The ISME Journal (2011) 5, 741–749

www.nature.com/ismej

ORIGINAL ARTICLE

BIPES, a cost-effective high-throughput method for assessing microbial diversity

Hong-Wei Zhou[1], Dong-Fang Li[2], Nora Fung-Yee Tam[3], Xiao-Tao Jiang[2], Hai Zhang[4], Hua-Fang Sheng[1], Jin Qin[5], Xiao Liu[2] and Fei Zou[1]

[1]*Department of Environmental Health, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou, Guangdong, China;* [2]*Beijing Genomics Institute, Shenzhen, Guangdong, China;* [3]*Department of Biology and Chemistry, City University of Hong Kong, Hong Kong SAR, China;* [4]*Network Center, Southern Medical University, Guangzhou, Guangdong, China and* [5]*Department of Biochemistry, Hong Kong University, Hong Kong SAR, China*

Pyrosequencing of 16S rRNA (16S) variable tags has become the most popular method for assessing microbial diversity, but the method remains costly for the evaluation of large numbers of environmental samples with high sequencing depths. We developed a barcoded Illumina paired-end (PE) sequencing (BIPES) method that sequences each 16S V6 tag from both ends on the Illumina HiSeq 2000, and the PE reads are then overlapped to obtain the V6 tag. The average accuracy of Illumina single-end (SE) reads was only 97.9%, which decreased from ~99.9% at the start of the read to less than 85% at the end of the read; nevertheless, overlapping of the PE reads significantly increased the sequencing accuracy to 99.65% by verifying the 3′ end of each SE in which the sequencing quality was degraded. After the removal of tags with two or more mismatches within the medial 40–70 bases of the reads and of tags with any primer errors, the overall base sequencing accuracy of the BIPES reads was further increased to 99.93%. The BIPES reads reflected the amounts of the various tags in the initial template, but long tags and high GC tags were underestimated. The BIPES method yields 20–50 times more 16S V6 tags than does pyrosequencing in a single-flow cell run, and each of the BIPES reads costs less than 1/40 of a pyrosequencing read. As a laborsaving and cost-effective method, BIPES can be routinely used to analyze the microbial ecology of both environmental and human microbiomes.

The ISME Journal (2011) **5,** 741–749; doi:10.1038/ismej.2010.160; published online 21 October 2010
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: 16S; BIPES; Illumina; microbial diversity; Solexa; V6

Introduction

Microbial communities are present in all described biomass. To fully understand the ecology of any system it is imperative that a complete and accurate description is made of the diversity and relative abundance of the organisms present (Fuhrman, 2009). Most of the current methods assessing microbial diversity within and between communities cannot determine high-throughput data with accurate taxonomy identification. For instance, cultivation methods are biased toward the isolation of cultivable strains; the denaturing gradient gel electrophoresis method is prone to identifying abundant microbes; and terminal-restriction fragment polymorphism is a high-throughput method that detects many operational taxonomic units, but does not precisely identify the taxonomies of the operational taxonomic units (Schutte *et al.*, 2008). In comparison, the sequencing of 16S rRNA (16S) clone libraries that contain thousands of clones (Tringe and Hugenholtz, 2008) and the use of microarrays (DeSantis *et al.*, 2007) produce high-throughput data and provide sufficient taxonomic information; however, the high cost of these methods prevents their routine use.

The development of 454 (Roche, Branford, CT, USA) pyrosequencing to detect short 16S tags is a great leap forward for microbial ecology studies (Sogin *et al.*, 2006; Tringe and Hugenholtz, 2008). This method determines ~400 000–1 000 000 reads in a single run, and the tag sequence contains adequate information for a taxonomic assignment (Liu *et al.*, 2008). The barcode primer technique further reduces the cost of each sample (Hamady *et al.*, 2008). The use of this method has resulted in many novel observations, both in the human microbiome and in environmental microbial communities (Sogin *et al.*, 2006; Huber *et al.*, 2007;

Correspondence: H-W Zhou, Department of Environmental Health, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou, Guangdong 510515, China.
E-mail: biodegradation@gmail.com
Received 13 April 2010; revised 13 September 2010; accepted 13 September 2010; published online 21 October 2010

Roesch *et al*., 2007; Quince *et al*., 2008; Costello *et al*., 2009; Turnbaugh *et al*., 2009). Nevertheless, the cost of a pyrosequencing run is not trivial, and a significant cost can be incurred when this method is used for large numbers of environmental samples with diverse microbial communities.

The Illumina (Solexa, San Diego, CA, USA) instrument, another type of next-generation sequencing system, can generate up to 50 times more reads than can pyrosequencing in a single flow cell run; however, the 35-bp read length of the Solexa Genome Analyzer is too short for the determination of 16S variable tag sequences. Recently, the read length of the Illumina systems was upgraded to 75 bp (Genome Analyzer II), and more recently, it was upgraded to over 100 bp (HiSeq 2000) (Kircher and Kelso, 2010). Furthermore, Illumina utilizes a unique paired-end (PE) sequencing strategy; each DNA molecule can be sequenced from both ends, and the PE reads may be overlapped to sequence tags that are longer than 100 bp. In this study, we developed a barcoded Illumina PE sequencing method (BIPES) that reads 100 bp of 16S V6 PCR amplicons from both ends on a HiSeq 2000 system. A bioinformatics pipeline was developed to analyze the BIPES data. The sequencing accuracy was evaluated using mock libraries with known 16S V6 fragments as templates. The results suggest that BIPES is a highly accurate technique to identify V6 tags, and the method is very cost effective and can be used for a broad range of microbial diversity studies.

