J. Microbiol. Biotechnol. (2015), 25(6), 765–770
http://dx.doi.org/10.4014/jmb.1409.09057

jmb

Bioinformatic Suggestions on MiSeq-Based Microbial Community Analysis S

Tatsuya Unno*

Faculty of Biotechnology, College of Applied Life Sciences, Jeju National University, Jeju 690-756, Republic of Korea

Received: September 18, 2014
Revised: November 18, 2014
Accepted: January 7, 2015

First published online
January 7, 2015

**Corresponding author*
Phone: +82-64-754-3354;
Fax: +82-64-756-3351;
E-mail: tatsu@jejunu.ac.kr

Supplementary data for this paper are available on-line only at http://jmb.or.kr.

pISSN 1017-7825, eISSN 1738-8872

Recent sequencing technology development has revolutionized fields of microbial ecology. MiSeq-based microbial community analysis allows us to sequence more than a few hundred samples at a time, which is far more cost-effective than pyrosequencing. The approach, however, has not been preferably used owing to computational difficulties of processing huge amounts of data as well as known Illumina-derived artefact problems with amplicon sequencing. The choice of assembly software to take advantage of paired-end sequencing and methods to remove Illumina artefacts sequences are discussed. The protocol we suggest not only removed erroneous reads, but also dramatically reduced computational workload, which allows even a typical desktop computer to process a huge amount of sequence data generated with Illumina sequencers. We also developed a Web interface (http://biotech.jejunu.ac.kr/~abl/16s/) that allows users to conduct fastq-merging and mothur batch creation. The study presented here should provide technical advantages and supports in applying MiSeq-based microbial community analysis.

Keywords: 16S rRNA, bioinformatics, microbial community, mothur, pear

Introduction

Recent development in sequencing technology has been revolutionizing the fields of various study areas, and more and more microbial-ecology-related research outcomes are being reported. Next-generation sequencing (NGS) is a term used for non-Sanger-based sequencing technology that allows high-throughput sequencing of millions/billions of DNA fragments without a need for cloning [18]. NGS has been widely applied to various types of research involving genome, metagenome, and PCR amplicon sequencing. There are several platforms currently considered to be NGS, which includes Roche 454, Illumina, Ion Torrent, and ABI SOLiD technologies. Different platforms have different accuracy, coverage rate, and systematic biases [7]. Among NGS platforms, 454 pyrosequencing (GS FLX Titanium) has been preferably used for 16S rRNA gene-based microbial community analysis, because of its relatively long read length (>450 bp) and high consensus accuracy (99.995%).

The length of the 16S rRNA is approximately 1,500 bp, which is not within a range of any NGS platform's read length. Therefore, hypervariable regions of the 16S rRNA gene are commonly used to study bacterial ecology. Youssef *et al*. [19] suggested to use V4, V4+V5, or V6+V7 regions, because other combinations of variable regions resulted in either over or under estimation of species richness. On the other hand, Liu *et al*. [12] reported that analysis of short reads (*i.e*., 100 bp) captures the same microbial community differences (beta-diversity) as done by full-length 16S rRNA gene sequencing, suggesting that NGS platforms that provide short read lengths (*i.e*., Illumina) can be used for microbial community analysis. Recently, the Illumina-based microbial community analysis based on V6 sequences has been reported to be cost-effective, in which results were also comparable to that of the 454 pyrosequencing [3, 5, 21], although the sequence of the V6 region is highly divergent and short (60 bp). Illumina-based microbial community analysis, however, has major drawback suchs as artefacts that often appear as erroneous reads and imperfect operational taxonomic units (OTUs) [3]. For this reason, the Illumina sequencing platform

has not been preferably used for microbial community analyses, especially ones targeting rare taxa [3, 15].

Recently, Illumina's MiSeq platform has become capable of producing paired 250–300 bp reads with high sequencing capacity (7.5–8.5 Gb), equivalent to maximum 25 million paired-end reads. The use of MiSeq for microbial community analysis was first reported by Caporaso *et al*. [2], in which 1.5 Gb (5 million 150 bp paired-end reads) were generated per day. Furthermore, Kozich *et al*. [9] reported the dual-index sequencing strategy, allowing to multiplex 384 samples using only 40 primers (16 × 24) with an average of 180,000–200,000 reads per sample. Although the aforementioned drawback remains unsolved, microbial community analysis using MiSeq is getting more attentions owing to its longer read length, cost-effectiveness, and high multiplicity, which has never been achieved with other NGS platforms known to date.

