APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May 2006, p. 3291–3301 Vol. 72, No. 5
0099-2240/06/$08.00+0 doi:10.1128/AEM.72.5.3291–3301.2006

Environmental Whole-Genome Amplification To Access Microbial Populations in Contaminated Sediments

Carl B. Abulencia,[1,2] Denise L. Wyborski,[1,2] Joseph A. Garcia,[1] Mircea Podar,[1] Wenqiong Chen,[1,2] Sherman H. Chang,[1] Hwai W. Chang,[1] David Watson,[3] Eoin L. Brodie,[2,4] Terry C. Hazen,[2,4] and Martin Keller[1,2]*

Diversa, San Diego, California 92121[1]; Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831[3]; Lawrence Berkeley National Laboratory, Berkeley, California 94720[4]; and Virtual Institute for Microbial Stress and Survival[2]†

Received 10 December 2005/Accepted 26 February 2006

Low-biomass samples from nitrate and heavy metal contaminated soils yield DNA amounts that have limited use for direct, native analysis and screening. Multiple displacement amplification (MDA) using ϕ29 DNA polymerase was used to amplify whole genomes from environmental, contaminated, subsurface sediments. By first amplifying the genomic DNA (gDNA), biodiversity analysis and gDNA library construction of microbes found in contaminated soils were made possible. The MDA method was validated by analyzing amplified genome coverage from approximately five *Escherichia coli* cells, resulting in 99.2% genome coverage. The method was further validated by confirming overall representative species coverage and also an amplification bias when amplifying from a mix of eight known bacterial strains. We extracted DNA from samples with extremely low cell densities from a U.S. Department of Energy contaminated site. After amplification, small-subunit rRNA analysis revealed relatively even distribution of species across several major phyla. Clone libraries were constructed from the amplified gDNA, and a small subset of clones was used for shotgun sequencing. BLAST analysis of the library clone sequences showed that 64.9% of the sequences had significant similarities to known proteins, and "clusters of orthologous groups" (COG) analysis revealed that more than half of the sequences from each library contained sequence similarity to known proteins. The libraries can be readily screened for native genes or any target of interest. Whole-genome amplification of metagenomic DNA from very minute microbial sources, while introducing an amplification bias, will allow access to genomic information that was not previously accessible.

Recent studies have demonstrated that natural attenuation and bioremediation of metals, radionuclides, and organic contaminants cannot be effectively applied at many sites until we have a better understanding of the physiology, ecology, and phylogeny of microbial communities at contaminated sites (12, 38, 57). However, the success of many monitored natural attenuation and bioremediation approaches depends largely on our understanding of regulatory mechanisms and cellular responses to different environmental factors affecting the contaminant degradation or metal reduction activity in situ. Microorganisms are often exposed to multiple stress conditions in situ, and the microbial community structure is most likely affected by many different abiotic and biotic variables in a nonlinear fashion (58). Given the extreme stresses that the microbial community is under at these contaminated sites, an accurate assessment of the microbial community structure and the ecogenomics is critical (17, 20).

Despite the dominance of microorganisms in the biosphere, relatively little is known about the majority of environmental microorganisms, largely because of their resistance to culture under standard laboratory conditions (59). As such, alternative approaches are required to assess the large amount of information in the environmental metagenome. Environmental sequencing projects targeted at small-subunit (SSU) rRNA are a popular method to assess the phylogenetic diversity of uncultured organisms (26, 30, 31). Cloning and sequencing of PCR-amplified functional genes from environmental samples are also powerful tools for investigating the ecology and role of microorganisms (1, 5, 18, 39, 47). More recently, sequencing of environmental genomic DNA has furthered our understanding of the metabolic potential of microorganisms occupying various environmental niches. Sequencing efforts include analyzing individual large-insert bacterial artificial chromosome (BAC) clones, small-insert libraries made directly from environmental DNA, and high-throughput shotgun sequencing, which provides a global view of the environmental community (18, 22, 35, 41, 52, 54, 55). These techniques have advanced our understanding of the types of microorganisms and biodegradation/biotransformation capabilities found in various habitats.

To understand how biogeochemical processes affect microbial community structure and bioremediation, the U.S. Department of Energy (DOE) Natural and Accelerated Bioremediation Research (NABIR) program has established a Field Research Center (FRC) on the DOE Oak Ridge Reservation (http://www.esd.ornl.gov/nabirfrc/) in eastern Tennessee. The FRC is heavily contaminated with nitrate, heavy metals, radionuclides, and halogenated organics (58). Characterization of indigenous subsurface microbial populations from this site has mainly focused on microbes collected from groundwater (27, 58) or after biostimulation (26, 31). However, it is recognized that the planktonic microbial population in groundwater may

* Corresponding author. Mailing address: Diversa, 4955 Directors Place, San Diego, CA 92121. Phone: (858) 526-5162. Fax: (858) 526-5662. E-mail: mkeller@diversa.com.

† http://vimss.lbl.gov.

not be representative of the active population residing as biofilms on sediment surfaces (15). In fact, the biofilm mode of growth confers resistance to environmental stresses such as heavy metals (50).

The extremely low cell densities within these types of biofilms in combination with high clay content and heavy metal/radionuclide contamination can inhibit many standard molecular approaches, including shotgun and BAC library generation. Isolation of DNA for library construction is complicated and would require large quantities of subsurface material. The current minimum of DNA to construct a library used for shotgun sequencing is around 0.5 to 4 μg of DNA, which can be obtained from a minimum of 0.5 g of microbe-rich material (53). In low-biomass subsurface environments with cell densities as low as 10^4 per g, 11 to 88 kg of sediment would be required, assuming that an average cell contains 4.5 fg of DNA. Since the quantity of DNA needed for BAC library construction is 20-fold greater, it is clear that current approaches are not viable for such low-biomass environments. A recent approach to amplify small amounts of viral DNA by random PCR prior to shotgun library construction has enabled genomic analysis of viral communities and the sequencing of phage genomes (4, 34). The effectiveness of using this method on environmental microbial genomes has not been studied.

One technique used for molecular surveys of microorganisms from environmental samples is to amplify DNA by PCR using primers to highly conserved positions in specific SSU rRNA genes (11, 22). However, low-cell-count environments often do not yield enough DNA for PCR (31). In addition, DNA amplification by PCR is inherently biased because not all SSU rRNA genes amplify with the same "universal primers," reducing the effectiveness of this approach to survey such environments. Therefore, new methods are required to combine environmental whole genome amplification (WGA) with library construction for metagenome analyses of low-cell-density environments.

Here we report the construction of whole-genome-amplified environmental libraries for biodiversity assessment of low biomass contaminated subsurface sediment cores.

MATERIALS AND METHODS

DNA amplification. Genomic DNA (gDNA) was amplified by multiple displacement amplification (MDA) using the GenomiPhi DNA amplification kit (Amersham Biosciences, Piscataway, NJ). Amplification was carried out according to the protocol with a modification in reaction incubation time. We added 1 μl of template to 9 μl of sample buffer and heated the mixture to 95°C for 3 min to denature the template DNA. The sample was cooled and mixed with 9 μl of reaction buffer and 1 μl of enzyme mix and then incubated at 30°C for 3 to 6 h. After amplification the DNA polymerase was heat inactivated during a 10-min incubation at 65°C. Each sample was amplified in triplicate. The three MDA reaction products per sample were then combined before further processing.

GeneChip analysis. MDA genome coverage was analyzed by using the Affymetrix *Escherichia coli* genome GeneChip array (Affymetrix, Inc., Santa Clara, CA). The chip contains 7,231 probe sets spanning the entire *E. coli* genome. The *E. coli* strain used was XL1-Blue MR (Stratagene, La Jolla, CA). Genomic DNA extracted from 1 ml of an overnight *E. coli* culture (10^9 cells) was used as a positive control. This was compared to MDA products amplified from gDNA extracted from two dilutions of the culture (5,000 and five cells). Cell culture concentration was estimated by a spectrophotometer reading (optical density at 600 nm), and serial dilutions were made to 5,000 and 5 cells (cell numbers were estimated based on the dilution series). Dilutions were grown on solid agar plates to verify the cell counts. The gDNA was extracted from the overnight culture and the dilutions by first encasing the cells in agarose and then completing the extractions as described below. The three sets of DNA were concentrated by ethanol precipitation and fragmented with DNase I (0.6 U per μg of genomic DNA) at 37°C for 10 min. After inactivating DNase I at 100°C for 10 min., the fragmented DNA was end labeled with biotin-ddATP and hybridized to the *E. coli* genome array using Affymetrix standard protocols for RNA. The probe array was scanned twice, and the intensities were averaged with a GeneArray Scanner (Hewlett-Packard, Palo Alto, CA). Scanned images were processed and quantified by using GeneChip Suite 5.0 (Affymetrix).

