Journal of Microbiology (2015) Vol. 53, No. 1, pp. 60–69
Copyright ⓒ 2015, The Microbiological Society of Korea

DOI 10.1007/s12275-015-4601-y

Improved pipeline for reducing erroneous identification by 16S rRNA sequences using the Illumina MiSeq platform[§]

Yoon-Seong Jeon[1,2], Sang-Cheol Park[3], Jeongmin Lim[1], Jongsik Chun[1,2,3], and Bong-Soo Kim[1,4*]

[1]ChunLab, Inc., Seoul National University, Seoul 151-742, Republic of Korea
[2]Interdisciplinary Graduate Program in Bioinformatics, Seoul National University, Seoul 151-742, Republic of Korea
[3]School of Biological Sciences and Bioinformatics Institute, BIO-MAX, Seoul National University, Seoul 151-742, Republic of Korea
[4]Department of Life Sciences, Hallym University, Chuncheon, Gangwon-do 200-702, Republic of Korea

(Received Oct 21, 2014 / Revised Nov 19, 2014 / Accepted Nov 25, 2014)

The cost of DNA sequencing has decreased due to advancements in Next Generation Sequencing. The number of sequences obtained from the Illumina platform is large, use of this platform can reduce costs more than the 454 pyrosequencer. However, the Illumina platform has other challenges, including bioinformatics analysis of large numbers of sequences and the need to reduce erroneous nucleotides generated at the 3′-ends of the sequences. These erroneous sequences can lead to errors in analysis of microbial communities. Therefore, correction of these erroneous sequences is necessary for accurate taxonomic identification. Several studies that have used the Illumina platform to perform metagenomic analyses proposed curating pipelines to increase accuracy. In this study, we evaluated the likelihood of obtaining an erroneous microbial composition using the MiSeq 250 bp paired sequence platform and improved the pipeline to reduce erroneous identifications. We compared different sequencing conditions by varying the percentage of control phiX added, the concentration of the sequencing library, and the 16S rRNA gene target region using a mock community sample composed of known sequences. Our recommended method corrected erroneous nucleotides and improved identification accuracy. Overall, 99.5% of the total reads shared 95% similarity with the corresponding template sequences and 93.6% of the total reads shared over 97% similarity. This indicated that the MiSeq platform can be used to analyze microbial communities at the genus level with high accuracy. The improved analysis method recommended in this study can be applied to amplicon studies in various environments using high-throughput reads generated on the MiSeq platform.

***Keywords*:** 16S rRNA gene, MiSeq, identification

Introduction

Next Generation Sequencing (NGS) has been widely used to study microbes in various environments and has provided a large amount of microbiome information. The taxonomic composition of microbes in nature have been analyzed using 16S rRNA genes, which are the molecular chronometers of bacteria and archaea (Woese, 1987). Roche 454 pyrosequencing has been used in numerous studies of microbes due to its relatively long read length (Oh *et al*., 2012). However, recently, the Illumina and IonTorrent platforms have been used in environmental microbiome studies, improving their performance (Gloor *et al*., 2010; Caporaso *et al*., 2012; Degnan and Ochman, 2012; Junemann *et al*., 2012; Bell *et al*., 2013). The Illumina platform produces a large number of sequences; however, its relatively short read length was a constraint for microbial community studies. This limitation was overcome by elongating sequence reads, up to 250 bp or 300 bp paired reads, with the MiSeq platform. Longer sequences lead to more accurate taxonomic assignment (Wang *et al*., 2007); thus, the Illumina platform can now be used for microbial community studies. Furthermore, the high throughput sequences obtained using the MiSeq platform can reduce costs and increase the depth of sequencing per sample in a community analysis.

Application of the MiSeq platform to microbial community studies requires the development of a bioinformatics process for handling large numbers of paired sequences. Several studies have analyzed Illumina-based sequences using different methods. In some studies, beta diversity among communities was compared by examining extensive sequence curation (Caporaso *et al*., 2011; Werner *et al*., 2012). In another study, quality scores were used to trim sequence reads (Bokulich *et al*., 2013). The Illumina platform is known to produce substitution errors in sequencing reactions, and attempts to correct these erroneous sequences have also been reported (Claesson *et al*., 2010; Gloor *et al*., 2010; Nakamura *et al*., 2011; Kozich *et al*., 2013). In addition, erroneous sequences, including chimeric sequences and PCR bias, can also be generated during amplification due to PCR drift (Wagner *et al*., 1994), non-specific primer hybridization (Kurata *et al*., 2004), low annealing temperature (Ishii and Fukui, 2001), and template reannealing (Suzuki and Giovannoni, 1996). Correction of these erroneous sequences is necessary for accurate analysis of environmental microbial communities. A curation method for mismatched sequences within overlapping regions of paired sequences has been proposed (Gloor

*For correspondence. E-mail: bkim79@hallym.ac.kr; Tel.: +82-33-248-2093; Fax: +82-33-256-3420
[§]Supplemental material for this article may be found at http://www.springerlink.com/content/120956.

et al., 2010). However, this method cannot correct erroneous sequences present in non-overlapping regions. Resequencing of a mock community sample and a sequence assembly method were used to reduce overall error rates with a 250 bp paired-end MiSeq platform (Kozich *et al.*, 2013). Unfortunately, 98-271 operational taxonomic units (OTUs) were obtained in the V4/V5 sequence results after removing chimeras and sequencing errors using the mock community, which was originally composed of 20 OTUs. This could lead to an erroneous community composition. Therefore, improved methods that reduce erroneous sequences in the final results are required to enable use of the MiSeq platform for accurate microbial community analysis.

Nine different hypervariable regions are present in the 16S rRNA gene, and certain regions have better discriminative ability for a particular bacterial lineage (Kumar *et al.*, 2011). However, no region is known to be the best for community analysis, and studies on different samples have recommended different target regions (Wang *et al.*, 2007; Huse *et al.*, 2008; Liu *et al.*, 2008). Community composition is influenced more by regional variation than by other variables, such as DNA extraction or PCR-related bias (Engelbrektson *et al.*, 2010; Kumar *et al.*, 2011). Targeted variable regions are also an important consideration for community analysis with the increased read lengths on the MiSeq platform. Different variable regions were targeted in previous studies of 100 bp and 150 bp paired sequences using the Illumina platform. The V6 and V3 regions have been used in studies with 100 bp based sequences (Gloor *et al.*, 2010; Bartram *et al.*, 2011; Degnan and Ochman, 2012). The V3/V4 and V4/V5 regions were used in a study based on 150 bp sequences (Claesson *et al.*, 2010), and the V3/V4, V4, and V4/V5 regions were used in a study with 250 bp sequences (Kozich *et al.*, 2013).

