AEM
Journals.ASM.org

Novel High-Rank Phylogenetic Lineages within a Sulfur Spring (Zodletone Spring, Oklahoma), Revealed Using a Combined Pyrosequencing-Sanger Approach

Noha Youssef, Brandi L. Steidley,* and Mostafa S. Elshahed

Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, Oklahoma, USA

The utilization of high-throughput sequencing technologies in 16S rRNA gene-based diversity surveys has indicated that within most ecosystems, a significant fraction of the community could not be assigned to known microbial phyla. Accurate determination of the phylogenetic affiliation of such sequences is difficult due to the short-read-length output of currently available high-throughput technologies. This fraction could harbor multiple novel phylogenetic lineages that have so far escaped detection. Here we describe our efforts in accurate assessment of the novelty and phylogenetic affiliation of selected unclassified lineages within a pyrosequencing data set generated from source sediments of Zodletone Spring, a sulfide- and sulfur-rich spring in southwestern Oklahoma. Lineage-specific forward primers were designed for 78 putatively novel lineages identified within the pyrosequencing data set, and representative nearly full-length small-subunit (SSU) rRNA gene sequences were obtained by pairing those primers with reverse universal bacterial primers. Of the 78 lineages tested, amplifiable products were obtained for 52, 32 of which had at least one nearly full-length sequence that was representative of the lineage targeted. Analysis of phylogenetic affiliation of the obtained Sanger sequences identified 5 novel candidate phyla and 10 novel candidate classes (within *Fibrobacteres*, *Planctomycetes*, and candidate phyla BRC1, GN12, TM6, TM7, LD1, WS2, and GN06) in the data set, in addition to multiple novel orders and families. The discovery of multiple novel phyla within a pilot study of a single ecosystem clearly shows the potential of the approach in identifying novel diversities within the rare biosphere.

The utilization of high-throughput short-read sequencing technologies in 16S rRNA-based diversity surveys (32) allowed an exponential increase in the sampling depths and number of samples analyzed for a fraction of the cost of Sanger sequencing, leading to what has been described as renaissance for the16S rRNA-based diversity analysis (71). Pyrosequencing- and, more recently, Illumina-based surveys have provided insights into spatiotemporal variations of microbial communities in various habitats as well as into the impact of various environmental factors on microbial community structure, composition, and diversity (2, 23, 24, 26, 46, 49, 60, 61). Efforts aimed at standardization of amplification and sequencing procedures (51), quality control (40, 47, 64, 79), and development of microbial community analysis programs tailored for high-throughput sequencing surveys (8, 11, 28) have resulted in the widespread democratization and adaptation of these approaches by the majority of microbial ecologists (4, 7, 9, 10, 43, 45, 53–55, 78, 80).

One of the important observations obtained from currently published high-throughput diversity surveys is that a significant fraction of the obtained sequences (for example, 32% in reference 80, 13% in reference 49, 10% in reference 74, 25% to 47% in reference 36, and 15% in reference 66) is often unclassifiable; i.e., the sequence similarity to the closest classifiable relative in databases is lower than a certain preset sequence similarity threshold, e.g., 80%. A fraction of sequences belonging to such unclassifiable fractions could putatively be representatives of truly novel high-rank taxonomic lineages, where the depth achieved by high-throughput surveys allowed access to hitherto untapped reservoirs of diversity. Indeed, a significant fraction of unclassifiable sequences often belonged to rare operational taxonomic units (OTUs) (see, for example, references 5 and 24). On the other hand, a fraction of these sequences could putatively be unclassifiable due to the constraints of sequence technology (i.e., short read lengths of approximately 250 and 450 bp in FLX and Titanium pyrosequencing technology, respectively [1, 72]) and sequence analysis (i.e., classification based on identifying the nearest neighbor in databases, rather than detailed phylogenetic analysis). The relative contribution of each of these two possible scenarios to the overall proportion of unclassified sequences in a specific data set is unclear and probably depends on multiple factors such as the level of diversity in an ecosystem, length of amplicons amplified, region of 16S rRNA gene amplified, alignment and classification tool and database utilized, and set thresholds for classification.

Given the unprecedented scope of the high-throughput sequencing surveys that have been conducted (2, 23, 24, 26, 27, 46, 49, 60, 61), or are currently being conducted or are planned to be conducted in the near future (25), gauging an accurate estimate of the nature and depth of novel diversities within various ecosystems is crucial. Here, using sediments harboring an extremely diverse microbial community from Zodletone Spring, a sulfide- and sulfur-rich spring in southwestern Oklahoma, we developed and tested an approach for the experimental evaluation of the phylogenetic affiliation of selected rare members of a microbial

Received 3 January 2012 Accepted 27 January 2012
Published ahead of print 3 February 2012
Address correspondence to Mostafa S. Elshahed, Mostafa@okstate.edu.
* Present address: Department of Microbiology, Ohio State University, Columbus, Ohio, USA.
Supplemental material for this article may be found at http://aem.asm.org/.

doi:10.1128/AEM.00002-12

community. This report demonstrates for the first time the feasibility of conducting an ecosystem-wide evaluation of the phylogenetic affiliations of a large number of unclassified lineages from a pyrosequencing data set. The proposed approach combines the high-throughput capabilities of pyrosequencing and the read length and phylogenetic resolution of Sanger sequencing. Our results indicate that, while a large fraction of the lineages could accurately be binned into existing lineages (classes, orders, or families) once representative longer sequences are obtained, several of the unclassifiable lineages represent novel, previously undocumented bacterial phyla and classes, hence clearly demonstrating that the rare biosphere is indeed a yet-untapped reservoir of novel phylogenetic diversity.

MATERIALS AND METHODS

Site description and sampling. Sediments were collected from the source area of Zodletone Spring, a sulfide- and sulfur-rich spring in southwestern Oklahoma, in May 2009. Details of the spring geochemistry (18, 67, 80) have been described before. The source of the spring is a contained area (1 m^2) in which anaerobic, biomass-laden, and sulfide-rich black viscous sediments are covered by an anoxic, sulfide-rich (8.4 mM) 40-cm-deep water column (18, 67). The sampling procedure has been described before (80). Sediments were stored undisturbed on ice for approximately 3 h until transferred to the laboratory, where they were stored at −20°C for 3 days prior to DNA extraction.

Pyrosequencing and identification of putatively novel lineages within the pyrosequencing data set. Pyrosequencing, sequence analysis, binning, and alignment, as well as various aspects of the phylogenetic diversity and community structure of the pyrosequencing data set, have been detailed in a previous publication (80). Briefly, DNA was extracted from 4 adjacent (within 1 mm of each other) samples (1 g each) by the use of a FastDNA spin kit for soil (MP Biomedicals, Solon, OH). The V1 and V2 regions of the 16S rRNA gene were amplified using primer pair 8F and 338R. While this pair is not the best choice for maximizing phylogenetic capability for identification (51), or for the most accurate species richness estimates (21, 79), the choice was aimed at maximizing the size of the product obtained when paired with a universal reverse primer (see below). Overall, pyrosequencing yielded a total of 292,130 high-quality reads (83.5% of all reads obtained), with an average read length of 263 ± 17 bases. Using the Greengenes alignment and classification system, 32% of sequences (50,675 sequences in 13,619 $\text{OTUs}_{0.03}$) remained unclassified using a criterion of less than 85% similarity to a closest relative in Greengenes database (80). These unclassified sequences formed 7,884, 7,339, and 6,189 distinct lineages at 8% (putative family), 10% (putative order), and 15% (putative class) sequence divergence cutoffs, respectively (13, 14).

Our current study utilized 78 such lineages for detailed pyrosequencing-Sanger analysis. These lineages were chosen by insertion of representative $\text{OTUs}_{0.03}$ of the unclassified sequences (13,619 OTUs) into the ARB program (52) for preliminary phylogenetic inferences. ARB putatively grouped the majority of sequences into established higher taxonomic cutoffs, i.e., previously known divisions and candidate divisions. The remaining sequences (2,014 sequences binned into 385 OTUs and 78 lineages) were considered putatively novel and chosen for further analysis. Therefore, it is important to note that this effort represented a pilot study that did not target all novel groups in Zodletone Spring. Moreover, the census of the bacterial community in Zodletone Spring is far from complete. The pyrosequencing data set utilized in this study fails to identify all sequences within Zodletone Spring, as shown by the rarefaction curve and coverage analysis (80). Relative abundances of chosen lineages ranged between 0.0003% and 0.3% of the total pyrosequencing data set.

Sequences that were considered of poor quality as evaluated using the quality control procedures employed with the original pyrosequencing data set were removed from the data set based on the following criteria: average quality score of 25 as suggested before (see reference 39), an incorrect primer sequence, one or more ambiguous bases, a homopolymer stretch longer than 8 bases, and/or length shorter than 80 bp (80). However, recently, newer approaches have been implemented to reduce the effect of sequencing errors on downstream analysis (62) and to improve the process of quality filtering of pyrosequences. To avoid any effect that pyrosequences with suboptimal quality might have on the downstream process of identifying novel lineages, we applied the new quality-filtering approach (using a minimum average quality score of 35 over a window of 50 bases instead of an average quality score of 25 as previously applied to the entire pyrosequencing data set), as suggested in reference 62, on the 2,014 potentially novel sequences. While the more stringent approach removed 22 additional sequences and hence decreased the number to 1,992 high-quality pyrosequences, the number of $\text{OTUs}_{0.03}$ and lineages did not change. This is due to the fact that the trimmed sequences belonged to OTUs with two or more sequences and that other sequences in the same OTU were of high quality. The 78 putatively novel lineages (now composed of 1,992 high-quality pyrosequences) were subsequently targeted in an effort to obtain nearly full-length sequences representative of such lineages.

Designing lineage-specific primers. Short pyrosequences from each lineage were used to design forward lineage-specific primers by the use of the probe design function of ARB (52), with selected criteria of a length of 18 bp and a GC content within the 35% to 65% range. ARB-generated primers were further evaluated for specificity using the Probebase program of RDP (11) and through BlastN comparisons (42). Primers were designed so that they would have a difference of at least 1 bp from the closest database relative, provided that no more than 10 sequences of various affiliations with such criteria were present in public databases, and at least two mismatches to large members of all coherent phylogenetic lineages. In general, 1 to 3 forward lineage-specific primers were designed for each lineage, such that a PCR product of at least 1,100 bp would be obtained upon amplification. A list of all primers designed is shown in Table S1 in the supplemental material.