Materials and methods

The BIPES workflow

In the BIPES procedure (Supplementary Figure S1), the 16S V6 tag of each sample is amplified with a barcoded primer, and all the PCR products of various samples are pooled as one sample for PE sequencing using the Illumina sequencing instrument. After sequencing, the PE reads are overlapped to construct full-length V6 tags, which are further separated into their original samples according to the barcode sequences. After obtaining the V6 tag sequences, the downstream analysis pipeline is the same as that used for pyrosequencing (Dethlefsen *et al*., 2008; Huse *et al*., 2008, 2010). The taxon richness and the community structure are analyzed using the V6 tag data.

Generation of a known V6 amplicon library

We extracted DNA from a mangrove sediment sample using the Powersoil DNA Kit (Mobio, Solana Beach, CA, USA) and amplified the 16S V6 fragment using the 985F (CNACGCGAAGAACCT TANC) and 1046R (CGACAGCCATGCANCACCT) primers. The PCR products were cloned using the pGEM-T easy kit (Promega, Madison, WI, USA) and sequenced using the 3730 DNA Analyzer (Applied Biosystems, Foster City, CA, USA). Nine different plasmids were extracted that contained V6 fragments (Supplementary Table S1) ranging from 106 to 129 bp (including the primers) using a plasmid extraction kit (Tiangeng, Beijing, China). The plasmids were digested with *Eco*RI (Promega), and a small fragment of ~110 bp was cut from the gel and recovered using a DNA Gel Purification kit (Tiangeng). The recovered mixed fragments were used as the template for the subsequent BIPES analysis.

PCR amplification and sequencing using HiSeq 2000

In the BIPES procedure, each sample was amplified using a forward primer with a unique barcode sequence. We used an eight-digit error-correcting barcode as described by Hamady *et al*. (2008). In addition, a 2-bp GT linker was added between the barcode and the 5′ end of the 985F primer to avoid a potential match between the barcode and the target 16S sequences. Therefore, the forward primer was a 29-base barcode-GT-985F, in which the barcode indicates the eight barcode sequences that are specific to the different samples. In this study, GCGGATAA was used for the replicate sample R1 and GCTTAACG for R2. The reverse primer was 1046R (indicated above). The high-fidelity ExTaq (Takara, Dalian, China) cocktail was used to amplify the 16S V6 tags. The PCR conditions comprised of an initial denaturation at 94 °C for 2 min; 25 cycles of 94 °C 30 s, 57 °C 30 s and 72 °C 30 s, and a final extension at 72 °C for 5 min.

The barcode-tagged 16S V6 PCR products were pooled with the other samples and sequenced using HiSeq 2000 at the Beijing Genomics Institute (Shenzhen, China). The sample was purified using the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany). The DNA was end-repaired, A-tailed and PE-adapter ligated using the Paired-end Library Preparation Kit (Illumina, San Diego, CA, USA). After ligation of the adapters, the sample was purified and dissolved in 30 μl elution buffer, and 1 μl of the mixture was used as a template for 12 cycles of PCR amplification. The PCR product was gel purified using the QIAquick Gel Extraction Kit (Qiagen) and sequenced using the 100-bp PE strategy on the Illumina HiSeq 2000 according to the manufacturer's instructions. A base-calling pipeline (Sequencing Control Software, SCS; Illumina) was used to process the raw fluorescent images and the call sequences. These sequence data have been submitted to the GenBank databases under accession nos. HQ180225–180234 and to the NCBI Sequence Read Archive SRA023706.

Bioinformatics analysis

During the PE sequencing process, each cluster was read from both ends, which generated two sequencing files that corresponded to the positive and negative strand sequences. The barcode sequence

could be present in either sequencing file 1 or 2. Therefore, we wrote a Perl script to screen the two sequencing files simultaneously. Both reads were collected if an exact match of the barcode sequence was detected at the 5′ end of either file. Finally, we obtained two sequencing files for each sample that contained the barcode sequence in one of the two single-end (SE) reads.

The PE reads were overlapped to assemble the final V6 tag sequences, which is a step that is unique to the BIPES method. We used the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970) embedded in Merger (http://emboss.sourceforge.net/apps/release/6.2/emboss/apps/merger.html) to merge the two SE reads. We found that a gap-opening penalty of 50 and a gap-extension penalty of 0 were appropriate for connecting the PE reads. We selected overlapping tags with merged e-values of 150 or higher for the subsequent analysis. During the overlap step, if any mismatch was detected between the paired reads, the base that was closer to the 5′ end of its read was used, as we found that the sequencing accuracy decreased with the sequence length. If the two distances were the same, the base was selected according to the sequencing quality score.