In this study, we improved the analysis method for 16S rRNA gene sequences obtained with the MiSeq platform to reduce erroneous identifications. In addition, we compared the variable regions of the 16S rRNA gene to select the best region in a community analysis using the MiSeq platform and determined the optimal phiX percentage and sequencing library concentration for optimal sequencing. The improved analysis method for sequence curation recommended in this study will be helpful in future studies for microbial community analyses using the MiSeq platform.

Materials and Methods

Preparation of mock community and DNA extraction

The mock community consisted of 47 known sequences that were obtained from previous soil clone libraries (Ahn *et al.*, 2006; Kim *et al.*, 2008). A list of the 47 clones and their sequences are shown in Supplementary data Table S1. Plasmid DNAs were extracted from each clone using the QIAGEN Plasmid mini kit (Qiagen) as previously described (Ahn *et al.*, 2006) and pooled at different concentrations. The concentrations of the extracted DNAs were quantified using the PicoGreen dsDNA Assay kit (Invitrogen). To evaluate the improved sequencing method and bioinformatics process, genomic DNA was extracted from porcine fecal samples using the FastDNA SPIN extraction kit (MP Biomedicals,). Fecal samples were obtained from the National Institute of Animal Science in the Republic of Korea, and our use was approved by the Institutional Animal Care and Use Committee of the National Institute of Animal Science (No. 2013-075).

Sequencing of the amplicon library

To determine which target sites to use for amplicon sequencing, an *in silico* test was performed by combining two variable regions (Table 1). In the *in silico* test, the average length of the detectable sequences and the proportion of detectable sequences in the database amplified by each primer were calculated using the EzTaxon-e database (Kim *et al.*, 2012). Extracted DNAs from the mock community and fecal samples were amplified with ExTaq polymerase (TaKaRa) using

Table 1. The comparison of simulated amplicon by different combinations of two variable regions. Primer sequences were obtained from previous reports (Claesson *et al.*, 2010; LaTuga *et al.*, 2011; Berry *et al.*, 2012).

Combined region	Amplicon size (bp)	Detectable sequences in database (%)[a]	Primer	Sequence (5′→3′)
V1_V2	314.5 ± 29.6	43.87	V1_forward	AGAGTTTGATCCTGGCTCAG
			V2_reverse	TACGGYAGGCAGCAG
V2_V3	388.8 ± 19.2	68.47	V2_forward	AGYGGCGNACGGGTGAGTAA
			V3_reverse	CCAGCAGCCGCGGTAAT
V3_V4	416.8 ± 11.2	75.55	V3_forward	ACTCCTACGGRAGGCAGCAG
			V4_reverse	GGATTAGATACCCTGGTAGTC
V4_V5	372.0 ± 7.4	86.05	V4_forward	AYTGGGYDTAAAGNG
			V5_reverse	AAACTRAAAYYAATTGACGG
V5_V6	249.6 ± 7.4	84.64	V5_forward	RGGATTAGATACCC
			V6_reverse	AGGTGNTGCATGGRRGTCG
V6_V7	193.7 ± 14.0	57.90	V6_forward	AACGCGAAGAACCTTAC
			V7_reverse	GGAAGGTGGGGATGACGT
V7_V8	277.3 ± 6.9	73.92	V7_forward	GYAACGAGCGCAACCC
			V8_reverse	GRACWCACCGCCCGTC
V8_V9	301.2 ± 41.6	50.39	V8_forward	GAGGAAGGTGKGGAYG
			V9_reverse	AGTCGTAACAAGGTAN

[a] Detectable sequences indicates the coverage of primer sets to sequences in the EzTaxon-e database (Kim *et al.*, 2012).

a C1000 Touch thermal cycler (Bio-Rad). The amplification conditions were previously described (Jeon *et al.*, 2013). Amplified products were purified with the QIAquick PCR purification kit (Qiagen) and quantified using the PicoGreen dsDNA Assay kit (Invitrogen). Then, 1 μg of purified amplicon was used to construct a sequencing library with the Truseq library kit (Illumina) according to the manufacturer's instructions. Library concentrations were measured by quantitative real-time PCR with Illumina adapters, targeting primers, and SYBR Green (KAPA SYBR FAST Universal qPCR kit; KAPA Biosystems) using a CFX96 Real-time PCR Detection system (Bio-Rad). Libraries were mixed with phiX control (Illumina) at various percentages and denatured using NaOH according to the manufacturer's instructions. We compared different concentrations (4, 6, and 8 pM) of amplicon libraries and different percentages (50%, 10%, and 5%) of control phiX added to the libraries to determine the optimal conditions for 16S rRNA amplicon sequencing using the MiSeq platform.

Merging sequences of paired reads

Reads containing more than one ambiguous base (N) and low-quality sequences (average Q < 25) in any paired reads were removed before paring. For quality filtering, we compared the average quality score four of different determination sizes (whole sequence, 10 bp-, 20 bp-, and 30 bp-removed) in each read. Generally, low quality scores were detected at the 3′-end of sequences from Miseq. Thus, we used the average quality score of whole reads for quality filtering. Paired sequences obtained from the MiSeq platform were merged, and overlapping regions were identified in each read. Pairwise alignment (Miller and Myers, 1988) was used to find mismatched sequences within the primer and overlapping regions. Primer sequences in both paired reads were removed for further analysis. The pairwise similarity of the overlapping sequences was calculated, and the correlations between the mismatched nucleotides and corresponding quality scores were analyzed. The Q score decreased toward the 3′-end of reads with increasing sequence length, and higher error rates were reported for reverse reads than for forward reads (Claesson *et al.*, 2010; Nakamura *et al.,* 2011; Kozich *et al.*, 2013). Two different methods could be used to solve this problem, and these two methods were compared in this study. The first method is trimming the low quality sequences at the 3′-end of reads, and correcting mismatched sequences within overlapping regions according to Q scores (Zhou *et al.*, 2011). When a mismatched sequence is found in an overlapping region, the sequence closer to the 5′-end is chosen between the pairs, or the higher quality nucleotide is selected if the position from the 5′-end is the same. The second method is the removal of a specific sequence length (e.g., 10, 20, or 30 bp) from the 3′-end to eliminate sequences. Error rates were calculated using read sequences of the mock community sample by the following steps: 1) a BLAST search of the merged reads against a database of the 47 known sequences, 2) pairwise alignment of the merged reads to the most similar sequences obtained in the BLAST search (>1e^{-5}), 3) The lengths of the merged reads varied. Thus, the reads were divided into 10 deciles for the comparison, 4) calculation of the error rate per read for each decile. Error rates were calculated using the equation:

$$\text{Error rates in each decile per read} = \text{En}/\text{Tmr}$$

where En is the sum of the nucleotides that are mismatched compared to the reference sequence within each decile, and Tmr is the total number of merged reads.

