APPLIED AND ENVIRONMENTAL MICROBIOLOGY, July 2006, p. 5069–5072 Vol. 72, No. 7
0099-2240/06/$08.00+0 doi:10.1128/AEM.03006-05

Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB

T. Z. DeSantis,[1] P. Hugenholtz,[2] N. Larsen,[3] M. Rojas,[4] E. L. Brodie,[1] K. Keller,[5] T. Huber,[6] D. Dalevi,[7] P. Hu,[1] and G. L. Andersen[1]*

Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 70A-3317, Berkeley, California 94720[1]; Microbial Ecology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Bldg. 400-404, Walnut Creek, California 94598[2]; Danish Genome Institute, Gustav Wieds vej 10 C, DK-8000 Aarhus C, Denmark[3]; Department of Bioinformatics, Baylor University, P.O. Box 97356, 1311 S. 5th St., Waco, Texas 76798-7356[4]; Department of Bioengineering, University of California, Berkeley, California 94720[5]; Departments of Biochemistry and Mathematics, The University of Queensland, Brisbane, Queensland 4072, Australia[6]; and Department of Computer Science, Chalmers University of Technology, SE-412 96 Göteborg, Sweden[7]

Received 20 December 2005/Accepted 15 April 2006

A 16S rRNA gene database (http://greengenes.lbl.gov) addresses limitations of public repositories by providing chimera screening, standard alignment, and taxonomic classification using multiple published taxonomies. It was found that there is incongruent taxonomic nomenclature among curators even at the phylum level. Putative chimeras were identified in 3% of environmental sequences and in 0.2% of records derived from isolates. Environmental sequences were classified into 100 phylum-level lineages in the *Archaea* and *Bacteria*.

Comparative analysis of 16S small-subunit rRNA genes is commonly used to survey the constituents of microbial communities (4, 13, 23, 24), to infer bacterial and archaeal evolution (14, 19), and to design monitoring and analysis tools, such as microarrays (5, 10, 17, 20, 29, 30). Because the rate of production of 16S small-subunit rRNA gene sequence records for uncultured organisms now exceeds the rate of production for their cultured counterparts, taxonomic placement of sequences lags behind. In fact, 43% of full-length 16S small-subunit rRNA gene records in the GenBank database are amalgamated into the pseudodivisions "environmental samples" and "unclassified." Annotation styles are inconsistent, creating barriers for computational categorization of biological sources. Furthermore, since rRNA genes from environmental DNA are usually PCR amplified, it is suspected that many clandestine chimeric sequences are intercalated into the public databases. For a small sample of 1,399 sequence records from known phyla, it was estimated that 3% of the public data might contain chimeras (2). The effect of these poor-quality data, exacerbated by barriers in exchanging nomenclature, has led to several conflicting taxonomies. The probability of mistakenly adopting a chimeric sequence in a phylogenetic inference or as a reference for probe/primer design is increasing noticeably. Finally, ARB (21) database administration needs to be streamlined for workers who maintain 16S small-subunit rRNA gene collections on their local computers.

Greengenes addresses these concerns by providing four features: a standardized set of descriptive fields, taxonomic assignment, chimera screening, and ARB compatibility. Heuristics are used to consider the author's annotations and categorize each source as a named or unnamed isolate, an unnamed symbiont, or an uncultured organism. Other standard descriptors include sequence quality measurements, authors, and a "study_id" that links all the records associated with a project. Greengenes maintains a consistent multiple-sequence alignment (MSA) of both archaeal and bacterial 16S small-subunit rRNA genes to facilitate taxonomic placement. Taxonomy proposed by independent curators, including the NCBI, the Ribosomal Database Project (RDP) (Bergey's) (7), Wolfgang Ludwig (21), Phil Hugenholtz (16), and Norman Pace (23), is tracked to promote user awareness of several estimations of phylogenetic descent, allowing a balanced approach to node nomenclature when dendrograms are generated. Comprehensive chimera assessment is a distinguishing characteristic of the Greengenes data assembly process. Each sequence is scored for chimeric potential, a breakpoint is estimated, and parent sequences are identified. Furthermore, since biologists often collect and visualize 16S small-subunit rRNA gene relationships using the freely available ARB software, Greengenes simplifies the chore of keeping a research group's private ARB database current by providing standardized alignments and an import filter (greengenes.ift) that imports the alignment and other standardized fields from 16S small-subunit rRNA gene records vetted weekly from GenBank.

