Capturing Greater 16S rRNA Gene Sequence Diversity within the Domain Bacteria

T. Winsley, J. M. van Dorst, M. V. Brown, and B. C. Ferrari
School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia

A large proportion of “universal” 16S PCR primers lack sequence homology to many of the “candidate” divisions, severely limiting bacterial diversity assessments. We designed a primer set that offers a 50% increase in silico in coverage of the domain Bacteria over the commonly used primer combination 27F/519R. Comparisons using pyrosequencing on soil environments showed a significant increase in recovery of taxonomic diversity with around a 3-fold increase in recovery of sequences from candidate divisions.

Over 3 decades have passed since Woese and Fox first utilized the small subunit of the ribosome to define the three domains of life (27). Since then, the utilization of the PCR, particularly in surveys of DNA extracted directly from environmental samples, has led to a dramatic increase in our knowledge of the diversity associated with these domains, particularly among the prokaryotes (1, 22). Indeed the number of phyla within the domain Bacteria, defined using phylogenetic analysis of 16S rRNA gene sequences, has increased from the original 12 defined by Woese to 92 currently listed in the NCBI databanks. Advancements in sequencing technology, oligonucleotide synthesis, and data processing have all contributed to this surge in cataloguing of bacterial diversity (19, 24). The recent widespread availability and affordability of 454 pyrosequencing to survey ribosomal gene tags, has enabled the generation of hundreds of thousands of reads from a single sample. Together, these advances in sequencing and computing power have resulted in databases, such as the Ribosomal Database Project (RDP), increasing significantly from ~500 16S rRNA gene sequences in 1992 to 1,613,063 sequences today (1).

The vast majority of 16S rRNA gene sequences within these repositories are the outcome of diversity studies on a wide range of environments. However, there are a number of taxon-targeted studies that also significantly contribute to these databases (6, 25). The bulk of diversity assessments are performed with purported “universal” 16S rRNA gene PCR primers, yet one of the more widely used PCR primer sets (27F/519R) was designed 30 years ago (11, 12, 16). At that time, there were relatively few phylogenetically established phyla, with Proteobacteria, Firmicutes, Bacteroidetes, and Actinobacteria accounting for the bulk of the 16S rRNA gene sequence data and isolates (1, 16, 27). Not surprisingly, it is possible that these primers lack 16S rRNA gene sequence homology to some of the “newer” bacterial phyla, particularly the yet-to-be cultured “candidate” divisions that comprise about 40% of the domain Bacteria. Hence, diversity assessments of many environments are potentially underestimated and the prospect for many candidate divisions to be excluded from 16S rRNA gene libraries remains a problem.

As we learn of new divisions and expand environmental studies further, we now know that commonly used primer sets need updating (2, 10). Several attempts have been made to redesign and optimize the “universal” bacterial 16S primer sets, with little success and without adoption by the research community (3, 20, 26). Here we describe a novel set of 16S rRNA gene PCR primers designed with particular emphasis on obtaining greater sequence homology to the neglected candidate divisions.

**Primer design and in silico testing.** Bacterial and archaeal 16S rRNA gene sequences were obtained from the RDP, and alignments were constructed using ClustalX software (5, 18) and manually curated. Alignments consisted of members of every defined bacterial phylum. Additionally, several archaeal phyla were included to prevent recognition of this domain. Highly conserved regions within the Bacteria were identified, and candidate primers were assessed first with RDP’s probe match function to identify the breadth of homology and then with Primer3 software (5, 23) to determine suitability for use in PCR amplification. In silico performance of our selected primer set, 356F (5′-ACWCCTACGGGWGCGWC-3′) and 1064R (5′-AYTCTACGRCACGCAGCTGAC-3′), was tested by correlating accession numbers matched in the probe match function against several commonly used “universal” 16S PCR primers (see Table S1 in the supplemental material). A few PCR primers that were popular in the literature were selected for in silico comparison: these included 27F (16), 63F (20), 519R (17), 530F (16), 787R (9), 910R (13), 1100R (16), 1392R (16), and 1492R (16). As sequences submitted to repositories are usually trimmed to remove the primer from the sequence, extrapolation to a “best case scenario” was done by considering the number of accessions available at each particular position on the 16S rRNA gene within the RDP. While this was a crude approach, it provided a comparable situation for primers across the length of the whole gene. This in silico analysis showed that the commonly used primers missed up to 92% (63F) of diversity when surveying the domain Bacteria. Additionally, primers (519R and 530F) that covered a greater portion of the Bacteria also resulted in homology to a high number of archaeal sequences. By comparison, 365F and 1064R obtained 99% and 95% coverage of the Bacteria, respectively, according to the extrapolated values (i.e., no archaeal sequences that were homologous to our candidate primer set were
correlated to both the forward and reverse sequences). Combined in silico performance of our candidate primer set covered 85% of the domain Bacteria, compared to an extrapolated 35% of the 27F/519R set (data not shown).