The error rates patterns were compared for different library concentrations in the same sequencing run and the same concentration of library in different runs. The error rates for the different concentrations of sequencing library and phiX were calculated after merging to correct overlapping sequences.

Correcting erroneous nucleotides for taxonomic analysis

The effects of erroneous sequences on the identification of microbial composition were analyzed by a scatter plot using R software (ver. 3.0.3). Pairwise similarities between merged reads and the corresponding sequences of the mock community were used to evaluate the effects of erroneous sequences. Decreased quality scores for 3′-end sequences were related to heterogeneous sequences in the overlapping regions of paired sequences. The correlation between the number of heterogeneous nucleotides in the overlapping regions and the similarity of merged reads to reference sequences are shown in a scatter plot. Mismatched sequences after merging were compared to the mock community sequences by using the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) and displayed using the Integrative Genome Viewer (IGV) (Robinson *et al.*, 2011). To correct erroneous nucleotides in the reads, similar sequences (97% sequence similarity cutoff) were clustered with the USEARCH program (Edgar, 2010), and consensus sequences were produced in each cluster. The consensus sequence was generated by selecting the major nucleotide in each mismatched column within a cluster after pairwise alignment of all sequences against the longest one in the cluster. This step was repeated until the number of clusters did not change. Clusters including reads with fewer than 0.005% of all sequences were removed for further analyses as described in a previous report (Bokulich *et al.*, 2013). Corrected sequences were identified using the EzTaxon-e database (Kim *et al.*, 2012), and chimeras were detected with the UCHIME program (Edgar *et al.*, 2011). A chimera check was conducted twice, once after the first round of clustering and again after the final round of clustering. After the first round of clustering, consensus sequences still contained many chimeric raw reads. Therefore, removing chimeric reads after the first round of clustering could reduce the number of reads and computation for subsequent analyses. Detecting chimeras and removing them after the final clustering step is necessary to obtain the final consensus reads that assign taxonomic positions. Determinations of OTUs were performed using the USEARCH program with a 97% similarity cut-off value. Sequence reads generated in this study are available at the EMBL SRA database (study accession number PRJEB5083) (http://www.ebi.ac.uk/ena/data/view/PRJEB5083).

Statistical analyses

Statistical analyses were performed to evaluate the significance of community differences and the correlation of quality scores with error rates. Bacterial community differences were evaluated by Fisher's exact test (Fisher, 1922), and multiple comparisons were conducted using Bonferroni correction (Dunnett, 1955). The correlation between error rates and quality scores were analyzed using R software. The error frequency in each region was calculated as the percentage of mismatched sequences to corresponding reference sequences.

Results and Discussion

In silico analysis to determine primer sets

To determine which 16S rRNA gene target regions to use for amplicon sequencing using MiSeq 250 bp paired reads, combinations of two variable regions were analyzed for their average amplicon length, and the proportion of detectable sequences in the database amplified by each primer set was calculated (Table 1). The amplicon size varied based on the combination of regions. The shortest reads were obtained using the V6/V7 combination (193.7 ± 14.0 bp), whereas the longest reads were generated by using the V3/V4 regions (416.8 ± 11.2 bp). However, the V3/V4 region was reported to generate significant amplification bias in a previous study (Claesson *et al.*, 2010); therefore, the V3/V4 region was excluded from the candidate target regions. In the 250 bp-based MiSeq platform, sequences with lengths shorter than 250 bp led to reduced advantages of paired read sequences and the accuracy of identification (Janda and Abbott, 2007). Amplified regions shorter than 250 bp (V5/V6 and V6/V7) were also excluded. The highest sequence coverage was observed using the V4/V5 combination, for which 86.05% of the sequences were detectable in the database. Conversely, there was relatively low detection of the V1/V2 regions (43.87%). This low detection was likely due to a lack of sequences, specifically primer sequences, in the databases for the V1 region, as most sequences in public databases do not contain sequence information for the V1 forward primer region. The V8/V9 region has also low coverage in databases; thus, the primer set for this region was excluded. Two target regions, V2/V3 (388.8 ± 19.2 bp) and V4/V5 (372.0 ± 7.4 bp), were selected for evaluation because they had longer amplified sizes than V1/V2 (314.5 ± 29.6 bp) and relatively high sequence coverage in the databases.

Comparison of the percentage of phiX added and amplicon library concentration

In the previous MiSeq platform, amplicon libraries were mixed with 50% phiX to increase genetic diversity. In the current system, the Real Time Analysis (RTA) in the MiSeq Control software (MCS) has been improved. This improvement reduced the amount of phiX added to the sequencing libraries and increased data quality in samples with low genetic diversity, such as 16S rRNA gene amplicons. The performance of three different percentages of phiX in sequencing conditions was compared (Fig. 1). The results showed that the number of target sequence reads was increased by reducing the amount of phiX in the libraries (1.9×10^6 at 50% phiX and 10.03×10^6 at 5% phiX), and the number of reads ≥ Q30 was also increased by reducing the percentage of phiX (Fig. 1B). This indicated that a lower percentage of phiX in the libraries resulted in more reads, and thus could be applied to further amplicon sequencing. This result was consistent with a previous report, which compared different percentages of phiX (Kozich *et al.*, 2013). They obtained 9.0×10^6 paired reads, 80.1% of which were ≥Q30 using 8.0% phiX and 10.5×10^6 paired reads, 74.6% of which were ≥Q30 using 6.2% phiX.

Different amplicon library concentrations were compared

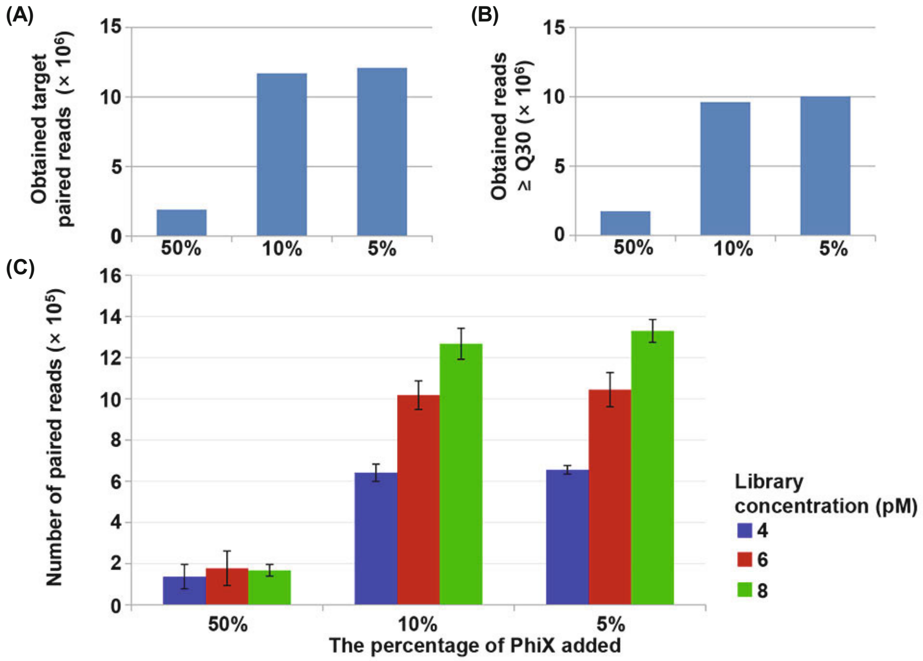

Fig. 1. The comparison of different concentrations of library and percentages of phiX added. (A) the numbers of target reads, (B) the proportions of greater than Q30 among the different percentages of phiX added, (C) the numbers of obtained reads from different concentrations of sequencing library were compared.