PCR amplification, sequencing, and assembly. For each lineage, a specific forward primer was paired with a reverse universal bacterial primer, 1391R (5′-GACGGGCGGTGWGTRCA-3′), 1492R (5′-GGTTACCTTGTTACGACTT-3′), or 1525R (5′-AAGGAGGTGWTCCARCC-3′). The choice of the reverse primer to pair with the specific forward primer was based on GC content and the absence of primer dimers or other secondary structures. If no product was obtained using the first-choice reverse primer, we proceeded to the second-choice primer and so on. PCRs were conducted in a 25-μl volume. Each reaction mixture contained approximately 10 ng of the extracted DNA, 1× PCR buffer (Promega, Madison, WI), 2.5 mM $MgSO_4$, a mixture of 0.2 mM deoxynucleoside triphosphates (dNTPs), 0.5 U of GoTaq Flexi DNA polymerase (Promega, Madison, WI), and the forward and the reverse primers (10 μM [each]). A similar PCR protocol was used for all reactions, except for the change in the annealing temperature utilized. The PCR protocol was as follows: initial denaturation at 95°C for 5 min, followed by 40 cycles of denaturation at 95°C for 45 s, annealing at 46 to 54°C for 45 s, and elongation at 72°C for 90 s. A final elongation step at 72°C for 20 min was included. The highest possible annealing temperatures for the reverse primers were 54°C, 52°C, and 50°C for 1391R, 1492R, and 1525R, respectively. A trial-and-error approach was used to obtain a single-band PCR product with the expected size. For each lineage, the theoretical possible range of annealing temperatures was calculated as 2 to 10 degrees below the primer melting temperature (T_m −2 to T_m −10). For those forward primers that can theoretically anneal at temperatures above 54°C (the highest possible annealing temperature of all the reverse primers), the annealing temperature range was truncated to 54°C to T_m −10. PCR was first conducted at the highest possible annealing temperature using one reverse primer at a time. Temperatures were decreased sequentially, and the reverse primer was changed until a single-band PCR product was obtained at the correct size. In some instances, PCR products were not

TABLE 1 Novel phyla identified in this study[a]

	Pyrosequencing lineage			Sanger-generated sequence result[b]		qPCR quantification	
		Forward primer[c]					
Novel phylum	No.[e]	Position	Sequence	No. obtained	% similarity to closest GenBank relative	No. of DNA copies/mg	% of total 16S rRNA copies
ZDP1	19	188	GGCTGACTGATAAAAGGG	1	81	$9.6 \times 10^4 \pm 4.2 \times 10^3$	0.5 ± 0.03
	20	177	AATAGCATTGGAGAGTCG	6	80–82		
ZDP2	24[d]	169	TAATCCCGCATGTGCTCT	5	86–88	$4.9 \times 10^3 \pm 9 \times 10^2$	0.026 ± 0.002
ZDP3	36[d]	35	TTGGCGGTGCGTCTTAGA	2	81–82	$5.2 \times 10^2 \pm 1 \times 10^2$	0.002 ± 0.001
ZDP4	52[d]	216	CGGTCGCCATCGGATGAG	4	83–84	$1.2 \times 10^5 \pm 1.5 \times 10^3$	0.62 ± 0.06
	53[d]	191	CTTCCATTGATGAAAGGC	7	83–84		
ZDP5	65[d]	63	GTACGGAACTATGCTAGC	2	77–84	$7.7 \times 10^4 \pm 3.4 \times 10^3$	0.4 ± 0.06

[a] Data represent group-specific forward primer sequences, percent similarity to GenBank first hits for the Sanger-generated sequences, and results of qPCR quantification for members of the proposed phyla.

[b] Data represent Sanger-generated sequences obtained. No., number of sequences obtained for each lineage; % similarity, range of % sequence similarities to GenBank first hits for the sequences affiliated with each group.

[c] Forward primer used in the PCR. Position corresponds to numbering of the *E. coli* 16S rRNA gene.

[d] Primer may not be specific. Some of the Sanger-generated sequences in this group did not belong to that novel phylum.

[e] Pyrosequencing lineage numbers refer to the number given to each of the 78 pyrosequencing lineages investigated for novelty in this study as described in detail in Table S1 in the supplemental material.

obtained despite our best efforts; i.e., after all possible forward-reverse primer pair combinations were tried at all possible annealing temperatures, no PCR product, a multiple-band PCR product, or a single-band PCR product of the incorrect size was obtained. For the rest of the lineages, a single-band PCR product of the correct size was obtained. The obtained PCR products were cloned using a TOPO TA cloning kit (Invitrogen Corp., Carlsbad, CA) according to the manufacturer's instructions. For each lineage, 8 clones were sequenced at the Oklahoma State core facility using both M13f and M13r primers to obtain a nearly full-length sequence. Forward and reverse sequences were assembled using DNA-baser V2 sequence assembly software (HeracleBioSoft, Pitesti, Romania). Chimeras were first checked using the command chimera.slayer (30) within MOTHUR (65). After removing potential chimeras, a second round of chimera checking was conducted on the remaining sequences using the Bellerophon program, available on the Greengenes website (14). The two-round chimera checking was essential to ensure that sequences were not considered novel based on a chimeric origin.

Phylogenetic analysis. To determine whether the Sanger-generated sequences of a certain lineage were representative of the targeted short pyrosequencing-generated sequences, both the long and short sequences from each lineage were aligned using ClustalX (48), followed by trimming both the 5′ end (base pairs between 8f and the forward lineage-specific primer in the pyrosequencing reads) and the 3′ end (base pairs in Sanger-generated sequences past the 3′ end of the short sequences) in Jalview (73). The aligned truncated sequences were then analyzed using MOTHUR (65) to create a distance matrix followed by clustering them into OTUs. We considered a certain lineage to be successfully targeted when the Sanger-generated sequences and the short pyrosequences belonging to that lineage clustered in an operational taxonomic unit, with percent sequence divergence less than or equal to 15%. This rather stringent percent sequence divergence cutoff (15%, equivalent to a putative class level) was chosen as the criterion for defining target sequences to avoid any spurious effect a higher cutoff might have on falsely identifying target sequences.

Preliminary evaluation of phylogenetic affiliations of sequences obtained was conducted as follows: (i) comparison to the GenBank nr database using BlastN (42); (ii) alignment to the global 7,682-character Greengenes NAST-aligner (14) and subsequent classification by the use of the Greengenes taxonomic framework; and (iii) alignment to the global 7,682-character Greengenes NAST-aligner, import of these alignments into the Greengenes October 2010 database in the ARB software package (version 5.1-private-6215 M) (52), and determination of their positions after parsimony insertion into the universal ARB phylogenetic tree.

The combined output of such methods was used to identify lineages putatively representing novel phylogenetic lineages at the phylum or class levels. For detailed phylogenetic assessments, target sequences and their closest relatives (if any) were aligned to a collection of reference sequences representing 17 phyla and to 8 candidate phyla using ClustalX (48). The phylogenetic position of target sequences was evaluated using distance, parsimony, maximum-likelihood, and Bayesian approaches. Parsimony and distance neighbor-joining analyses were conducted using PAUP 4.0b (Sinauer Associates, Sunderland, MA) with the appropriate distance substitution model determined using ModelTest 3.7 (58). Maximum-likelihood analysis was conducted using RAxML 7.0 (68), and Bayesian analysis was conducted using MrBAyes 3.1 (37). Sequences were deemed representative of a new phylum if two or more distinct sequences remained reproducibly monophyletic and formed a bootstrap-supported independent cluster by the use of various character inputs, i.e., with or without a Lane mask, upon applying various tree-building algorithms as well as upon varying the composition and size of the data set used for phylogenetic analysis (12). The exact phylogenetic affiliation of novel lineages at the subphylum level was evaluated with special insertion into detailed phylum trees to determine the exact phylogenetic novelty of the lineage within this phylum.

Estimation of 16S rRNA gene copy numbers of proposed novel bacterial phyla by the use of qPCR. The total copies of bacterial 16S rRNA genes per milligram of template DNA were quantified using universal primer pair 27F (5′-AGAGTTTGATCCTGGCTCAG-3′) and 338R (5′-GCTGCCTCCCGTAGGAGT-3′) (80). To quantify members belonging to the novel phyla proposed in this study, we used quantitative PCR (qPCR) with a forward primer specific to each phylum and the universal reverse primer 338R. Two of the proposed novel phyla (ZDP2 and ZDP5) were composed of a single pyrosequence lineage each, and all corresponding Sanger sequencing products obtained were representative of such lineages. Therefore, for ZDP2 and ZDP5, the same forward group-specific primer used for amplification of longer reads for Sanger sequencing (see above and Table 1) was utilized for and paired with the universal reverse primer 338R in the quantitative PCR analysis of the respective novel phylum. On the other hand, two of the proposed novel phyla (ZDP1 and ZDP4) were composed of 2 lineages each. One novel phylum (ZDP3) was composed of a single lineage, but that lineage also harbored sequences shown to belong to 2 distinct novel classes as well. Accordingly, for these

3 novel phyla (ZDP1, ZDP3, and ZDP4), designing separate qPCR phylum-specific forward primers was essential. The PRIMROSE program (3) was used to design such primers. The specificity was checked using RDP (11) against the probebase database and BLAST (42) against the nr database. The designed qPCR primers and the amplicon size expected when paired with the 338R universal reverse primer are shown in Table S2 in the supplemental material.

qPCR was conducted using a MyIQ thermocycler (Bio-Rad Laboratories, Hercules, CA) and B-R SYBR green Supermix for IQ (Quanta Biosciences Inc., Gaithersburg, MD). pCR4-TOPO (Life Technologies, Grand Island, NY) plasmids carrying phylum-specific 16S rRNA gene inserts were used both as qPCR-positive controls and standards for the corresponding novel phylum. Plasmids were extracted from clones grown overnight on LB plus kanamycin using a PureLink Quick Plasmid Miniprep kit (Life Technologies, Grand Island, NY) according to the manufacturer's instructions. Quantification of plasmid DNA concentrations was carried out using a Quant-iT double-stranded DNA (dsDNA) assay kit and a Qubit fluorometer (Life Technologies, Grand Island, NY). To avoid nonspecific target amplification, we conducted PCR at a range of annealing temperatures corresponding to 2 to 10°C below the theoretical primer melting temperature prior to the qPCR analysis. The highest annealing temperature resulting in a single-band PCR product of the correct size was chosen for downstream qPCR analysis. The specificity of primer pairs designed for qPCR (i.e., the ZDP1, ZDP3, and ZDP4 phylum-specific forward primers and the 338R universal reverse primer) was examined by cloning PCR products and subjecting 8 clones to Sanger sequencing. All clones were affiliated with the corresponding novel phylum (data not shown). qPCR was conducted using a 25-μl reaction mixture containing the forward phylum-specific primer (0.3 μM [each]) and 338R, 4.25 ng of template DNA, and 12.5 μl of B-R SYBR green Supermix for IQ. Since the expected amplicons ranged in size between 134 to 320 bp, a 3-step qPCR protocol was used as follows. Reactions were heated at 95°C for 3 min followed by 50 cycles of 20 s at 95°C, 45 s at the annealing temperature, and 45 s at 72°C.