The data were normalized by setting the mean hybridization signal for each sample equal to 100. The absolute call represents a qualitative indication of whether or not a transcript is detected within a sample. These calls are determined by using the following metrics: (i) the ratio of the number of positive probe pairs to the number of negative probe pairs (known as the positive/negative ratio), (ii) the fraction of positive probe pairs (positive fraction), and (iii) the average across the probe set of each probe pair's log ratio of positive intensity over negative intensity (log average ratio) (25).

Amplification and analysis of mixed isolates. Eight isolates with fully sequenced genomes were chosen to represent a range of genome sizes (2.5 to 8.7 Mb) and G+C content (32 to 72%). The isolates were *Deinococcus radiodurans* ATCC 13939, *Desulfovibrio vulgaris* ATCC 29579, *Geobacter sulfurreducens* ATCC 51573, *Mesorhizobium loti* ATCC 35173, *Nitrosomonas europaea* ATCC 19718, *Shewanella oneidensis* ATCC 700550, *Staphylococcus epidermidis* ATCC 35984, and *Streptomyces coelicolor* ATCC 10147 (American Type Culture Collection, Manassas, VA). The gDNA from the eight isolates was mixed and normalized based on genome size. The resulting mix had a concentration of 60 ng/μl. A small insert library was made from 4 μg (66 μl) of the mixed DNAs, as described below. The mix was then diluted 100- and 10,000-fold to 600 and 6 pg/μl, respectively. A total of 1 μl of the diluted gDNA was then amplified by MDA. The DNA from each dilution was amplified to greater than 4 μg. Small insert libraries were created from 4 μg of the unamplified mixed DNA and from the amplified DNA as described below. Random clones from each of the three libraries were end sequenced. Isolate representation within each library was determined by a BLASTN search of the end sequences against the NCBI genome database. The ratio of sequences from each isolate within the libraries were compared (unamplified versus amplified). If we assume unbiased library construction, an ideal, unbiased MDA amplified library should contain the same ratio of sequences as the unamplified library.

Site and sample description. Soil core samples were collected from contaminated subsurface sediments at the U.S. DOE NABIR FRC, located at the Y-12 plant within the Oak Ridge Reservation in Oak Ridge, Tenn. Approximately 10 million liters of liquid nitric acid and uranium bearing wastes were discarded at this site per year for 30 years, until it was closed in 1984. The site's groundwater plumes originate from the former waste disposal ponds at the Y-12 plant. Nine soil samples were taken from five areas surrounding the S-3 waste ponds. The samples were from sites of various distances to the S-3 ponds and from core depths ranging from 3.5 to 9 m, with each sample containing different levels of contamination of uranium(VI), nitrate, plutonium, technetium, toxic metals (nickel, aluminum, barium, chromium, mercury), chelating agents (EDTA), chlorinated hydrocarbons (trichloroethylene and tetrachloroethylene), polychlorinated biphenyls, and fuel hydrocarbons (toluene, benzene). A full description of the area can be obtained at the FRC Web site (http://www.esd.ornl.gov/nabirfrc/).

Of the nine contaminated samples, the following three were analyzed in more detail: (i) FB075, area 3, which is adjacent to the west side of the ponds, core segment depth of 8.4 to 9 m (sample 1, library 1); (ii) FB076, area 3, core segment depth of 3.9 to 4.5 m (sample 3, library 3); and (iii) FB078, area 2, which is >200 m to the southwest of the ponds, core segment depth of 6.1 to 6.4 m (sample 5, library 5). Sediment cores were sampled with an Acker Drill Co. Holegator track drill equipped with polyurethane sleeves lining the corer. The cores were anaerobically sealed in argon and shipped on ice within 24 h of sampling, and DNA extractions were done on arrival in a radiation control area at the Lawrence Berkeley National Laboratory.

DNA isolation. Each soil sample was removed from the core sleeve and mixed manually, and 50 g of soil per sample was used for gDNA isolation. DNA was isolated directly from cells separated from the environmental matrix (2, 32, 42, 43). Highly purified suspensions of microbial consortia were obtained by isopycnic density gradient centrifugation with Nycodenz (Sigma-Aldrich, St. Louis, MO). The resulting cell pellet was immobilized in an agarose plug and lysed by enzymatic and chemical digestions (46). The isolated gDNA was then used directly in the amplification reaction, as described above.

SSU rRNA gene analysis. Bacterial SSU rRNA genes were amplified from the MDA DNA products by using the universal primers 8F (5′-AGAGTTTGATC CTGGCTCAG-3′) and 1492R (5′-GGTTACCTTGTTACGACTT-3′) and the Roche Expand Long Template PCR system (Roche Applied Science, Indianapolis, IN). SSU rRNA clone libraries were generated by using the TOPO TA

TABLE 1. Amplification of mixed isolates to test for MDA bias[a]

Isolate	Size (Mb)	%G+C	Pre-MDA		10^{-2} dilution (post MDA)		10^{-4} dilution (post MDA)	
			No. of sequences	%	No. of sequences	%	No. of sequences	%
Deinococcus radiodurans	3.3	66	47	9.2	5	1.2	12	3.3
Desulfovibrio vulgaris	3.8	60	44	8.6	5	1.2	8	2.2
Geobacter sulfurreducens	3.8	61	29	5.7	54	12.8	58	16.2
Mesorhizobium loti	7.6	62	123	24.1	15	3.6	12	3.3
Nitrosomonas europaea	2.8	50	4	0.8	42	10.0	46	12.8
Shewanella oneidensis	5.1	45	97	19.0	277	65.8	192	53.5
Staphylococcus epidermidis	2.5	32	26	5.1	21	5.0	30	8.4
Streptomyces coelicolor	8.7	72	140	27.5	2	0.5	1	0.3
Total			510		421		359	

[a] Classification of sequences to their respective isolate determined by BLASTN searches against the NCBI genome database.

Cloning kit (Invitrogen, Carlsbad, CA). Sequencing was performed by using an Applied Biosystems 3730xl DNA analyzer (Applied Biosystems, Foster City, CA), and the individual clone reads were assembled by using Sequencher (Gene Codes Corp., Ann Arbor, MI).

The assembled sequences were checked for potential chimeric artifacts by using the Bellerophon program and the chimeric sequences were discarded. The final sequence datasets were aligned by CLUSTAL W with the closest sequence relatives from the Ribosomal Database Project (6) and GenBank databases. The alignments were manually curated by using BioEdit (16).

To analyze the diversity of the different sequence libraries, a Jukes-Cantor corrected distance matrix was calculated by using DNADIST (PHYLIP) (13, 16). The matrix was used as input into the program DOTUR (37) using the furthest neighbor algorithm to obtain a variety of diversity richness estimators at different genetic distance values (rarefaction curve, bias corrected Chao1 richness, abundance-based coverage estimator ACE, and Shannon-Weaver diversity index). Phylogenetic analysis was conducted by using PAUP* (48) with the distance criterion (Jukes Cantor), followed by bootstrapping (1,000 replicates). Trees were visualized by using Tree Explorer (http://evolgen.biol.metro-u.ac.jp/TE/TE_man.html).

Library construction, sequencing, and analysis. Unamplified gDNA (mixed isolate DNA experiment) and MDA-amplified DNA (sample numbers 1, 3, and 5 and mixed isolate DNA) were mechanically sheared and used to generate libraries in ZAP-based lambda phage cloning vectors according to the manufacturer's protocol. Phagemid libraries were produced from the parental lambda clones through the in vivo excision properties of ZAP-based cloning vectors and used to infect *E. coli* host cells (44). The average insert sizes were 2 to 4 kb. Plasmid inserts from randomly picked colonies were end sequenced by using an Applied Biosystems 3730xl DNA analyzer with the primers T3 (5′-AATTAACCCTCACTAAAGGG-3′) and T7 (5′-GTAATACGACTCACTATAGGGC-3′). The end sequences of the random clones were analyzed against the NCBI protein database by using BLASTX. Only hits with e-values of <1e-10 were considered for further protein analysis. Sequences were assembled into contigs by using Vector NTI Contig Express. For the MDA bias analysis, paired clone-end singletons were discarded.