Table 2. **The numbers of paired reads by different trimming lengths at each 3′-end of paired reads in different target regions**

Trimming length at each read of pairs (bp)	Target region									
	V2/V3					V4/V5				
	0	10	20	30	40	0	10	20	30	40
Number of reads with identical sequences in overlap region	338,288	492,620	660,071	712,577	670,371	213,206	270,621	335,317	410,010	539,858
Number of correcting reads by overlap sequences	831,636	677,388	509,335	381,920	346,122	585,783	528,818	464,747	390,546	258,653
Number of total merged reads	1,169,924	1,170,008	1,169,406	1,094,497	1,016,493	798,989	799,439	800,064	800,556	798,511
Average length of merged reads (bp)	380.0	380.0	379.9	376.5	370.9	372.0	372.0	372.0	372.0	371.9

in each run and the average number of obtained reads in each sample is shown in Fig. 1C. In the 50% phiX sequencing runs, the numbers of reads obtained with different library concentrations were similar, whereas in the 10% and 5% phiX runs, the number of obtained reads increased with increased library concentration. More than 6 × 10^5 reads per sample were generated using 4 pM library, and more than 12 × 10^5 reads were obtained using 8 pM library. The increase in the number of reads from 6 pM to 8 pM (2.7 × 10^5) was less than the increase from 4 pM to 6 pM (3.8 × 10^5). The number of reads obtained using a 10 pM library in a previous report (Kozich *et al.*, 2013) was similar to the read numbers obtained using an 8 pM library in this study. In a previous study, runs with 10 pM library generated 2 × 10^5 reads more than runs with 5 pM library. Determination of the appropriate library concentration is necessary to obtain high quality sequence reads. Therefore, for raw sequence reads, 8 pM library containing 10% phiX was used to improve the sequencing analysis method, and these improved methods were evaluated in the presence study.

Merging paired sequences

Comparisons of the average quality score in each read for the four different determination sizes (whole sequence of read, and short window of 10 bp, 20 bp, and 30 bp) are shown in Supplementary data Table S2. Most sequences (over 50% of the total reads) were removed using an average score in a short window of 10–30 bp, because more low quality scores were detected in short windows towards the 3′-end of a read. Therefore, we used the average quality score of whole sequences in each read for quality filtering. After quality filtering, merging of paired reads can elongate the length of amplicon sequences. The average lengths of the expected amplicons selected using an *in silico* test were 389 bp (V2/V3) and 372 bp (V4/V5). Primer sequences (approximately 20 bp) at the 5′-end of each read were trimmed for taxonomic analysis. The length of the sequences overlapping between the forward and reverse reads was 80 bp,

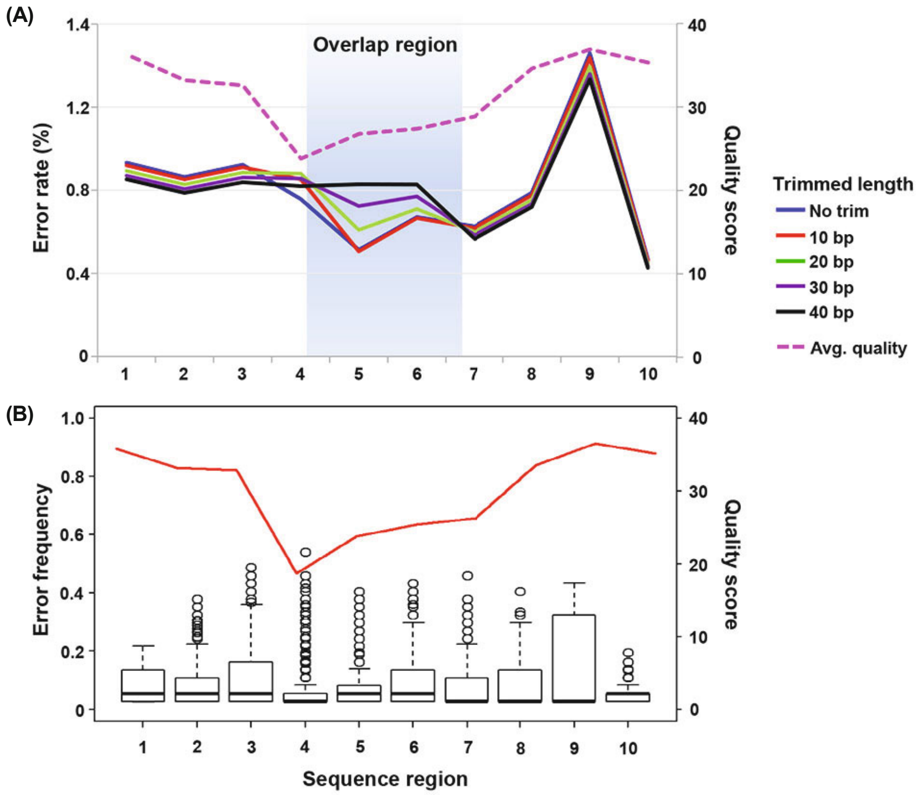

Fig. 2. **The distributions of the calculated error rates for the 10 deciles.** (A) The effects of curating sequences within the overlapping region and different trimming sequences length were compared. The correlations of quality score with error rates at each decile were analyzed in the 8 pM V4/V5 library with 10% phiX. The highest error rate was detected at 9th decile; however, the quality score at this region was over 35. (B) The correlation of error frequency with quality score at each decile was analyzed using boxplot by R software. The red line indicates the average quality scores.