Prediction of novel phylum 16S rRNA molecule secondary structure. Small-subunit rRNA molecules are known to fold into conserved secondary and tertiary structures to facilitate interaction with ribosomal proteins (57). Early work by Woese and colleagues established the secondary structure of 16S rRNA (29, 56, 77) that was subsequently used as a model for its three-dimensional (3-D) folding (69). To confirm that the sequences representing each of the proposed novel phyla could be folded into credible secondary structures, a sequence representative of each novel phylum was first aligned with the *Escherichia coli* strain K-12 subsp. MDS42 16S rRNA gene sequence (GenBank accession number AP012306) by the use of ClustalX (48). Alignment was essential before attempting secondary-structure prediction to account for any insertions or deletions in the sequences. Alignments were visualized using Jalview (73). Since the bases corresponding to 1:X (where X denotes the position of the specific forward primer according to *E. coli* numbering) in sequences belonging to novel phyla were truncated, pairing of those bases with each other (e.g., bases 9 to 13 with bases 21 to 25 and bases 61 to 82 with bases 87 to 106) or with other bases (e.g., bases 17 to 19 with bases 916 to 918, bases 27 to 37 with bases 547 to 556, bases 39 to 46 with bases 395 to 403, bases 52 to 58 with bases 354 to 359, bases 113 to 115 with bases 312 to 314, and bases 122 to 142 with bases 221 to 239) could not be predicted in sequences belonging to most of the novel phyla. As a result, prediction of the secondary structure of the whole 16S rRNA molecule was not possible. To overcome this problem, guided by the alignment to the *E. coli* 16S rRNA gene sequence, we divided the representative 16S rRNA sequence belonging to each novel phylum into subsequences corresponding to *E. coli* base numbers 61 to 350 (only for the 2 novel phyla ZDP3 and ZDP5), 179 to 220 (only for the 2 novel phyla ZDP1 and ZDP2), 240 to 311 and 316 to 350 (only for the 3 novel phyla ZDP1, ZDP2, and ZDP4), and 367 to 393, 404 to 546, 556 to 915, and 920 to 1502 (for all novel phyla). Each of these subsequences was then submitted for secondary-structure prediction using the Mfold web server (82), followed by comparison of the minimum energy structure predicted to the conserved secondary structure of the *E. coli* 16S rRNA molecule.

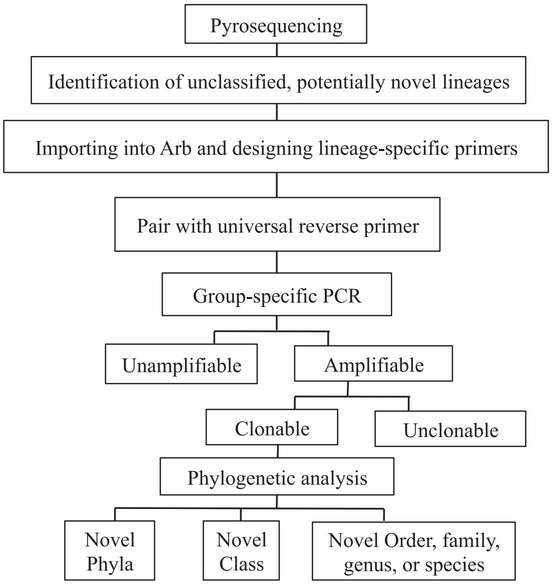
Pyrosequencing
Identification of unclassified, potentially novel lineages
Importing into Arb and designing lineage-specific primers
Pair with universal reverse primer
Group-specific PCR
Unamplifiable
Amplifiable
Clonable
Unclonable
Phylogenetic analysis
Novel Phyla
Novel Class
Novel Order, family, genus, or species

FIG 1 Flowchart of the combined pyrosequencing-Sanger approach for novel lineage identification.

Nucleotide sequence accession numbers. The Sanger sequences obtained in this study have been deposited in GenBank under accession numbers JN387275 to JN387595.

RESULTS

Utility of a combined pyrosequencing-Sanger approach in obtaining representative sequences of target lineages. Seventy-eight putatively novel lineages (1,992 sequence binned into 385 OTUs$_{0.03}$) were determined based on automated classification of pyrosequences and subsequent insertion into an ARB global phylogenetic tree. Figure 1 represents an outline of the combined pyrosequencing-Sanger process. Details on the primers designed, reverse primer utilized, and annealing temperatures are presented in Table S1 in the supplemental material. Of the 78 lineages, 19 gave no clonable PCR products despite our best efforts, 7 gave a clonable product but no inserts were obtained upon sequencing, and 52 were successfully amplified and cloned and eight clones were sequenced per lineage. After checking and removing potential chimeras (8 sequences), 322 nearly full-length sequences were obtained and subjected to further analysis.

Sanger sequences were compared to their respective pyrosequencing reads used for primer design to determine whether the Sanger sequence obtained represented the targeted pyrosequencing lineage according to the criteria outlined above. Of the 322 nearly full-length sequences analyzed, 124 sequences belonging to 32 lineages could confidently be labeled as target sequences, i.e., were less than 15% divergent from their target short pyrosequences. Within these 32 lineages, 17 were less than 5% divergent, 9 were more than 5% but less than 10% divergent, and 6 were more than 10% but less than 15% divergent from their target short pyrosequences. The large number of putative nontarget sequences is partly due to the fact that, while the designed primers were specific to the target pyrosequencing reads used for primer design and did contain mismatches to all publicly available sequences in ARB and RDP databases, 16S rRNA gene sequences that have no mismatches to the designed primer but do not belong to the target lineage are present within the extremely diverse source sediments in Zodletone Spring. Indeed, a significant fraction of nontarget sequences were not phylogenetically novel (see below) but belonged to well-described bacterial lineages. On the other hand, note that in some cases, confidently determining whether the obtained Sanger sequences belonged to target lineages was not feasible due to inadequate overlap between short and long sequences, since the designed primers were close to the 3′ end of the short pyrosequencing reads. Moreover, in some cases, Sanger-generated sequences were considered putative nontarget sequences due to a value of divergence from pyrosequencing-generated sequences only slightly higher than 15%, a threshold that could potentially be considered conservative, knowing the hypervariable nature of the overlap region.

Within the 32 lineages for which representative nearly full-length sequences were obtained, Sanger sequencing yielded exclusively target sequences in 13 lineages, while the nearly full-length sequences obtained for 19 lineages had both target and putative nontarget sequences (see Table S3 in the supplemental material).

Phylogenetic affiliations of the nearly full-length sequences obtained. All nearly full-length sequences obtained were subjected to detailed phylogenetic analysis. Results identified multiple novel lineages at all high-rank taxonomic levels (phylum, class, order), as well as many representing novel families, genera, or species within identified lineages. Within the absolute majority of novel phyla and classes identified in this study, the nearly full-length sequences obtained were indeed representatives of the original pyrosequences used for primer design. In addition, in several cases, nearly full-length sequences representing two or more distinct pyrosequencing lineages clustered in a single, distinct, phylogenetically coherent monophyletic group. The fact that multiple short pyrosequencing lineages could cluster together to form one phylum is not surprising, since our initial identification of lineages in the pyrosequencing data set was based on a 15% sequence divergence cutoff and since the proportion of hypervariable bases in the V1-plus-V2 region is higher than that in nearly full-length fragments, thus inflating levels of sequence divergence in short fragments. Below, we present a detailed analysis of the phylogenetic affiliations of the nearly full-length sequences obtained. A list of all novel lineages encountered in this study is shown in Table S4 in the supplemental material.

Novel candidate phyla. Detailed phylogenetic analysis identified five novel candidate phyla within the Zodletone Spring source community labeled ZDP1 to ZDP5 (Table 1). In total, 27 sequence representatives belonging to these 5 phyla were identified. The intralineage dissimilarity within each novel phylum ranged from 4% to 12%. These sequences originated from 7 different pyrosequencing lineages (Table 1). Sequences belonging to each of the novel phyla remained monophyletic and reproducibly formed a bootstrap-supported independent branch upon application of various tree-building algorithms (a distance neighbor-joining tree is shown in Fig. 2; parsimony, maximum-likelihood, and Bayesian trees are shown in Fig. S1 in the supplemental material).

Of the 27 sequences representing those novel phyla, only 3 were putative nontarget sequences (had more than 15% divergence from the corresponding short pyrosequences). In general, sequences belonging to these lineages have low sequence similarity to their closest relatives in public databases. Further, the closest relatives (e.g., the first five BLAST hits) of such sequences often belonged to multiple phylogenetic lineages, further suggesting that these novel sequences are not affiliated with any currently recognized phyla. Owing to the low percent similarity of sequences belonging to the proposed novel phyla to database-deposited 16S rRNA gene sequences, it was essential to ensure that those sequences were indeed representing small-subunit rRNA. We used Mfold to predict whether sequences belonging to the proposed novel phyla could fold into a credible small-subunit rRNA secondary structure. Characteristic secondary-structure features predicted for the proposed 5 novel phyla compared to those of *E. coli* are shown in Table S5 in the supplemental material. In general, all sequences representing novel phyla had the minimal features of small-subunit rRNA as described in reference 77. Variations from the conserved *E. coli* secondary structure were mainly in bases 144 to 219 (a multiple loop with 3 helices-hairpin loops corresponding to helices H8, H9, and H10 [76]) and 437 to 497 (a helix-internal loop-hairpin loop combination corresponding to helix H17 [76]) due to insertion or deletion in the corresponding bases in sequences belonging to the novel phyla. These helices were shown to interact with ribosomal proteins S4, S5, S12, S16, S17, and S20 (76). The base insertions or deletions led to a slightly shorter or slightly longer helix, internal loop, or hairpin loop (see Table S5 in the supplemental material). The only exceptions were sequences belonging to novel phylum ZDP2, where helix H10 was absent (see Table S5 in the supplemental material). However, according to reference 76, helix H10 does not seem to interact with any small subunit ribosomal protein, and so its presence might be dispensable.