To determine the reference sequence source of each overlapped BIPES read we ran a pairwise comparison of each read using nine reference sequences and their reverse complimentary sequences with the Needleman–Wunsch method. We calculated the identity between each read and the reference sequences to determine the reference sequence to which each read mapped most closely. All of the subsequent error calculations were based on a comparison of the reads to their assigned reference sequences, as described by Huse *et al.* (2007).

For all of the data sets, we removed all sequences that contained one or more ambiguous reads (*N*'s), those that did not have a recognizable reverse primer sequence, and those that contained more than six errors in the primers. Similarly to Huse *et al.* (2010), we compared all of the reads to V6ref using GAST (Huse *et al.*, 2008) and removed contaminated reads that demonstrated a best match to a nontarget sequence that was at least 10% better than the match to the nearest template sequence. We also removed reads that either did not demonstrate any match or did not have a match that spanned at least 80% of their length. The remaining reads were stored as clean reads and used for subsequent analyses of the sequencing accuracy.

We compared each read to the relevant set of template sequences using the Needleman–Wunsch module (with options −g5.75 −e2.75). The error rate was calculated as the number of individual insertions, deletions and substitutions divided by the length of the template sequences. The final base error rate was determined using the total number of incorrect bases divided by the total number of bases in the sample. In comparison, the tag error rate was calculated as the number of incorrect bases in the V6 variable region of the read. All types of errors, error positions and sequencing quality scores were stored in a mySQL database for the final analysis.

Results

Sequencing accuracy of Illumina reads

The Illumina PE sequencing method determines two SE reads for every cluster in the flow cell, and each member of a paired read has the same ID number with a suffix of /1 or /2. Therefore, every PE sequencing sample has two SE sequence files, both of which contain the same number of short reads. In this study, we obtained 173 219 and 134 394 clean PE reads for the replicate samples R1 and R2, respectively (Table 1 and Supplementary Table S2). A comparison of the SE reads with their corresponding reference sequences showed that the sequencing accuracy diminished toward the end of each read, and there were good correlations between the sequencing length and error rate for all four SE read files (Figure 1). The error rate at each sequencing length was generally low for lengths of less than 50 bases, but it increased exponentially for lengths from 60 to 80 bases and finally increased to over 10%. The overall error rate for four SE sequencing files was 2.10 ± 0.19% (1 289 382 base errors out of 61 522 600 total bases), which is much higher than the rate reported for Solexa GA (0.6–1.0%) (Dohm *et al.*, 2008) and pyrosequencing (0.49%) (Huse *et al.*, 2007).

Most of the errors in the SE sequencing files were substitutions (67.8–73.8%), and deletions and insertions accounted for 16.2–19.6% and 9.9–12.6% of the errors, respectively (Supplementary Table S2). Ambiguous bases (*N*) were rare (less than 0.01%). The insertion and deletion (indel) and *N* percentages were much lower compared with those determined for pyrosequencing (Huse *et al.*, 2007). A specific analysis of the sequences obtained from template A033, which contains a seven-guanine homopolymer, revealed an error rate of 0.84% for the G7 site. A further analysis of the substitution types showed that T > G (26.9 ± 1.2%) and A > C (22.6 ± 5.1%) transversions were the two major error types (Supplementary Figure S2). The sequence contexts of the incorrect base calls demonstrated that G-error-G was the most frequent context (23.4 ± 1.1%), and G was the most frequent base before an error (∼44%), which is consistent with a previous analysis (Dohm *et al.*, 2008).

Because the 100-bp read length of HiSeq 2000 might include the V6 highly variable region, we determined whether the SE reads could be used to determine the V6 tags. The SE reads were compared with their reference sequences using the Needleman–Wunsch algorithm, and the primer sequences were removed accordingly. All nine types of template V6 fragments could be determined using SE reads but with very low tag accuracy.

Table 1 Sequencing accuracy of BIPES reads

	R1		*R2*	
	Number of occurrences	*Error rate (%)*	*Number of occurrences*	*Error rate (%)*
Base errors				
SE1	377 669	2.18	313 771	2.33
SE2	329 025	1.90	268 917	2.00
PE	46 022	0.24	35 996	0.24
PE_trim	10 568	0.07	7638	0.06
	Number of occurrences	*Percent of read (%)*	*Number of occurrences*	*Percent of read (%)*
Read errors (V6 fragment)				
SE1				
0 errors	59 961	34.6	44 134	32.8
1 error	35 355	20.4	26 643	19.8
>1 errors	77 963	45.0	63 617	47.4
SE2				
0 errors	69 578	40.2	52 124	38.8
1 error	36 179	20.9	27 491	20.5
>1 errors	67 552	38.9	54 779	40.7
PE				
0 errors	146 052	84.3	113 575	84.5
1 error	17 889	10.3	13 489	10.0
>1 errors	9338	5.4	7330	5.5
PE_trim				
0 errors	128 708	93.6	99 562	94.0
1 error	7871	5.7	5603	5.3
>1 errors	1020	0.7	746	0.7

Abbreviations: PE, paired end, overlapped sequence with SE1 and SE2; PE-trim, PE reads with only 0–1 mismatches within 40–70 bp and 0 errors in the primers; SE1, single-end sequencing file 1; SE2, single-end sequencing file 2.

