Journal of Microbiology (2014) Vol. 52, No. 7, pp. 566–573
DOI 10.1007/s12275-014-3485-6

# Regional Effects on Chimera Formation in 454 Pyrosequenced Amplicons from a Mock Community

**Sunguk Shin, Tae Kwon Lee[#], Jung Min Han[†], and Joonhong Park[*]**

*School of Civil and Environmental Engineering and WCU Center for Green Metagenomics, Yonsei University, Seoul 120-749, Republic of Korea*
*[#]Present address: Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, University of Vienna, Althanstrasse 14, Wien, Austria*
*[†]Present address: Department of Dental Hygiene, Yonsei University Wonju College of Medicine, Wonju-si, Gangwon-do 120-752, Republic of Korea*

(Received Sep 23 2013 / Revised Feb 4, 2014 / Accepted Mar 12, 2014)

**Chimeras are a frequent artifact in polymerase chain reaction and could be the underlying causes of erroneous taxonomic identifications, overestimated microbial diversity, and spurious sequences. However, little is known about the regional effects on chimera formation. Therefore, we investigated the chimera formation rates in different regions of phylogenetically important biomarker genes to test the regional effects on chimera formation. An empirical study of chimera formation rates was performed using the Roche GS-FLX$^{TM}$ system with sequences of the V1/V2/V3 and V4/V5 regions of the 16S rRNA gene and sequences of the *nifH* gene from a mock microbial community. The chimera formation rates for the 16S V1/V2/V3 region, V4/V5 region, and *nifH* gene were 22.1–38.5%, 3.68–3.88%, and 0.31–0.98%, respectively. Some amplicons from the V1/V2/V3 regions were shorter than the typical length (~7–31%), reflecting incomplete extension. In the V1/V2/V3 and V4/V5 regions, conserved and hypervariable regions were identified. Chimeric hot spots were located in parts of conserved regions near the ends of the amplicons. The 16S V1/V2/V3 region had the highest chimera formation rate, likely because of long template lengths and incomplete extension. The amplicons of the *nifH* gene had the lowest frequency of chimera formation most likely because of variations in their wobble positions in triplet codons. Our results suggest that the main reasons for chimera formation are sequence similarity and premature termination of DNA extension near primer regions. Other housekeeping genes can be a good substitute for 16S rRNA genes in molecular microbial studies to reduce the effects of chimera formation.**

***Keywords*:** chimera, pyrosequencing, regional effect, mock community

## Introduction

Chimeric sequences have been a major problem in the public databases of 16S rRNA (16S) gene sequences (Ashelford *et al.*, 2005). These chimeras can lead to incorrect taxonomic identification, overestimated richness, and artificial entities. Many quantitative studies on various factors affecting chimera formation (Qiu *et al.*, 2001) have identified sequence similarity as one of the most important factors. Co-amplification of two nearly identical 16S genes can generate chimeras at a frequency of up to 30%, while the frequency of chimeras decreases as template similarity diminishes (Wang and Wang, 1996). Interestingly, the 16S gene has an alternating pattern of conserved and hypervariable areas, reflecting the secondary structure that is important for its biological function (Neefs *et al.*, 1991; Chakravorty *et al.*, 2007). The bacterial 16S genes contain nine hypervariable regions and nine conserved regions. For example, the V1/V2/V3 region includes the C2 region, and the V4/V5 region includes the C4 region (Petrosino *et al.*, 2009). Because the sequence similarity of the 16S genes can vary among regions, the structure of the 16S gene can affect the chimera formation rates. Although a few studies have evaluated the chimera formation rates for different regions of the 16S genes, these studies could not identify a significant difference in chimera formation rates for the different regions, most likely because of the differently optimized polymerase chain reaction (PCR) conditions and similar template lengths for different regions (Haas *et al.*, 2011).

In addition to the 16S gene, certain functional genes can complement the taxonomical information provided by the 16S gene (Case *et al.*, 2007). However, relatively few studies have examined chimera formation during PCR in functional gene regions, and functional genes do not have highly conserved areas because of the wobble positions in triplet codons (Crick, 1966). These findings led us to expect less frequent chimera formation in the amplicons of functional genes. However, no study to date has directly compared chimera formation rates between functional genes and the 16S gene in a mock community.

Massively parallel pyrosequencing using the Roche GS system (Margulies *et al.*, 2006) is a popular method to extensively assess molecular diversity and taxonomy in microbial communities without the cultivation of microbes. Despite its advantages of high-throughput analysis and low cost per sequence read, pyrosequencing has an important limitation-namely, the limited read lengths. Only some local regions of the full-length 16S gene or some long functional genes can be chosen because the pyrosequencing read length is limited to ~500 base pairs by manufacturers in the case of

[*]For correspondence. E-mail: parkj@yonsei.ac.kr; Tel.: +82-2-2123-5798; Fax: +82-2-312-5798