Sequences belonging to novel phylum ZDP1 were most closely similar (80% to 82%, depending on the sequence compared) to those of uncultured microorganisms encountered in multiple habitats. Examples of the closest relatives were clones from the rhizosphere of phragmites at Sosei River in Sapporo, Japan (GenBank accession numbers AB240249, AB240377, and AB240493), clones from rice paddy soil (GenBank accession numbers AB486974 and AB487197) (41), and clones from microbial mats from Lower Kane Cave (Wyoming) (20) (GenBank accession number AM490643).

Sequences belonging to novel phylum ZDP2 were most closely similar (86% to 88%, depending on the sequence compared) to those of two clones from hydrothermally active sediments of the Guaymas Basin (GenBank accession number AF419661) (70) and from Lake Waiau sediment from the Hawaiian Archipelago (GenBank accession number AY345499) (16). Both clones do not belong to any of the recognized phyla within the Greengenes taxonomic schemes. Subsequent closest GenBank relatives were only 79% to 82% similar to ZDP2 sequences and were encountered in multiple habitats, e.g., an undisturbed tall grass prairie preserve in Kessler Farm Field Laboratory Biological Research Station in central Oklahoma (clones with GenBank accession numbers EU135313, EU135307, EU134920, EU134918, EU134910, and EU134671) (19) and a mud volcano sediment layer in the Neapolitan region of the eastern Mediterranean (GenBank accession number AY592608; see reference 34).

Sequences belonging to novel phylum ZDP3 were most closely similar (76% to 82%) to uncultured microorganisms encountered in multiple habitats. Examples include a clone from three microbial mat samples collected from volcano 1 on the Tonga-Kerma-

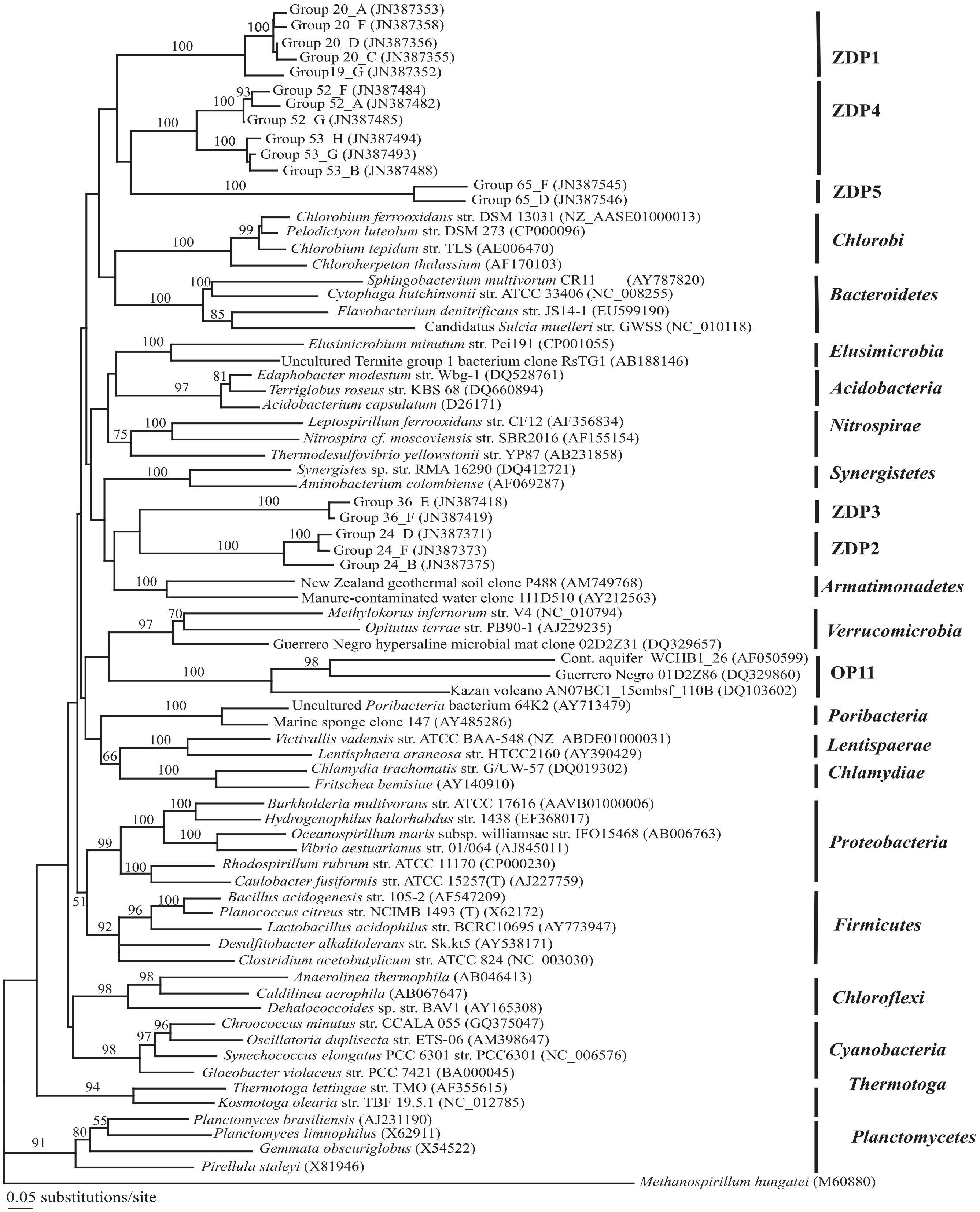
Group 20_A (JN387353)
100
Group 20_F (JN387358)
100
Group 20_D (JN387356)
Group 20_C (JN387355)
Group19_G (JN387352)
ZDP1
93
Group 52_F (JN387484)
100
Group 52_A (JN387482)
Group 52_G (JN387485)
100
100
Group 53_H (JN387494)
Group 53_G (JN387493)
Group 53_B (JN387488)
ZDP4
100
Group 65_F (JN387545)
Group 65_D (JN387546)
ZDP5
Chlorobium ferrooxidans str. DSM 13031 (NZ_AASE01000013)
99
Pelodictyon luteolum str. DSM 273 (CP000096)
100
Chlorobium tepidum str. TLS (AE006470)
Chloroherpeton thalassium (AF170103)
Chlorobi
100
Sphingobacterium multivorum CR11 (AY787820)
Cytophaga hutchinsonii str. ATCC 33406 (NC_008255)
100
85
Flavobacterium denitrificans str. JS14-1 (EU599190)
Candidatus Sulcia muelleri str. GWSS (NC_010118)
Bacteroidetes
100
Elusimicrobium minutum str. Pei191 (CP001055)
Uncultured Termite group 1 bacterium clone RsTG1 (AB188146)
Elusimicrobia
81
Edaphobacter modestum str. Wbg-1 (DQ528761)
97
Terriglobus roseus str. KBS 68 (DQ660894)
Acidobacterium capsulatum (D26171)
Acidobacteria
100
Leptospirillum ferrooxidans str. CF12 (AF356834)
75
Nitrospira cf. moscoviensis str. SBR2016 (AF155154)
Thermodesulfovibrio yellowstonii str. YP87 (AB231858)
Nitrospirae
100
Synergistes sp. str. RMA 16290 (DQ412721)
Aminobacterium colombiense (AF069287)
Synergistetes
100
Group 36_E (JN387418)
Group 36_F (JN387419)
ZDP3
100
Group 24_D (JN387371)
100
Group 24_F (JN387373)
Group 24_B (JN387375)
ZDP2
100
New Zealand geothermal soil clone P488 (AM749768)
Manure-contaminated water clone 111D510 (AY212563)
Armatimonadetes
70
Methylokorus infernorum str. V4 (NC_010794)
97
Opitutus terrae str. PB90-1 (AJ229235)
Guerrero Negro hypersaline microbial mat clone 02D2Z31 (DQ329657)
Verrucomicrobia
98
Cont. aquifer WCHB1_26 (AF050599)
100
Guerrero Negro 01D2Z86 (DQ329860)
Kazan volcano AN07BC1_15cmbsf_110B (DQ103602)
OP11
100
Uncultured Poribacteria bacterium 64K2 (AY713479)
Marine sponge clone 147 (AY485286)
Poribacteria
100
Victivallis vadensis str. ATCC BAA-548 (NZ_ABDE01000031)
66
Lentisphaera araneosa str. HTCC2160 (AY390429)
Lentispaerae
100
Chlamydia trachomatis str. G/UW-57 (DQ019302)
Fritschea bemisiae (AY140910)
Chlamydiae
100
Burkholderia multivorans str. ATCC 17616 (AAVB01000006)
100
Hydrogenophilus halorhabdus str. 1438 (EF368017)
100
Oceanospirillum maris subsp. williamsae str. IFO15468 (AB006763)
99
Vibrio aestuarianus str. 01/064 (AJ845011)
100
Rhodospirillum rubrum str. ATCC 11170 (CP000230)
Caulobacter fusiformis str. ATCC 15257(T) (AJ227759)
Proteobacteria
51
100
Bacillus acidogenesis str. 105-2 (AF547209)
96
Planococcus citreus str. NCIMB 1493 (T) (X62172)
92
Lactobacillus acidophilus str. BCRC10695 (AY773947)
Desulfitobacter alkalitolerans str. Sk.kt5 (AY538171)
Clostridium acetobutylicum str. ATCC 824 (NC_003030)
Firmicutes
98
Anaerolinea thermophila (AB046413)
98
Caldilinea aerophila (AB067647)
Dehalococcoides sp. str. BAV1 (AY165308)
Chloroflexi
96
Chroococcus minutus str. CCALA 055 (GQ375047)
97
Oscillatoria duplisecta str. ETS-06 (AM398647)
98
Synechococcus elongatus PCC 6301 str. PCC6301 (NC_006576)
Gloeobacter violaceus str. PCC 7421 (BA000045)
Cyanobacteria
94
Thermotoga lettingae str. TMO (AF355615)
Kosmotoga olearia str. TBF 19.5.1 (NC_012785)
Thermotoga
55
Planctomyces brasiliensis (AJ231190)
80
Planctomyces limnophilus (X62911)
91
Gemmata obscuriglobus (X54522)
Pirellula staleyi (X81946)
Planctomycetes
Methanospirillum hungatei (M60880)
0.05 substitutions/site

FIG 2 Distance dendrogram based on the 16S rRNA Sanger-generated sequences affiliated with novel phyla (ZDP1 to ZDP5) encountered in the Zodletone Spring source sediment clone libraries. The tree was obtained using a Tamura-Nei substitution model with a proportion of invariable sites = 0.0606 and a variable site gamma distribution shape parameter = 0.8004. Bootstrap values (in percentages) are based on 1,000 replicates and are shown for branches with more than 50% bootstrap support.