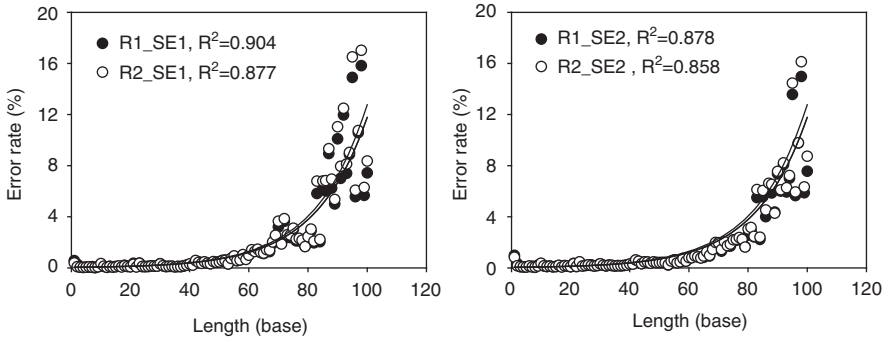

Figure 1 The error rate for single-end reads increased from the start to the end of the read. Each sample (R1 and R2) has two single-end sequencing files (SE1 and SE2). The error rate is the percentage of incorrect bases divided by the total number of bases at the specific length. Exponential fit curves and R^2 values are provided.

Only 57.0 ± 3.8% of these SE V6 tags showed zero to one errors compared with their reference sequence, whereas 21.1 ± 0.9% of them had two to three errors and 21.9 ± 3.0% had more than three errors (Table 1 and Supplementary Table S2).

Overlapping of paired-end reads

The most critical step of the BIPES procedure is the overlapping of PE reads, which differs from the single sequencing process used in pyrosequencing. In this study, the PCR products from the nine templates ranged from 106 to 129 bp, and therefore, 71 to 94 bp of overlapping sequence were obtained between the PE reads (Figure 2a). We used the Needleman–Wunsch algorithm embedded in Merger to overlap and assemble the PE reads. As described above, indels were infrequent in the Illumina sequencing reads, and an internal gap penalty value of 50 and an extension gap penalty value of 0 were found to be suitable for merging the PE reads.

The overlap step not only enabled the determination of sequences that were longer than the SE read, but it also significantly increased the sequencing accuracy. During the overlap step, sequences at the end of a SE read with high error rates were

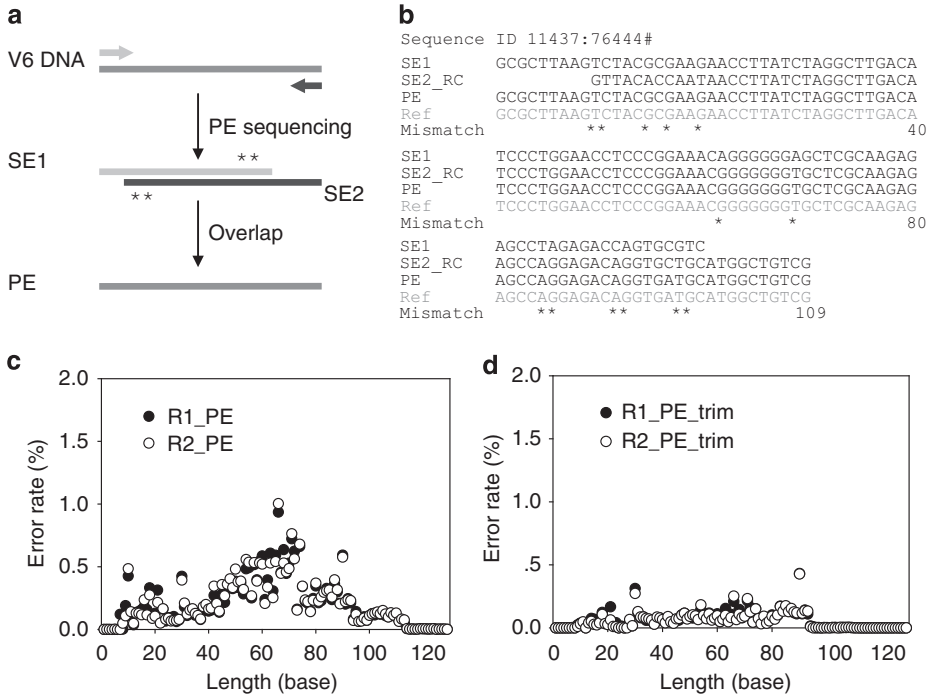

Figure 2 The overlapping of paired-end reads increased the sequencing accuracy. Schematic (**a**) showing the workflow of the overlap step. The 3′ end of single-end reads (labeled with stars) indicates a high error rate site, which is corrected by the complementary sequence. (**b**) shows an example of the overlap result. SE2_RC is the reverse complementary sequence of single-end read 2. The reference is the template sequence. The stars in the mismatch line show mismatches between SE1 and SE2. All of the errors found in SE1 and SE2 were corrected in the PE read. (**c**) is the error rate of the overlapped PE read. Notice that the maximum of the *y* axis is 2%; in contrast, a maximum of 20% is shown in Figure 1. (**d**) is the error rate of the trimmed paired-end reads with only zero to one mismatches within 40–70 bp and zero errors in the primers.

overlapped with highly accurate sequences at the start position of the other paired read (Figure 2a). In the case of a mismatch between the overlapped PE reads, the base closer to its read start was selected, and this procedure increased the overall base calling accuracy (Table 1). We also attempted to use the sequencing quality score to determine the final sequence, but the resultant tag was less accurate than that obtained using the previous method.