Mothur [17] and Qiime [1] are the most frequently used free bioinformatics software for NGS output for 16S rRNA sequence-based microbial community analysis. Whereas Qiime runs on the Linux environment only, Mothur is an OS-independent software, and therefore it could be a likely choice for most non-bioinformatics experts. In addition, the developers of Mothur have established a standard operation protocol for MiSeq-based microbial community analysis, so-called MiSeq SOP (http://www.mothur.org/wiki/MiSeq_SOP), which not only provides bioinformatics approaches but also includes the preparation of samples such as DNA extraction and PCR. In this study, we sequenced the V4 region of the 16S rRNA gene amplified from 52 samples to investigate the applicability of MiSeq-based analysis. Here, we discuss the choice of the assembly software, computational processing workload, and Illumina artefacts removal.

Methods

Sequencing Data Preparation

We sequenced a total of 281 samples, including soils, feces, and freshwater. DNA was extracted using a MOBIO Power Soil DNA isolation kit (MO BIO Laboratories Inc., CA, USA). Whereas soil and feces were directly processed for DNA extraction, 0.45 μm pore-size membranes (Advantec, Japan) were used to filter 500 ml of freshwater and sliced with sterile blades prior to the DNA extraction. MiSeq SOP was applied for sequencing sample preparations. Briefly, 2 μl of the total DNA from each sample was used as a template with primers containing the Illumina adaptor sequence and universal V4 region of the 16S rRNA gene (Table S1), and amplification was done in triplicates using the Maxime PCR PreMix Kit (iNtRON Biotechnology Inc., Republic of Korea) with the following conditions: 95ºC for 2 min; 30 cycles: 95ºC for 20 sec, 55ºC for 15 sec, 72ºC for 1 min; 72ºC for 5 min. The PCR products obtained were further gel-purified using an AccuPrep Gel Purification kit (Bioneer Inc., Republic of Korea). All obtained DNA were quantified using Qubit (Invitrogen, CA, USA), and equimolar purified amplicons were pooled and stored at −20°C until sequenced. Amplicons were sequenced using the Illumina MiSeq platform at Macrogen Inc. (Seoul, Republic of Korea) according to the manufacturer's instructions. A total of 52 samples (Table S2), including 6 paddy soils and 18 fecal samples obtained from 3 pigs, 3 beef cows, 3 milk cows, 3 ducks, 3 chickens, and 3 humans; and 28 freshwater, were selected to investigate the applicability of the processing protocol with Mothur.

Sequence data used in this study were deposited to Sequence Read Archive (SRA) with the accession number PRJNA258105.

Sequence Processing and Analysis

Mothur pipeline was used for the entire sequence data processing according to the Mothur SOP. Briefly, we conducted error removals through screening sequences that did not align to Silva database [16], preclustering to merge rare sequences into larger sequences if the difference is within one or two base pairs. Chimeric sequences were removed by using *uchime* [6]. Taxonomic classification was done based on Ribosomal Database Project [4] training set ver. 9, followed by non-bacterial sequence removal. OTUs were calculated at distance 0.03 using the Mothur subroutine cluster.split. Microbial community dissimilarity was analyzed based on the Yue and Clayton theta coefficient calculated by the tree.shared Mothur subroutine.

Results and Discussion

Paired-End Assembly

Table 1 compares assembly results between Mothur and PEAR for each type of sample. Whereas similar deviation was obtained regardless of the assembly software, a greater percentage of assembled sequences was observed for PEAR. Although this did not affect the overall taxonomic classification results (Fig. S1), it would be critical if only fewer sequences were recovered. It should be noted that Mothur assembled only 31.8% of sequences, whereas 91.2% was assembled by PEAR (Table 1).

Mothur's subroutine make.contigs was used to assemble paired-end (PE) reads. The make.contigs subroutine first aligns PE sequences, and it identifies any positions where the two reads disagree. Mismatches are screened based on quality score (25 for a gap/base and 6 or more for base/base mismatches), otherwise assigns "N." In our data set, Mothur assigned "N" to 619,941 reads (28%), which needs to be removed from further analysis. Among the 52 samples, one sample (SoilF12) was selected to investigate how a relatively large portion of the sequence data contained

Table 1. Comparison of assembly results between Mothur and PEAR for each type of samples.