COG analysis. COG (for clusters of orthologous groups) functional assignment of proteins predicted from DNA sequences from different MDA-amplified genomic libraries was done by using the library clone DNA end-sequences to BLAST against an expanded COG database from http://string.embl.de/ (49, 56), which covers 26,201 protein families, using BLASTX. Filtering criteria were set to discard the nonspecific BLAST hits: (i) if the aligned amino acid sequence length is ≥100 amino acids (aa), the BLAST e-value should be <1e-10 and (ii) if the aligned amino acid sequence length is ≥30aa but <100 aa, the amino acid percent identity should be >30%.

The filtered BLAST results were parsed to associate the COG identification numbers (IDs) with the individual library end sequences using a custom Perl script. In the case of a particular sequence which had multiple hits that belonged to different COG classifications, we allowed multiple COG assignment only when the bit scores from the second or third COG classification were no less than 3-fold different than the bit score from the top COG classification.

Nucleotide sequence accession numbers. The reported SSU rRNA sequences and library clone end sequences are listed with their respective GenBank accession numbers, DQ404590 to DQ404652, DQ404654 to DQ404938, and DX385314 to DX389173.

RESULTS

GeneChip analysis of *E. coli* MDA coverage. The amount of *E. coli* gDNA extracted from approximately 5,000 and 5 cells was very minute and could not be detected by the Affymetrix *E. coli* genome GeneChip after hybridization. This low number of cells was used to simulate the low abundance of organisms found in contaminated soils. Genomic DNA extracted from ca. 10^9 cells of an overnight *E. coli* culture elicited positive signals for all probe sets on the *E. coli* GeneChip. gDNA extracted from approximately 5,000 and 5 cells, after amplification by MDA, resulted in 99.94% and 99.2% of the probe sets called "present," respectively.

We also utilized the quantitative information from the GeneChip experiments to assess if there was over- or underamplification of regions of the gDNA during the MDA reaction. Over- or underamplification is defined as a detection value greater or less than three times the positive control value. For the 5,000-cell amplification, 0.4% of probe set regions was overamplified and 0.9% was underamplified. For the five-cell amplification, 0.6% was overamplified and 4.6% was underamplified. The regions of over- and underamplification are distributed across the entire genome and appear to be random.

Amplification of mixed isolates to test for MDA bias. In order to have complete and adequate representation of genomes in an environmental sample, WGA must amplify all of the genomes present with minimal bias. To determine whether any MDA bias exists when amplifying from the environment, gDNAs from eight different known isolates were mixed to simulate an environmental sample. We analyzed 510, 421, and 359 end sequences from the unamplified library, the 100-fold dilution-amplified library, and the 10,000-fold dilution-amplified library, respectively, and binned them to the corresponding isolate. Comparison of the different libraries showed that there is an MDA bias. BLAST analyses of the end-sequences from random clones from each of the three libraries revealed that *Shewanella oneidensis*, *Nitrosomonas europaea*, and *Geobacter sulfurreducens* were amplified preferentially. *Deinococcus radiodurans*, *Desulfovibrio vulgaris*, *Mesorhizobium loti*, and *Strepto-*

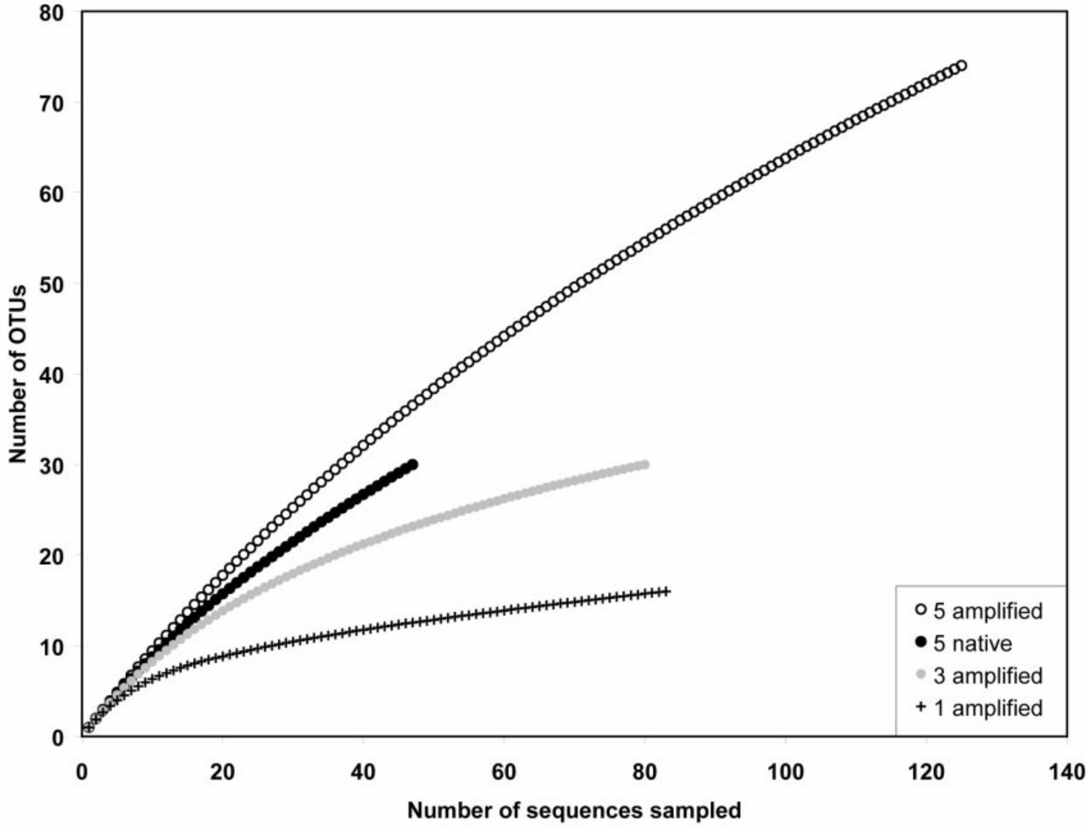

FIG. 1. Rarefaction curves for SSU rRNA genes from amplified samples 1, 3, and 5 and native sample 5 at the 98% sequence identity level.

myces coelicolor all had less representation in the amplified DNA library than in the unamplified DNA library, with *S. coelicolor* having the least representation. *Staphylococcus epidermidis* showed no bias between unamplified and amplified DNA. However, in this small sample set of sequences, all isolates were represented in the MDA-amplified libraries regardless of bias (Table 1).

Microbial diversity analysis of the soil libraries. DNA was isolated from soil sample cores according to the described protocol. Three different samples were used for further studies: samples FB075 (sample 1), FB076 (sample 3), and FB078 (sample 5). Extractions from samples 1 and 3 resulted in DNA concentrations that were not sufficient to obtain SSU rRNA gene PCR products. The only sample that contained a sufficient level of DNA template for SSU rRNA gene PCR was sample 5 (<5 ng/μl). A native SSU rRNA library (native library 5) was constructed from this sample. In addition, the DNA from this sample was amplified by MDA with >6 μg of DNA yield. An SSU rRNA library was constructed from the amplified product (amplified library 5). The same MDA amplification method was used on the DNA obtained from samples 1 and 3 to successfully yield >6 μg of DNA each, even though SSU rRNA gene PCR on the DNA isolated from these samples was negative. SSU rRNA libraries were constructed from these amplified products (amplified library 1 and amplified library 3).

To compare the microbial diversity before and after MDA, 47 SSU rRNA gene clones from native library 5 were sequenced and compared to 125 SSU rRNA gene clones derived from amplified library 5. Microbial diversity differences between deep (8.4 to 9 m, sample 1) and shallow (3.9 to 4.5 m, sample 3) soil samples from area 3 were analyzed by comparing the SSU rRNA libraries from amplified samples 1 and 3. To estimate the effectiveness of sampling of microbial diversity within these samples, rarefaction curves using a 2% difference in SSU rRNA genes were used for species separation. The two libraries prepared from sample 5, native and amplified, indicated a higher diversity of the amplified library (149 species versus 88 species for the native library), with upper estimates reaching an excess of 200 species. The Shannon-Weaver index value also indicates higher species diversity in the amplified library prepared from sample 5 (Fig. 1 and Table 2).

