APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Sept. 2008, p. 5422–5428 Vol. 74, No. 17
0099-2240/08/$08.00+0 doi:10.1128/AEM.00410-08

Novelty and Uniqueness Patterns of Rare Members of the Soil Biosphere▽†

Mostafa S. Elshahed,[1]* Noha H. Youssef,[1] Anne M. Spain,[2] Cody Sheik,[2] Fares Z. Najar,[3] Leonid O. Sukharnikov,[3] Bruce A. Roe,[3] James P. Davis,[1] Patrick D. Schloss,[4] Vanessa L. Bailey,[5] and Lee R. Krumholz[2]

Department of Microbiology and Molecular Genetics, Oklahoma State University, 1110 S. Innovation Way, Stillwater, Oklahoma 74074[1]; Department of Botany and Microbiology and Institute for Energy and The Environment, University of Oklahoma, 770 Van Vleet Oval, Norman, Oklahoma 73019[2]; Department of Chemistry and Biochemistry and the Advanced Center for Genome Technology, University of Oklahoma, 101 David L. Boren Blvd., Norman, Oklahoma 73019[3]; Department of Microbiology, University of Massachusetts, 639 North Pleasant Street, Amherst, Massachusetts 01003[4]; and Microbiology Division, Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, Washington 99354[5]

Received 18 February 2008/Accepted 26 June 2008

Soil bacterial communities typically exhibit a distribution pattern in which most bacterial species are present in low abundance. Due to the relatively small size of most culture-independent sequencing surveys, a detailed phylogenetic analysis of rare members of the community is lacking. To gain access to the rarely sampled soil biosphere, we analyzed a data set of 13,001 near-full-length 16S rRNA gene clones derived from an undisturbed tall grass prairie soil in central Oklahoma. Rare members of the soil bacterial community (empirically defined at two different abundance cutoffs) represented 18.1 to 37.1% of the total number of clones in the data set and were, on average, less similar to their closest relatives in public databases when compared to more abundant members of the community. Detailed phylogenetic analyses indicated that members of the soil rare biosphere either belonged to novel bacterial lineages (members of five novel bacterial phyla identified in the data set, as well as members of multiple novel lineages within previously described phyla or candidate phyla), to lineages that are prevalent in other environments but rarely encountered in soil, or were close relatives to more abundant taxa in the data set. While a fraction of the rare community was closely related to more abundant taxonomic groups in the data set, a significant portion of the rare biosphere represented evolutionarily distinct lineages at various taxonomic cutoffs. We reason that these novelty and uniqueness patterns provide clues regarding the origins and potential ecological roles of members of the soil's rare biosphere.

Compared to multicellular plants and animals, surveying the biodiversity of microorganisms represents a unique challenge. Analysis of species distribution patterns usually indicates that while a significant fraction of bacterial biomass belongs to a relatively small number of species, the majority of bacterial species within a complex microbial community are present in extremely low numbers (2, 24, 32). Consequently, statistical approaches that estimate species richness of complex environments from sampled data sets often provide estimates that are orders of magnitude higher than the observed number of identified species within the data set (9, 13, 27, 29). This fraction that represents a minority of the biomass yet a majority of the genomic diversity has recently been described as the rare biosphere (32).

Within virtually all ecosystems, little information is currently available on the composition, origins, dynamics, and ecological roles of members of the rare biosphere. To investigate these issues, a detailed analysis examining the presence and prevalence of previously unidentified lineages within the rare biosphere (i.e., novelty) and the phylogenetic and evolutionary relationships between rare and abundant members within a specific microbial community (i.e., uniqueness) is needed. Examination of the novelty and uniqueness of the rare biosphere in a specific habitat will obviously require extensive sampling efforts to access this fraction of the community. Pyrosequencing and other sequence tag-based approaches produce hundreds of thousands of short sequences (100 to 250 bp in length) (2, 14, 27, 32). As such, these data sets are helpful for comparative analysis of bacterial communities (19), species richness, and coverage estimates, as well as for an overall description of phylum- and class-level diversity. However, accurate sorting into bins of a specific sequence is often unreliable upon using short fragments, except for queries with high similarity to sequences currently available in public databases. The short amplicon size produced severely limits the utility of these sequences in satisfactorily documenting the presence of novel lineages, and hence sequences with low database similarity are usually sorted into "unclassified" or "other" categories in these studies (14, 27). In addition, the level of sequence divergence, and hence operational taxonomic unit (OTU) assignments, at various taxonomic cutoffs is not always comparable between pyrosequenced fragments and near-full-length 16S rRNA gene sequences (see File S1 in the supplemental material).