DNA sources	Percentage averaged resulting sequences (standard deviation)	
	Mothur	PEAR
Beef cow feces	78.6 (0.6)	97.8 (1.2)
Chicken feces	73.4 (6.2)	94.4 (6.9)
Duck feces	75.6 (5.9)	91.4 (7.2)
Freshwater	75.9 (4.4)	95.8 (3.7)
Human feces	55.5 (18.0)	73.1 (21.3)
Milk cow feces	77.1 (1.9)	96.7 (1.6)
Pig feces	74.8 (3.0)	94.9 (3.3)
Soil	31.8 (4.8)	91.2 (2.4)
Total	69.6 (15.6)	93.8 (7.8)

uncalled bases, "N." There were 51,307 reads for SoilF12, and 146 and 7,865 reads contained uncalled bases in the forward and reverse sequence data, respectively. Whereas forward sequencing results showed a relatively high quality score between 1 and 150 bp, the reverse showed a low quality score in the same area (Fig. S2). It is expected that PE merging is done by taking up bases with higher quality score and simply disregarding low bases with low quality score. Against our expectations, the resulting assembled reads contained more uncalled bases (38,692 reads). The purpose of paired-end reading is to correct errors on one another; however, our results show that PE assembly increased the number of reads with uncalled bases.

Assembling PE reads may seem simple; however, there exist several algorithms used in various assemblers such as PANDAseq [14], FLASH [13], and COPE [11]. Recently, PE ReAd mergeR, PEAR [20], was developed for PE reads assembly, which is implemented with the Bowtie2 [10] alignment algorithm against reference genomes and had shown a lower false-positive rate when compared with other assembly software. PEAR was designed to maximize assembly scores by choosing called bases when end-to-end alignment shows mismatches between bases and uncalled bases. In this way, if one of the PE reads showed a low quality score or an uncalled base, the assembled reads will take called bases; thus PE reads can compensate each other unlike Mothur's make.contigs. Illumina sequencers are not designed for PCR amplicon sequencing, and one of the benefits of using paired-end reading is to compensate the low sequence quality, especially near the 3′ end. Therefore, the PE read assembler should maximize merging qualities. Whereas Mothur resulted in assembling 38,692 reads with uncalled bases, PEAR assembly showed only 1,280 reads with uncalled bases. The same trends were observed for other samples (Table S2). As suggested in MiSeq SOP, we applied sequence alignment, chimera check, and non-bacterial/archaeal sequence removal. Starting from 2,213,683 raw sequences, Mothur resulted in removing nearly 40% of the sequence, whereas PEAR removed only 13% (Table 2).

Table 2. Number of reads remained at each error removal step.

Error removal steps	Number of reads	
	Mothur	PEAR
PE assembly	2,213,683	2,088,290
Length (<275), "N"	1,573,667	2,077,181
Alignment	1,570,940	2,064,210
Pre-clustering	1,409,135	2,064,210
Chimera check	1,388,921	1,945,712
Non-bacterial taxa	1,333,967	1,918,081

Clustering

Mothur offers two choices of clustering (*i.e.*, cluster and cluster.split subroutines). Since the cluster subroutine requires a high RAM resource, it might be difficult to cluster huge data sets generated with MiSeq. On the other hand, the cluster.split subroutine splits the data set according to the taxonomy level, and thus minimizes the number of sequences to be clustered and reduces the computational resource requirement. In this study, we used a Dell Web server equipped with 24 CPU cores and 98 GB RAM, and it took nearly 24 h to cluster 236,595 unique reads.

As a result of clustering, we obtained a total of 83,234 OTUs. Rarefaction analysis was conducted for fecal DNA samples (Fig. 1A). In this study, more than 9,000 reads were obtained for each fecal sample, which should be enough sequence-depth as compared with previous studies published to date. However, results from the rarefaction curve analysis indicated lack of sequencing efforts. The results of the diversity estimation seemed too high for fecal bacteria communities.