To illustrate the utility of the Greengenes data assembly process and to examine the validity of prokaryotic candidate phyla, we aligned and chimera checked more than 90,000 public 16S small-subunit rRNA gene sequences. Taxonomic classifications from the major curators were used when such classifications were available. Sequence data were imported from NCBI for complete or nearly complete gene sequences (length, >1,250 nucleotides) deposited as of 2 April 2006. Alignment of both archaeal and bacterial sequences was performed with the NAST aligner (8) against a "Core Set" of templates selected from a phylogenetically broad collection (16). The resulting MSA was formatted so that each sequence occupied a consistent 7,682 characters or 4,182 characters; the latter allowed

* Corresponding author. Mailing address: Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 70A-3317, Berkeley, CA 94720. Phone: (510) 495-2795. Fax: (510) 486-7152. E-mail: GLAndersen@lbl.gov.

compatibility with RDP v8.1 (22) alignments. Both these formats were concise enough for browsing in common MSA graphical interfaces, such as ClustalX (28), MEGA (18), and the platform-independent interface Jalview (6), as well as ARB. Other standard expansions, such as the >20,000-character Ludwig alignment, are alternate formats that will be available in future releases to give maximum flexibility to researchers.

For high-throughput chimera screening of the aligned sequences, the program Bellerophon (15) was used with two modifications. First, the algorithm was modified to reduce the number of potential parents considered in the partial trees, which allowed run time to scale linearly rather than logarithmically with the count of candidate sequences in a collection. Second, a new metric was implemented, which weighted the likelihood of a sequence being chimeric according to the similarity of the parent sequences. The more distantly related the parent sequences were to each other relative to their divergence from the candidate chimeric sequence, the greater the likelihood that the inferred chimera was real. This metric, called the divergence ratio, used the average sequence identity between the two fragments of the candidate and the corresponding parent sequences as the numerator and the sequence identity between the parent sequences as the denominator. All calculations were restricted to 1,287 conserved columns of aligned characters using a 300-bp window on either side of the most likely breakpoint. A divergence ratio of >1.1 and fragment-to-parent levels of similarity of >90% were required for classifying sequences as putatively chimeric.

Taxonomy was linked to each record by various methods. NCBI taxonomic nomenclature and RDP taxonomic nomenclature were extracted directly from the corresponding GenBank-formatted records. The Pace and Ludwig annotations were exported from curated ARB databases. The Hugenholtz taxonomy was also derived from a curated ARB database in which tree topologies had been verified using RAxML-VI (27) for maximum likelihood inference. The general time-reversible model of evolution was applied together with optimization of substitution rates and site-specific rates according to a gamma distribution. Different search algorithms were considered depending on the run time of the standard hill climb (SHC) search method. If the running time was less than 8 h, simulating annealing (SA) was processed with the default starting temperature and a termination time set at approximately 24 h. If simulating annealing was not used and SHC terminated within 24 h, SHC was used. Furthermore, rapid hill climb was used in all other cases when the running time was less than 24 h. If rapid hill climb did not terminate within the set limit, the number of taxa was reduced. After 100 bootstrap replications, a consensus tree was calculated using Consense (12) and imported into ARB. This database (greengenes.arb) is available for download through Greengenes and is updated periodically.

Of the 90,000 NCBI records analyzed, 54% were derived from uncultured organisms, the majority of which were deposited in the last 5 years (Fig. 1). Only three studies have submitted more than 1,000 full-length clones; however, we expect the number of large 16S small-subunit rRNA gene surveys to increase due to the availability and falling cost of high-throughput sequencing. Bellerophon detection of putative chimeras in

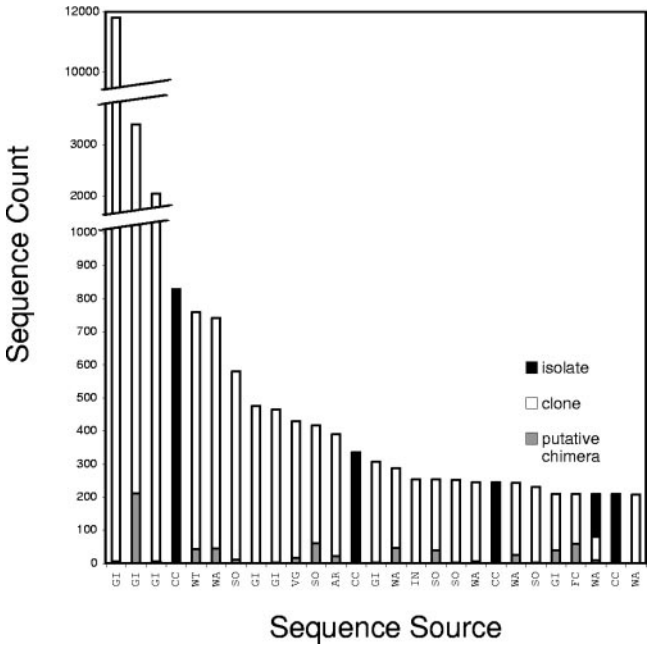

FIG. 1. 16S rRNA gene sequencing projects that produced more than 200 full-length records. All projects were submitted to GenBank between October 2000 and February 2006. Sequences were generated from gastrointestinal (GI), soil (SO), vaginal (VG), aerosol (AR), culture collection (CC), insect (IN), water (WA), waste treatment (WT), and fecal (FC) sources as indicated on the *x* axis. The projects are ordered by sequence count.