To avoid these limitations, we chose to utilize a capillary sequencing-based approach to investigate the nature of the rare biosphere in soil. Soil is an extremely valuable ecosystem for economic sustainability as well as for global nutrient cycling. The soil microbial community is extremely complex, and

* Corresponding author. Mailing address: Oklahoma State University, Department of Microbiology and Molecular Genetics, 1110 S. Innovation Way, Stillwater, OK 74074. Phone: (405) 744-3005. Fax: (405) 744-1112. E-mail: Mostafa@okstate.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 7 July 2008.

rRNA gene clone library-based estimates of species diversity range between 3,000 and 52,000 (27, 35). Soil is also one of the most intensively sampled ecosystems: the RDP project II release 9.56 (November 8, 2007) lists at least 77,692 16S rRNA gene sequences originating from soil. However, the vast majority of soil surveys generated small-sized clone libraries (less than 500 clones), with the exception of a composite library of 1,700 clones and a 1,033-clone library from Minnesota farm soil and Alaskan soil, respectively (29, 35). As such, the current collection of long (>300-bp) 16S rRNA gene sequences available in public databases can be regarded as a vast global survey of numerically abundant microorganisms in various soils as well as other habitats.

Here, we report our analysis of 13,001 near-complete 16S rRNA gene clones from an undisturbed tall grass prairie soil in central Oklahoma. The data set is one-fourth the size of the largest pyrosequencing-based soil data set recently reported from a boreal forest in northwestern Canada (27) and eight times the size of the largest published near-complete 16S rRNA gene clone library, which was derived from Minnesota farm soil (35). The analysis describes the novelty and uniqueness patterns observed within the community and suggests how these observed patterns could hold clues regarding the origins and potential ecological roles of rare members of the soil biosphere.

MATERIALS AND METHODS

Study site. Samples were collected from an undisturbed tall grass prairie preserve in Kessler Farm Field Laboratory Biological Research Station in central Oklahoma in November 2005. The research station is located in McClain County, OK (34°58′31.74″N, 97°31′18.05″W), approximately 40 km southwest of the University of Oklahoma campus in Norman. The field site is an old-field tall grass prairie abandoned from agriculture 30 years ago and has not been grazed for 20 years. Kessler Farm soil (KFS) is dominated by three C4 grass species (*Schizachyrium scoparium* [Little bluestem], *Sorghastrum nutans* [Indian grass], and *Eragrostis curvula* [Weeping lovegrass]) and two C3 grass species (*Ambrosia psilostachya* [Western ragweed] and *Xanthocephalum texanum* [Texas snakeweed]) (37). The mean annual temperature is 16.3°C, with January being the coldest month (3.3°C) and July the warmest (28.2°C). The soils are associated with the Nash-Lucien complex, which is characterized by a low permeability rate, high available water holding capacity, and moderately penetrable root zone (22). The soil was analyzed at the Soil, Water, and Forage Analytical Laboratory at Oklahoma State University in Stillwater. The sample had a pH of 7.5. Nitrate, sulfate, bicarbonate, chloride, boron, sodium, calcium, magnesium, potassium, iron, zinc, and copper concentrations were (in ppm) 1, 13, 131, 7, 0.031, 4, 44, 11, 0, 6.8, 2.33, and 6.81, respectively. The total organic matter in the KFS sample was 0.88%, and the total N was 0.07%.