When Illumina sequencers were first used for PCR amplicon analysis, it was reported that Illumina artefacts appeared to be "rare taxa," which inflates microbial communities [3]. In addition, these artefacts pass through a series of error detections such as alignment and chimeric sequence removals. To remove such artefacts, we applied sequence abundance thresholds and investigated the number of resulting reads and OTUs. Fig. 2 describes the number of reads and OTUs according to the sequence

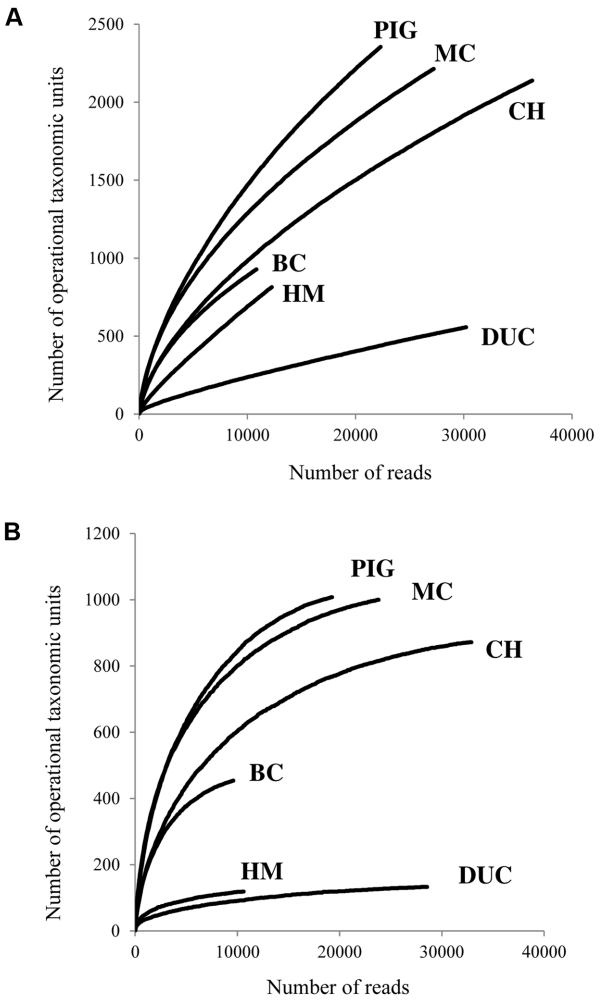

Fig. 1. Rarefaction analysis for the subset data containing fecal sample 16S rRNA gene sequences before (**A**) and after (**B**) abundance threshold was applied.
CH, BC, MC, HM, DUC, and PIG indicate chicken, beef cow, milk cow, human, duck, and pig, respectively.

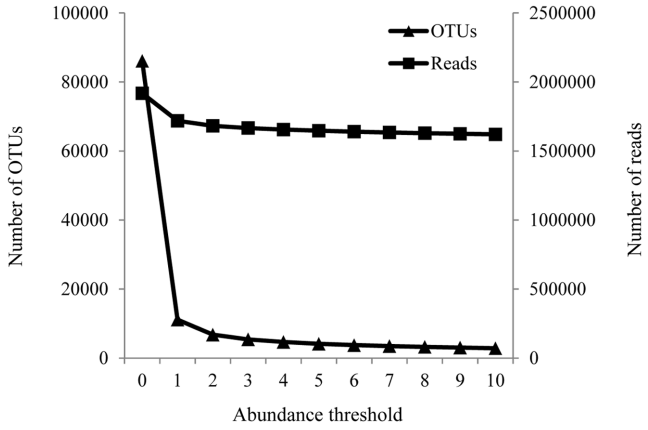

Fig. 2. Effects of abundance threshold on total number of reads and operational taxonomic units (OTUs).

abundance threshold. When abundance threshold "1" was applied, the number of OTUs decreased dramatically. On the other hand, the total number of reads was relatively maintained through the 10 different abundance thresholds, suggesting the majority of sequences exist in at least 10 replications. While most of the sequences exist in such high replicates, the numbers of singleton reads appeared to account for more than 85% of OTUs (Fig. 2). Considering PCR amplification logic, it is not likely that such non-amplified singletons exist in high diversity; therefore, these singletons were possible artefacts and should be removed.