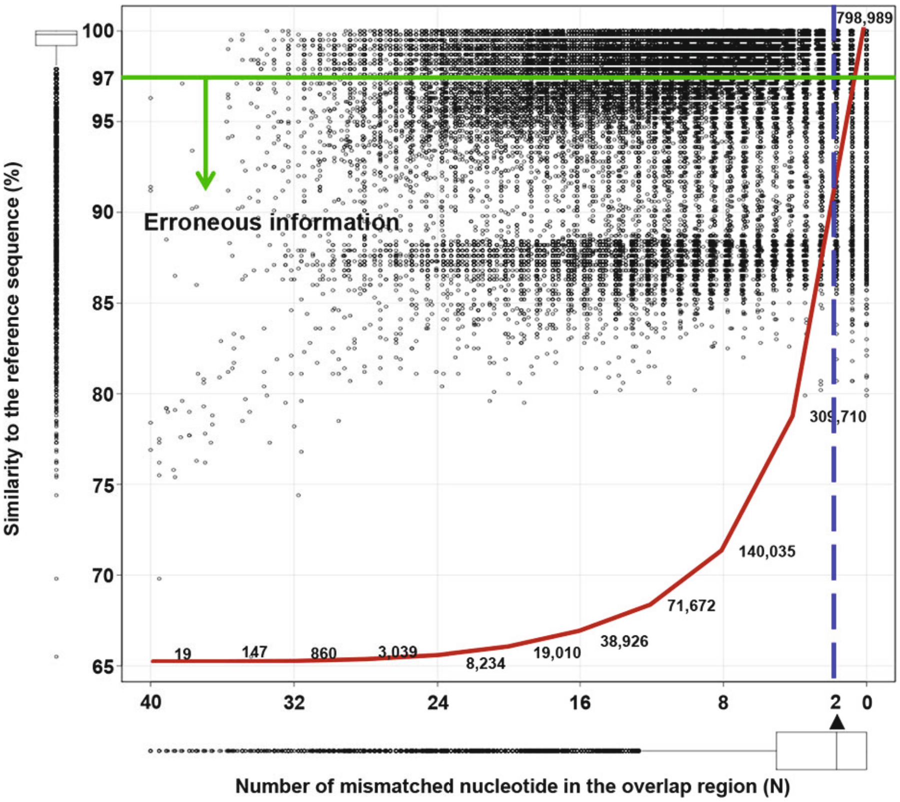

Fig. 3. The correlation between the number of mismatched nucleotides in overlapping regions and the similarity of merged reads to the corresponding sequences of the mock community were analyzed by scatter plot and median plot. Erroneous information of community composition can be generated by below 97% similarity reads to the reference sequence. The red plot indicates the accumulated numbers of merged reads, and the blue plot indicates the median value of mismatched nucleotides in the overlapping region.

assuming the target length was 380 bp (Supplementary data Fig. S1).

Two different methods for merging paired sequences were compared using known mock community sequences. Trimming over 40 bp sequences at the 3′-end of both reads of a pair cannot produce overlapping sequences due to the length of the overlapping region (80 bp). The number of reads obtained by trimming different sequence lengths from each 3′-end of paired reads was compared (Table 2). The number of paired reads containing identical sequences within overlapping regions increased as more sequence was trimmed from the 3′-end of both reads of a pair. This also showed that more erroneous nucleotides were generated toward the 3′-ends. Reads with identical sequences in the overlap region with high quality score can be used for analyses without correction. The number of reads with identical overlap sequences increased greatly when 30 bp and 40 bp were trimmed for the V2/V3 and V4/V5 combinations, respectively. This means that trimming 30 bp from the 3′-end of V2/V3 reads and trimming 40 bp from the 3′-end of V4/V5 reads can remove most of the erroneous sequences within the overlap region of paired reads. The number of identical overlapping sequences targeted to the V2/V3 region was reduced when 40 bp were trimmed from the 3′-end of each sequence because the average length of V2/V3 was 389 bp. Thus, sequences trimmed by 40 bp produced no overlap sequences (Supplementary data Fig. S1). The number of sequences that were correctable by comparing overlapping sequences between paired reads increased without trimming, indicating that trimming sequences at the 3′-end eliminated the possibility of correcting sequences during merging. Although the total number of merged reads corrected without trimming the sequences was similar to the number of reads trimmed by a specific length, correction of erroneous sequences could improve the accuracy of sequences. For example, the number of total merged reads after trimming 40 bp from the 3′-end of V4/V5 reads (798,511) was similar to the number of sequences corrected without trimming (798,989); however, correcting sequences within the overlap region can increase the number of reads in the analysis.

The accuracy of the sequenced reads after correcting heterogeneous nucleotides between paired sequences was evaluated by comparing the corrected sequences with the corresponding reference sequences in the mock community (Fig. 2). In overlapping regions, correction of heterogeneous nucleotides between paired reads (no trim) decreased the error rates, indicating that correcting erroneous sequences within the overlap region increased the accuracy of the sequences. We compared the error rates of each region with the average Q score of the corresponding region (Fig. 2A). The highest error rate was observed after 30 bp from the reverse primer sequences (9th decile). However, the average Q score of this region was higher than 35. No correlation between erroneous reads and quality score was observed in this study, and the error frequency in each decile was also not correlated with quality score (Fig. 2B). This result is contrary to that of a previous report in which errors generated by MiSeq were correlated with a low Q score (Kozich *et al.*, 2013). This difference may be due to difference in the sequencing conditions, such as primers and library composition for sequencing runs. However, a previous study also showed an overestimated OTU number after removing chimeras and cor-

Table 3. The comparison of cluster numbers after each round of clustering

Mock community sample library (n=47)	The number of Clusters				
	1st Round	After chimera removed	2nd Round	3rd Round	4th Round
8 pM V4/V5 region with 10% phiX	2,803	2,054	94	75	75
6 pM V4/V5 region with 10% phiX	1,710	1,249	86	73	72
8 pM V2/V3 region with 10% phiX	3,502	1,387	194	154	148
8 pM V4/V5 region with 5% phiX	9,282	3,801	416	337	333

recting sequencing errors in all tested regions (Kozich *et al.*, 2013). When 21 isolates were mixed for the mock community, 20 OTUs were expected in the final analysis result. However, more than 98 OTUs were obtained for the V4/V5 regions. This indicates that erroneous sequences with high quality scores could be obtained without an error correction step. This overestimated information can lead to erroneous microbial compositions in the analysis of amplicon sequencing.

Patterns and effects of erroneous sequences

We compared the error rates obtained using different concentrations of sequencing library and percentage of phiX added to analyze the correlation between the generation of erroneous sequences and the sequenced regions (Supplementary data Fig. S2). The error rates of 6 pM library were similar to those of 8 pM library in the same run (10% phiX added run). However, the error rates patterns for 6 pM library were different to those for the 8 pM library (Fig. 2). The error rate for the 3th decile was the highest among the 10 deciles analyzed for 6 pM library, and the second highest error rate was found at the 9th decile. The pattern of error rates obtained during different sequencing runs (5% phiX) was different from the patterns for the two sequencing results described above. These results showed that the erroneous sequences were not generated at a specific region, but were randomly generated. The community composition obtaining using V4/V5 region sequences in this study was more similar to the original mock community than the V2/V3 region (*P*<0.05). Therefore, we analyzed using the V4/V5 region in subsequent analyses (Supplementary data Fig. S3).