Sampling, DNA extraction, PCR amplification, library construction, and sequencing. The top 5 cm of soil was scooped using a sterile spatula into a sterile 50-ml Falcon tube, stored on dry ice, and transferred to the laboratory, where it was stored at −20°C. The sample did not contain any grass or apparent root structures. DNA was extracted using a FastDNA spin kit for soil (Bio 101 Corp., Vista, CA). A near-complete 16S rRNA gene fragment was amplified using the primer pair 27F (AGAGTTTGATCCTGGCTCAG) and 1391R (GACGGGCG GTGWGTRCA) (17) in a 50-μl reaction mixture containing (final concentrations) 2 μl of extracted DNA, 1× PCR buffer (Invitrogen), 2.5 mM $MgSO_4$, 0.2 mM deoxynucleoside triphosphate mixture, 2.5 U of platinum *Taq* DNA polymerase (Invitrogen), and 10 μM of each of the forward and reverse primers. PCR amplification was carried out according to the following protocol: initial denaturation for 5 min at 95°C, followed by 20 cycles of denaturation at 95°C for 45 s, annealing at 55°C for 45 s, and elongation at 72°C for 1.5 min, and a final elongation step at 72°C for 15 min was included. PCR products obtained were cloned into a TOPO-TA cloning vector according to the manufacturer's instructions (Invitrogen Corp., Carlsbad, CA) and sequenced at the Department of Energy Joint Genome Institute (Walnut Creek, CA) as previously described (35).

Phylogenetic analysis. The data set was initially run through the RDP classifier (36), and each clone was sorted into bins based on the resulting preliminary taxonomic affiliations. Each RDP classifier-generated bin was treated as a single data set and aligned using Greengenes NAST-aligner to a 7,682-character global alignment (7). Sequences were assigned to phyla and candidate phyla according to the Hugenholtz taxonomy framework (8), as well as by importing them to the Greengenes May 2007 ARB database in the ARB software package (version 06.03.22) (21), and determining their position after parsimony insertion into the universal ARB dendrogram. Sequences with less than 90% sequence identity with their closest relatives were further probed by comparing them to entries in the GenBank nr database using the BLASTn function (1). The combined use of Greengenes classifier, BLAST, and ARB resulted in sorting of all sequences into phyla and candidate phyla except for 15 sequences putatively identified as members of novel candidate phyla. Potential chimeric sequences within the data sets were identified from NAST-aligned sequences using the program Mallard (3), and chimeric sequences were removed from the data set. Distance matrices from aligned, chimera-checked sequences were generated using the "create distance matrix" function on the Greengenes web server. The resulting distance matrices were used to generate OTUs at different taxonomic cutoffs using the DOTUR program (28). The rarefaction curve for the entire KFS data set was constructed using the Analytic Rarefaction software available from the University of Georgia Stratigraphy Laboratory with the cumulative DOTUR output of all bacterial phyla. Chao and ACE estimates of species richness were calculated using the program EstimateS (5).

Phylogenetic trees were constructed by importing KFS NAST-aligned sequences to the May 2007 ARB database in the ARB software package (21). Sequences were initially inserted to the universal ARB dendrogram using the ARB parsimony function, and phylogenetic trees were subsequently constructed from KFS sequences and closely related sequences using the ARB neighbor-joining (ARB-NJ) method with a Lane mask filter (17). Phylogenetic affiliations of novel phyla and subphyla lineages were further evaluated by exporting aligned sequences of KFS as well as their closest relatives from the ARB database into the PAUP 4.01 software package (Sinauer Associates, Sunderland, MA). Evolutionary distance trees and maximum parsimony trees were constructed from the data set, and the bootstrap values (100 replicates) were determined. Novel phyla described in this study remained monophyletic, with >50% bootstrap support upon using all previously described tree-building approaches, as well as the alteration of the composition and size of the data set used for phylogenetic analysis. Each new candidate phylum had at least two sequences and was unaffiliated with any of the previously described bacterial phyla and candidate phyla (15).