Removing singleton reads resulted in nearly 85% of OTUs loss, which only removed approximately 10% of the total number of reads. After the removal of singletons, the rarefaction curve became saturated (Fig. 1B), which seems to be more comparable to what has been published previously. Although not instructed in the MiSeq SOP, we recommend the use of abundance threshold to remove possible Illumina artefacts. We also suggest that singleton removal should be done after pre-clustering. Pre-clustering was first suggested by Huse *et al*. [8] to convert rare sequence types into more frequent sequence types by adjusting a few base pairs. Mothur can do this with the pre.cluster subroutine, which adjusted nearly 50% of the unique sequences to the major sequence types (519,468 to 267,407 reads) in this study. If the abundance threshold is applied before pre-clustering, we might lose those adjustable erroneous data. In addition, the use of abundance threshold using the Mothur split.abund subroutine not only removed possible artefacts but also dramatically reduced the computational workload. As aforementioned, clustering 1.9 million reads took nearly 24 h; however, it took only 5 min to cluster 1.7 million reads after split.abund was applied. This also reduced the time of chimeric sequence detection dramatically (data not shown).

Lastly, we recommend the use of sub.sample for the normalization of sequence abundance. Since MiSeq provides a substantial amount of sequence data, the number of sequences could vary across samples. In this study, we made a subset of fecal, soil, and freshwater samples (total 52 out of 281 samples). The DNA extraction efficiency could vary depending on types of samples, but we observed no significant correlation between the number of reads and sample types (Fig. S3). Nevertheless, the resulting number of reads per sample ranged from 9,776 to 114,664 reads. It has been said that increasing the sequencing depth is