To investigate the effect of erroneous nucleotides on microbial community analysis, the similarity of merged reads to the corresponding reference sequences and the similarity of overlapping sequences between paired reads were compared in a scatter plot (Fig. 3). The correlation between the similarity of overlapping sequences and merged reads to reference sequences was analyzed because the number of mismatches in the overlapping regions was also used to determine qualified reads in a previous report (Gloor *et al.*, 2010). The median plots were used to determine the median value of each axis. There was no correlation between the number of mismatches in overlapping regions and erroneous sequences in merged reads. The median value of mismatched nucleotides in overlapping regions was 2. This means most of the sequencing reads (about 490,000 out of 798,989 reads) have less than 2 nucleotide mismatches in their overlapping region. However, the similarity values of these reads to corresponding sequences in the mock community were less than 97%, and similarity values higher than 97% were found in reads that included more than 2 mismatched nucleotides within the overlapping region. This is because erroneous sequences were generated randomly throughout all regions of a read (Fig. 2). The merged reads that had less than 97% similarity to the corresponding reference sequences could be potentially erroneous information about the microbial composition. Even if reads containing more than 2 mismatched sequences in the overlapping regions were removed from the analysis, an erroneous result of community structure could still be generated. There are no known reference sequences for community composition in environmental samples; thus, the correction of this erroneous information is difficult and necessary for accurate analysis. Correction of erroneous nucleotide in whole regions of reads is necessary to improve these problems.

Correcting erroneous nucleotides in all regions of reads

Clustering can reduce erroneous sequences, and the effect

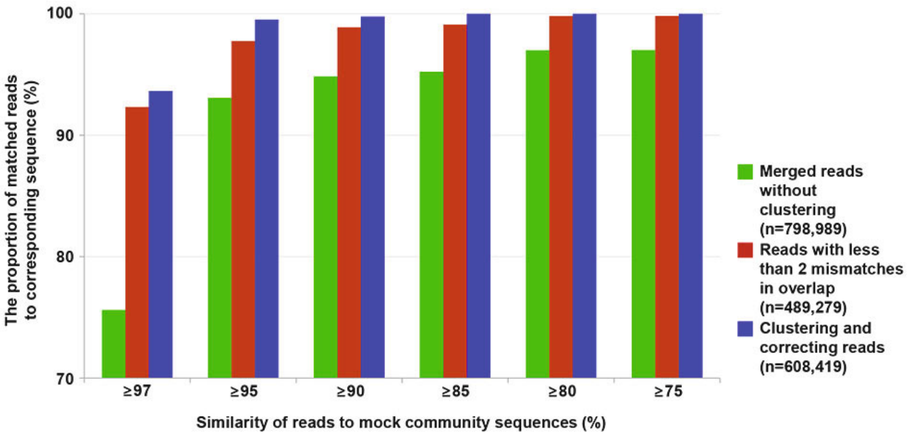

Fig. 4. The proportions of matched reads to the corresponding reference sequence at each similarity criteria. The mock community sample was used to determine the proportion of matched reads. The three different methods, merged reads without clustering, removed reads with more than two mismatched nucleotides in the overlapping regions, and corrected sequences using the clustering process, were compared. The total numbers of analyzed read in each method are presented below each method.

of clustering on the analysis of NGS reads has been reported (Schloss *et al.*, 2011; Kozich *et al.*, 2013). Randomly generated erroneous sequences can be corrected by clustering (97% similarity cutoff) using the consensus sequences of clusters. Consensus sequences were generated by selecting the most common sequences in heterogeneous columns of multiple alignments in the cluster (Supplementary data Fig. S4). The reduced number of clusters was compared to those for different sequencing conditions of mock community sample (Table 3). In general, the number of clusters does not change after the third clustering step in all of the samples (different target region, library concentration, and sequencing run). Over 1,000 clusters were obtained after the first clustering step in all of the samples, while the number of clusters was reduced after every iterated clustering step. The numbers of clusters from the V4/V5 region amplicon (75 clusters with 8 pM library and 72 clusters with 6 pM library) were lower than those for the V2/V3 region amplicon (148 clusters) and the V4/V5 region amplicon from a different sequencing run (mixed with 5% phiX; 333 clusters). The number of merged reads obtained from the V4/V5 sample in the 5% phiX mixed run (1,014,630 reads) was greater than that of the same library from the 10% phiX mixed run (798,989 reads). This suggests that increasing the number of reads lead to more erroneous information; therefore, it is necessary to determine the proper concentration of phiX.

The effects of clustering and correcting sequences were evaluated by comparing the ratios of matched reads to the corresponding reference sequences of the mock community. Two other methods of merging reads without clustering and discarding reads with more than two mismatches in the overlapping regions were compared (Fig. 4). In the merged reads without clustering, the proportion of total reads with greater than 97% similarity to the corresponding reference sequence was 75%, and the proportion of total reads with greater than 95% similarity to the reference sequences was 93%. These similarities cutoff can be used to determine the phylotypes of sequencing reads (Yarza *et al.*, 2014). With this method, erroneous analyses could be performed on reads with less than 97% similarity to mock community sequences (Fig. 3). More than 7% of the community composition could be erroneous when using results generated by discarding reads with more than two mismatched nucleotides within the overlapping regions. In contrast, when the clustering correction method was used, 99.5% of the total reads showed greater than 95% similarity to the reference sequences, and 93.7% of the reads showed greater than 97% similarity to the reference sequences. Although 6.3%