TABLE 2 Group name, primer sequence, and percent similarity and accession numbers of GenBank first hits for the Sanger-generated sequences belonging to putative novel classes identified in this study

		Pyrosequencing lineage			Sanger-generated sequence result[b]		
			Forward primer[c]			Closest relative in GenBank	
Novel class name	Phylum[a]	No.[e]	Position	Sequence	No. obtained	% similarity	GenBank accession no.
ZDC1	*Planctomycetes*	4[d]	198	GGATTTTCGGACCTTCTG	2	89	BX294875
		77[d]	177	GATGTGACCACACTGGCG	1	79	EU287119
ZDC2	*Fibrobacteres*	54	176	GGATATTGTGGAGCATCG	8	82–83	FJ716839
ZDC3	BRC1	36[d]	35	TTGGCGGTGCGTCTTAGA	3	78–80	GU172181
ZDC4	GN12	49	190	CGGCGATGAGCAAAGATG	4	80–83	GQ472374, EU048619
ZDC5	TM6	58	173	ACAGCATACGTCTTTTCG	3	82–84	GQ246408, FJ264771
ZDC6	LD1	67[d]	125	GGGTACTTGCCCTCGACT	7	78–80	HQ174951
ZDC7	TM7	75	173	ACTCCATGTGGTCTTACG	8	81–85	FJ542972
ZDC8	GN06	7	66	TGCGAGGCGGTTCTTCGG	4	86–87	GQ246356, FJ264777
ZDC9	GN06	9[d]	55	TGCAAGTCGGATGCGAAA	3	89–90	EU245453
		10	176	GCATACGCTTGTCTCTGT	7	87–92	EU245453, FJ516883, EU245159
ZDC10	WS2	33[d]	162	GCGCCGCTAATACCGGGT	4	90–91	EU385957, DQ787710
		36[d]	35	TTGGCGGTGCGTCTTAGA	1	90	EU385957

[a] Phylum to which the putative novel class in column 1 belongs.

[b] Sanger-generated sequences obtained. No., number of sequences obtained for each lineage, % similarity, range of percent sequence similarities to GenBank first hits for the sequences affiliated with each group. GenBank bank accession numbers of those first hits are shown.

[c] Forward primer used in the PCR. Position corresponds to numbering of the *E. coli* 16S rRNA gene.

[d] Primer may not be specific. Some of the Sanger-generated sequences in this group did not belong to that novel class.

[e] Pyrosequencing lineage numbers refer to the number given to each of the 78 pyrosequencing lineages investigated for novelty in this study as described in detail in Table S1 in the supplemental material.

dec arc (GenBank accession number HQ153894) (S. Murdock, H. Johnson, N. Forget, and K. S. Juniper, presented at the 4th International Symposium on Chemosynthesis-Based Ecosystems—Hydrothermal Vents, Seeps and Other Reducing Habitats, Roscoff, France, 2010), sequences from subseafloor sediment of the South China sea (GenBank accession number EU385810), and a clone with GenBank accession number AM490651 from the microbial mats in the Lower Kane Cave (20).

Sequences belonging to novel phylum ZDP4 were most closely similar (81% to 84%, depending on the sequence compared) to those of uncultured microorganisms from multiple habitats. The first 100 GenBank hits were all within the same range of percent similarity to ZDP4 sequences (79% to 84%) and corresponded to clones from 10 different studies. Examples include a clone from a biogas fermentation enrichment culture (GenBank accession number GU476604), a clone from a mesophilic biogas digester treating pig manure (GenBank accession number EU358689) (50), and a clone from water samples of Lake Cadagno, an alpine meromictic lake located in the Piora valley in the Southern Alps of Switzerland (GenBank accession number FJ502268) (31).

Finally, sequences belonging to novel phylum ZDP5 were most closely similar (77% to 84%) to a number of sequences retrieved from a few environmental surveys. These include clones from a mesophilic anaerobic reactor fed with effluent from the chemical industry in Mexico City, Mexico (GenBank accession numbers FJ462088, FJ462086, FJ462085, and FJ462087) and a clone from Fe-Mn nodules and surrounding soil in Wuhan, Hubei Province, Central China (GenBank accession number EF492917) (33).

As a further proof of their presence in Zodletone Spring sediments, we quantified members belonging to the above 5 proposed novel phyla by the use of qPCR. Results of qPCR quantification are shown in Table 1. Members belonging to novel phyla ZDP1 to ZDP5 represent 0.002% to 0.6% of the 16S rRNA genes identified by qPCR using universal primers.

Novel classes. In addition to the novel phyla identified, multiple targeted lineages formed unique monophyletic clusters within existing phyla and candidate phyla. Ten novel classes are proposed from 13 different pyrosequencing lineages (Table 2). Of the 55 sequences representing those novel classes, 17 were putative nontarget sequences with >15% sequence divergence from their corresponding short sequences. The intralineage diversity within each novel class ranged from 4% to 9%. A distance neighbor-joining tree of novel classes is presented in Fig. 3 (parsimony, maximum-likelihood, and Bayesian trees are shown in Fig. S2 in the supplemental material). As shown, novel classes in 2 previously recognized phyla (*Planctomycetes* and *Fibrobacteres*) and 7 previously recognized candidate phyla (BRC1, GN06, GN12, LD1, TM6, WS2, and TM7) were identified.

Planctomycetes class ZDC1 formed a monophyletic lineage independent of the 10 recognized classes within the phylum according to the Greengenes taxonomy (classes *Kuenenia*, *Phycisphaerae*, *Planctomycea*, C6, FFCH393, Koll-18, ODP123, PW285, agg27, and VadinHA49). *Fibrobacteres* class ZDC2 formed a monophyletic cluster independent of the 2 recognized classes within the phylum *Fibrobacteres* (*Fibrobacteres* and *Fibrobacteres* 2). Similar patterns of forming novel class-level independent lineages with low sequence similarity to closest relatives were observed with classes ZDC3, ZDC4, ZDC5, ZDC6, ZDC7, ZDC8, ZDC9, and ZDC10 within the candidate phyla BRC1, GN12, TM6, LD1, TM7, GN06 (for two classes), and WS2 respectively.

Novel orders. We identified 19 novel orders belonging to 21 lineages within previously described classes in 4 phyla (*Planctomyce-*