Rarefaction curves calculated for the 2% distance show signs of leveling of the number of novel operational taxonomic units

TABLE 2. Sample diversity

Sample	No. of sequences	Chao estimate (95% confidence interval)	ACE estimate	Shannon diversity index (95% confidence interval)
1	83	24 (17–92)	27	2.30 (2.12–2.49)
3	80	45 (33–102)	47	3.10 (2.91–3.29)
5 (amplified)	125	149 (107–248)	163	4.10 (3.95–4.25)
5 (native)	47	88 (41–330)	82	3.21 (2.97–3.46)

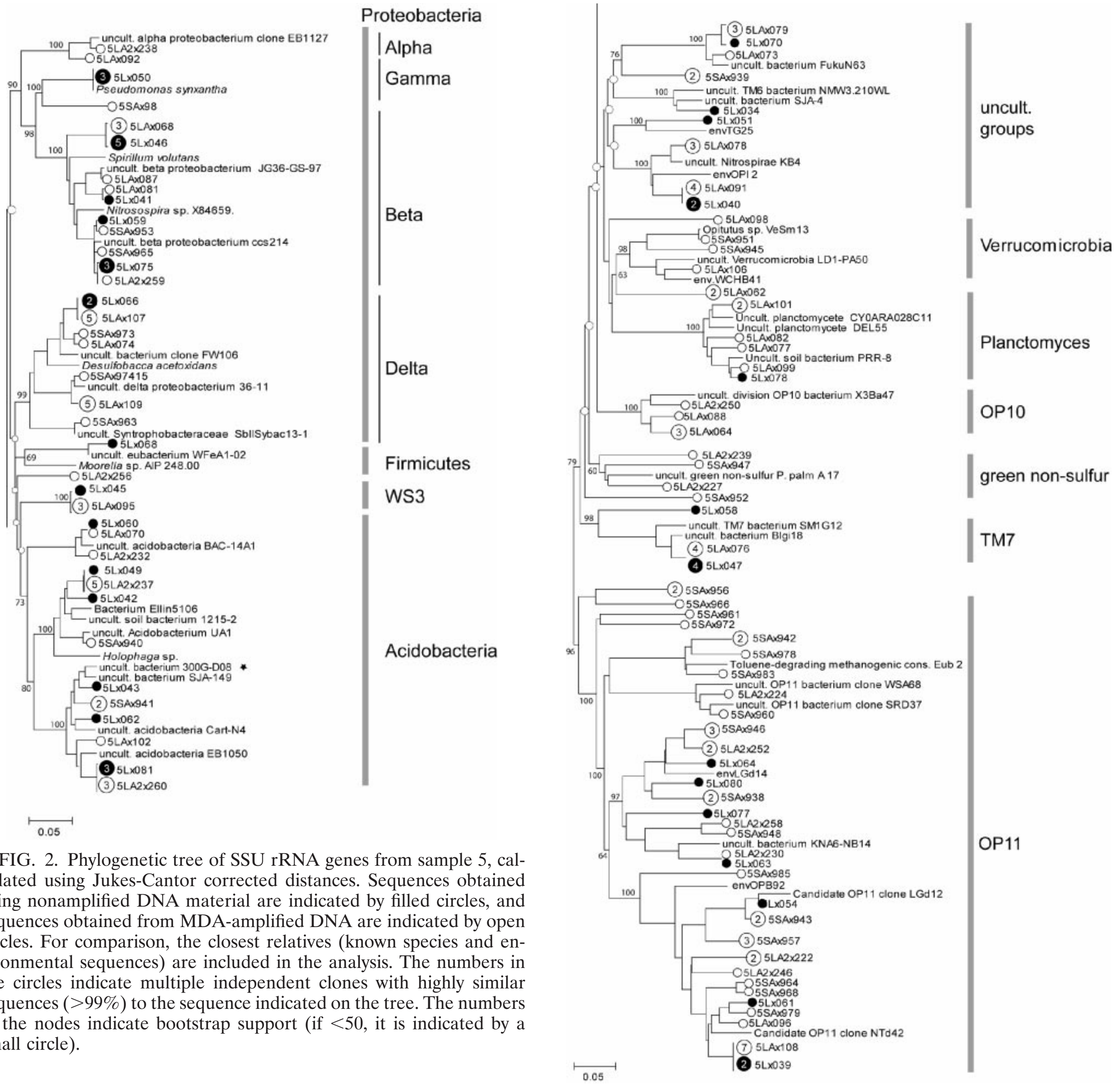

FIG. 2. Phylogenetic tree of SSU rRNA genes from sample 5, calculated using Jukes-Cantor corrected distances. Sequences obtained using nonamplified DNA material are indicated by filled circles, and sequences obtained from MDA-amplified DNA are indicated by open circles. For comparison, the closest relatives (known species and environmental sequences) are included in the analysis. The numbers in the circles indicate multiple independent clones with highly similar sequences (>99%) to the sequence indicated on the tree. The numbers at the nodes indicate bootstrap support (if <50, it is indicated by a small circle).

(OTU) (species) identified by increasing sequence sampling of sample 3 and especially sample 1. The Chao 1 estimator predicts a minimum number of 24 species (17 to 92 at the 95% confidence interval) for sample 1 and 45 species (33 to 102 at the 95% confidence interval) for sample 3, similar estimates being obtained by using ACE (Table 2).

The taxonomic diversity at higher levels is revealed by the phylogenetic trees constructed based on the sequences from the three different samples as well as the histogram figures summarizing the diversity at phylum or class level for each SSU rRNA library (Fig. 2, 3, and 4). The availability of sequences from both native and amplified community gDNA for sample 5 allowed us to investigate possible biases introduced by the amplification step. In general, there is good agreement between the taxonomic groups identified in the two samples with some differences, especially for groups that have low representation (e.g., alpha-*Proteobacteria*, *Verrucomicrobia* and some of the uncultured groups). Overall, however, no single phylum appears to be strongly dominating sampling either native or amplified sample 5. *Acidobacteria* and the candidate division OP11 are most abundant in both samples in terms of species.

Sample 3, while of higher diversity in comparison to sample 1, reveals a relatively even distribution of species across several major phyla, including *Proteobacteria*, *Actinobacteria*, *Planctomycetes*, *Verrucomicrobia*, and the candidate division OP11. In sample 1, however, there is a dominance of gamma-*Proteobacteria* (~35%), equally distributed between a close relative of *Pseudomonas synxantha* (>50% of sequences) and a *Legionella*-

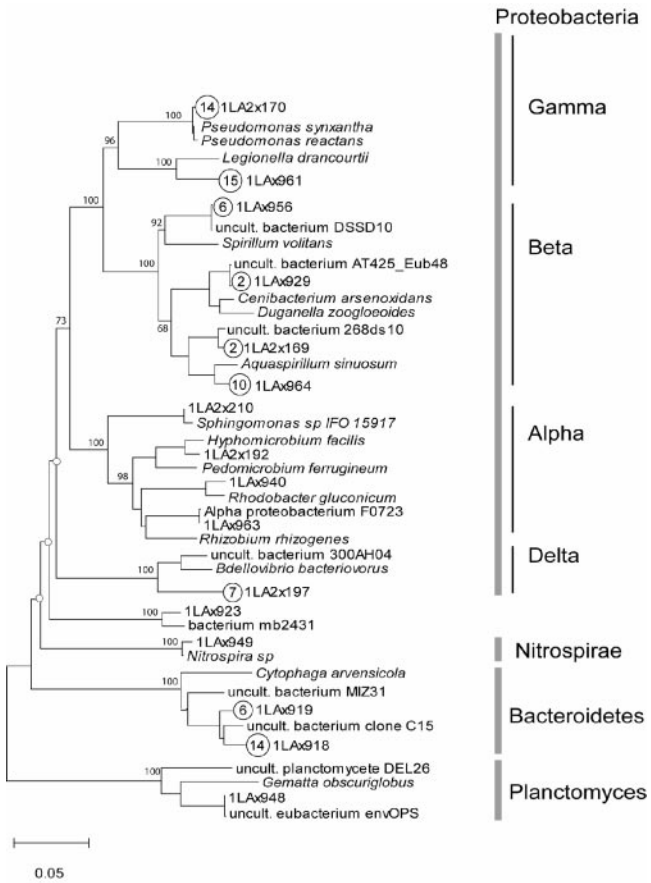

FIG. 3. Phylogenetic tree of SSU rRNA genes from sample 1, calculated using Jukes-Cantor corrected distances. For comparison, the closest relatives (known species and environmental sequences) are included in the analysis. The numbers in the circles indicate multiple independent clones with highly similar sequences (>99%) to the sequence indicated on the tree. The numbers at the nodes indicate bootstrap support (if <50, it is indicated by a small circle).

type species. The next most represented groups (25% each) are *Cytophaga*/*Bacteroidetes*, represented by 27 sequences and approximately four species-level OTUs, and beta-*Proteobacteria* (several taxa, including *Spirillum*, *Aquaspirillum*, and *Cenibacterium* relatives).