454 GS FLX (Engelbrektson *et al.*, 2010).

Therefore, testing the regional effects on chimera formation by sequence similarity in the pyrosequencing system is scientifically and practically important. The chimera formation rates of the 16S V1/V2/V3 region, V4/V5 region, and the well-known functional gene *nifH* [which encodes nitrogenase reductase (Maverech *et al.*, 1980)] were identified. In addition to the rates of chimera formation, the positions of chimera formation were scrutinized to identify the effects of the structure of the 16S gene on chimera formation in a mock community. In addition, chimeric amplicons contribute to spurious sequencing errors (Ewing and Green, 1998), although more detailed studies concerning the effects of chimeras on sequencing errors are required. To analyze the occurrence of spurious sequencing errors in response to the regions of amplicons, we analyzed ambiguous base (N), substitution, insertion, and deletion errors between our amplicon reads and reference sequences and mapped spurious sequencing errors according to their positions using BLAST (Altschul *et al.*, 1990). Many chimera detection programs have been developed, including Chimera Slayer (CS), Wigeon (Haas *et al.*, 2011), Bellerophon (Huber *et al.*, 2004), and UCHIME (Edgar *et al.*, 2011). Because Wigeon is adapted to full-length 16S rRNA gene sequences, CS and UCHIME were used to test our sequences from the 454 pyrosequencing system.

## Materials and Methods

### Mock community construction

Defined mock community DNA was constructed from the genomes of the following 20 bacterial isolates, which have been genome-sequenced except *Ralstonia pickettii* PKO1: *Xanthomonas campestris* ATCC 33913 (GenBank accession number gi: 21229478), *Sphingobium yanoikuyae* B1 (gi: 123967428), *Nostoc* sp. PCC 7120 (gi: 17227497), *Bacillus cereus* ATCC 14579 (gi: 30018278), *Chromobacterium violaceum* ATCC 12472 (gi: 34495455), *Staphylococcus epidermidis* ATCC 12228 (gi: 27466918), *Corynebacterium glutamicum* ATCC 13032 (gi: 58036263), *Rhodospirillum rubrum* ATCC 11170 (gi: 83591340), *Burkholderia xenovorans* LB400 (gi: 91685338, 91689770, and 91692731), *Roseobacter denitrificans* OCh 114 (gi: 110677421), *Rhodococcus jostii* RHA1 (gi: 111017022), *Polaromonas naphthalenivorans* CJ2 (gi: 121602919), *Burkholderia vietnamiensis* G4 (gi: 134137285, 134135188, and 134134073), *Pseudomonas putida* F1 (gi: 148545259), *Ochrobactrum anthropi* ATCC 49188$^T$ (gi: 151559234 and 151561966), *Ralstonia pickettii* PKO1 (gi: 567855), *Desulfitobacterium hafniense* DCB-2 (gi: 22728), *Rhodobacter sphaeroides* KD131 (gi:21808), *Escherichia coli* K12 substr. W3110 (gi: 11048), and *Neisseria sicca* ATCC 29256 (gi: 217378). Genomic DNA was extracted using the PowerSoil DNA isolation kit (MoBio, USA) and quantified using Nanodrop ND-1000 spectrophotometer (Thermo Scientific, USA).

### PCR and GS FLX titanium pyrosequencing

Barcoded primers were used in multiplex amplicon sequencing. The primer regions were V1-9F (GAGTTTGATCMTGGCTCAG) and V3-541R (WTTACCGCGGCTGCTGG) (Chun *et al.*, 2010) for the 16S V1/V2/V3 region, F1 (AYTGGGYDTAAAGNG) and R5 reverse (CCGTCAATTYYTTTRAGTTT) (Marsh *et al.*, 2013) for the 16S V4/V5 region, and PolF (TGCGAYCCSAARGCBGACTC) and PolR (ATSGCCATCATYTCRCCGGA) (Poly *et al.*, 2002) for the *nifH* gene. PCR was conducted using AccuPrime$^{TM}$ *Taq* DNA Polymerase High Fidelity (Invitrogen, USA). All DNAs were amplified using the same PCR conditions (35 cycles consisting of 1 min at 94°C, 1 min at 55°C, and 2 min at 72°C), and 50 μl of each mix of PCR components was prepared according to the AccuPrime$^{TM}$ *Taq* DNA protocol with the addition of 0.5 μl of $MgSO_4$. Our optimal PCR conditions were the same for all genes except the 16S V1/V2/V3 region. The optimal PCR conditions used for the 16S V1/V2/V3 region in our recent studies were 40 cycles consisting of 30 sec at 94°C, 1 min at 60°C, and 2 min 30 sec at 72°C without the addition of $MgSO_4$. After products of the correct lengths were excised, the products were extracted and purified using the QIAquick Gel Extraction Kit (Qiagen, USA). An additional purification was performed using the QIAquick PCR Purification Kit (Qiagen). The amplified products were quantified using a Nanodrop ND-1000 Spectrophotometer (Thermo Scientific). The products were pooled and concentrated using the MinElute PCR Purification Kit (Qiagen) after purification and quantification. Emulsion PCR was performed according to the manufacturer's instructions. The Roche 454 GS (FLX Titanium) pyrosequencer was used, and pyrosequencing was run on 1/8 of a sequencing plate.