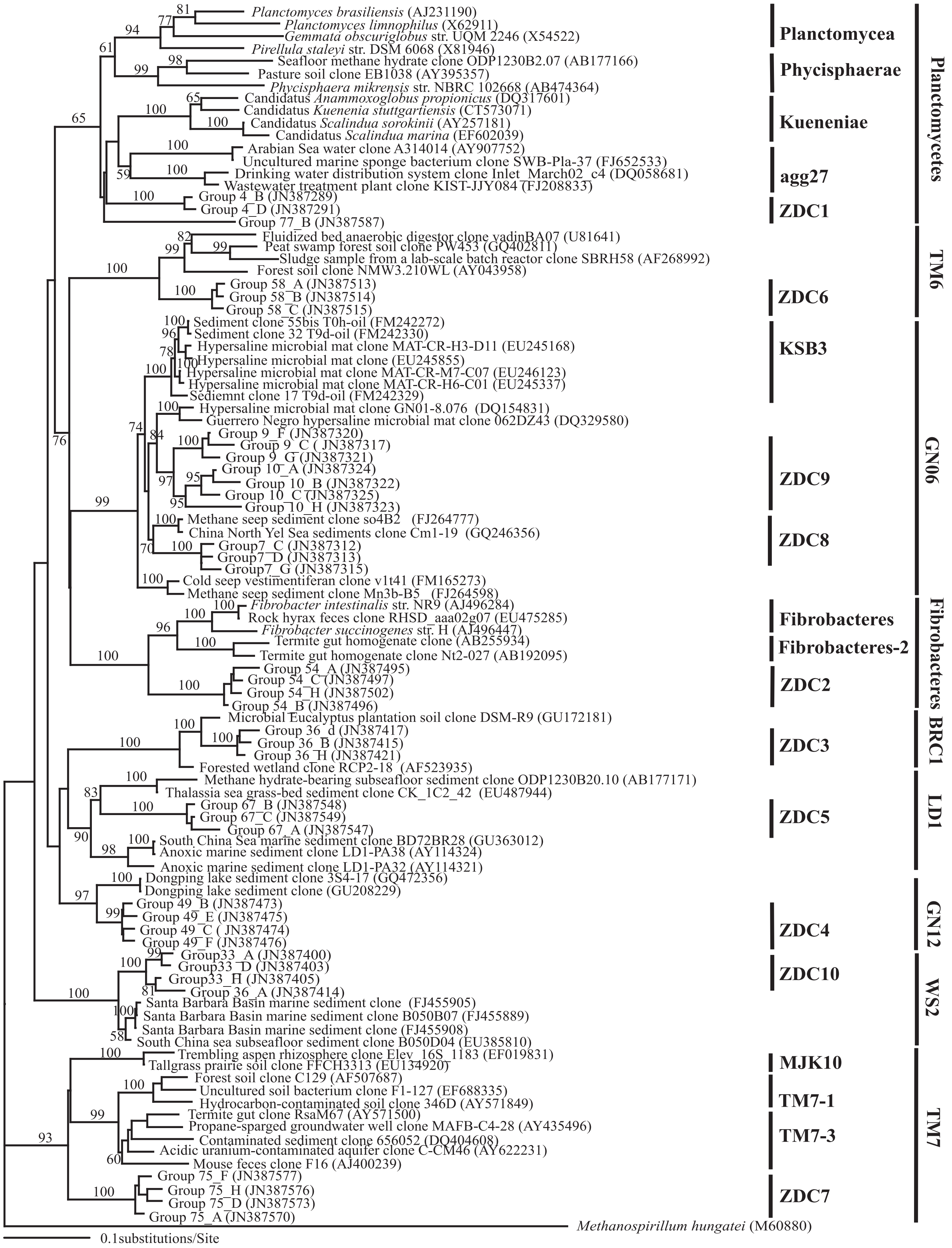
Planctomyces brasiliensis (AJ231190)
Planctomyces limnophilus (X62911)
Gemmata obscuriglobus str. UQM 2246 (X54522)
Pirellula staleyi str. DSM 6068 (X81946)
Seafloor methane hydrate clone ODP1230B2.07 (AB177166)
Pasture soil clone EB1038 (AY395357)
Phycisphaera mikrensis str. NBRC 102668 (AB474364)
Candidatus Anammoxoglobus propionicus (DQ317601)
Candidatus Kuenenia stuttgartiensis (CT573071)
Candidatus Scalindua sorokinii (AY257181)
Candidatus Scalindua marina (EF602039)
Arabian Sea water clone A314014 (AY907752)
Uncultured marine sponge bacterium clone SWB-Pla-37 (FJ652533)
Drinking water distribution system clone Inlet_March02_c4 (DQ058681)
Wastewater treatment plant clone KIST-JJY084 (FJ208833)
Group 4_B (JN387289)
Group 4_D (JN387291)
Group 77_B (JN387587)
Fluidized bed anaerobic digestor clone vadinBA07 (U81641)
Peat swamp forest soil clone PW453 (GQ402811)
Sludge sample from a lab-scale batch reactor clone SBRH58 (AF268992)
Forest soil clone NMW3.210WL (AY043958)
Group 58_A (JN387513)
Group 58_B (JN387514)
Group 58_C (JN387515)
Sediment clone 55bis T0h-oil (FM242272)
Sediment clone 32 T9d-oil (FM242330)
Hypersaline microbial mat clone MAT-CR-H3-D11 (EU245168)
Hypersaline microbial mat clone (EU245855)
Hypersaline microbial mat clone MAT-CR-M7-C07 (EU246123)
Hypersaline microbial mat clone MAT-CR-H6-C01 (EU245337)
Sediemnt clone 17 T9d-oil (FM242329)
Hypersaline microbial mat clone GN01-8.076 (DQ154831)
Guerrero Negro hypersaline microbial mat clone 062DZ43 (DQ329580)
Group 9_F (JN387320)
Group 9_C (JN387317)
Group 9_G (JN387321)
Group 10_A (JN387324)
Group 10_B (JN387322)
Group 10_C (JN387325)
Group 10_H (JN387323)
Methane seep sediment clone so4B2 (FJ264777)
China North Yel Sea sediments clone Cm1-19 (GQ246356)
Group7_C (JN387312)
Group7_D (JN387313)
Group7_G (JN387315)
Cold seep vestimentiferan clone v1t41 (FM165273)
Methane seep sediment clone Mn3b-B5 (FJ264598)
Fibrobacter intestinalis str. NR9 (AJ496284)
Rock hyrax feces clone RHSD_aaa02g07 (EU475285)
Fibrobacter succinogenes str. H (AJ496447)
Termite gut homogenate clone (AB255934)
Termite gut homogenate clone Nt2-027 (AB192095)
Group 54_A (JN387495)
Group 54_C (JN387497)
Group 54_H (JN387502)
Group 54_B (JN387496)
Microbial Eucalyptus plantation soil clone DSM-R9 (GU172181)
Group 36_d (JN387417)
Group 36_B (JN387415)
Group 36_H (JN387421)
Forested wetland clone RCP2-18 (AF523935)
Methane hydrate-bearing subseafloor sediment clone ODP1230B20.10 (AB177171)
Thalassia sea grass-bed sediment clone CK_1C2_42 (EU487944)
Group 67_B (JN387548)
Group 67_C (JN387549)
Group 67_A (JN387547)
South China Sea marine sediment clone BD72BR28 (GU363012)
Anoxic marine sediment clone LD1-PA38 (AY114324)
Anoxic marine sediment clone LD1-PA32 (AY114321)
Dongping lake sediment clone 3S4-17 (GQ472356)
Dongping lake sediment clone (GU208229)
Group 49_B (JN387473)
Group 49_E (JN387475)
Group 49_C (JN387474)
Group 49_F (JN387476)
Group33_A (JN387400)
Group33_D (JN387403)
Group33_H (JN387405)
Group 36_A (JN387414)
Santa Barbara Basin marine sediment clone (FJ455905)
Santa Barbara Basin marine sediment clone B050B07 (FJ455889)
Santa Barbara Basin marine sediment clone (FJ455908)
South China sea subseafloor sediment clone B050D04 (EU385810)
Trembling aspen rhizosphere clone Elev_16S_1183 (EF019831)
Tallgrass prairie soil clone FFCH3313 (EU134920)
Forest soil clone C129 (AF507687)
Uncultured soil bacterium clone F1-127 (EF688335)
Hydrocarbon-contaminated soil clone 346D (AY571849)
Termite gut clone RsaM67 (AY571500)
Propane-sparged groundwater well clone MAFB-C4-28 (AY435496)
Contaminated sediment clone 656052 (DQ404608)
Acidic uranium-contaminated aquifer clone C-CM46 (AY622231)
Mouse feces clone F16 (AJ400239)
Group 75_F (JN387577)
Group 75_H (JN387576)
Group 75_D (JN387573)
Group 75_A (JN387570)
Methanospirillum hungatei (M60880)
0.1substitutions/Site
Planctomycea
Phycisphaerae
Kueneniae
agg27
ZDC1
ZDC6
KSB3
ZDC9
ZDC8
Fibrobacteres
Fibrobacteres-2
ZDC2
ZDC3
ZDC5
ZDC4
ZDC10
MJK10
TM7-1
TM7-3
ZDC7
Planctomycetes
TM6
GN06
Fibrobacteres
BRC1
LD1
GN12
WS2
TM7

FIG 3 Distance dendrogram based on the 16S rRNA Sanger-generated sequences affiliated with novel classes (ZDC1 to ZDC10) encountered in Zodletone Spring source sediment clone libraries. The tree was obtained using a Tamura-Nei substitution model with a proportion of invariable sites = 0.0908 and a variable site gamma distribution shape parameter = 0.7571. Bootstrap values (in percentages) are based on 1,000 replicates and are shown for branches with more than 50% bootstrap support.

TABLE 3 Group name, primer sequence, and percent similarity and accession numbers of GenBank first hits for the Sanger-generated sequences belonging to putative novel orders identified in this study

			Pyrosequencing lineage			Sanger-generated sequences[c]		
				Forward primer[d]			Closest relative in GenBank	
Novel order name	Phylum[a]	Class[b]	No.[f]	Position	Sequence	No.	% similarity	GenBank accession no.
ZDO1	*Chloroflexi*	*Chloroflexi*	39[e]	183	CCACAGAGTCTTCGGGCT	2	91	AB473922
ZDO2		*Anaerolineae*	13[e]	120	CACGTGGGTCATTTGCCC	1	86	DQ811881
ZDO3	*Proteobacteria*	Delta	33[e]	162	GCGCCGCTAATACCGGGT	2	84–85	AB630482
ZDO4		Delta	56[e]	36	GGGCCGGGCGTGCTTAAC	1	81	HQ397064
ZDO5		Delta	67[e]	125	GGGTACTTGCCCTCGACT	1	93	AY499723
ZDO6	*Verrucomicrobia*	R76–B18	27	274	GGCTTACGGGTAGTTGGT	5	91–93	EU287160, FJ416080
ZDO7	*Planctomycetes*	*Phycisphaerae*	3	129	AACGTACCCATAGCACGG	5	83–88	FJ710716
ZDO8		*Phycisphaerae*	61[e]	154	GTCCCGAAAGGGGCGGTA	3	90–93	HQ405605, HQ183982, EF125448
ZDO9		VadinHA49	4[e]	198	GGATTTTCGGACCTTCTG	4	85–89	AB231079, AB522154, CU919930, FJ374260
ZDO10		VadinHA49	5	197	GGTTGGGATCTATGGACC	7	84–86	AF149887, AB234542
ZDO11	TG3	TG3–1	42[e]	230	AGCTTGCGGGCCCATTAG	3	86–90	GU476604
			44[e]	183	GATCGGCTTGGCGCATGT	3	87–89	GQ458248
			45	66	CGTGAAGCTCAGGCAACT	7	88	GQ458248, GU476604
ZDO12	WS6	SC72	6[e]	230	GGATTGCGTCCTATCAGC	1	84	GU363049
ZDO13	OP8	OP8–2	63	209	ATTAGGATCAAAGGGGGG	2	87	EF688191
ZDO14	OP3	BD4–9	6[e]	230	GGATTGCGTCCTATCAGC	1	89	AY792312
ZDO15		PBS–25	13[e]	120	CACGTGGGTCATTTGCCC	1	90	AJ390463
ZDO16		Koll–11	6[e]	230	GGATTGCGTCCTATCAGC	1	86	AM991240
ZDO17	OP11	OP11–2	78[e]	85	GTATCGCGCTCTTAGCGG	3	85–87	AY218580, AF419686
ZDO18		WCHB1–64	6[e]	230	GGATTGCGTCCTATCAGC	1	80	AB510995
ZDO19	GN06	KSB3	60[e]	232	CCCGCGGACTATTAGTTA	1	89	FM242438

[a] Phylum to which the putative novel order in column 1 belongs.

[b] Class to which the putative novel order in column 1 belongs.

[c] Sanger-generated sequences obtained. No., number of sequences obtained for each lineage, % similarity, range of percent sequence similarities to GenBank first hits for the sequences affiliated with each group. GenBank bank accession numbers of those first hits are shown.

[d] Forward primer used in the PCR. Position corresponds to numbering of the *E. coli* 16S rRNA gene.

[e] Primer may not be specific. Some of the Sanger-generated sequences in this group did not belong to that novel class.

[f] Pyrosequencing lineage numbers refer to the number given to each of the 78 pyrosequencing lineages investigated for novelty in this study as described in detail in Table S1, in the supplemental material.

tes, *Proteobacteria*, *Chloroflexi*, and *Verrucomicrobia*) and 6 candidate phyla (GN06, OP3, OP8, OP11, TG3, and WS6) according to the criteria described above. Details are shown in Table 3. Nine of 21 lineages were target sequences. Novel *Planctomycetes* orders were within classes *Phycisphaerae* (ZDO1, ZDO2) and VadinHA49 (ZDO3, ZDO4); novel *Chloroflexi* orders were within classes *Chloroflexi* (ZDO5) and *Anaerolineae* (ZDO6); novel *Proteobacteria* orders were within class *Deltaproteobacteria* (ZDO7, ZDO8, ZDO9); and novel *Verrucomicrobia* orders were within classes R76 to B18 (ZDO10). Similarly, patterns of forming novel order-level independent lineages with low sequence similarity to closest relatives were observed with orders ZDO11 to ZDO19 within previously described classes of candidate phyla TG3, WS6, OP8, OP3 (3 novel orders), OP11 (2 novel orders), and GNO6, respectively.