Partial end sequencing and COG analysis of environmental clones from MDA-amplified genomic libraries 1, 3, and 5. Although a SSU rRNA gene PCR product was successfully amplified from the sample 5 gDNA, the amount of gDNA obtained was not sufficient for phagemid library construction. Extracted DNA from all three samples was amplified by MDA to create a sufficient quantity of DNA (2 μg) to construct genomic libraries. To check the quality and diversity of environmental libraries 1, 3, and 5, constructed from amplified DNA, 960, 864, and 864 random clones from each were sequenced, respectively. A total of 4,021 sequences of 400 nucleotides or longer were generated from both ends of the clones. Of these sequences, 2,610 (64.9%) showed similarities to known proteins in the database. Further analysis showed that 2,420 (60.2%) sequences had similarity to bacterial proteins, 134 (3.3%) had similarity to archaeal proteins, and 51 (1.3%) had similarity to eukaryal proteins. A contig assembly of the end sequences was used to analyze the extent of MDA bias on the amplified DNA used to construct the libraries. A high proportion of sequences that form contigs possibly indicate a bias in amplification of these sequence regions. Libraries 1, 3, and 5 had 370 sequences that formed 101 contigs, 152 sequences that formed 53 contigs, and 141 sequences that formed 54 contigs, respectively (Table 3). The sizes of the largest contig formed in libraries 1, 3, and 5 were 1,799, 2,373, and 2,749 nucleotides, respectively.

In order to explore the possible functions of the predicted proteins from the sequences collected in each soil library, COG analysis with the random sequence reads was performed. More than half of the sequences from each library showed sequence similarity to the entries in the STRING COG database. For example, for library 1, among the 1,394 random sequences, 1,154 of them were assigned to 674 distinct COG IDs. For library 3, among the 1,118 sequences obtained, 782 were assigned to 561 distinct COG IDs. For library 5, among the 1,509 sequences obtained, 1,126 were assigned to 800 COG IDs. As shown in Table 4, the three environmental libraries contain similar distributions for most of the COG functional categories. There are a few exceptions, such as coenzyme transport and metabolism, which has twofold more representation in library 1 than in library 3 and library 5, and secondary metabolites biosynthesis, transport, and catabolism, which has twofold less representation in library 5 than in libraries 1 and 3. The COG IDs that are present at higher frequencies in each library belong mostly to hypothetical proteins. In library 1, the top five most frequent COGs are as follows: COG3762, hypothetical protein; COG0642, sensor histidine kinase; COG0438, hypothetical protein PF1364; COG0745, putative two-component regulator; and COG3696, cation efflux system protein. In library 3, the top three most frequent COGs are as follows: COG0642, sensor histidine kinase; COG0463, putative glycosyl transferase; and KOG1869, annotation not available. In library 5, the top three COGs that are present most frequently are as follows: COG0642, sensor histidine kinase; KOG1869, annotation not available; and COG0457, tetratricopeptide repeat domain 13 (Table 4). The differences between some of the functional classifications of MDA-amplified libraries 1, 3, and 5 and the average of the published bacteria (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) can be due to the limited number of environmental sequences, an MDA bias, or the environment itself. Further sequence analysis will be necessary to answer this question.

DISCUSSION

The economics of sequencing is rapidly changing due to the improvement of tools for sequencing and assembly. This has provided a significant boost to the field of environmental genomics (24). Shotgun sequencing provides a wealth of biomarkers that can be used to assess the phylogenetic diversity of a sample with more power than conventional PCR-based SSU rRNA studies allow (55). In addition, the sequencing of environmental samples has provided valuable insights into the lifestyles and metabolic capabilities of uncultured organisms occupying various environmental niches (53). Information derived from comparative metagenomic analyses may be used to

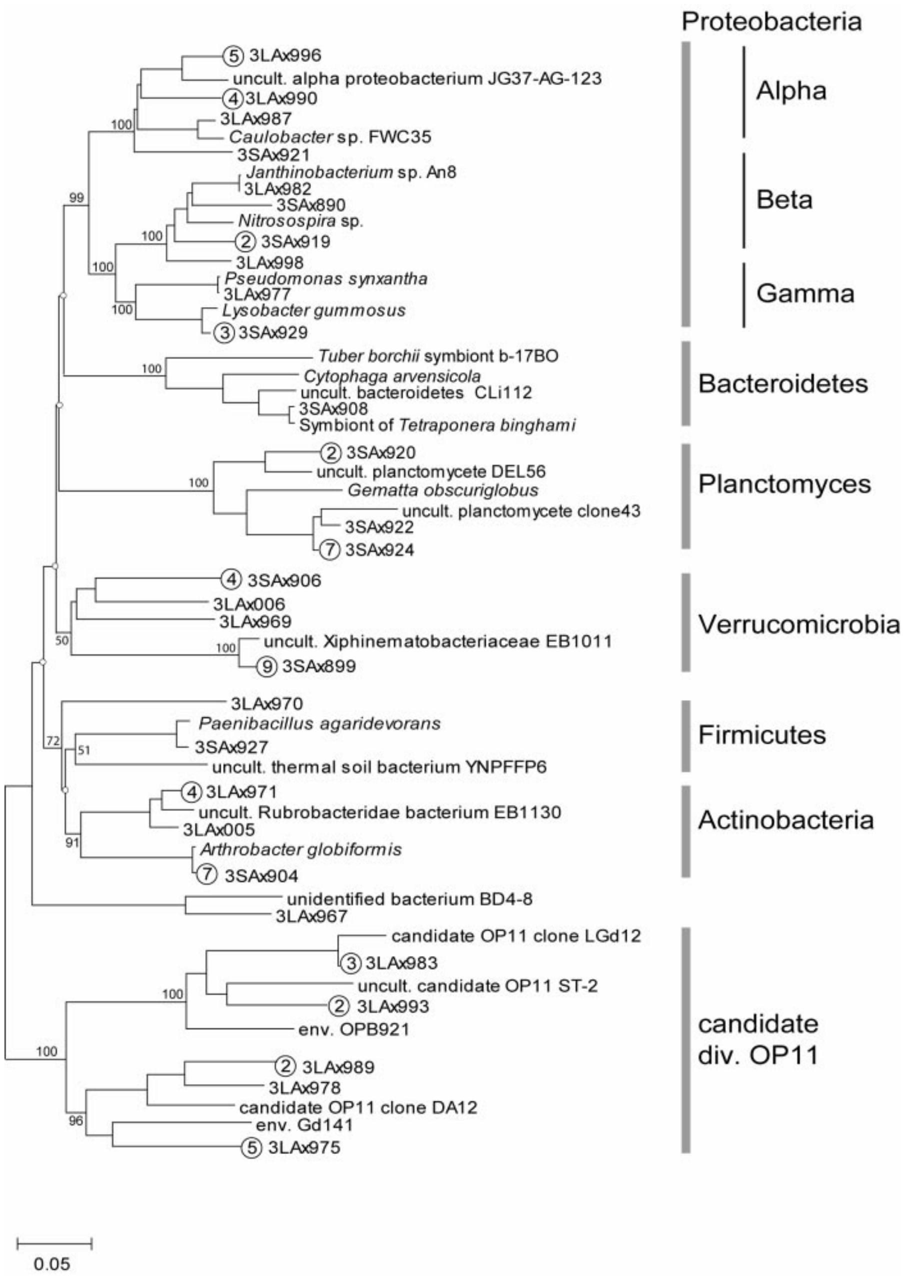

FIG. 4. Phylogenetic tree of SSU rRNA genes from sample 3, calculated using Jukes-Cantor corrected distances. For comparison, the closest relatives (known species and environmental sequences) are included in the analysis. The numbers in the circles indicate multiple independent clones with highly similar sequences (>99%) to the sequence indicated on the tree. The numbers at the nodes indicate bootstrap support (if <50, it is indicated by a small circle).

predict features of the sampled environments, such as elemental recycling, conversion of biomass, bioremediation, and stress response (24). Thus far, however, only environments with relatively high biomass, such as biofilms, open ocean water, agricultural soils, and forest soils have been studied (52–55). Samples with extremely low cell counts, such as contaminated subsurface sediment cores, are usually not accessible for an environmental sequencing study. A significant, and typically impractical, amount of contaminated sediment would be necessary to isolate enough DNA for traditional library construction, and in the case of radionuclide-contaminated soils, it would pose the added problem of secondary waste generation. Possible solutions include the development of new methods to decrease the amount of DNA needed for library construction or methods to amplify environmental DNA to obtain enough for library construction. PCR-based methods have been used for whole-genome amplification; however, they have been shown to exhibit a high amplification bias and do not amplify genomes in their entirety (7, 21).