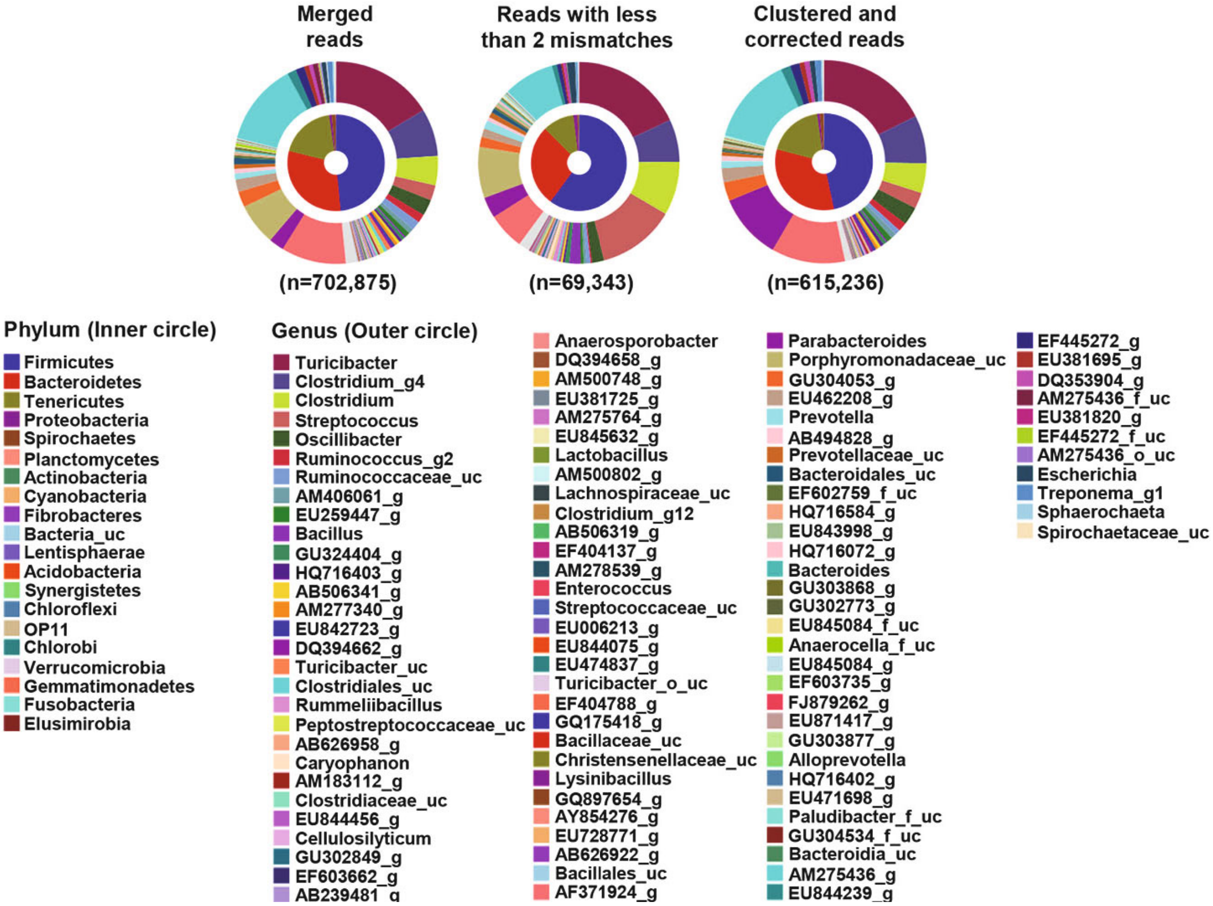

Fig. 5. **The bacterial compositions of the same fecal sample analyzed by three different methods were compared using double pie charts of phylum and genus.** The total numbers of analyzed read are presented below each pie, and each color is defined below the figure. The nomenclature of phylotype is based on the EzTaxon-e database (Kim *et al.*, 2012).

of the reads differed from the corresponding reference sequences, highly accurate information was obtained, at over 95% similarity. This similarity was consistent with the taxonomic assignment of the reads at the genus level (Tindall *et al.*, 2010). Therefore, clustering and correcting steps can reduce erroneous information about community composition up to the genus level using 250 bp paired reads from the MiSeq platform.

Comparison of community compositions obtained by using different primer sets and different analysis methods

Amplified products produced by using two primer sets selected from the *in silico* test were assigned based on their taxonomic composition, and compared to the original mock community at both the phylum and genus levels (Supplementary data Fig. S3). In a UPGMA tree based on Fast UniFrac distances among communities, the amplified product of V4/V5 was more similar to the mock community composition than the amplicons of V2/V3. The proportions of each genus were different from that of the original template because the efficiency of amplification for each genus was different. However, the V4/V5 primer set yielded results that were similar to the original community, and the highest classification accuracies were also reported for this region across RDP-Classifier and MEGAN data based on BLAST searches (Claesson *et al.*, 2010). The amplified V2/V3 region even differed at the phylum level. Therefore, the V4/V5 target region is considered to be suitable for amplicon sequencing based on 250 bp paired sequences using the MiSeq platform.

The bacterial community compositions in pig fecal samples obtained using different analysis methods were compared (Fig. 5). The phylum composition determined by the total merged reads analysis was similar to the composition determined by the clustered and corrected reads analysis. However, the phylum composition determined by including reads with fewer than two mismatches in the overlapping region was different. This difference could be caused by discarding most of the reads within the overlapping region when using the two mismatch rule. The total number of analyzed reads after discarding mismatched reads (69,343 reads; 10% of total reads) was smaller than that analyzed in the total merged reads (702,875 reads) and corrected reads (615,236 reads) methods. Reduced number of analyzed reads by two mismatches rule can provide erroneous information about the community composition. Analysis of the clustered and corrected reads allowed for analysis of most merged reads (87.5% of merged reads), and this method did correct erroneous information (Fig. 3). The number of observed phyla in the analysis of the total merged reads was 19, whereas 11 phyla were detected in the clustered and corrected reads analysis. The number of genera in the analysis of merged reads was 537, while 173 genera were detected in the clustered and corrected reads analysis. This suggests that the overestimated diversity could be corrected by using this improved analysis method. We applied our method to published data (Nelson *et al.*, 2014), and compared the community results (Supplementary data Fig. S5). There were unexpected taxa in a previous study. In contrast, we obtained reference taxa using our method. This result indicates that the correcting and clustering method described in this study can improve the accuracy of identification.

In this study, we evaluated various sequencing conditions for the MiSeq platform and suggested a method for correcting erroneous reads for accurate identification. Generation of a large number of sequences (more than 10,000,000 paired reads) per run by using the MiSeq platform can provide high sequence depth as previously reported (Kozich *et al.*, 2013). However, this large number of sequences can also lead to erroneous information about community composition (Figs. 3 and 4). Therefore, correction of erroneous nucleotides is necessary to obtain accurate identification. Amplification of the V4/V5 variable region can yield the most accurate microbial composition information in both the *in silico* and community analyses. On the MiSeq platform, the best results were obtained with 8 pM sequencing library and 10% phiX. The improved analysis pipeline, using clustering and correcting steps, reduced the overestimated community composition of the remaining reads. This method also reduced the erroneous analysis of the mock community data (Fig. 4 and Table 3). We recommend that other users apply this clustering and correcting step for their microbial community studies when using the MiSeq platform. The methods for correcting erroneous sequences and determining sequencing conditions described in this study will be useful to develop longer sequencing reads on the Illumina platform.

Acknowledgements

This work was supported by Industrial Strategic Technology Development Program (No. 10040176) funded by the Korean Ministry of Knowledge Economy (MKE) and Hallym University Research Fund, 2014 (HRF-201409-002). We thank Dr. Seung-Hak Yang at National Institute of Animal Science, Republic of Korea for providing pig fecal sample and advice.

