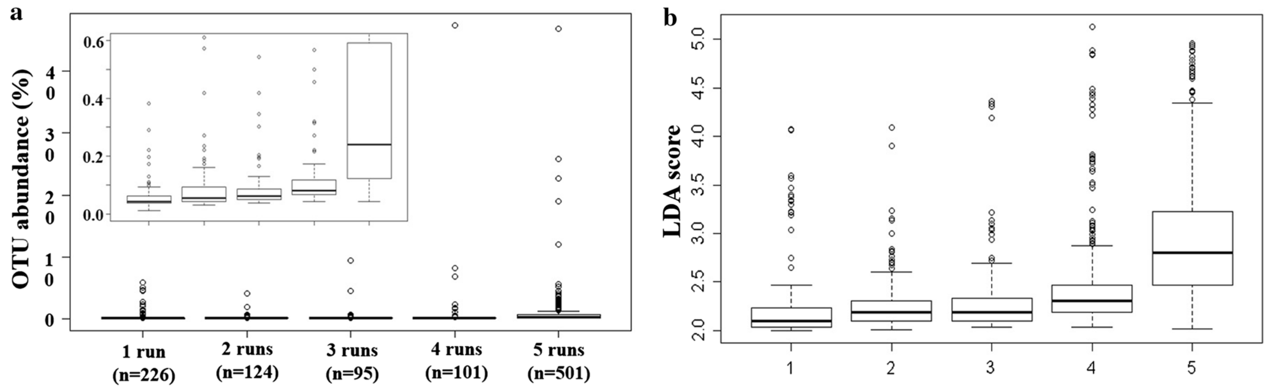

Fig. 2 Number and abundance of biomarker OTUs consensually identified among MiSeq runs. Inlet graph is the same graph scaled up to 0.6% OTU abundance (**a**), and distribution of the linear discriminant analysis (LDA) score for each differential OTUs consensually identified among MiSeq runs (**b**)

runs) were low in abundance (< 0.7%) except for one OTU in Human#2 and Pig#3, which was 1.9% and 1.1% abundant, respectively. Therefore, our results imply that applying a 0.7% abundance threshold will help to remove most of the artificial OTUs. On the other hand, the number of shared ESVs showed higher portion of run-specific ESVs and lower portion of consensus ESVs compared to those of OTUs, suggesting that ESVs were more prone to sequence artefacts than OTUs (Additional file 1: Fig. S3).

To investigate how these artificial OTUs affected the selection of biomarkers, the data set obtained from each MiSeq run was subjected to differential abundance tests to identify biomarker OTUs that represent the sample types used in this study. The numbers of OTUs identified through the MiSeq runs and their abundance are summarized in Fig. 2. Four runs exhibited a wide range of OTU abundance (> 40%) for non-consensus biomarker OTUs (Fig. 2a). Linear discriminant analysis (LDA) scores for these non-consensus biomarker OTUs ranged from 2.0 to 5.0; thus, artificial biomarker OTUs cannot be screened out based on LDA scores (Fig. 2b). While removing low abundance OTUs (i.e., 1%) reduces a substantial amount of sequence artefacts, a large portion of non-artificial biomarker OTUs would be lost as well. Thus, simply applying an abundance threshold will lead to loss of important findings about consensus biomarker OTUs. However, our results indicate that applying lower abundance (i.e., 0.1%) may remove most of the artificial biomarker OTUs identified up to four runs, while retaining more than 75% of consensus biomarker OTUs.

In summary, the reproducibility of MiSeq was investigated through five runs of the 12 DNA samples isolated from soil, groundwater, and human and pig feces. While beta-diversity analysis did not show significant difference among runs, some samples with higher richness showed slight variation in alpha-diversity analysis. The number of run-specific OTUs was more than half the total OTUs and that the environmental samples tend to suffer higher OTU-inflation. A large part of biomarker OTUs identified through the differential abundance test comprised the possible sequence artefacts, but our results indicate that a 0.1% abundance threshold reduced the false identification of biomarker OTUs. Since a differential abundance test for biomarker OTU identification is one of the commonly applied approaches in MiSeq-based microbial community analysis, our results indicate that biomarker OTUs with extremely low abundance should be interpreted with caution. Therefore, if results from the differential abundance test includes rare OTUs, we suggest that verification of the results by quantification methods such as qPCR or colony counting may be applicable if species-specific primers and selective agars are available.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13765-019-0467-8.

Additional file 1: Table S1. Description of DNA samples used in this study. **Table S2.** Maximum read abundance (%) of shared OTUs between multiple MiSeq runs. **Fig. S1.** Differences in species richness (**A**), evenness (**B**), and beta-diversity (**C**) among multiple MiSeq runs. **Fig. S2.** Tree-based beta-diversity comparison between multiple MiSeq runs. Sample names indicate type of samples (i.e., soil, human, groundwater, and pig) followed by the number indicating different run of MiSeq and sample IDs (i.e., 1, 2, and 3). **Fig. S3.** Number of shared exact sequence variants (ESVs) between multiple MiSeq runs.

Acknowledgements

We are grateful to Sustainable Agriculture Research Institute (SARI) in Jeju National University for providing the experimental facilities.

Authors' contributions

HJ, JH carried out the experiment and wrote the manuscript. TU supervised this work. All authors discussed the results and commented on the manuscript. Both authors read and approved the final manuscript.

Funding

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through Agricultural Microbiome R&D Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (918021-4).

Competing interests

The authors declare that they have no competing interests.

Received: 16 August 2019 Accepted: 14 October 2019
Published online: 29 October 2019

References

1. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI et al (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10:57–59
2. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ et al (2016) DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583
3. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA et al (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 108(Suppl 1):4516–4522
4. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T et al (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596
5. Rognes T, Flouri T, Nichols B, Quince C, Mahe F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584
6. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541
7. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L et al (2011) Metagenomic biomarker discovery and explanation. Genome Biol 12:R60
8. Unno T (2015) Bioinformatic suggestions on Miseq-based microbial community analysis. J Microbiol Biotechnol 25:765–770
9. Westcott SL, Schloss PD (2017) OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. mSphere. 2:e00073-17

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.