ϕ29 DNA polymerase is an enzyme which is widely used for

TABLE 3. Statistics on library end sequences

Parameter	Library 1		Library 3		Library 5		Overall total (%)
	Total	%	Total	%	Total	%	
No. of clones sequenced	960		864		864		
No. of sequences generated	1,920		1,728		1,728		
No. of quality sequences[a]	1,394	100	1,118	100	1,509	100	4,021 (100)
No. of sequences that form contigs	370	26.5	152	13.6	141	9.3	663 (16.5)
No. of contigs assembled	101		53		54		208
No. of sequences with similarities to known proteins[b]	928	66.6	692	61.9	990	65.6	2,610 (64.9)
No. of sequences with highest similarity to bacterial protein	901	64.6	629	56.3	890	59.0	2,420 (60.2)
No. of sequences with highest similarity to archaeal protein	12	0.9	43	3.8	79	5.2	134 (3.3)
No. of sequences with highest similarity to eukaryotic protein	12	0.9	18	1.6	21	1.4	51 (1.3)

[a] Sequences >400 nucleotides in length.

[b] e-values of <1e-10 from BLASTX searches against the NCBI protein database.

rolling-circle amplification of plasmids and circular DNA templates (8, 10). The strand-displacing enzyme has proofreading activity, is extremely sensitive, and has been shown to amplify DNA of up to 70 kb (3). This polymerase has also been used for whole-genome amplification of bacterial isolates (10, 29). The sensitivity of this method is problematic; therefore, reactions are performed under exceptionally clean conditions. In addition, it is important to perform negative control experiments, which help to detect potential contamination of the reagents through foreign DNA. To evaluate ϕ29 DNA polymerase MDA for whole-genome amplification from low-biomass samples, we first performed MDA on as few as five *E. coli* cells. GeneChip data demonstrated that 99.2% of the *E. coli* genome was detectable, showing no significant areas of over- or underrepresentation. Previous analysis of *Xyella fastidiosa* libraries constructed from amplified and unamplified genomic DNA also showed similar genome coverage from both DNA sources (10). However, the mixing of eight sequenced bacterial strains, normalized according to their genome size, revealed the preferred amplification of certain strains. This bias is not thought to be based on genome accessibility due to G+C content or secondary structures (3, 21). The introduction of this bias may be due to the sizes of DNA templates (possibly sheared during extraction and mixing), random primer availability, and stochastic effects of amplifying from very low concentrations of template (29). Another possibility may be the differential cloning efficiency of the amplified DNA in comparison to that of the unamplified DNA. Although the MDA approach may be the only way to access some environments, the bias introduced by MDA is a primary concern, and the possibility of underrepresenting or overlooking species cannot be discounted. Analysis of the library constructed from the unamplified DNA also demonstrated some bias (particularly with *N. europaea* and *G. sulfurreducens*) possibly introduced through different cloning efficiencies which can be affected by G+C content, repeats native to the genome sequence, and

TABLE 4. COG analysis

Function classification	Library 1		Library 3		Library 5		Bacteria (%)
	No. of sequences	%	No. of sequences	%	No. of sequences	%	
Amino acid transport and metabolism	64	4.06	47	4.13	75	4.63	9.14
Carbohydrate transport and metabolism	187	11.85	184	16.18	232	14.33	6.39
Cell cycle control, cell division, and chromosome partitioning	16	1.01	11	0.97	21	1.30	1.38
Cell motility	4	0.25	4	0.35	7	0.43	1.97
Cell wall/membrane/envelope biogenesis	117	7.41	88	7.74	122	7.54	5.47
Coenzyme transport and metabolism	68	4.31	23	2.02	31	1.91	3.66
Defense mechanisms	42	2.66	28	2.46	43	2.66	1.60
DNA replication, recombination, and repair	85	5.39	61	5.37	84	5.19	6.58
Energy production and conversion	216	13.69	162	14.25	246	15.19	5.62
Function unknown	117	7.41	47	4.13	67	4.14	7.66
General function prediction only	119	7.54	100	8.80	130	8.03	12.93
Inorganic ion transport and metabolism	88	5.58	57	5.01	44	2.72	5.93
Intracellular trafficking, secretion, and vesicular transport	12	0.76	12	1.06	15	0.93	2.43
Lipid transport and metabolism	36	2.28	12	1.06	22	1.36	3.25
Nucleotide transport and metabolism	30	1.90	21	1.85	47	2.90	2.26
Posttranslational modification, protein turnover, chaperones	190	12.04	146	12.84	196	12.11	3.68
RNA processing and modification	1	0.06	0	0.00	3	0.19	0.02
Secondary metabolites biosynthesis, transport and catabolism	26	1.65	21	1.85	13	0.80	2.52
Signal transduction mechanisms	80	5.07	31	2.73	91	5.62	4.79
Transcription	37	2.34	30	2.64	47	2.90	7.07
Translation, ribosomal structure and biogenesis	43	2.73	52	4.57	83	5.13	5.67
Total	1,578		1,137		1,619		

toxicity, among others. However, even from the limited number of MDA-amplified library clones sampled by sequencing, clones from all of the bacterial strains could be identified within the MDA-amplified constructed library.

It has been demonstrated that core samples obtained from the NABIR site are very low in bacterial cell counts (26). Most of the microbial diversity studies on samples from these sites have been performed on groundwater or on biostimulated samples (27, 31, 58). One study that attempted SSU rRNA PCR from DNA extracts of contaminated sediments was unsuccessful (31). In the current study, an SSU rRNA gene PCR product from native gDNA was obtained only from one sample out of nine. Therefore, we used this sample to study the possible bias on microbial diversity introduced through MDA. The comparison of the microbial diversity of the native and amplified SSU rRNA library showed small changes within the taxonomic groups. Distinguishing traditional taxonomic units based on SSU rRNA gene sequences is controversial (36). Typically, a 2 to 3% difference at the level of SSU rRNA genes is used to define OTU equivalent to bacterial species, while 5% could be used to distinguish genera. The rarefaction curves calculation used for these samples revealed a higher proposed species count for the MDA-amplified library (149 versus 88 species). This can be due to a more sensitive amplification of underrepresented species through the MDA process in comparison to the PCR amplification directly from environmental DNA. This might be in agreement with Gonzalez et al. (14), who demonstrated a more effective amplification of species through MDA combined with SSU rRNA specific PCR versus direct SSU rRNA specific amplification from environmental samples. The differences between the taxonomic groups identified in the two libraries, which occur for groups that have low representation (e.g., alpha-*Proteobacteria*, *Verrucomicrobia*, and some of the uncultured groups), may be due to limited sequence sampling, especially for the native DNA library. The sequence quality of the SSU rRNA amplified library was in general good, and there was no evidence for chimera formation introduced through the MDA amplification.

The SSU rRNA library from sample 3, while of lower diversity, reveals a relatively even distribution of species across several major phyla, including *Proteobacteria*, *Actinobacteria*, *Planctomycetes*, *Verrucomicrobia*, and the candidate division OP11. The lower diversity in library 1 compared to that in library 3 can be explained by the increased depth of sample 1 and differences in the contamination composition. However, we cannot exclude that the extremely small amount of DNA obtained from these samples biases the total number of predicted species. Both library 1 and library 3 contain sequences closely related to *Pseudomonas synxantha*, which is a known reducer of Cr(VI) as well as a hydrocarbon degrader, pointing toward a potential involvement in bioremediation (23, 28, 51). Overall, our data show that MDA can be used to obtain enough DNA from low-biomass samples to evaluate their microbial diversity. For samples 1 and 3, microbial diversity could be analyzed only after MDA. Trace amounts of DNA template in an environmental sample may not be enough to generate an SSU rRNA gene PCR product, and PCR may also be inhibited by chemical inhibitors found in the soils (14). Only after first MDA amplifying the gDNA in the soil samples is PCR possible.