**PCR optimization and practical validation.** The synthesized primers were optimized for PCR against genomic DNA from 8 bacterial isolates spanning 4 different phyla: (i) Actinobacteria, Micrococcus luteus and Microbacterium ginsengisoli; (ii) Bacteroidetes, Chitinophaga sp. nov.; (iii) Firmicutes, Lactococcus lactis; and (iv) Proteobacteria, Methylobacterium radiotolerans, Sphingomonas melonis, Escherichia coli, and Pseudomonas aeruginosa. The PCR program consisted of 95°C for 5 min and then 35 cycles of; 95°C for 30 s, 60°C for 30 s, and 72°C for 60 s, followed by a final step of 72°C for 5 min. A 50-µl reaction mixture consisted of 3 mM MgCl₂, 800 µM deoxynucleoside triphosphates (dNTPs), 5 µg bovine serum albumin (BSA), 10 pmol each primer, and between 2 and 20 ng of genomic DNA. Subsequently, the primer pair was validated in triplicate by bar-coded amplicon pyrosequencing on a Roche 454 Titanium instrument (Roche, Branford, CT) (7).

Genomic DNA was extracted from soils originating from the Antarctic, sub-Antarctic, and Australian Desert regions using the FastDNA spin kit for soil (MP Biomedicals, Seven Hills, New South Wales, Australia). To compare and benchmark the performance of the new primer set, the same soils were also assessed with the “universal” primer set 27F and 519R (16,17). This primer set was chosen for the practical comparison due to the in silico result, performance in amplicon pyrosequencing, and popularity throughout the literature.

**Pyrosequencing data analysis.** Sequence data were processed with the mothur software package (24). This involved quality screening of sequences, denoising, and chimera removal via the Chimera Uchime algorithm contained within mothur (8), followed by distance-based clustering of sequences and binning into operational taxonomic units (OTU). Since primer set 27F/519R spans hypervariable regions V1 to V3 and 356F/1064R spans regions V6 to V9, different OTU definitions were required to call species-level assignments. These dissimilarities were 0.04 and 0.02, respectively, as determined by Kim et al. (15). For the purpose of standardizing sampling effort, the number of reads for each environment was normalized by randomly subsampling from the larger group to the number of reads of the smallest group. Taxonomy was assigned from the GreenGenes database with a bootstrap cutoff of 80% (21). Rarefaction data were generated via a sampling without replacement method using the mothur package. Sample-by-OTU abundance data matrices from mothur were subsequently transposed, and multivariate analysis was performed with the PRIMER (Plymouth Routines in Multivariate Ecology Research) software package (4).

Preliminary alpha-diversity analysis across the samples immediately highlighted the increased richness obtained by the 356F/1064R primer set over 27F/519R. The numbers of species-level OTU observed for 27F/519R were 566, 888, and 330, compared to 1,116, 1,249, and 532 in Antarctic, sub-Antarctic, and Australian Desert samples, respectively, using 356F/1064R. When considering the observed number of OTU compared to the sampling effort, the rarefaction curves showed that samples surveyed with the 356F/1064R primer set will reach an asymptote much later than those interrogated with 27F/519R (Fig. 1). Chao1 and abundance-based coverage estimation (ACE) estimates for total species richness showed similar improvements in the overall diversity captured with the 356F/1064R primer set.

A significant difference in the abundance of phyla across the different environments was observed between the primer sets (Fig. 2). More phyla were detected using the primer set 356F/1064R.
and the relative abundances of dominant well-characterized phyla were reduced as diverse phyla emerged (Fig. 2). The most significant differences were up to 3-fold increases in the abundance of the candidate divisions: 1.33% to 4.78% for the Antarctic soil (with 27F/519R compared to 356F/1064R), 1.47% to 3.7% for the sub-Antarctic soil, and 2.55% to 7.08% for the Australian Desert soil (see Table S2 in the supplemental material). SIMPER analysis in PRIMER was used to assess the difference between primer sets on environments by determining the Bray-Curtis dissimilarity between samples; The average dissimilarities in the observed phyla obtained by both 27F/519R and 356F/1064R across Antarctic, sub-Antarctic, and Australian Desert soils were 45.41, 19.78, and 21.39, respectively (see Table S3 in the supplemental material). The analysis confirmed that less abundant phyla contributed to the diversity obtained, with 356F/1064R creating dissimilarity to the samples when surveyed using 27F/519R. Over-all comparison of the two primer sets showed that the 356F/1064R set recovered a similar diversity composition at the phylum level across environments with greater species richness and evenness of taxonomic breadth. In contrast, the 27F/519R primer set displayed a limited recovery of richness and evenness with a bias toward abundant taxa.

Current sequencing and computing technology has made the acquisition of large DNA sequence data sets tractable to any laboratory. With this access to rapid and easily obtainable sequence data, researchers are interrogating more diverse environments (14, 19). However, without access to primer sets that reflect a greater range of the domain Bacteria, these studies will remain limited in accuracy of the assessment of true diversity. The primer set we have developed appears to present a greater reflection of the diversity of microbial consortia within soil samples. In the future, when pyrosequencing platforms enable increased read lengths, even more phylogenetic information will be available by using the primer set 356F/1064R owing to the larger size of the amplicon. Until sequencing technology completely negates the use of PCR for assaying microbial diversity, primer sets will need to be updated and optimized as we learn more of the ever-expanding domain Bacteria. Pending such advancements, we propose primer set 356F/1064R as a suitable candidate for more accurate assessments of bacterial diversity in microbial ecology investigations.

ACKNOWLEDGMENTS

This work was funded by the University of New South Wales (UNSW). Soil samples were kindly donated by Ian Snape and Rachael Anderson of the Australian Antarctic Division and Malcolm Walter of the Australian Centre for Astrobiology at UNSW. Sequencing was performed by Scot Dowd at the Research and Testing Laboratory, Lubbock, Texas.

REFERENCES

3. Ben-Dov E, Shapiro OH, Siboni N, Kushmaro A. 2006. Advantage of