References

Ahn, J.H., Kim, M.S., Kim, M.C., Lim, J.S., Lee, G.T., Yun, J.K., Kim, T., Kim, T., and Ka, J.O. 2006. Analysis of bacterial diversity and community structure in forest soils contaminated with fuel hydrocarbon. *J. Microbiol. Biotechnol.* **16**, 704–715.

Bartram, A.K., Lynch, M.D.J., Stearns, J.C., Moreno-Hagelsieb, G., and Neufeld, J.D. 2011. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Appl. Environ. Microbiol.* **77**, 5569–5569.

Bell, T.H., Yergeau, E., Maynard, C., Juck, D., Whyte, L.G., and Greer, C.W. 2013. Predictable bacterial composition and hydrocarbon degradation in arctic soils following diesel and nutrient disturbance. *ISME J.* **7**, 1200–1210.

Berry, D., Schwab, C., Milinovich, G., Reichert, J., Ben Mahfoudh, K., Decker, T., Engel, M., Hai, B., Hainzl, E., Heider, S., *et al.* 2012. Phylotype-level 16S rRNA analysis reveals new bacterial indicators of health state in acute murine colitis. *ISME J.* **6**, 2091–2106.

Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D.A., and Caporaso, J.G. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **10**, 57–59.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., *et al.* 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., and Knight, R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. USA* **108**, 4516–4522.

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., and O'Toole, P.W. 2010. Comparison of two Next-Generation Sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res.* **38**, e200.

Degnan, P.H. and Ochman, H. 2012. Illumina-based analysis of microbial community diversity. *ISME J.* **6**, 183–194.

Dunnett, C.W. 1955. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Ass.* **50**, 1096–1121.

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200.

Engelbrektson, A., Kunin, V., Wrighton, K.C., Zvenigorodsky, N., Chen, F., Ochman, H., and Hugenholtz, P. 2010. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J.* **4**, 642–647.

Fisher, R.A. 1922. On the interpretation of χ2 from contingency tables, and the calculation of P. *J. Royal Statist. Soc.* **85**, 87–94.

Gloor, G.B., Hummelen, R., Macklaim, J.M., Dickson, R.J., Fernandes, A.D., MacPhee, R., and Reid, G. 2010. Microbiome profiling by Illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS One* **5**, e15406.

Huse, S.M., Dethlefsen, L., Huber, J.A., Welch, D.M., Relman, D.A., and Sogin, M.L. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* **4**, e1000255.

Ishii, K. and Fukui, M. 2001. Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. *Appl. Environ. Microbiol.* **67**, 3753–3755.

Janda, J.M. and Abbott, S.L. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, Perils, and Pitfalls. *J. Clin. Microbiol.* **45**, 2761–2764.

Jeon, Y.S., Chun, J., and Kim, B.S. 2013. Identification of household bacterial community and analysis of species shared with human microbiome. *Curr. Microbiol.* **67**, 557–563.

Junemann, S., Prior, K., Szczepanowski, R., Harks, I., Ehmke, B., Goesmann, A., Stoye, J., and Harmsen, D. 2012. Bacterial community shift in treated periodontitis patients revealed by ion torrent 16S rRNA gene amplicon sequencing. *PLoS One* **7**, e41606.

Kim, M.C., Ahn, J.H., Shin, H.C., Kim, T., Ryu, T.H., Kim, D.H., Song, H.G., Lee, G.H., and Kai, J.O. 2008. Molecular analysis of bacterial community structures in paddy soils for environmental risk assessment with two varieties of genetically modified rice, Iksan 483 and Milyang 204. *J. Microbiol. Biotechnol.* **18**, 207–218.

Kim, O.S., Cho, Y.J., Lee, K., Yoon, S.H., Kim, M., Na, H., Park, S.C., Jeon, Y.S., Lee, J.H., Yi, H., *et al.* 2012. Introducing EzTaxon-e: A prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int. J. Syst. Evol. Microbiol.* **62**, 716–721.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. 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.

Kumar, P.S., Brooker, M.R., Dowd, S.E., and Camerlengo, T. 2011. Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PLoS One* **6**, e20956.

Kurata, S., Kanagawa, T., Magariyama, Y., Takatsu, K., Yamada, K., Yokomaku, T., and Kamagata, Y. 2004. Reevaluation and reduction of a PCR bias caused by reannealing of templates. *Appl. Environ. Microbiol.* **70**, 7545–7549.

LaTuga, M.S., Ellis, J.C., Cotton, C.M., Goldberg, R.N., Wynn, J.L., Jackson, R.B., and Seed, P.C. 2011. Beyond bacteria: A study of the enteric microbial consortium in extremely low birth weight infants. *PLoS One* **6**, e27858.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**, 1754–1760.

Liu, Z.Z., DeSantis, T.Z., Andersen, G.L., and Knight, R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* **36**, e120.

Miller, W. and Myers, E.W. 1988. Sequence comparison with concave weighting functions. *Bull. Math. Biol.* **50**, 97–120.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., *et al.* 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**, e90.

Nelson, M.C., Morrison, H.G., Benjamino, J., Grim, S.L., and Graf, J. 2014. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS One* **9**, e94249.

Oh, J., Kim, B.K., Cho, W.S., Hong, S.G., and Kim, K.M. 2012. Pyrotrimmer: A software with GUI for pre-processing 454 amplicon sequences. *J. Microbiol.* **50**, 766–769.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. 2011. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26.

Schloss, P.D., Gevers, D., and Westcott, S.L. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**, e27310.

Suzuki, M.T. and Giovannoni, S.J. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **62**, 625–630.

Tindall, B.J., Rossello-Mora, R., Busse, H.J., Ludwig, W., and Kampfer, P. 2010. Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol.* **60**, 249–266.

Wagner, A., Blackstone, N., Cartwright, P., Dick, M., Misof, B., Snow, P., Wagner, G.P., Bartels, J., Murtha, M., and Pendleton, J. 1994. Surveys of gene families using polymerase chain-reaction - PCR selection and PCR drift. *Syst. Biol.* **43**, 250–261.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. 2007. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267.

Werner, J.J., Zhou, D., Caporaso, J.G., Knight, R., and Angenent, L.T. 2012. Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *ISME J.* **6**, 1273–1276.

Woese, C.R. 1987. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271.

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzeby, J., Amann, R., and Rossello-Mora, R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645.

Zhou, H.W., Li, D.F., Tam, N.F.Y., Jiang, X.T., Zhang, H., Sheng, H.F., Qin, J., Liu, X., and Zou, F. 2011. Bipes, a cost-effective high-throughput method for assessing microbial diversity. *ISME J.* **5**, 741–749.