From our validation of the MDA protocol, we have observed very high genome coverage from an isolate and a representative DNA library from a mix of known genomes. A shotgun sequencing approach was used to further test the quality of the libraries constructed from MDA-amplified gDNA. From the sampling of sequences of library clone ends a number of sequences were assembled into contigs. The fact that contigs were revealed within a shallow sampling of library sequences might suggest that there is some amplification bias introduced by MDA. The assembled sequences possibly represent the gDNA areas of overamplification. The great majority of the contigs were comprised of only two sequences. Although some contigs were comprised of more than two sequences, the longest of these contigs contained 10 sequences and was 2,700 nucleotides long (data not shown). This indicates that, whereas there is some bias introduced (as observed in the GeneChip data), there is no major area of overamplification. It is known that degraded or damaged template DNA can decrease yield and increase bias of MDA reactions (7). The damage to the gDNA due to the environmental soil conditions of these samples is not known. To decrease any further shearing of the DNA, an extraction method that maintains high-molecular-weight DNA was used. To further decrease the effect of MDA bias on library cloning, each sample was amplified in triplicate. Each sample's amplified products were then mixed before any further processing or analysis. In this way, any random over- or underamplification in one reaction should be balanced out by the random over- and underamplifications in the other two reactions, the end result being representative genome coverage. Only by sequencing to a far greater depth can areas of underamplification be analyzed. Analysis by BLAST showed that 35.1% of the library sequences had no sequence similarity to any of the known proteins in the database (e-values of >1e-10). These sequences are common in normal MDA-independent genomic libraries (data not shown) (52) and could contain noncoding, intergenic regions, or ORFans (45). Within MDA libraries they could also be the result of MDA artifacts such as primer-derived multimers (similar to PCR primer-dimer artifacts) and direct and inverted sequence repeats. Because the sensitivity of the MDA reaction is known to create DNA products even in the absence of added cells (29), it can be expected to find sequences with low sequence similarity and sequences containing combinations of sequence repeats in MDA-amplified libraries. The data from the limited number of clones sequenced reveal diverse environmental libraries that can be readily screened for native or novel sequences. Analysis of the environmental amplified-DNA libraries showed that 64.9% of the sequences had significant similarities to known proteins.

The similarity of the COG analysis from area 3 (libraries 1 and 3) suggests that the environment may influence the "functional" profile of a community, which had been hypothesized previously (53). The significance of the sensor histidine kinase and two-component regulator COGs present in all contaminated sediments examined and their relationship to stress response are being analyzed further. Overall, the sequence analysis demonstrates that MDA introduces some artifacts but allows the creation of libraries for shotgun sequencing from these extremely low biomass-containing samples. Further studies may reveal whether it is possible to contig whole pathways or genomes, although significant information about the environment can be obtained by identifying environment-specific genes (53).

Our limited sequencing analysis of these contaminated environments is not intended to define the soil metagenome but to give insight into the possibilities provided by WGA using ϕ29 DNA polymerase. By amplifying the extracted DNA from a sample and constructing environmental libraries from the amplified DNAs, it is now possible to access the genome information from contaminated environments that was not previously accessible and is therefore an additional step forward of the work published by Gonzalez et al. (14). Downstream analyses could involve functional screening for active clones or sequence-based screening using probes homologous to known genes. Cell-free cloning may enable the sequencing of toxic genes and sequences recalcitrant to cloning (19). In the future, MDA, combined with the recent advancements in nonelectrophoretic sequencing methods (33, 40), may greatly increase the accessibility of metagenomes from these environments.

Genomic analyses of uncultured microbes can provide significant insight into the biological properties of individuals within microbial populations (9). By examining convergent and divergent sets of proteins and regulatory elements, within niches and across niches, we can better understand subsurface mobilization and immobilization of radionuclides and metals. This will help to manipulate, stabilize, and predict the long-term stabilities of these contaminants and their relative risks.

ACKNOWLEDGMENTS

This study was part of 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 Program: GTL through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy and by the Office of Science (BER), U.S. Department of Energy, Genomics: GTL program, grant DE-FG02-04ER63771.

We thank the sequencing team at Diversa for excellent sequencing support. We also thank Sharon Borglin, Dominique Joyner, Rick Huang, and Tamas Torok at the Lawrence Berkeley National Laboratory for help with the initial sample processing. We thank Mel Simon and Keith Kretz for critical reading of the manuscript.

REFERENCES

1. **Baker, B. J., D. P. Moser, B. J. MacGregor, S. Fishbain, M. Wagner, N. K. Fry, B. Jackson, N. Speolstra, S. Loos, K. Takai, B. S. Lollar, J. Fredrickson, D. Balkwill, T. C. Onstott, C. F. Wimpee, and D. A. Stahl.** 2003. Related assemblages of sulphate-reducing bacteria associated with ultradeep gold mines of South Africa and deep basalt aquifers of Washington State. Environ. Microbiol. **5:**267–277.
2. **Berry, A. E., C. Chiocchini, T. Selby, M. Sosio, and E. M. Wellington.** 2003. Isolation of high molecular weight DNA from soil for cloning into BAC vectors. FEMS Microbiol. Lett. **223:**15–20.
3. **Blanco, L., A. Bernad, J. M. Lazaro, G. Martin, C. Garmendia, and M. Salas.** 1989. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase: symmetrical mode of DNA replication. J. Biol. Chem. **264:** 8935–8940.
4. **Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer.** 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. USA **99:**14250–14255.
5. **Calvó, L., M. Cortey, J. L. García-Marín, and L. J. Garcia-Gil.** 2005. Polygenic analysis of ammonia-oxidizing bacteria using 16S rDNA, *amoA*, and *amoB* genes. Int. Microbiol. **8:**103–110.
6. **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.
7. **Dean, F. B., S. Hosono, L. Fang, X. Wu, A. F. Faruqi, P. Bray-Ward, Z. Sun, Q. Zong, Y. Du, J. Du, M. Driscoll, W. Song, S. F. Kingsmore, M. Egholm, and R. S. Lasken.** 2002. Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. USA **99:**5261–5266.
8. **Dean, F. B., J. R. Nelson, T. L. Giesler, and R. S. Lasken.** 2001. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. **11:**1095–1099.
9. **DeLong, E. F.** 2002. Microbial population genomics and ecology. Curr. Opin. Microbiol. **5:**520–524.
10. **Detter, J. C., J. M. Jett, S. M. Lucas, E. Dalin, A. R. Arellano, M. Wang, J. R. Nelson, J. Chapman, Y. Lou, D. Rokhsar, T. L. Hawkins, and P. M. Richardson.** 2002. Isothermal strand-displacement amplification applications for high-throughput genomics. Genomics **80:**691–698.
11. **Dojka, M. A., P. Hugenholtz, S. K. Haack, and N. R. Pace.** 1998. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. Appl. Environ. Microbiol. **64:**3869–3877.
12. **Edwards, K. J., P. L. Bond, T. M. Gihring, and J. F. Banfield.** 2000. An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. Science **287:**1796–1799.
13. **Felsenstein, J.** 1989. PHYLIP: phylogeny inference package (version 3.2). Cladistics **5:**164–166.
14. **Gonzalez, J. M., M. C. Portillo, and C. Saiz-Jimenez.** 2005. Multiple displacement amplification as a pre-polymerase chain reaction (pre-PCR) to process difficult to amplify samples and low copy number sequences from natural environments. Environ. Microbiol. **7:**1024–1028.
15. **Haglund, A. L., E. Tornblom, B. Bostrom, and L. Tranvik.** 2002. Large differences in the fraction of active bacteria in plankton, sediments, and biofilm. Microb. Ecol. **43:**232–241.
16. **Hall, T. A.** 2005. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41:**95–98.
17. **Hazen, T. C., and H. H. Tabak.** 2005. Developments in bioremediation of soils and sediments polluted with metals and radionuclides. 2. Field research on bioremediation of metals and radionuclides. Rev. Environ. Sci. Biotechnol. **4:**157–183.
18. **Henne, A., R. Daniel, R. A. Schmitz, and G. Gottschalk.** 1999. Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Appl. Environ. Microbiol. **65:**3901–3907.
19. **Hutchison, C. A., III, H. O. Smith, C. Pfannkoch, and J. C. Venter.** 2005. Cell-free cloning using ϕ29 DNA polymerase. Proc. Natl. Acad. Sci. USA **102:**17332–17336.
20. **Koenigsberg, S. S., T. C. Hazen, and A. D. Peacock.** 2005. Environmental biotechnology: a bioremediation perspective. Remediation J. **15:**5–25.
21. **Lasken, R. S., and M. Egholm.** 2003. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. Trends Biotechnol. **21:**531–535.
22. **Liles, M. R., B. F. Manske, S. B. Bintrim, J. Handelsman, and R. M. Goodman.** 2003. A census of rRNA genes and linked genomic sequences within a soil metagenomic library. Appl. Environ. Microbiol. **69:**2684–2691.
23. **McLean, J. S., T. J. Beveridge, and D. Phipps.** 2000. Isolation and characterization of a chromium-reducing bacterium from a chromated copper arsenate-contaminated site. Environ. Microbiol. **2:**611–619.
24. **Nelson, K. E.** 2003. The future of microbial genomics. Environ. Microbiol. **5:**1223–1225.
25. **Nimgaonkar, A., D. Sanoudou, A. J. Butte, J. N. Haslett, L. M. Kunkel, A. H. Beggs, and I. S. Kohane.** 2003. Reproducibility of gene expression across generations of Affymetrix microarrays. BMC Bioinformatics **4:**27–38.
26. **North, N. N., S. L. Dollhopf, L. Petrie, J. D. Istok, D. L. Balkwill, and J. E. Kostka.** 2004. Change in bacterial community structure during in situ biostimulation of subsurface sediment cocontaminated with uranium and nitrate. Appl. Environ. Microbiol. **70:**4911–4920.
27. **Palumbo, A. V., J. C. Schryver, M. W. Fields, C. E. Bagwell, J. Z. Zhou, T. Yan, X. Liu, and C. C. Brandt.** 2004. Coupling of functional gene diversity and geochemical data from environmental samples. Appl. Environ. Microbiol. **70:**6525–6534.
28. **Pattanapipitpaisal, P., A. N. Mabbett, J. A. Finlay, A. J. Beswick, M. Paterson-Beedle, A. Essa, J. Wright, M. R. Tolley, U. Badar, N. Ahmed, J. L. Hobman, N. L. Brown, and L. E. Macaskie.** 2002. Reduction of Cr(VI) and bioaccumulation of chromium by gram-positive and gram-negative microorganisms not previously exposed to Cr-stress. Environ. Technol. **23:**731–745.
29. **Raghunathan, A., H. R. Ferguson, Jr., C. J. Bornarth, W. Song, M. Driscoll, and R. S. Lasken.** 2005. Genomic DNA amplification from a single bacterium. Appl. Environ. Microbiol. **71:**3342–3347.
30. **Rappe, M. S., and S. J. Giovannoni.** 2003. The uncultured microbial majority. Annu. Rev. Microbiol. **57:**369–394.
31. **Reardon, C. L., D. E. Cummings, L. M. Petzke, B. L. Kinsall, D. B. Watson, B. M. Peyton, and G. G. Geesey.** 2004. Composition and diversity of microbial communities recovered from surrogate minerals incubated in an acidic uranium-contaminated aquifer. Appl. Environ. Microbiol. **70:**6037–6046.
32. **Robertson, D. E., J. A. Chaplin, G. DeSantis, M. Podar, M. Madden, E. Chi,**