qPCR. DNA was extracted from the KFS soil samples in triplicate using the MoBio UltraClean soil DNA isolation kit and then purified with the MoBio Power Clean DNA cleanup kit (MoBio, Carlsbad, CA). Extracts were pooled into one sample prior to use in quantitative PCR (qPCR). qPCR was performed in triplicate using a MyIQ real-time PCR system (Bio-Rad, Hercules, CA). The general bacterial 16S rRNA gene primers EUB338F (5′-ACTCCTACGGGAG GCAGCA) and EUB518R (5′-ATTACCGCGGCTGCTGG) (11) were used for amplification. Each reaction mixture (25-μl total volume) consisted of 12.5 μl IQ SYBR Green Supermix (Bio-Rad), 8.5 μl of water, 0.75 μl of each primer (Invitrogen), and 2.0 μl of DNA. qPCR conditions were 5 min at 95°C, followed by 40 cycles of 95°C for 30 s, 54°C for 30 s, and 72°C for 30 s. Tenfold serial dilutions of pCR4-TOPO plasmid (Invitrogen Corp., Carlsbad, CA) containing a EUB338/EUB518 PCR fragment amplified from the *Escherichia coli* 16S rRNA gene (strain K-12 MG1655) were used to construct standard curves. The copy number in the KFS was determined from the standard curve and subsequently standardized to copy numbers per gram of dry soil.

Novelty estimates. Novelty (sequence divergence between a specific OTU and its closest relative in public databases) was determined by identifying the closest relative of each OTU within the chimera-checked, near-full-length 151,925 16S rRNA gene sequences available in the Greengenes database (August 2007) using the Classifier program (8). As well, novelty was also determined against the noncurated and frequently updated collection of complete and partial 16S rRNA gene sequences available in the GenBank nr database in September 2007 (see File S1 in the supplemental material).

Defining rare and abundant species in Kessler farm soil. We reasoned that OTUs labeled as rare in the KFS data set should represent OTUs with a low probability of being encountered in average-sized clone libraries. Using the formula $P = 1 - (1 - x)^y$, where P is the probability of detecting a species with relative abundance x in the large data set in a small data set of size y (33), we determined the probability of encountering OTUs with different occurrences in the KFS data set in smaller clone libraries. OTUs occurring once in the KFS data set have only a 0.77% probability of being encountered in a 100-clone library. Similarly, OTUs with two, three, four, and five clones have 1.53, 2.28, 3.03, and 3.77% chance of being encountered in a 100-clone library. Therefore, while not

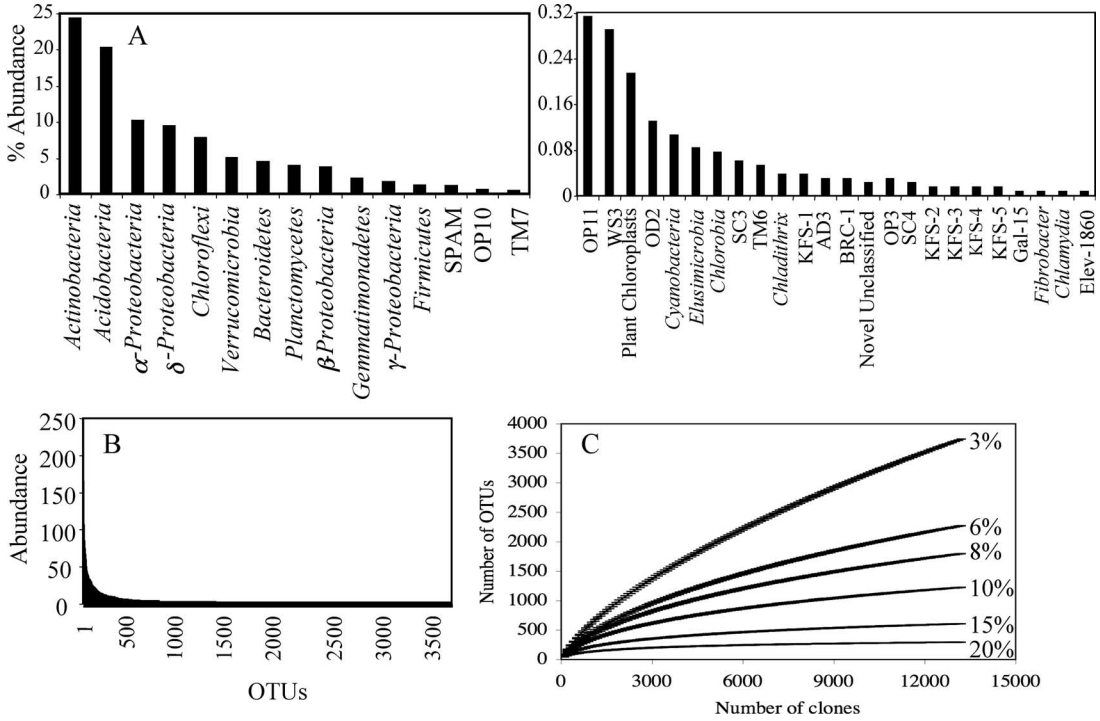

FIG. 1. Kessler Farm soil overall library composition. (A) Distribution of various phyla in KFS. (B) Species distribution pattern of KFS $OTU_{0.03}$ asignments. (C) Rarefaction curve at different taxonomic cutoffs.