### Reference sequences

Reference sequences were constructed as a stable reference for error detection, chimera detection, and the study of variable and conserved regions. Sometimes simply using dominant sequences as reference sequences is problematic because of contaminant sequences, various genome sizes, and different copy numbers of genes. To select the best reference sequences for our mock community, we used various methods. We downloaded genomic sequences for the 16S and *nifH* genes from NCBI and matched our primer sets with these sequences. To discard unmatched sequences, we checked all downloaded sequences using a hidden Markov model (HMM) method. HMM has been widely used in pattern recognition problems such as sequence alignment (Needleman and Wunsch, 1970), *in silico* gene detection (Krogh *et al.*, 1994), structure prediction (Bystroff *et al.*, 2000), and data mining literature. Many bioinformatics software programs use HMM, such as HMMSTR (Bystroff *et al.*, 2000), SAM (Hughey *et al.*, 2003), VEIL (Henderson *et al.*, 1997), and UGENE (Okonechnikov *et al.*, 2012). Additionally, we added a few dominant sequences as reference sequences after manual inspection of our sequencing data.

### Initial process and conservation analysis

We chose reads that had average quality scores greater than 20. Sequence contamination was removed through an RDP classifier at the order level. DNA reference sequences were aligned using MUSCLE (Edgar, 2004). The gap-treated Shannon entropy (H') at each alignment position was calculated

**Table 1.** **Chimera formation of the 16S and *nifH* gene amplicons with their average lengths**

| | No. of amplicons | Chimera frequency by CS | Chimera frequency by UCHIME | Average length of amplicons (bp) | Length of longest amplicon (bp) |
|---|---|---|---|---|---|
| V13r1 | 2359 | 33.0 | 38.5 | 499 | 525 |
| V13r2 | 1174 | 22.1 | 24.3 | 464 | 513 |
| V45r1 | 3896 | 3.82 | 3.88 | 330 | 339 |
| V45r2 | 3721 | 3.68 | 3.71 | 349 | 381 |
| *nifH*r1 | 2552 | 0.313 | 0.392 | 321 | 331 |
| *nifH*r2 | 1834 | 0.709 | 0.981 | 321 | 330 |

V13r1, The 16S region covering V1 to V3 from replicate sample 1; V13r2, The 16S region covering V1 to V3 from replicate sample 2; V45r1, The 16S region covering V4 to V5 from replicate sample 1; V45r2, The 16S region covering V4 to V5 from replicate sample 2; *nifH*r1, *nifH* from replicate sample 1; *nifH*r2, *nifH* from replicate sample 2.

as follows (Zhang *et al.*, 2007): where is the relative frequency of DNA at the alignment position, and represents the number of gaps at the alignment position *i* divided by the number of alignment sequences. We regarded positions below a Shannon entropy value of 0.2 in succession as conserved regions.

### Error calculation

All amplicons, chimeric amplicons, and nonchimeric amplicons after Q20 filtering were compared with reference sequences using BLAST (Altschul, 1990). The results from BLAST were parsed with Perl, and error rates were calculated using the numbers of errors and bases at the positions. Only chimeric and nonchimeric sequences from both CS and UCHIME were selected and used as chimeric and nonchimeric sequences, respectively, in the present study.

## Results

### Chimera rate

The chimera formation rates and average lengths of the 16S and *nifH* amplicons from our mock community are summarized in Table 1. The 16S V1/V2/V3 amplicons had the highest chimera formation rates (22.1–38.5%) and longest average lengths. The rates were higher than the rate (~17%) obtained in a recent quantitative study (Haas *et al.*, 2011). The average length of the V4/V5 amplicons was similar to that of the *nifH* amplicons. However, the V4/V5 region had much higher chimera formation rates than the *nif*H region. In Table 1, the abnormal, longest amplicon reads from V13r1, V13r2, and V45r2 were chimeric.

### Distribution of amplicons

In addition, we analyzed the length distribution of amplicons because we believed that sequence length and/or extension time during PCR can affect chimera formation and amplicon length distribution. In some studies, longer templates require longer extension times (Barnes, 1994), and chimera formation rates decrease as elongation times increase in PCR (Qiu *et al.*, 2001). Each sample has one length peak except the V1/V2/V3 amplicons, which have low peaks around 380 bp and 450 bp (Fig. 1).

### Conserved areas of reference sequences

Because sequence similarity can affect chimera formation, conserved and hypervariable regions of reference sequences were analyzed in detail without bias or error in multi-template PCR (Fig. 2). In addition, the highly conserved regions of reference DNA sequences were identified by calculating the Shannon entropy, $H'$, at each multi-alignment position (Table 2). The 16S gene reference sequences had highly variable and conserved regions, as shown in Table 2 and Fig. 2. The *nifH* gene reference sequences did not have conserved regions because of wobble positions.

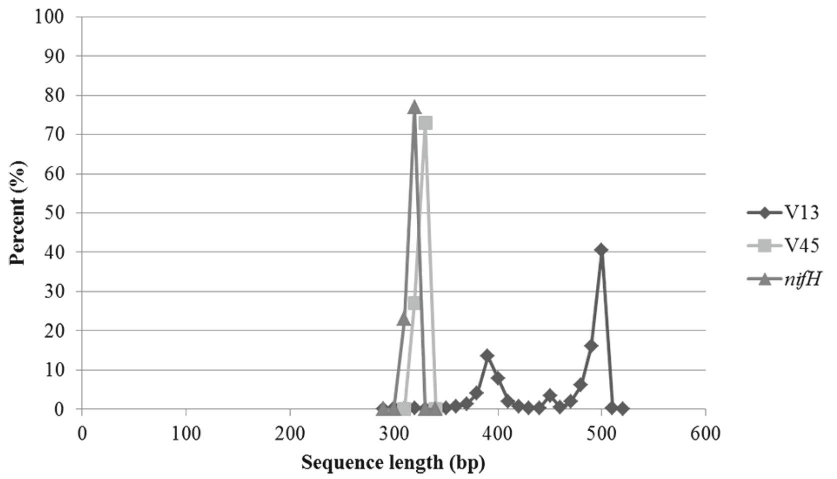

**Fig. 1.** **Distribution of sequence lengths. V13, The 16S region covering V1 to V3; V45, The 16S region covering V4 to V5; *nifH*, *nifH*.** The reads were divided into 10 bp segments.