Figure 2b shows an example of overlapping for read number 11 437:76 444. In this read, eight errors in the forward SE sequencing read and five errors in the reverse SE sequencing read can be observed. However, after the overlap, the merged tag was the same as the reference sequence A033. As shown in Figure 2c, the overlapped PE reads were far more accurate compared with the SE reads. After the overlap step, the overall base error rate was reduced significantly to $0.24 \pm 0.00\%$, which is close to the rate of pyrosequencing with low-quality data trimmed (0.25%) (Huse *et al.*, 2007). Accordingly, the V6 tags that were determined by overlapped PE reads demonstrated a significantly higher accuracy compared with those detected in the SE reads. Among the tags, $94.6 \pm 0.1\%$ had zero to one errors with respect to the theoretical sequences, $3.9 \pm 0.0\%$ had two to three errors, and only $1.5 \pm 0.0\%$ contained more than three errors (Table 1 and Supplementary Table S2).

We observed a peak in the error rate in the middle of the overlapped PE sequences (Figure 2c). This peak was due to location of the sequences, which were nearly 40–70 bases from the beginnings of both SE reads in which the sequencing accuracy started to decrease (Figure 1). Therefore, these base errors could not be corrected using the overlap approach. A further evaluation of the tag accuracy distribution patterns revealed that most of the reads with mismatches at the beginning or end of the read contained zero to one error compared with their reference sequence, but a large fraction of the reads with mismatches of approximately 55 bp displayed two or more errors (Figure 3). We found a good correlation between the error rate determined for the V6 tags and the number of mismatched bases within 40–70 bp of the SE reads (Figure 4). Because the error rate of PE sequences with two or more mismatched bases was higher than the overall error rate, the former sequences were excluded from the final sequencing results. In addition, we found that the error rate in the primer region also correlated with the accuracy determined for the whole tag (Figure 4). The overlapped tags with one or more errors in the primers demonstrated error rates that were higher than the overall rate, and therefore, they were also removed from the final data set. The elimination of tags that contained two or more mismatches within 40–70 bases and those with any

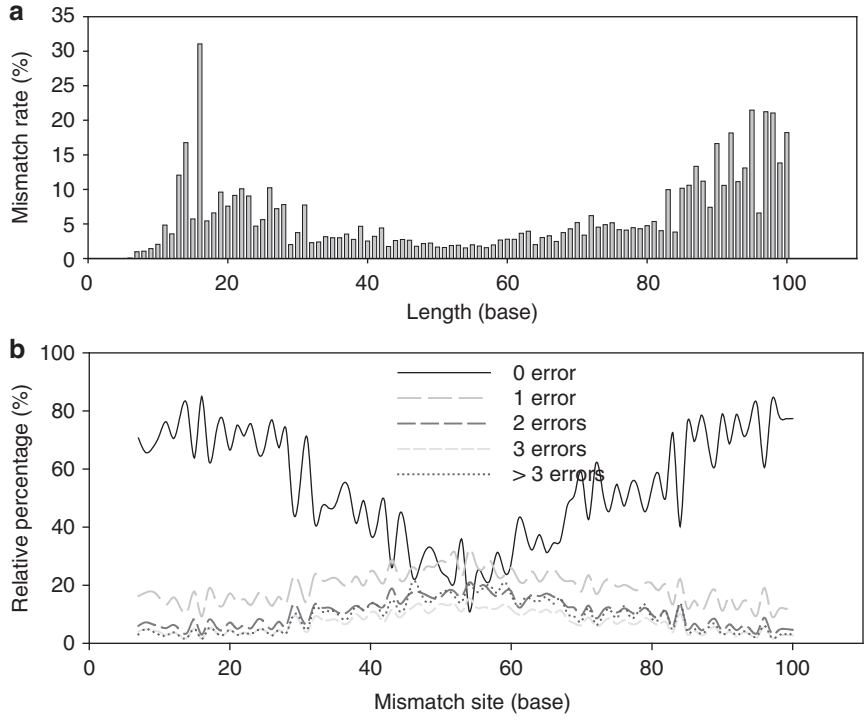

Figure 3 Mismatch and error rates for the tags with mismatches. (**a**) shows the percentage of mismatched bases divided by the total number of bases for each of the sequencing lengths. (**b**) shows the relative percentage of reads with mismatches at the sequencing length. The error number in (**b**) indicates the number of base errors in the V6 tag. For instance, 7.6% of the total PE reads have a mismatch at base 20; (**a**) among this 7.6%, 66.5% have no errors in the V6 tag, 17.2% have one error, 7.2% have two errors, 4.5% have three errors and 4.6% have more than three errors. (**b**) Although there are many mismatched bases at both ends of the reads, most of the tags with mismatches at these sites contain zero or one error; however, at the middle site (~base 55), a high percentage of tags with mismatches at these sites demonstrate two or more errors in the V6 tag, even though the number of mismatches is small.