counting each microorganism present in KFS soil, we decided to empirically define rare species at a lower cutoff of $n = 1$ and a higher cutoff of $n \leq 5$. In this sense, OTUs present at such low abundances have rarely been sampled from soils and other environments.

Nucleotide sequence accession numbers. Sequences obtained in this study have been deposited in the GenBank database under accession numbers EU131915 to EU135578.

RESULTS

Kessler Farm soil bacterial community composition. Out of the 13,486 sequences from KFS, 485 sequences were putatively identified as chimeras and were removed from the data set. Using sequence similarity thresholds of 97, 94, 92, 90, 85, and 80% (6), the KFS clone library contained 3,747, 2,314, 1,685, 1,231, 614, and 305 putative species, genera, families, orders, classes, and phyla, respectively. However, using these empirical thresholds for taxonomic placement greatly overestimates the number of taxa at higher taxonomic orders (classes and phyla), since the level of sequence divergence between well-established classes and phyla often exceeds the 15 and 20% empirical thresholds (see File S3 in the supplemental material).

Detailed phylogenetic analysis grouped KFS clones into 34 different phyla and subphyla (Fig. 1A). Of these, 15 phyla have cultured representatives, 14 are previously described candidate phyla with no cultured representatives, and 5 phyla are novel (novel candidate phyla KFS1 to KFS5; see below). The phylum-level diversity in KFS is much higher than in previously reported data sets, and even higher than the total number (32 phyla) collectively compiled by Janssen (16) from a large number of 16S soil studies. However, this is not necessarily a reflection of a higher-than-expected phylum-level diversity in KFS, since it could be attributed to the ability of larger data sets to identify phyla present in extremely low abundance. In KFS, 24 phyla were present at less than 1% abundance, 14 phyla were represented by less than five clones, and 4 phyla were represented by a single clone.

Proportion of rare clones in the KFS data set. The large KFS data set of near-complete 16S rRNA gene sequences represents a unique opportunity to compare and contrast the phylogenetic diversity, novelty, and sequence relationships between rare and abundant members of the KFS bacterial community. The species distribution pattern (Fig. 1B) indicates that a large number of the identified OTUs at the 0.03 sequence divergence cutoff ($OTU_{0.03}$) are present in low abundance, and rarefaction curve analysis (Fig. 1C), as well as ACE and Chao estimators of species diversity at 97% sequence similarity (8,654 ± 210, and 10,519 ± 85, respectively), suggest that the sampling effort did not identify all bacterial species in KFS. Therefore, low-abundance OTUs within the data set (even those occurring only once) do not represent the rarest species in KFS. qPCR analysis using bacteria-specific primers enumerated 5.5×10^7 cells/g of soil. As such, a 13,001-clone library samples 1 out of every 282 to 4,230 cells/g of soil (since rRNA operon copy numbers could range between 1 and 15 copies/genome) (34).

We used two empirically defined clone abundance cutoffs of $n = 1$ and $n \leq 5$ clones to define rare $OTU_{0.03}$ within the KFS data set (see Materials and Methods). Using these values, the percentage of rare species in KFS ranges between 18.1% ($n = 1$) and 37.1% ($n \leq 5$) of the total number of KFS clones (Table 1). Interestingly, the proportion of rare species varied widely among various major bacterial phyla in KFS (defined as phyla represented by >500 clones), with *Planctomycetes* having the

TABLE 1. Clones belonging to rare $OTU_{0.03}$s in the entire data set as well as in bacterial phyla represented by more than 500 clones in the data set