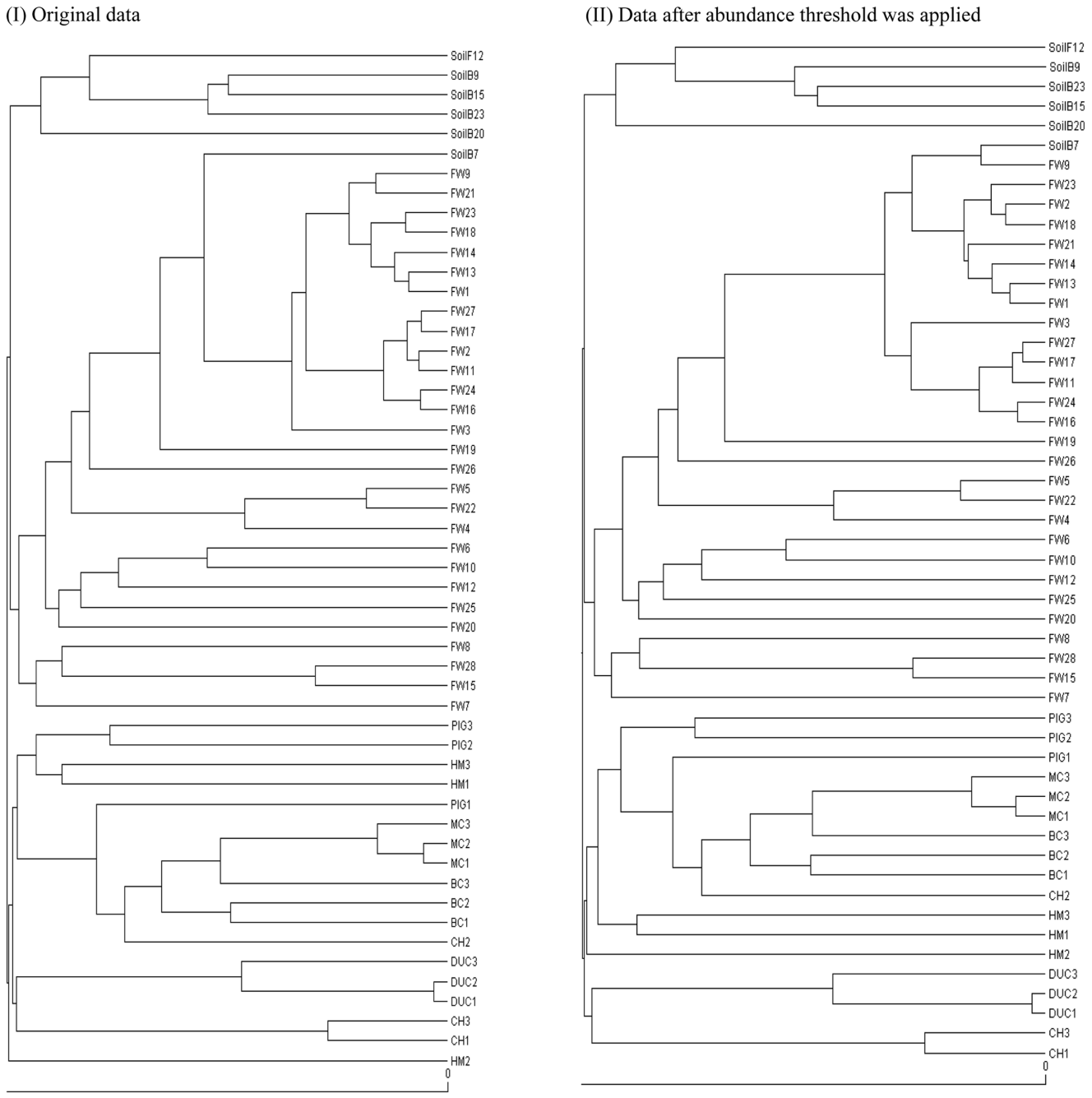

Fig. 3. Effects of abundant threshold and normalization on beta-diversity analysis.
Trees were generated using the raw data (I) and data were processed with split.abund and sub.sample Mothur subroutines (II). Soil, FW, PIG, HM, DUC, CH, BC, and MC denote soil, freshwater, pig feces, human feces, duck feces, chicken feces, beef cow feces, and milk cow feces, respectively.

not likely to provide additional insight [2]; therefore, normalization of the read number for each sample is recommended to not only reduce the tasks of converting the read number to "relative abundance," but also to add logical properness in comparison analyses. Results shown in Fig. 3 suggest that the abundance threshold and read normalization would not change the results of beta-diversity analysis.

Web Interface Support

Since PEAR is only supported to run on the UNIX-based OS and does not provide a "group file," which is required to process with Mothur, we developed a Web interface

(http://biotech.jejunu.ac.kr/~abl/16s) that allows users to merge two paired-end fastq files and create a group file. The Web interface consists of four pages: How to use; Fastq assembly; Group file creation; and Batch file creation. The Fastq assembly page allows you to upload one set of PE-fastq file, and then it sends a download-link to your fasta file *via* email. The Group file creation page allows you to upload one archived file containing merged fasta files, and then it sends a download link to your group file *via* email. Lastly, the Batch file creation page helps you to create a Mothur-batch file as suggested in this study. Links to reference files are also provided on the same page. The Web interface is available to use free of charge.

In summary, recent sequencing technology development offers reasonable per-sample sequencing price as long as data processing methods are well developed and easy to use. It may not be yet for many to apply MiSeq-based microbial community analysis; however, the study presented here should provide technical advantages and support in using MiSeq for microbial community analysis.

Acknowledgments

This work was carried out with the support of "Cooperative Research Program for Agriculture Science & Technology Development (Project PJ009782)," Rural Development Administration, Republic of Korea.

References

1. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7:** 335-336.
2. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, *et al.* 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6:** 1621-1624.
3. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, *et al.* 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. USA* **108 (Suppl 1):** 4516-4522.
4. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, *et al.* 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37:** D141-D145.
5. Degnan PH, Ochman H. 2012. Illumina-based analysis of microbial community diversity. *ISME J.* **6:** 183-194.
6. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194-2200.
7. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, *et al.* 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10:** R32.
8. Huse SM, Welch DM, Morrison HG, Sogin ML. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* **12:** 1889-1898.
9. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79:** 5112-5120.
10. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9:** 357-359.
11. Liu B, Yuan J, Yiu SM, Li Z, Xie Y, Chen Y, *et al.* 2012. COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics* **28:** 2870-2874.
12. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res.* **35:** e120.
13. Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27:** 2957-2963.
14. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. 2012. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13:** 31.
15. Mizrahi-Man O, Davenport ER, Gilad Y. 2013. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PLoS One* **8:** e53608.
16. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41:** D590-D596.
17. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al.* 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75:** 7537-7541.
18. Schuster SC. 2008. Next-generation sequencing transforms today's biology. *Nat. Methods* **5:** 16-18.
19. Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA, Elshahed MS. 2009. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl. Environ. Microbiol.* **75:** 5227-5236.
20. Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30:** 614-620.
21. Zhou HW, Li DF, Tam NF, Jiang XT, Zhang H, Sheng HF, *et al.* 2011. BIPES, a cost-effective high-throughput method for assessing microbial diversity. *ISME J.* **5:** 741-749.