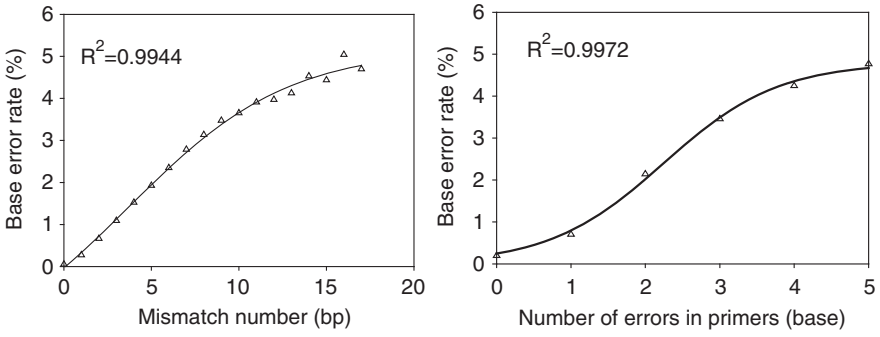

Figure 4 The mismatch number within the 40–70 bp region and the number of errors in the primers correlate with the sequencing error rate. The base error rate indicates the percentage of errors in the V6 tag sequences. The left figure shows the number of mismatches within the 40–70 bp fragment found during the overlap step. The right figure shows the number of errors in primers. The sigmoid fit curves and R^2 values are provided.

errors in the primers excluded ~20% of the total reads (35 620/1 73 219 for R1 and 28 483/1 34 394 for R2), but the sequencing error rate for BIPES was further reduced from 0.24% to 0.06% (Table 1 and Supplementary Table S2). The base error rate in the middle of the sequence decreased markedly (Figure 2d). Approximately 94% of the reads had no errors with respect to their reference sequence, 99.3% of the reads had zero to one error and only 0.2% of the reads contained three or more errors (Supplementary Table S2).

During our analysis, we also compared the relationship among the e-values determined for the overlap step and that between the sequencing quality scores of the reads and the read accuracy; however, no clear relationship was observed.

The chimera was analyzed using overlapped V6 tags with higher than 5% errors compared with reference sequences. In sample R1, there were 16 971 overlapped tags showing smaller than 95% identity with any reference sequences, in which only 480 tags fulfilled the requirement of 0–1

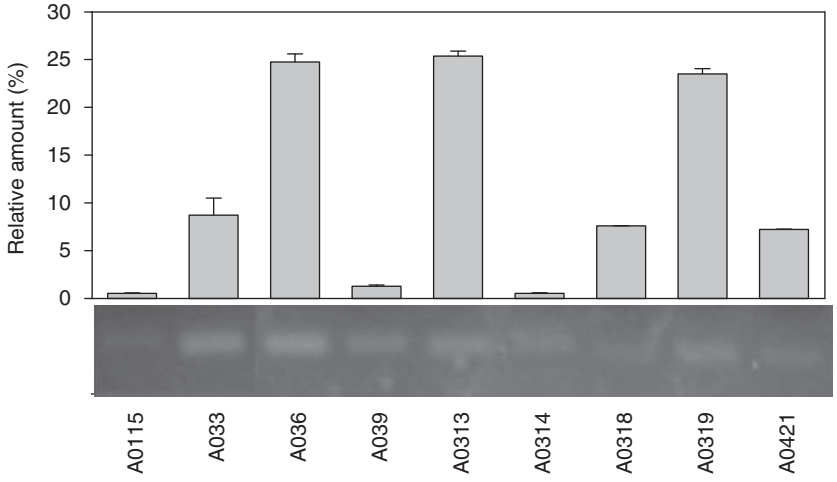

Figure 5 Quantitative determinations of the nine templates. The image showing the electrophoresed samples shows the relative amounts of the nine templates obtained during the initial DNA recovery step. The quantitative result was calculated using tags that were mapped to each one of the nine templates divided by the total number of identified tags.

mismatch within 40–70 bases and zero errors in primers. These 480 tags formed 255 unique sequences, and all of them were aligned with the most homologous reference tags. The chimeras were searched by visual examination for breakpoints and 11 potential chimera sequences were observed.

Quantification results

In addition to the sequencing accuracy, a quantitative determination is also critical for studying microbial diversity. As described in Materials and methods, the template used was a recovered DNA mixture of nine types of V6 tags consisting of different initial amounts (Figure 5). All of the determined PE reads could be mapped to one of the nine tags, and their relative amounts are compared in Figure 5. We observed good consistency between the initial template concentration and the final number of determined tags for six of the nine templates; three templates (A033, A039 and A0314) demonstrated fewer than the expected number of tags. Interestingly, these three templates showed unique properties: A033 had the highest GC percentage (62.9%) among the nine tags, whereas the other two templates (80 bp and 81 bp) were longer compared with the other tags (Supplementary Table S1).