Phylum or subphylum	% Clones belonging to rare OTUs[a]
Entire KFS data set	18.1–37.1
Proteobacteria	18.2–35.7
Actinobacteria	15.7–32.1
Acidobacteria	12.0–25.0
Alphaproteobacteria	18.9–37.4
Deltaproteobacteria	15.1–34.9
Chloroflexi	21.2–47.0
Bacteriodetes	14.9–33.4
Verrucomicrobia	13.4–31.6
Planctomycetes	33.4–77.8
Betaproteobacteria	17.7–28.9

[a] Rare OTUs were defined at two empirical cutoffs, $n = 1$ and $n \leq 5$.

highest percentage of rare species (34.4 to 77.8% of total *Planctomycetes* clones) and *Acidobacteria* the lowest (12.0 to 25.0% of total *Acidobacteria* clones).

Novelty of rare versus abundant members of the KFS bacterial community. We identified the closest relative of each $OTU_{0.03}$ within the data set among the 151,925 near-complete, chimera-checked sequences available in the Greengenes database (August 2007). The results showed that while the rare OTUs vary greatly in the degree of difference to their closest relatives, a general trend is observed in which the level of sequence divergence decreases as the number of clones per $OTU_{0.03}$ increases (Fig. 2A). Averaging the percent sequence divergence at each frequency of $OTU_{0.03}$ occurrence showed a direct relationship between rarity and novelty ($r = 0.57$) (Fig. 2B). A rare $OTU_{0.03}$ represented by a single clone had an average sequence divergence of 10.4% from its closest relative in the Greengenes database. Similarly, the $OTU_{0.03}$s with two, three, four, and five clones had an average sequence divergence of 8.5, 8.2, 7.6, and 7.4% from their closest Greengenes database relatives, respectively. On the other hand, abundant $OTU_{0.03}$s that were represented by >50 clones had 0.8 to 5.9% sequence divergence from their closest Greengenes database relatives. Three notable exceptions were $OTU_{0.03}$s FFCH5010, FFCH15698, and FFCH9674, which contained a large number of clones (65, 67, and 67, respectively) but had low similarities (15.7%, 11.9%, and 14.3%) to their closest relatives. Interestingly, these three OTUs belong to two novel lineages within the *Deltaproteobacteria* and are collectively responsible for the overrepresentation of this group in KFS compared to most of the previously reported soil data sets. A similar pattern was obtained when using the BLAST nr database (September 2007) (1) instead of Greengenes to plot the frequency of occurrence versus sequence divergence (see File S2 in the supplemental material).

Novel phylogenetic diversity in KFS rare biosphere. The community exhibited a fairly typical overall phylum-level distribution pattern, with the nine major phyla often encountered in soil (*Proteobacteria, Actinobacteria, Acidobacteria, Chloroflexi, Verrucomicrobia, Bacteroidetes, Planctomycetes, Gemmatimonadetes*, and *Firmicutes* [16]), representing 95.7% of the KFS clones. However, a meta-analysis of multiple soil clone libraries (16) and a recently published study that quantified six different major bacterial phyla in 71 unique soils by using group-specific qPCR primers (10) clearly indicate that the proportions of these phyla differ widely among different soils. Compared to data reported in these studies (10, 16), KFS appears to be comparatively rich in *Actinobacteria*, *Deltaproteobacteria*, and members of candidate division SPAM (see File S4 in the supplemental material).

Detailed phylogenetic analysis of the KFS data set identified multiple novel lineages at the phylum and subphylum level. Except for three OTUs within the *Deltaproteobacteria*, all novel phylum- and class-level lineages identified were present in low abundance (see Files S4 to S6 in the supplemental material). Fifteen clones (14 OTUs) could not be placed within any of the currently described phyla in the Hugenholtz taxonomy framework in the Greengenes database (8) and as such, whether alone or with few previously unaffiliated sequences, could be grouped into five novel candidate divisions designated KFS1 to KFS5 (Fig. 3). In addition, we speculate that the future availability of sequences related to OTUs FFCH894, FFCH16611, and FFCH9315 might result in recognizing them as three additional novel bacterial phyla (Fig. 3).