## Sequencing errors and chimeric sequences

As a common mistake, chimeric sequences can be identified as problematic sequences that contain more sequencing errors. In our study, most errors such as substitutions, insertions, and deletions, arose from chimeric sequences (Table 3). We selected chimeric and nonchimeric sequences from both CS and UCHIME and compared the sequences with reference sequences by BLAST analysis (Fig. 3). To analyze the effect of location, the sequences were divided into 10 bp segments. Chimeric sequences had higher spurious error rates, particularly at both ends of the matched sequences. True sequencing errors were more evenly distributed in nonchimeric sequences. However, nonchimeric sequences, particularly V1/V2/V3, also included a few chimeric sequences that were only identified by our manual inspection because the chimera detection programs have high selectivity and low sensitivity (Schloss *et al.*, 2011). Therefore,

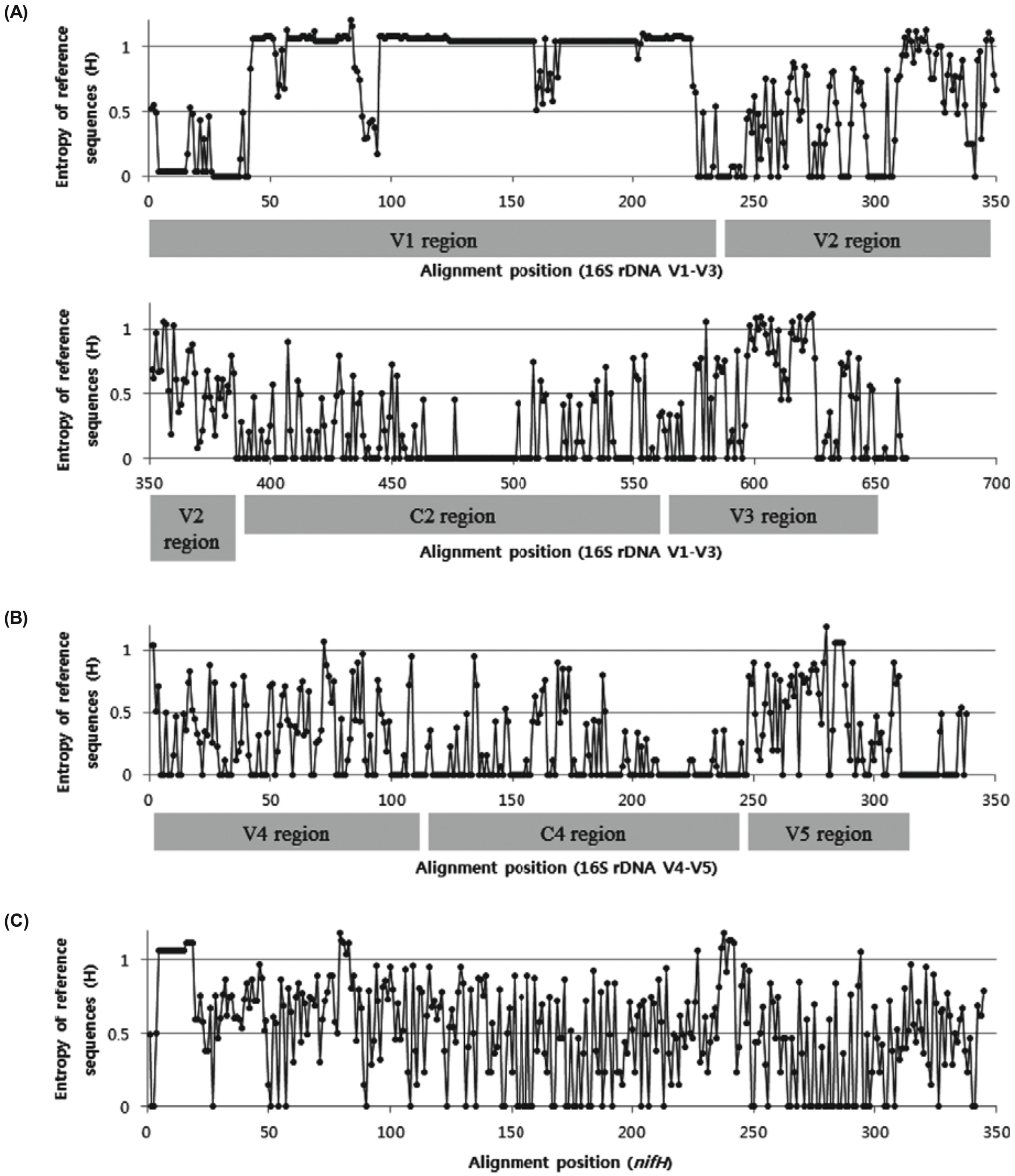

Fig. 2. **Shannon entropy, *H'*, of reference sequences.** Gap-treated Shannon entropy was calculated at each alignment position. (A) The 16S region covering V1 to V3. (B) The 16S region covering V4 to V5. (C) *nifH*.