Discussion

This study demonstrated the efficiency of using Illumina PE sequencing to determine microbial 16S variable tags. Among the three next-generation sequencing methods, 454 pyrosequencing provides the longest sequencing reads but the lowest throughput; in contrast, SOLiD has the highest throughput but provides the shortest sequencing lengths (Kircher and Kelso, 2010). At present, only the 454 system is widely used to sequence 16S tags, even though the Illumina system yields 20–50 times more reads at a much lower cost per read as compared with 454 system. Furthermore, with its upgraded read length and its unique PE sequencing strategy, the Illumina system can now be used to determine 16S variable tags that are longer than its sequencing threshold.

Currently, the V6 tag is the most suitable among the nine 16S variable regions for developing the BIPES method. An *in silico* analysis has shown that the lengths of V6 tags in the V6ref database range from 50 to 472 bp, with an average of 60 bp. Furthermore, 99.62% of the tags in the V6ref database are shorter than 77 bp (Supplementary Figure S3). In the latter group, the whole tag can be overlapped using the 100-bp PE sequencing strategy, and at least 27 bp can be overlapped for 75-bp reads (GAII). These results suggest that the V6 tags available in the V6ref database can be determined well using the BIPES method. In addition, the V6 tag sequence is sufficient to obtain a taxonomic assignment. The use of short next-generation sequencing read tags for taxonomic classification has been recently discussed (Huse *et al*., 2008; Liu *et al*., 2008; Youssef *et al*., 2009). Liu *et al*. (2008) suggested that reads of at least 250 bases could provide a satisfactory classification using Greengenes or the RDP classifier (http://rdp.cme.msu.edu/). Nevertheless, Huse *et al*. (2008) developed a GAST method and found that more than 97% of the V6 tags generated *in silico* yielded the same taxonomic assignment as the corresponding full-length 16S sequence, which suggests that the V6 tag is suitable for a taxonomic assignment. As the Illumina sequencing lengths increase, we can expect that longer variable tags, which could be analyzed using the RDP classifier, may be read using BIPES in the future. The third reason for using V6 tag is that the V6 tag is the most variable region in the 16S sequence (Hamp *et al*.,

2009), and therefore, it demonstrates a higher resolution compared with any other short-variable tag for distinguishing bacteria with different 16S sequences. Finally, the V6 tag has historically been the most widely used tag for microbial diversity analysis by pyrosequencing (Sogin *et al.*, 2006; Huse *et al.*, 2008; Turnbaugh *et al.*, 2009), which permits comparisons between studies.

The present results indicated that Illumina single sequencing had a relatively high error rate. The Illumina system employs four fluorophore-labeled deoxynucleotide triphosphates and deoxyribonucleotide triphosphates that function as reversible terminators. Therefore, only a single base is added per molecule in each cycle, which differs from the procedure performed during 454 pyrosequencing. The two platforms have been suggested to have similar sequencing accuracy, but Illumina yields fewer indels (Dohm *et al.*, 2008; Kircher and Kelso, 2010). Consistent with these reports, a low frequency of indels was found in our results, and even a homopolymer of seven guanines in template A033 demonstrated only a slightly lower accuracy compared with the overall sequence. Nevertheless, the HiSeq 2000 reads had a high error rate after 60 bases, and the overall error rate was even higher than that obtained using the Solexa GA system (Dohm *et al.*, 2008). We suggest that V6 tags determined using Illumina SE sequencing reads have too many errors, and microbial diversity studies using Solexa SE reads (Lazarevic *et al.*, 2009) should be analyzed with great care. The high error rate of HiSeq 2000 should also be noted for sequencing studies that lack overlap correction.

The overlapping of PE sequencing reads in the BIPES process significantly increased the overall sequencing accuracy. The rationale is quite sound, because the odds of obtaining the same incorrect base in both PE sequences are much lower than are those of obtaining a SE sequencing error. In the case of mismatches, the base closer to its read start (which demonstrated a higher accuracy rate) was selected; therefore, the final sequencing accuracy at each base was determined according to the more accurate of the two SE reads. In addition, the accuracy of the overlapped PE reads should be higher than that of sequencing one read twice from the same end, because the errors are related to the upstream sequence (Dohm *et al.*, 2008), which differed in most cases for the two paired sequences. The removal of sequences with two or more mismatches within 40–70 bp is a very useful criterion, because this region demonstrated relatively high error rates for both SE reads.