A detailed description of the phylogenetic affiliation of KFS OTUs belonging to previously recognized phyla and candidate phyla is presented as supplemental material (see Files S4 to S6 in the supplemental material). The analysis revealed an impressive level of novel subphylum- and order-level diversity among almost all major bacterial soil phyla (with the exception of the *Gammaproteobacteria* and the *Firmicutes*). The presence of novel KFS subphyla was not restricted to the relatively less-studied phyla that appear to be prevalent only in soils (*Acidobacteria*, *Verrucomicrobia*, and *Gemmatimonadetes*) but extended to other major bacterial phyla with global distributions and numerous cultured representatives (*Actinobacteria*, *Proteobacteria*, and *Bacteroidetes*).

In addition, within the rare members of the KFS data set, we identified several clones that, although belonging to previously described lineages, have rarely been encountered in soils. Examples include clones belonging to the phyla *Chlorobia, Calditrix*,

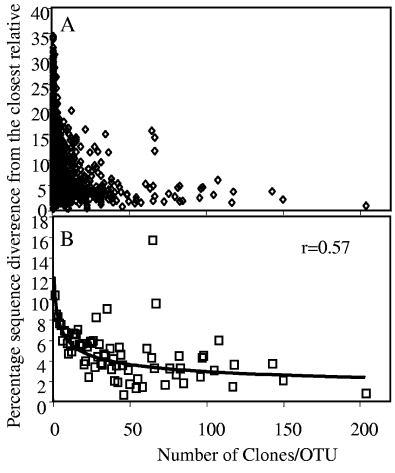

FIG. 2. (A) Correlation between novelty and abundance of clones within each $OTU_{0.03}$ identified in the KFS data set. (B) Correlation between average percent sequence divergence and abundance of clones within KFS $OTU_{0.03}$ assignments.

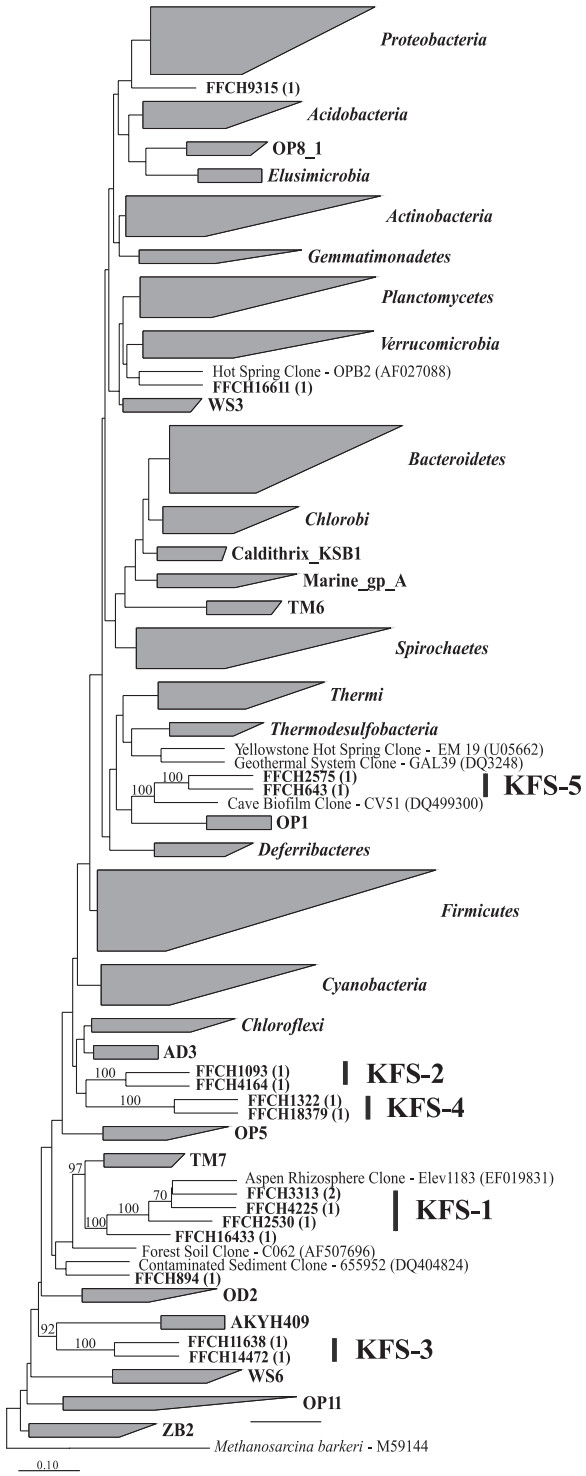

FIG. 3. Distance NJ tree highlighting the phylogenetic position of five novel candidate phyla identified in KFS data set. The tree was constructed from 1,643 aligned sequences using the ARB-NJ method with Olsen correction and a Lane mask filter. Bootstrap values are based on 1,000 replicates and are shown for novel candidate phyla branches.