**Table 2.** Number of highly conserved regions in reference sequences

| | V13 | V45 |
|---|---|---|
| 5–7 bp | 8 | 10 |
| 8–14 bp | 8 | 2 |
| 15–21 bp | 0 | 1 |
| 22–28 bp | 1 | 1 |
| Total | 17 | 14 |

V13, The 16S region covering V1 to V3; V45, The 16S region covering V4 to V5. We could not find conserved regions in *nifH*. We regarded positions below a Shannon entropy 0.2 in succession as conserved regions.

**Table 3.** Overall error rates of chimeric and nonchimeric amplicons

| | Error rate per nucleotide of chimeric amplicons | Error rate per nucleotide of nonchimeric amplicons |
|---|---|---|
| V13r1 | 0.04092 | 0.00400 |
| V13r2 | 0.03628 | 0.00719 |
| V45r1 | 0.04114 | 0.00154 |
| V45r2 | 0.04323 | 0.00386 |
| *nifH*r1 | 0.03086 | 0.00133 |
| *nifH*r2 | 0.04407 | 0.00163 |

V13, The 16S region covering V1 to V3; V45, The 16S region covering V4 to V5; *nifH*, *nifH*.

the actual error rates could be lower than our error rates in nonchimeric sequences because of undetected chimeric sequences that were not removed by CS and UCHIME.

## Position of chimera formation

The positions of chimera formation were analyzed using CS (Fig. 4) to identify the relationship between sequence sim-

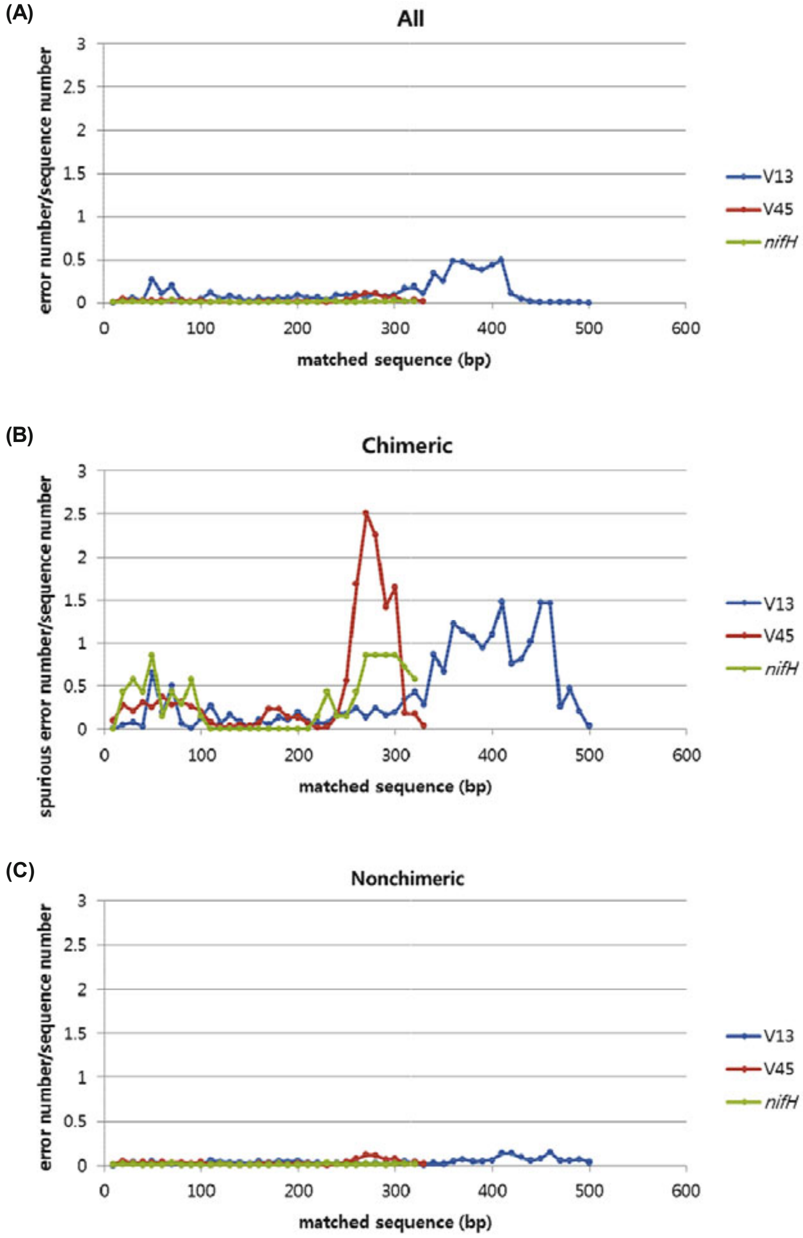

**Fig. 3. Error rates of all, chimeric, and nonchimeric sequences.** Error rates per sequence were calculated in the 16S V1/V2/V3 (V13), V4/V5 (V45), and *nifH* (*nifH*) reads. (A) All reads containing nonchimeric and chimeric sequences. (B) Chimeric sequences from both CS and UCHIME. (C) Nonchimeric sequences from both CS and UCHIME.