These two objective criteria are critical for improving the sequencing accuracy, although $\sim$20% of the total reads were removed during these steps. Pyrosequencing errors have been reported to cause significantly inflated estimates of microbial species, particularly for rare biospheres (Quince *et al.*, 2009; Reeder and Knight, 2009). Currently, PCR biases, sequencing errors and bioinformatics pipelines are the three major hurdles to the accurate assessment of microbial diversity (Hamp *et al.*, 2009; Huber *et al.*, 2009; Huse *et al.*, 2010; Kunin *et al.*, 2010). Our study encompassed both PCR and sequencing errors, but we believe that the sequencing errors were minimized. The overall accuracy was much higher than that reported for pyrosequencing, even though the BIPES procedure included 12 additional PCR cycles during the library preparation step. Future studies using BIPES without a PCR step and detailed comparisons of various PCR conditions will aid in reducing the PCR and sequencing errors, respectively. The present result proved that the chimera was rare, which is similar to those obtained in previous reports concerning the infrequent formation of chimeras for short V6 tags (Huse *et al.*, 2010). Nevertheless, we suggest that screening chimeras for V6 tags will further improve the accuracy for the estimation of the taxon richness.

This study demonstrated the feasibility and accuracy of sequencing V6 tags using the Illumina system. The BIPES procedure is generally similar to that used in pyrosequencing, except that the V6 tag sequences are determined after the overlapping of PE reads. The downstream bioinformatics pipeline of the analysis of V6 tags for operational taxonomic unit clustering, taxonomic assignment and α- and β-diversities are the same as those used in pyrosequencing (Supplementary Figure S1). The quantitative feature of the PCR-based high-throughput sequencing results has rarely been discussed. This study showed that tags with a high GC content and long tags could be underestimated, which is consistent with previous reports (Arezi *et al.*, 2003). However, because we used a relatively small number of different tags as templates, further studies are warranted to confirm these biases. At present, we suggest that BIPES tags obtained using the same PCR conditions could be used to evaluate changes in microbial communities at various spatial and time scales and in samples exposed to environmental perturbations.

In conclusion, the results of this study demonstrated the feasibility of determining 16S tags using the BIPES method. BIPES determines approximately 20–50 times more tags with a higher accuracy compared with pyrosequencing. This capacity allows the simultaneous quantification of highly abundant and rare microbes, which is particularly useful for the assessment of environmental samples (Fuhrman, 2009). In addition to its high-throughput capabilities, BIPES is affordable; the associated cost is less than one dollar per 2000 tags. With the development of next-generation sequencing techniques, the cost of each tag may further decrease and the read length may increase to encompass longer variable regions. BIPES provides new avenues of research to investigate many interesting microbial ecological questions, such as the universal patterns of microbes, taxa–area relationships and microbial

community networks (Fuhrman, 2009). In general, the BIPES method provides the possibility of modeling microbial community structure changes in response to health and environmental factors.

Acknowledgements

We acknowledge Guanhua Deng, Tengjiao Guo, Jian Zhang, Shaochuan Li, Ye Chen and Nan Qin for assistance with the Perl programming and Illumina sequencing. This study was partly supported by the PhD Programs Foundation of the Education Ministry of China (No. 20094433120017), the Natural Science Foundation of China (No. 30971193 and 31040013) and a research grant from the Shenzhen Science, Technology and Information Bureau, Shenzhen Government (Project No. (2008)121).

References

Arezi B, Xing W, Sorge JA, Hogrefe HH. (2003). Amplification efficiency of thermostable DNA polymerases. *Anal Biochem* **321**: 226–235.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**: 1177486.

DeSantis T, Brodie E, Moberg J, Zubieta I, Piceno Y, Andersen G. (2007). High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb Ecol* **53**: 371–383.

Dethlefsen L, Huse S, Sogin ML, Relman DA. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* **6**: e280.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**: e105.

Fuhrman JA. (2009). Microbial community structure and its functional implications. *Nature* **459**: 193–199.

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Meth* **5**: 235–237.

Hamp TJ, Jones WJ, Fodor AA. (2009). Effects of experimental choices and analysis noise on surveys of the 'rare biosphere'. *Appl Environ Microbiol* **75**: 3263–3270.

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Huber JA, Morrison HG, Huse SM, Neal PR, Sogin ML, Mark Welch DB. (2009). Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environ Microbiol* **11**: 1292–1302.

Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* **4**: e1000255.

Huse SM, Huber J, Morrison H, Sogin M, Welch D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**: R143.

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.

Kircher M, Kelso J. (2010). High-throughput DNA sequencing—concepts and limitations. *Bioessays* **32**: 524–536.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118–123.

Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Qsteras M *et al.* (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* **79**: 266–271.

Liu Z, DeSantis TZ, Andersen GL, Knight R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucl Acids Res* **36**: e120.

Needleman SB, Wunsch CD. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–453.

Quince C, Curtis TP, Sloan WT. (2008). The rational exploration of microbial diversity. *ISME J* **2**: 997–1006.

Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM *et al.* (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Meth* **6**: 639–641.

Reeder J, Knight R. (2009). The 'rare biosphere': a reality check. *Nat Methods* **6**: 636–637.

Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Schutte UM, Abdo Z, Bent SJ, Shyu C, Williams CJ, Pierson JD *et al.* (2008). Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. *Appl Microbiol Biotechnol* **80**: 365–380.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Tringe SG, Hugenholtz P. (2008). A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11**: 442–446.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

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.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)