3% of the sequences from uncultured organisms was not unexpected considering the initial estimates (2). Surprisingly, 0.2% of sequences derived from pure cultures were also determined to be putative chimeras. Multiple distinct 16S rRNA genes have been encountered when clone libraries have been created from colonies assumed to be pure cultures prepared from numerous third-party sources (Colleen Cavanaugh, personal communication). It is possible that isolated colonies contain symbiotic bacteria which increase PCR template complexity, enabling chimera formation. In addition, thousands of full-length 16S small-subunit rRNA gene-annotated GenBank records were only partially aligned using NAST. Future versions of NAST could be altered to allow alignment extensions across regions having low template similarity or to allow candidates to be aligned in sections using divergent templates. Both of these options may allow a greater abundance of chimeric data to be imported into Greengenes but perhaps would capture novel phyla from the public repositories. Alternately, manually aligned sequences from novel phyla can be offered from the user community for recruitment to the Core Set advocating periodic reevaluation of the partially aligned set.

Discovery of chimeras in 16S small-subunit rRNA gene data collections is crucial if the data set is going to be a foundation for applied bioinformatics. Chimeras are a fundamental problem when they are used as templates with probe selection software, a growing concern with the recent increase in 16S small-subunit rRNA gene microarray probe development (3, 8, 11). The 15 to 30 bases surrounding the chimeric breakpoint can appear to be sufficiently different from all other records in

a database to cause a probe selection algorithm to justifiably identify the region as a target's signature and suggest complementary probes that can be synthesized. These probes could appear to be very valuable considering their minimal mishybridization potential, but in fact, they would rarely be useful since they target nonexistent organisms. Chimera test results from Greengenes allow greater control over input to probe selection software, should aid in avoiding artificial terminal restriction fragment length polymorphism pattern predictions from ARB-compatible TRF-CUT (25), and can increase the accuracy of sampling rarefaction curves (26).

The fraction of putative chimeras in the deposited sequences from an individual study varies from none to more than 20% (Fig. 1), suggesting that chimera screening is still not being uniformly applied by sequence generators. The problem is exacerbated with sparsely populated candidate phyla. For instance, the bacterial phyla "SAM" and "5" and the class GN4 (*Proteobacteria*) may require reevaluation. Likewise, the genera *Tistrella*, *Caldotoga*, *Dehalobacterium*, and *Desulfovermiculus* are currently anchored by sequences with evidence of chimeric composition. Additional sequences could lead to empirical rejection of certain classifications or may aid in defining the true breadth of sequence variation for these taxa.

Comparison of five different taxonomies uncovered surprisingly great disparity between expert curators. Loosely interpreting a "phylum" to be any labeled group or division immediately subordinate to the domain *Archaea* or *Bacteria*, we compared the five curations in a Venn diagram (Fig. 2). The main source of the disparity is the discordant naming of novel candidate phyla or the absence of names for candidate phyla. For example, Pace and Hugenholtz have independently named more than 12 phylum-level lineages, many of which are the same lineages, and RDP has not named any of these lineages. This is a consequence of the huge number of environmental sequences in the public databases and the frequent redundant naming of environmental lineages in the literature. We hope that making multiple taxonomic classifications available through Greengenes will aid in standardizing classification, particularly classification of environmental lineages.

Greengenes is also a functional workbench to assist in analysis of user-generated 16S rRNA gene sequences. Batches of sequencing reads can be uploaded for quality-based trimming and creation of multiple-sequence alignments (9). Three types of non-MSA similarity searches are also available, seed extension by BLAST (1), similarity based on shared 7-mers by a tool called "Simrank," and a direct degenerative pattern match for probe/primer evaluation. Results are displayed using user-preferred taxonomic nomenclature and can be saved between sessions.