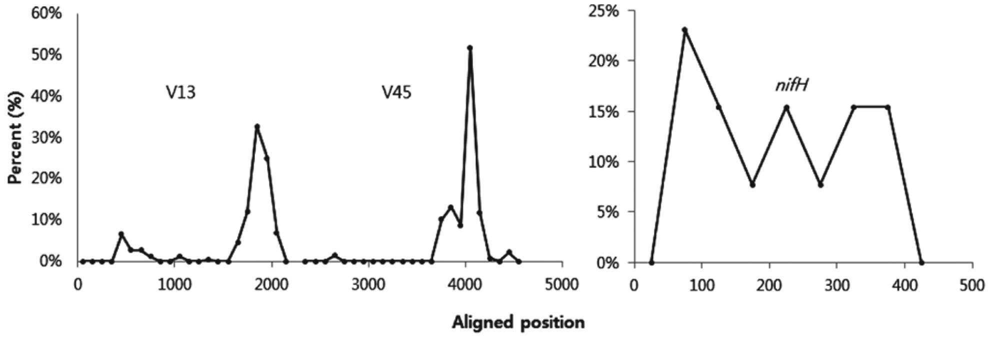

**Fig. 4. Frequency of chimera formation.** The frequencies of chimera formation were calculated in the 16S V1/V2/V3 (V13), V4/V5 (V45), and *nifH* (*nifH*) reads and were expressed as percentages. The 16S amplicon reads were aligned using greengenes gold alignment and divided into 100 bp segments. The *nifH* amplicon reads were aligned using MUSCLE and divided into 50 bp segments.

ilarity and chimera formation. The 16S reads were aligned using greengenes gold alignment, and the *nifH* reads were aligned using MUSCLE. The positions of chimera formation were very similar to those of sequencing errors. Most chimeric sequences occurred around both ends of the amplicons (Fig. 4). The positions with high percentages of chimeric sequences were 1,800–1,900 (16S V1/V2/V3), 4,000–4,100 (16S V4/V5), and 50–100 (*nifH*). These positions were consistent with 450–500, 200–250, and 25–50, in Fig. 3A, Fig. 3B, and Fig. 3C, respectively. In the 16S reference sequences, these areas were highly conserved.

### Comparison of amplicon lengths

The distribution of amplicon lengths was studied in both chimeric and nonchimeric sequences because some chimeric reads had longer lengths, as shown in Table 1. The distribution of amplicon lengths was slightly different in chimeric and nonchimeric sequences (Fig. 5). In the 16S V1/V2/V3 region, the distribution of chimeric read lengths was slightly wider, and a few very short reads were found among nonchimeric reads. In the 16S V4/V5 region, a few longer reads were found among the chimeric reads. In *nifH*, the chimeric reads were slightly shorter than the nonchimeric reads. The lengths of the 16S regions are slightly different depending on the microorganism. The longest amplicon reads were the chimeric sequences of the microorganisms that have longer 16S regions.

## Discussion

Massively parallel pyrosequencing underpins the era of metagenomics, allowing direct sequencing from environmental samples. However, uncertainty concerning the choice of a target region for amplification is due to sequence length limitations. In this study, we amplified two different regions of the 16S gene and one functional gene, *nifH*. Each region demonstrated very different chimera formation rates and characteristics, such as typical sequence length and sequence similarity. These characteristics, except the number of templates, may affect chimera formation because, in studies of chimera formation, chimera formation frequencies were similar regardless of the number of templates (Wang and Wang, 1996, 1997).

In the present study, the V1/V2/V3 amplicons had the highest chimera formation rates, and the rates were higher than those rates found in a recent quantitative study (Haas *et al.*, 2011), most likely because of different PCR conditions. In addition, in that study (Haas *et al.*, 2011), the V1/V2/V3 amplicons had chimera formation rates similar to those of the V3/V4/V5 amplicons, most likely because of the similar template lengths. In the current study, the 16S V4/V5 amplicons that were used are shorter than the 16S V3/V4/V5 amplicons. Interestingly, the 16S V1/V2/V3 amplicons demonstrated a much higher chimera formation rate than the V4/V5 amplicons, most likely because of the difference in template length. Although the V1/V2/V3 and V4/V5 reference sequences had similar numbers of conserved areas, as