Other sequences. The remaining sequences (n = 154) belonged to 19 distinct lineages (8 of the 19 lineages were target sequences) and formed 18 distinct families, 14 distinct genera, and 20 distinct species. Details are shown in Table S6 in the supplemental material.

Correlation between value of sequence similarity to closest relative of a pyrosequencing read and its "true" phylogenetic affiliation. Based on the results obtained, we sought to determine whether the percent similarity to closest relatives obtained for a specific pyrosequencing-generated read could accurately predict its putative phylogenetic novelty, i.e., whether such sequences truly represent a novel phylogenetic lineage when a nearly full-length sequence is obtained and analyzed. Such quantification could potentially contribute to better judgment in identification of putatively novel lineages within pyrosequencing data sets in similar future studies.

For each of the 32 lineages whose Sanger-generated sequences were indeed representative of their corresponding pyrosequencing reads, the original pyrosequencing reads used for the forward primer design were identified. These reads (1,566 sequences, 192 OTUs) were compared to the NCBI nr database using BLAST nr (42), and the percent sequence divergence from the nearest relative was determined. Within those lineages, 10 were proposed as members of novel phyla, 6 were proposed as members of novel classes, 9 were proposed as members of novel orders, and 6 were proposed as members of novel families, genera, or species. The correlation between the percent sequence similarity to closest relative and "true" phylogenetic affiliation obtained from examining the corresponding Sanger sequence was analyzed. Results (Table 4) indicate averages of 78.8%, 81.3%, and 86.3% for phyla, classes, and orders, respectively, with a low coefficient of variation (0% to 4.29%). Therefore, for an 8f-338R pyrosequencing-generated 16S

TABLE 4 Correlation between sequence similarity to closest relative value of a pyrosequencing read and its true phylogenetic affiliation

Taxonomic level[a]	Lineage (group)[b]	Average % similarity to closest database relative (CV)[c]	Average % similarity for that taxonomic level[d]
Phylum	19	81 (0)	78.8
	20	76 (one sequence)	
	24	82.5 (4.29)	
	52	78.9 (1.71)	
	53	77.5 (1.08)	
	65	77 (0)	
Class	4	81 (one sequence)	81.3
	7	84 (1.77)	
	9	80 (3.51)	
	10	80 (0)	
	33	80 (1.15)	
	36	80.5 (0.55)	
	49	86 (0)	
	54	79 (one sequence)	
	58	79 (0.65)	
	77	83 (0)	
Order	3	85 (0.51)	86.3
	6	88.7 (3.51)	
	13	80.4 (0.98)	
	27	85 (one sequence)	
	35	86 (one sequence)	
	42	87.3 (one sequence)	
	44	87.8 (one sequence)	
	45	89.3 (one sequence)	
	63	87.5 (3.4)	
Family (or lower)	23	85 (0)	88.3
	51	87 (one sequence)	
	59	90 (one sequence)	
	60	89 (one sequence)	
	62	90.5 (0.88)	
	76	83.9 (1.11)	
	78	93 (0.88)	

[a] Phylogenetic affiliation of only target Sanger-generated sequences belonging to lineages shown in column 2.

[b] Numbers refer to the pyrosequencing lineage numbers given to the 32 target lineages used for this correlation.

[c] Average percentage sequence similarity to the nearest GenBank relative for all pyrosequences belonging to the lineages shown in column 2. For each of the lineages shown, all pyrosequences were identified and compared to NCBI nr database using BLAST, and the average percent sequence similarity to the first hit (closest relative) was calculated. Numbers in parentheses represent the coefficients of variation of percent sequence similarities to closest relatives of all pyrosequences belonging to that lineage calculated as $CV = \frac{SD}{mean} \times 100$, where CV is the coefficient of variation, SD is the standard deviation, and mean is average percent similarity.

[d] For each taxonomic level, the number corresponds to the average of all percent similarities shown in column 3 for that taxonomic level.

rRNA data set, sequences with values around these thresholds should be used in similar future protocols to maximize the potential for identifying novel phyla, classes, and orders.

DISCUSSION

In this study, a direct, straightforward, and fairly cost-effective approach to assessment of the phylogenetic affiliations of multiple unclassified reads generated during 16S rRNA gene pyrosequencing surveys of specific ecosystems is presented. The combined Sanger-pyrosequencing approach proposed in this study produces high-quality sequences with read lengths of 1,100 to 1,500 bp, values that exceed the newest GS-FLX+ Titanium chemistry pyrosequencing read lengths (average read length, 700 bp [September 2011]). However, unlike the sequences with shorter read lengths generated by next-generation sequencing technologies, the Sanger-generated sequences would be available in curated databases for future reference and comparison, and sequencing error rates for Sanger-generated sequences (22, 35) are considerably lower than the sequencing error rates generated by pyrosequencing (35). Using this combined approach, we show that it is possible to retrieve nearly full-length sequences of targeted pyrosequencing lineages by pairing a lineage-specific primer to a universal reverse primer. We also show that, while the majority of examined pyrosequencing reads that are unclassified within our data set could be accurately assigned to known classes, orders, and families, a fraction of the unclassified sequences indeed represents previously unencountered novel phyla (ZDP1 to ZDP5) and novel classes (ZDC1 to ZDC10) within the domain *Bacteria*.

The rise and currently nearly exclusive utilization of high-throughput sequencing technologies is understandable due to the cost advantage and versatility compared to Sanger sequencing. Use of those technologies has resulted in valuable ecological insights, e.g., the more accurate estimates of species richness and evenness (59, 60, 63, 81) and the more detailed comparisons between communities at various temporal (17, 44, 46) and spatial (6, 17, 24, 49, 75) gradients. Unfortunately, the rise of such sequencing technologies has resulted in overlooking detailed phylogenetic analysis due to the short amplicon size and the sheer number of clone sequences analyzed. The approach outlined in this study thus offers a valid alternative for exploring the phylogenetic diversity of putatively novel members of the rare biosphere. To our knowledge, only one previous study has utilized a similar approach, on a much narrower scale, to explore the phylogenetic affiliation of selected marine sponge-affiliated tag sequences with <75% similarity to 16S rRNA sequences in the curated SILVA database (74). However, the study yielded only 4 Sanger sequences and provided no detailed methodological information regarding primer design, success rates in obtaining nearly full-length sequences representative of targeted lineages, and proportions of novel high-rank diversity (e.g., candidate phyla, classes, and orders) obtained as a fraction of the total number of lineages examined. Further, the study paired a 616F general bacterial primer with V6 lineage-specific reverse primers, resulting in relatively short sequences (approximately 900 bp) that would not be deposited and classified in curated databases, e.g., Greengenes and SILVA. The current report, on the other hand, represents an ecosystem-wide evaluation of the phylogenetic affiliation of a large number of unclassified lineages from a pyrosequencing data set. The report also provides detailed information on (i) criteria for the selection of pyrosequencing lineages for this type of analysis; (ii) the utility and value of utilizing various universal reverse primers in obtaining nearly full-length sequences of targeted lineages; (iii) criteria for lineage-specific primer design and evaluation; (iv) details on the success rate of such an approach; (v) evidence demonstrating the importance of sequencing multiple clones per targeted lineage; (vi) a detailed methodology regarding the detection and quantification of nontarget versus target amplification of nearly full-length 16S rRNA gene sequences; and (vii) detailed information regarding the correlation between the value of sequence similarity to the closest relative of a pyrosequencing read and its "true" phylogenetic affiliation.

The identification of five novel candidate phyla in a single study is a significant expansion of the scope of diversity within the domain bacteria. Currently, 84 phyla and candidate phyla have been described (according to the latest update of Greengenes taxonomy [July 2011]) [14]). While earlier diversity surveys (see, e.g., references 15 and 38) had often been successful in identifying multiple novel phyla within a single study using a relatively small number of sequences, the pace of discovery of novel phyla has subsequently slowed due to the wide utilization of 16S rRNA surveys in examining diversity in almost all accessible habitats on earth. The apparent saturation of phylum-level diversity, however, was challenged by the discovery of novel, putatively distinct lineages of rare members of microbial communities in recent high-throughput diversity surveys. This report demonstrates that a fraction of the unclassifiable members of the spring community that have not been accessed in Sanger surveys due to low abundance indeed belongs to novel microbial phyla and that a targeted approach that combines the high-throughput capabilities of pyrosequencing and the read length of Sanger sequencing can identify and accurately describe such lineages. Therefore, we reason that a similar approach, conducted on multiple ecosystems, and at more depth (more lineages per sample), could indeed greatly expand the scope of understanding of bacterial (and putatively archaeal) phylogenetic diversity on earth. The reason for the presence and maintenance of these lineages at low abundance, as well as their global distribution patterns and putative ecological roles (or lack thereof), would be an area of interest to microbial ecologists and evolutionary microbiologists.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation Microbial Observatories Program (grant EF0801858).