In summary, Greengenes offers annotated, chimera-checked, full-length 16S rRNA gene sequences in standard alignment formats. The relational database links taxonomies from multiple curators and multiple sequences from a single study. We found that there is incongruent taxonomic nomenclature among curators even at the phylum level. Bellerophon found putative chimeras in sequences derived from both uncultured and isolated organisms. The data set can be compared to user-provided sequences via a web interface or can be imported directly into ARB for advanced analyses. We anticipate

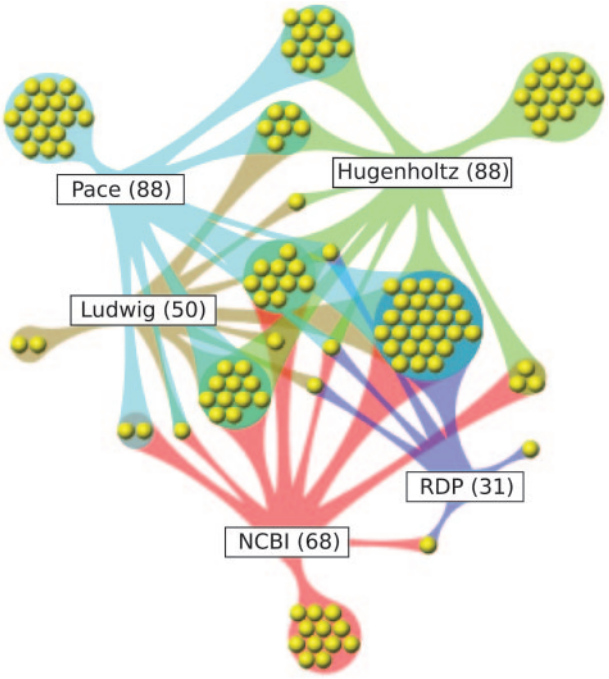

FIG. 2. Phylum-level nomenclature shared by independent curators represented as a five-way Venn diagram. Yellow spheres represent the 126 phylum or candidate division names encountered in at least one of the five taxonomy systems (Pace, Hugenholtz, Ludwig, RDP, or NCBI). The numbers in parentheses are the counts for phylum or candidate division names recognized by an individual curator. Clusters of yellow spheres connected by more than one colored web symbolize names recognized by multiple curators. The image was rendered by the AutoFocus software (Aduna B.V., The Netherlands). A complete table of phylum-level nomenclature comparisons is available at http://greengenes.lbl.gov/TaxCompare.

that Greengenes will be valuable to researchers conducting environmental surveys and for 16S rRNA microarray design.

In the immediate future, we plan to develop and implement a number of community curation tools. This should allow the user community to actively participate in improving the quality of the Greengenes database and should ensure that time-consuming manual improvements of sequence and sequence-associated data, including taxonomic corrections, are propagated for the benefit of the whole community. Specifically, five curation tools that should capture manual improvements are in development: (i) improvements in individual sequence alignments, (ii) manual verification of putative chimeras, (iii) recruitment of novel lineages to the Core Set, (iv) corrections in the Greengenes description (the abbreviated description of the record usually has the form [habitat] clone [clone name] for environmental sequences), and (v) updating taxonomic group names. One of the main challenges in the implementation of these tools is to ensure that only high-quality manual edits are incorporated into Greengenes. For example, for a suggested alignment alteration, the submitted sequence must (i) match the existing sequence, (ii) preserve the location of highly conserved positions in the 16S rRNA gene, and (iii) record the curator information as part of the update transaction. We recognize the desire of many users to contribute to a distrib-

uted curation effort, and we hope that Greengenes will become a resource to facilitate this desire.

We thank Kirk Harris and Norman Pace for sharing their ARB database and Richard Phan and Yvette Piceno for assistance with the web interface.

The computational infrastructure was provided in part by the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U.S. Department of Energy Office of Science Office of Biological and Environmental Research Genomics:GTL Program and the Natural and Accelerated Bioremediation Research Program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. Web application development was funded in part by the Department of Homeland Security under grant HSSCHQ04X00037.