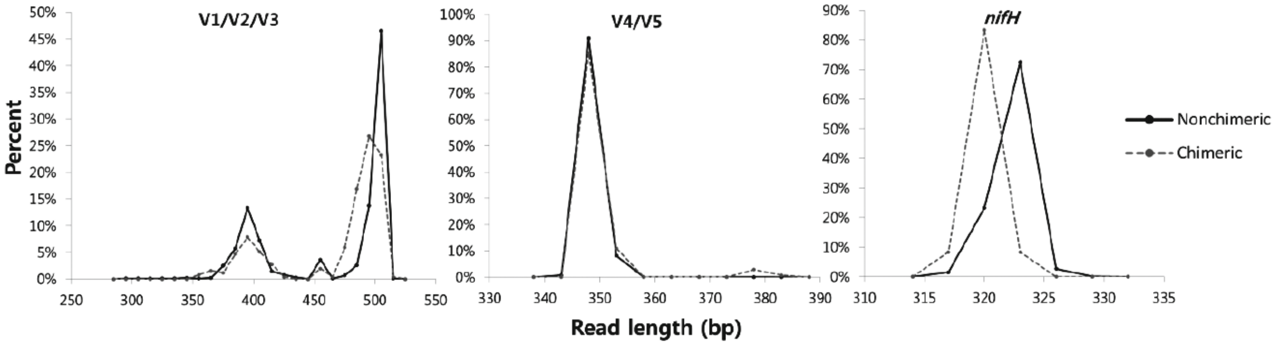

**Fig. 5. Frequency of read lengths.** The read lengths were divided into 10 bp, 5 bp, and 3 bp segments for the 16S V1/V2/V3 region, V4/V5 region, and *nifH* gene, respectively.

shown in Table 2, the V1/V2/V3 amplicons had the highest chimera formation rate, most likely because of the longer typical sequence length. Longer DNA templates require longer extension times. For the same extension time, longer template lengths might lead to incomplete DNA extension, producing shorter amplicons, as shown in Fig. 1, and more lesions in the DNA templates (Pääbo *et al.*, 1990). When the PCR extension time decreases, the percentage of chimeras can increase (Qiu *et al.*, 2001). Longer sequence lengths might increase the chimera formation rate in a similar way to the increase of chimera formation rate by a shorter PCR extension time. As a result, incomplete extension by dissociation of the polymerase during PCR near the end of the template DNA might contribute to frequent chimera formation in the 16S gene V1/V2/V3 region. According to Fig. 1 and Fig. 4, chimera hot spots exist in conserved regions near the ends of sequences in the 16S regions. In addition, in the *nifH* region, chimeras occurred slightly more frequently near the ends of amplicons. Furthermore, frequent spurious errors at both ends of the amplicons in Fig. 3 can indirectly reflect DNA extension during PCR that stops frequently near the ends of the templates. In Table 3, chimeric sequences can possess up to 30 times more errors than nonchimeric sequences. Other factors, such as the number of PCR cycles and the type of polymerases, can have effects on the different chimera formation rates of the 16S V1/V2/V3 and V4/V5 regions (Shafikhani, 2002; Acinas *et al.*, 2005); however, the above results suggest that indirect extension may be one of the important factors responsible for chimera formation. We suggest that only the PCR conditions, and not the pyrosequencing length limitation, affect the sequence length distribution because Fig. 1 shows a clear sequence distribution around 500 base pairs and because amplicons as long as 963 bp had quality-filtering pass rates >50% in other studies (Engelbrektson *et al.*, 2010). Based on our observations, we suggest that the position of chimera hot spots in the 16S region occurs in the conserved regions near the ends of amplicons (Fig. 6). Chimera formation was more frequent in the V4/V5 region than in *nifH* as we expected. $H'$ revealed that the 16S gene reference sequences contain highly conserved regions, whereas *nifH* reference sequences possess only slightly conserved regions (Fig. 2B, 2C, and Table 2) because of wobble positions in triplet codons. Conserved areas may increase the possibility of linking different DNA fragments and templates.

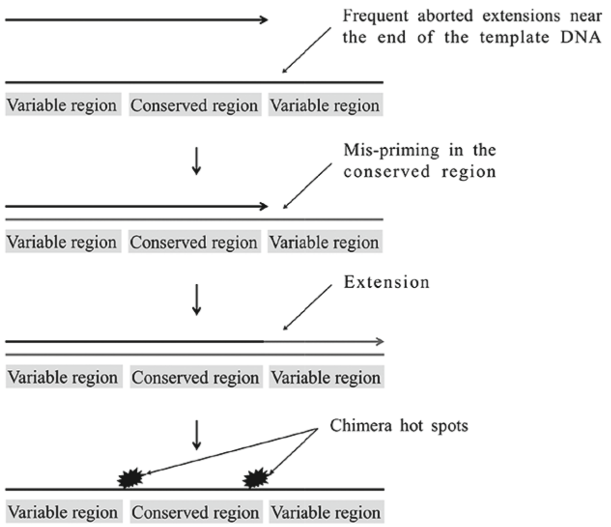

**Fig. 6. The positions of chimera hot spots in the 16S gene regions.** Frequent aborted extensions occur near the ends of template DNAs. Chimera hot spots occur in the conserved regions near the ends of amplicons from the 16S genes.

Although different microorganisms in our mock community have slightly different typical sequence lengths of the 16S gene reference sequences, our amplicons of the V1/V2/V3 region have three separate length peaks around 380 bp, 450 bp, and 500 bp in Fig. 1. These peaks might indicate more frequent incomplete extensions in the V1/V2/V3 region under the same PCR conditions. Shorter or longer chimeric amplicons occurred particularly in the 16S gene region because the lengths of the 16S regions are slightly different depending on the specific microorganism and region. For example, as shown in Table 1, the longest, abnormal sequences of the 16S region were chimeric. Our manual inspection revealed that abnormally shorter/longer amplicon reads can be formed if the chimeric sequences of the microorganisms have short/long parts of the 16S regions. Particularly in the 16S V1/V2/V3 region, the distribution of amplicon lengths was wider in chimeric sequences than in nonchimeric sequences (Fig. 5).