REFERENCES

1. **Ahn J, et al.** 2011. Oral microbiome profiles: 16S rRNA pyrosequencing and microarray assay comparison. PLoS One **6**:e22788.
2. **Andersson AF, Riemann L, Bertilsson S.** 2009. Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. ISME J. **4**:171–181.
3. **Ashelford KE, Weightman AJ, Fry JC.** 2002. PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. Nucleic Acids Res. **30**:3481–3489.
4. **Bibby K, Viau E, Peccia J.** 2010. Pyrosequencing of the 16S rRNA gene to reveal bacterial pathogen diversity in biosolids. Water Res. **44**:4252–4260.
5. **Brazelton WJ, et al.** 2010. Archaea and bacteria with surprising microdiversity show shifts in dominance over 1,000-year time scales in hydrothermal chimneys. Proc. Natl. Acad. Sci. U. S. A. **107**:1612–1617.
6. **Brown MV, et al.** 2009. Microbial community structure in the North Pacific ocean. ISME J. **3**:1374–1386.
7. **Callbeck C, et al.** 2011. Microbial community succession in a bioreactor modeling a souring low-temperature oil reservoir subjected to nitrate injection. Appl. Microbiol. Biotechnol. **91**:799–810.
8. **Caporaso JG, et al.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods **7**:335–336.
9. **Caporaso JG, et al.** 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A. **108**: 4516–4522.
10. **Claesson MJ, et al.** 2011. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. Proc. Natl. Acad. Sci. U. S. A. **108**:4586–4591.
11. **Cole JR, et al.** 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. **37**:D141–D145.
12. **Dalevi D, Hugenholtz P, Blackall LL.** 2001. A multiple-outgroup approach to resolving division-level phylogenetic relationships using 16S rDNA data. Int. J. Syst. Evol. Microbiol. **51**:385–391.
13. **DeSantis T, et al.** 2007. High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microb. Ecol. **53**:371–383.
14. **DeSantis TZ, et al.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. **72**:5069–5072.
15. **Dojka MA, Hugenholtz P, Haack SK, Pace NR.** 1998. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. Appl. Environ. Microbiol. **64**:3869–3877.
16. **Donachie SP, et al.** 2004. The Hawaiian archipelago: a microbial diversity hotspot. Microb. Ecol. **48**:509–520.
17. **Dumbrell AJ, et al.** 2011. Distinct seasonal assemblages of arbuscular mycorrhizal fungi revealed by massively parallel pyrosequencing. New Phytol. **190**:794–804.
18. **Elshahed MS, et al.** 2003. Bacterial diversity and sulfur cycling in a mesophilic sulfide-rich spring. Appl. Environ. Microbiol. **69**:5609–5621.
19. **Elshahed MS, et al.** 2008. Novelty and uniqueness patterns of rare members of the soil biosphere. Appl. Environ. Microbiol. **74**:5422–5428.
20. **Engel AS, et al.** 2009. Linking phylogenetic and functional diversity to nutrient spiraling in microbial mats from Lower Kane Cave (U. S. A.). ISME J. **4**:98–110.
21. **Engelbrektson A, et al.** 2010. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J. **4**:642–647.
22. **Ewing B, Hillier L, Wendl MC, Green P.** 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. **8**:175–185.
23. **Fulthorpe RR, Roesch LFW, Riva A, Triplett EW.** 2008. Distantly sampled soils carry few species in common. ISME J. **2**:901–910.
24. **Galand PE, Casamayor EO, Kirchman DL, Lovejoy C.** 2009. Ecology of the rare microbial biosphere of the Arctic Ocean. Proc. Natl. Acad. Sci. U. S. A. **106**:22427–22432.
25. **Gilbert J, et al.** 2010. Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project. Stand. Genomic Sci. **3**:243–248.
26. **Gilbert JA, et al.** 2009. The seasonal structure of microbial communities in the western English Channel. Environ. Microbiol. **11**:3132–3139.
27. **Gilbert JA, et al.** 2011. Defining seasonal marine microbial community dynamics. ISME J. **6**:298–308.
28. **Giongo A, et al.** 2010. PANGEA: pipeline for analysis of next generation amplicons. ISME J. **4**:852–861.
29. **Gutell RR, Weiser B, Woese CR, Noller HF.** 1985. Comparative anatomy of 16-S-like ribosomal RNA. Nucleic acid res. **32**:155–216.
30. **Haas BJ, et al.** 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. **21**: 494–504.
31. **Halm H, et al.** 2009. Co-occurrence of denitrification and nitrogen fixation in a meromictic lake, Lake Cadagno (Switzerland). Environ. Microbiol. **11**:1945–1958.
32. **Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R.** 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods **5**:235–237.
33. **He J, Zhang L, Jin S, Zhu Y, Liu F.** 2008. Bacterial communities inside and surrounding soil iron-manganese nodules. Geomicrobiol. J. **25**: 14–24.
34. **Heijs SK, Laverman AM, Forney LJ, Hardoim PR, Van Elsas JD.** 2008. Comparison of deep-sea sediment microbial communities in the Eastern Mediterranean. FEMS Microbiol. Ecol. **64**:362–377.
35. **Hoff KJ.** 2009. The effect of sequencing errors on metagenomic gene prediction. BMC Genomics **10**:520–528.
36. **Hollister EB, et al.** 2010. Shifts in microbial community structure along an ecological gradient of hypersaline soils and sediments. ISME J. **4**:829–838.
37. **Huelsenbeck JP, Ronquist F.** 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754–755.
38. **Hugenholtz P, Pitulle C, Hershberger KL, Pace NR.** 1998. Novel division level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. **180**:366–376.
39. **Huse SM, Welch DBM.** 2011. Accuracy and quality of massively parallel

DNA pyrosequencing, p 149–155. *In* F. J. de Bruijn (ed), Handbook of molecular microbial ecology I: metagenomics and complementary approaches. John Wiley & Sons, Inc., Hoboken, NJ.

40. **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.
41. **Ishii S, et al.** 2009. Microbial populations responsive to denitrification-inducing conditions in rice paddy soil, as revealed by comparative 16S rRNA gene analysis. Appl. Environ. Microbiol. **75**:7070–7078.
42. **Johnson M, et al.** 2008. NCBI BLAST: a better web interface. Nucleic Acids Res. **36**:W5–W9.
43. **Jones RT, et al.** 2009. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. ISME J. **3**:442–453.
44. **Kataoka T, Hodoki Y, Suzuki K, Saito H, Higashi S.** 2009. Tempospatial patterns of bacterial community composition in the western North Pacific Ocean. J. Marine Syst. **77**:197–207.
45. **Kim Y-S, et al.** 2011. Analyses of bacterial communities in meju, a Korean traditional fermented soybean bricks, by cultivation-based and pyrosequencing methods. J. Microbiol. **49**:340–348.
46. **Kirchman DL, Cottrell MT, Lovejoy C.** 2010. The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. Environ. Microbiol. **12**:1132–1143.
47. **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.
48. **Larkin MA, et al.** 2007. Clustal W and Clustal X version 2.0. Bioinformatics **23**:2947–2948.
49. **Lauber CL, Hamady M, Knight R, Fierer N.** 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl. Environ. Microbiol. **75**:5111–5120.
50. **Liu FH, et al.** 2009. The structure of the bacterial and archaeal community in a biogas digester as revealed by denaturing gradient gel electrophoresis and 16S rDNA sequencing analysis. J. Appl. Microbiol. **106**:952–966.
51. **Liu Z, DeSantis TZ, Andersen GL, Knight R.** 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. **36**:e120.
52. **Ludwig W, et al.** 2004. ARB: a software environment for sequence data. Nucleic Acids Res. **32**:1363–1371.
53. **Manter D, Delgado J, Holm D, Stong R.** 2010. Pyrosequencing reveals a highly diverse and cultivar-specific bacterial endophyte community in potato roots. Microb. Ecol. **60**:157–166.
54. **Mobberley JM, Ortega MC, Foster JS.** 2011. Comparative microbial diversity analyses of modern marine thrombolitic mats by barcoded pyrosequencing. Environ. Microbiol. **14**:82–100. doi:10.1111/j.1462-2920.2011.02509.x.
55. **Monchy S, et al.** 2011. Exploring and quantifying fungal diversity in freshwater lake ecosystems using rDNA cloning/sequencing and SSU tag pyrosequencing. Environ. Microbiol. **13**:1433–1453.
56. **Noller HF.** 1984. Structure of ribosomal RNA. Annu. Rev. Biochem. **53**:119–162.
57. **Noller HF, Woese CR.** 1981. Secondary structure of 16S ribosomal RNA. Science **212**:403–411.
58. **Posada D.** 2006. ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. Nucleic Acids Res. **34**:W700–W703.
59. **Quince C, Curtis TP, Sloan WT.** 2008. The rational exploration of microbial diversity. ISME J. **2**:997–1006.
60. **Roesch LFW, et al.** 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. **1**:283–290.
61. **Rousk J, et al.** 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J. **4**:1340–1351.
62. **Schloss PD, Gevers D, Westcott SL.** 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One **6**:e27310.
63. **Schloss PD, Handelsman J.** 2006. Toward a census of bacteria in soil. PLoS Comput. Biol. **2**:786–793.
64. **Schloss PD, Westcott SL.** 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl. Environ. Microbiol. **77**:3219–3226.
65. **Schloss PD, et al.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. **75**:7537–7541.
66. **Schütte UME, et al.** 2010. Bacterial diversity in a glacier foreland of the high Arctic. Mol. Ecol. **19**:54–66.
67. **Senko JM, et al.** 2004. Barite deposition mediated by photootrophic sulfide-oxidizing bacteria. Geochim. Cosmochim. Acta **68**:773–780.
68. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688–2690.
69. **Stern S, Powers T, Changchien L-M, Noller HF.** 1989. RNA-protein interactions in 30S ribosomal subunits: folding and function of 16S rRNA. Science **244**:783–790.
70. **Teske A, et al.** 2002. Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities. Appl. Environ. Microbiol. **68**:1994–2007.
71. **Tringe SG, Hugenholtz P.** 2008. A renaissance for the pioneering 16S rRNA gene. Curr. Opin. Microbiol. **11**:442–446.
72. **Uroz S, Buée M, Murat C, Frey-Klett P, Martin F.** 2010. Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. Environ. Microbiol. Rep. **2**:281–288.
73. **Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ.** 2009. Jalview version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics **25**:1189–1191.
74. **Webster NS, et al.** 2010. Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. Environ. Microbiol. **12**:2070–2082.
75. **Will C, et al.** 2010. Horizon-specific bacterial community composition of German grassland soils, as revealed by pyrosequencing-based analysis of 16S rRNA genes. Appl. Environ. Microbiol. **76**:6751–6759.
76. **Wimberly BT, et al.** 2000. Structure of the 30S ribosomal subunit. Nature **407**:327–339.
77. **Woese Gutell RR, Gupta R, Noller HF.** 1983. Detailed analysis of the higher-order structure of the 16{S}-like ribosomal ribonucleic acids. Microbiol. Rev. **47**:621–669.
78. **Ye L, Shao M-F, Zhang T, Tong AHY, Lok S.** 2011. Analysis of the bacterial community in a laboratory-scale nitrification reactor and a wastewater treatment plant by 454-pyrosequencing. Water Res. **45**:4390–4398.
79. **Youssef N, et al.** 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.
80. **Youssef NH, Couger MB, Elshahed MS.** 2010. Fine-scale bacterial beta diversity within a complex ecosystem (Zodletone Spring, OK, U. S. A.): the role of the rare biosphere. PLoS One **5**:e12414.
81. **Youssef NH, Elshahed MS.** 2008. Species richness in soil bacterial communities: a proposed approach to overcome sample size bias. J. Microbiol. Methods **75**:86–91.
82. **Zuker M.** 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31**:3406–3415.