T. Richardson, A. Milan, M. Miller, D. P. Weiner, K. Wong, J. McQuaid, B. Farwell, L. A. Preston, X. Tan, M. A. Snead, M. Keller, E. Mathur, P. L. Kretz, M. J. Burk, and J. M. Short. 2004. Exploring nitrilase sequence space for enantioselective catalysis. Appl. Environ. Microbiol. **70:**2429–2436.

33. **Rogers, Y. H., and J. C. Venter.** 2005. Genomics: massively parallel sequencing. Nature **437:**326–327.
34. **Rohwer, F., V. Seguritan, D. H. Choi, A. M. Segall, and F. Azam.** 2001. Production of shotgun libraries using random amplification. BioTechniques **31:**108–118.
35. **Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman.** 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. **66:**2541–2547.
36. **Schloss, P. D., and J. Handelsman.** 2004. Status of the microbial census. Microbiol. Mol. Biol. Rev. **68:**686–691.
37. **Schloss, P. D., and J. Handelsman.** 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol. **71:**1501–1506.
38. **Senko, J. M., J. D. Istok, J. M. Suflita, and L. R. Krumholz.** 2002. In-situ evidence for uranium immobilization and remobilization. Environ. Sci. Technol. **36:**1491–1496.
39. **Seow, K. T., G. Meurer, M. Gerlitz, E. Wendt-Pienkowski, C. R. Hutchinson, and J. Davies.** 1997. A study of iterative type II polyketide synthases, using bacterial genes cloned from soil DNA: a means to access and use genes from uncultured microorganisms. J. Bacteriol. **179:**7360–7368.
40. **Shendure, J., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church.** 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science **309:**1728–1732.
41. **Short, J. M.** 1997. Recombinant approaches for accessing biodiversity. Nat. Biotechnol. **15:**1322–1323.
42. **Short, J. M.** September 1999. Protein activity screening of clones having DNA from uncultivated microorganisms. U.S. patent 5,958,672.
43. **Short, J. M.** August 2001. Gene expression library produced from DNA from uncultivated organisms and methods for making the same. U.S. patent 6,280,926.
44. **Short, J. M., J. M. Fernandez, J. A. Sorge, and W. D. Huse.** 1988. Lambda ZAP: a bacteriophage lambda expression vector with in vivo excision properties. Nucleic Acids Res. **16:**7583–7600.
45. **Siew, N., and D. Fischer.** 2003. Twenty thousand ORFan microbial protein families for the biologist? Structure **11:**7–9.
46. **Stein, J. L., T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong.** 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol. **178:**591–599.
47. **Steward, G. F., J. P. Zehr, R. Jellison, J. P. Montoya, and J. T. Hollibaugh.** 2004. Vertical distribution of nitrogen-fixing phylotypes in a meromictic, hypersaline lake. Microb. Ecol. **47:**30–40.
48. **Swofford, D. L.** 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.
49. **Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41–54.
50. **Teitzel, G. M., and M. R. Parsek.** 2003. Heavy metal resistance of biofilm and planktonic *Pseudomonas aeruginosa*. Appl. Environ. Microbiol. **69:**2313–2320.
51. **Thomassin-Lacroix, E. J., Z. Yu, M. Eriksson, K. J. Reimer, and W. W. Mohn.** 2001. DNA-based and culture-based characterization of a hydrocarbon-degrading consortium enriched from Arctic soil. Can. J. Microbiol. **47:**1107–1115.
52. **Treusch, A. H., A. Kletzin, G. Raddatz, T. Ochsenreiter, A. Quaiser, G. Meurer, S. C. Schuster, and C. Schleper.** 2004. Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. Environ. Microbiol. **6:**970–980.
53. **Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin.** 2005. Comparative metagenomics of microbial communities. Science **308:**554–557.
54. **Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37–43.
55. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66–74.
56. **von Mering, C., L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork.** 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. **33:**D433–D437.
57. **Vrionis, H. A., R. T. Anderson, I. Ortiz-Bernad, K. R. O'Neill, C. T. Resch, A. D. Peacock, R. Dayvault, D. C. White, P. E. Long, and D. R. Lovley.** 2005. Microbiological and geochemical heterogeneity in an in situ uranium bioremediation field site. Appl. Environ. Microbiol. **71:**6308–6318.
58. **Yan, T., M. W. Fields, L. Wu, Y. Zu, J. M. Tiedje, and J. Zhou.** 2003. Molecular diversity and characterization of nitrite reductase gene fragments (*nirK* and *nirS*) from nitrate- and uranium-contaminated groundwater. Environ. Microbiol. **5:**13–24.
59. **Zengler, K., G. Toledo, M. Rappe, J. Elkins, E. J. Mathur, J. M. Short, and M. Keller.** 2002. Cultivating the uncultured. Proc. Natl. Acad. Sci. USA **99:**15681–15686.