Our analysis of the 16S and *nifH* gene sequences revealed regional effects on chimera formation. Sequence similarity and premature termination during amplification appear to be the major cause of frequent chimera formation. Chimera hot spots in the 16S regions occurred in the parts of the conserved regions near the ends of amplicons. Wobble positions in functional genes might decrease the chimera formation rates relative to the 16S genes. Housekeeping genes should have low chimera formation rates and may be good substitutes for the 16S gene with regard to chimera formation. We recommend that DNA amplification for pyrosequencing on the Roche GS be performed with less conserved regions, that the DNA extension time not be short, and that template lengths not be too long.

## Acknowledgements

This subject is supported by Korea Ministry of Environment as "The GAIA Project" and the Students' Association of the Graduate School of Yonsei University funded by the Graduate School of Yonsei University.

## References

**Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., and Polz, M.F.** 2005. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol.* **71**, 8966–8969.

**Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J.** 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

**Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., and Weightman, A.J.** 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* **71**, 7724–7736.

**Barnes, W.M.** 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl. Acad. Sci. USA* **91**, 2216–2220.

**Bystroff, C., Thorsson, V., and Baker, D.** 2000. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* **301**, 173–190.

**Case, R.J., Boucher, Y., Dahllof, I., Holmström, C., Doolittle, W.F., and Kjelleberg, S.** 2007. Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* **73**, 278–288.

**Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D.** 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods* **69**, 330–339.

**Chun, J., Kim, K.Y., Lee, J.H., and Choi, Y.** 2010. The analysis of oral microbial communities of wild-type and toll-like receptor 2-deficient mice using a 454 GS FLX Titanium pyrosequencer. *BMC Microbiol.* **10**, 101.

**Crick, F.H.C.** 1966. Codon-anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* **19**, 548–555.

**Edgar, R.C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acid Res.* **66**, 1–40.

**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., 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.

**Ewing, B. and Green, P.** 1998. Base-calling of automated sequencer traces usingPhred II. Error probabilities. *Genome Res.* **8**, 186–194.

**Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., and *et al.*** 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* **21**, 494–504.

**Henderson, J., Salzberg, S., and Fasman, K.H.** 1997. Finding genes in DNA with a hidden Markov model. *J. Comput. Biol.* **4**, 127–141.

**Huber, T., Faulkner, G., and Hugenholtz, P.** 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317–2319.

**Hughey, R., Karplus, K., and Krogh, A.** 2003. SAM: sequence alignment and modeling software system. Technical report UCSC-CRL-99-11 (Report). University of California, Sa, Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics.* **20**, 2317–2319.

**Krogh, A., Brown, M., and Haussler, D.** 1994. A hidden markov model that finds genes in *E. coli* DNA. *Nucleic Acid Res.* **22**, 4768–4778.

**Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., and *et al.*** 2006. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.

**Marsh, A.J., O'Sullivan, O., Hill, C., Ross, R.P., and Cotter, P.D.** 2013. Sequencing-based analysis of the bacterial and fungal composition of kefir grains and milks from multiple sources. *PloS One* **8**, e69371.

**Maverech, M., Rice, D., and Haselkorn, R.** 1980. Nucleotide sequence of cyanobacterial *nifH* gene coding for nitrogenase reductase. *Proc Natl. Acad. Sci. USA* **77**, 6476–6480.

**Needleman, S.B. and Wunsch, C.D.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

**Neefs, J.M., Van de Peer, Y., De Rijk, P., Goris, A., and De Wachter, R.** 1991. Compilation of small ribosomal subunit RNA sequences. *Nucleic Acid Res.* **18**, 2237–2317.

**Okonechnikov, K., Golosova, O., and Fursov, M.** 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167.

**Pääbo, S., Irwin, S.D.M., and Wilson, A.C.** 1990. DNA damage promotes jumping between templates during enzymatic amplification. *J. Biol. Chem.* **265**, 4718–4721.

**Petrosino, J.F., Highlander, S., Luna, R.A., Gibbs, R.A., and Versalovic, J.** 2009. Metagenomic pyrosequencing and microbial identification. *Clin. Chem.* **55**, 856–866.

**Poly, F., Gros, R., Jocteur-Monrozier, L., and Perrodin, Y.** 2002. Short-term changes in bacterial community fingerprints and potential activities in an alfisol supplemented with solid waste leachates. *Environ. Sci. Technol.* **36**, 4729–4734.

**Qiu, X., Wu, L., Huang, H., McDonel, P.E., Palumbo, A.V., Tiedje, J.M., and Zhou, J.** 2001. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl. Environ. Microbiol.* **67**, 880–887.

**Scholss, 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.

**Shafikhani, S.** 2002. Factors affecting PCR-mediated recombination. *Environ. Microbiol.* **4**, 482–486.

**Wang, G.C. and Wang, Y.** 1996. The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* **142**, 1107–1114.

**Wang, G.C. and Wang, Y.** 1997. Frequency of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl. Environ. Microbiol.* **63**, 4645–4650.

**Zhang, S.W., Zhang, Y.L., Pan, Q., Cheng, Y.M., and Chou, K.C.** 2007. Estimating residue evolutionary conservation by introducing von Neumann entropy and a novel gap-treating approach. *Amino Acids* **35**, 495–501.