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman.** 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.
3. **Ashelford, K. E., A. J. Weightman, and J. C. Fry.** 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. **Brodie, E., S. Edwards, and N. Clipson.** 2002. Bacterial community dynamics across a floristic gradient in a temperate upland grassland ecosystem. Microb. Ecol. **44:**260–270.
5. **Castiglioni, B., E. Rizzi, A. Frosini, K. Sivonen, P. Rajaniemi, A. Rantala, M. A. Mugnai, S. Ventura, A. Wilmotte, C. Boutte, S. Grubisic, P. Balthasart, C. Consolandi, R. Bordoni, A. Mezzelani, C. Battaglia, and G. De Bellis.** 2004. Development of a universal microarray based on the ligation detection reaction and 16S rRNA gene polymorphism to target diversity of cyanobacteria. Appl. Environ. Microbiol. **70:**7161–7172.
6. **Clamp, M., J. Cuff, S. M. Searle, and G. J. Barton.** 2004. The Jalview Java alignment editor. Bioinformatics **20:**426–427.
7. **Cole, J. R., B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje.** 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. **33:**D294–D296.
8. **DeSantis, T. Z., I. Dubosarskiy, S. R. Murray, and G. L. Andersen.** 2003. Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. Bioinformatics **19:**1461–1468.
9. **DeSantis, T. Z., P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen.** NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res., in press.
10. **DeSantis, T. Z., C. E. Stone, S. R. Murray, J. P. Moberg, and G. L. Andersen.** 2005. Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. FEMS Microbiol. Lett. **245:**271–278.
11. **Emrich, S. J., M. Lowe, and A. L. Delcher.** 2003. PROBEmer: a web-based software tool for selecting optimal DNA oligos. Nucleic Acids Res. **31:**3746–3750.
12. **Felsenstein, J.** 1989. PHYLIP—Phylogeny Inference Package (version 3.65). Cladistics **5:**164–166.
13. **Harris, J. K., S. T. Kelley, and N. R. Pace.** 2004. New perspective on uncultured bacterial phylogenetic division OP11. Appl. Environ Microbiol. **70:**845–849.
14. **Harris, J. K., S. T. Kelley, G. B. Spiegelman, and N. R. Pace.** 2003. The genetic core of the universal ancestor. Genome Res. **13:**407–412.
15. **Huber, T., G. Faulkner, and P. Hugenholtz.** 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics **20:**2317–2319.
16. **Hugenholtz, P.** 2002. Exploring prokaryotic diversity in the genomic era. Genome Biol. **3:**1–8.
17. **Kelly, J. J., S. Siripong, J. McCormack, L. R. Janus, H. Urakawa, S. El Fantroussi, P. A. Noble, L. Sappelsa, B. E. Rittmann, and D. A. Stahl.** 2005. DNA microarray detection of nitrifying bacterial 16S rRNA in wastewater treatment plant samples. Water Res. **39:**3229–3238.
18. **Kumar, S., K. Tamura, and M. Nei.** 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. **5:**150–163.
19. **Lane, D. J., A. P. Harrison, Jr., D. Stahl, B. Pace, S. J. Giovannoni, G. J. Olsen, and N. R. Pace.** 1992. Evolutionary relationships among sulfur- and iron-oxidizing eubacteria. J. Bacteriol. **174:**269–278.
20. **Lehner, A., A. Loy, T. Behr, H. Gaenge, W. Ludwig, M. Wagner, and K. H. Schleifer.** 2005. Oligonucleotide microarray for identification of *Enterococcus* species. FEMS Microbiol. Lett. **246:**133–142.
21. **Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer.** 2004. ARB: a software environment for sequence data. Nucleic Acids Res. **32:**1363–1371.
22. **Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker, Jr., P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje.** 2001. The RDP-II (Ribosomal Database Project). Nucleic Acids Res. **29:**173–174.
23. **Pace, N. R.** 1997. A molecular view of microbial diversity and the biosphere. Science **276:**734–740.
24. **Radosevich, J. L., W. J. Wilson, J. H. Shinn, T. Z. DeSantis, and G. L. Andersen.** 2002. Development of a high-volume aerosol collection system for the identification of air-borne micro-organisms. Lett. Appl. Microbiol. **34:**162–167.
25. **Ricke, P., S. Kolb, and G. Braker.** 2005. Application of a newly developed ARB software-integrated tool for in silico terminal restriction fragment length polymorphism analysis reveals the dominance of a novel *pmoA* cluster in a forest soil. Appl. Environ. Microbiol. **71:**1671–1673.
26. **Schloss, P. D., and J. Handelsman.** 2004. Status of the microbial census. Microbiol. Mol. Biol. Rev. **68:**686–691.
27. **Stamatakis, A., T. Ludwig, and H. Meier.** 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics **21:**456–463.
28. **Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins.** 1997. The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:**4876–4882.
29. **Webster, G., C. J. Newberry, J. C. Fry, and A. J. Weightman.** 2003. Assessment of bacterial community structure in the deep sub-seafloor biosphere by 16S rDNA-based techniques: a cautionary tale. J. Microbiol. Methods **55:**155–164.
30. **Wilson, K. H., W. J. Wilson, J. L. Radosevich, T. Z. DeSantis, V. S. Viswanathan, T. A. Kuczmarski, and G. L. Andersen.** 2002. High-density microarray of small-subunit ribosomal DNA probes. Appl. Environ. Microbiol. **68:**2535–2541.