Elusimicrobia, candidate phylum BRC-1, clones affiliated with the genus *Salinibacter* within the *Bacteroidetes*, and *Clostridiales*-affiliated clones, as well as clones belonging to the Sup-05 lineage within the *Gammaproteobacteria.* Interestingly, many of these

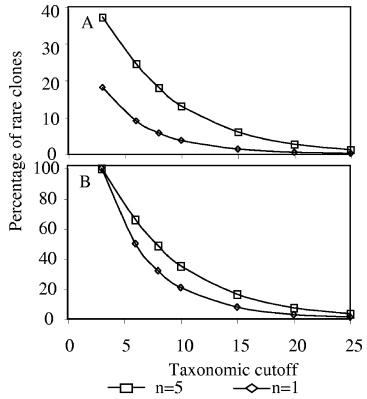

FIG. 4. Quantification of the proportion of unique clones within the KFS rare biosphere. Rare OTUs are defined at two cutoffs: those containing a single clone ($n = 1$) and those containing five or fewer clones ($n \leq 5$). (A) Percentage of clones belonging to rare OTUs at different taxonomic cutoffs expressed as a fraction of the total number of clones in the KFS data set (13,001). (B) Percentage of clones belonging to rare OTUs at different taxonomic cutoffs expressed as a fraction of the number of clones belonging to rare OTUs at a 97% taxonomic cutoff ($OTU_{0.03}$).

clones belonged to lineages requiring specific environmental conditions (strict anaerobic conditions, high salt, high temperature) that are usually not prevalent in soil ecosystems.

Proportion of unique clones among rare members of the KFS bacterial community. We quantified the percentage of clones that belong to rare OTUs ($n = 1$ and $n \leq 5$) within the KFS clone library at different taxonomic cutoffs (3, 6, 8, 10, 15, 20, and 25% sequence divergence). The proportion of clones that remains assigned to rare OTUs at higher sequence divergence cutoffs represents unique clones with no close relatives among more abundant members of the community. Similarly, the drop in the number of clones assigned to rare OTUs at higher sequence divergence cutoffs represents the fraction of rare clones with close relatives among more abundant members of the community. Clones within OTUs identified as rare at a putative genus level (6% sequence divergence) represented 9.1 to 24.6% of the total KFS clones (Fig. 4A) and 50.1 to 66.1% (at $n = 1$ and $n \leq 5$) of the rare clones at the putative species level ($OTU_{0.03}$) (Fig. 4B). At the putative class level (15% sequence divergence), clones within OTUs identified as rare represented only 1.4 to 6.1% of the total KFS clones (Fig. 4A) and only 7.9 to 16.3% of the rare clones at the putative species level (Fig. 4B). These results indicate that while a fraction of the rare biosphere is closely related to more abundant species, a significant fraction is unique and represents evolutionary distinct lineages within KFS biosphere.

DISCUSSION

In this study, we examined the phylogenetic diversity in a 13,001-clone library derived from an undisturbed tall grass prairie site in central Oklahoma. We used the data set to access low-abundance, rarely sampled microbial species in soil and examined their phylogenetic affiliations, similarity to current global 16S rRNA gene inventories, and relationship to more abundant members of the soil microbial community.

Based on our evaluation of the novelty of rare clones (Fig.

2), their phylogenetic affiliations (Fig. 3; see also Files S4 to S6 in the supplemental material), as well as their relationships to more abundant members of the community (Fig. 4), we broadly identify two main groups within the KFS rare biosphere: those with close relatives among the more abundant members of the KFS bacterial community, and those that belong to unique, phylogenetically distinct lineages with no close sequence similarity to more abundant members of KFS. Using 15% sequence divergence from the closest abundant relative within the KFS data set as an empirical "uniqueness" cutoff, members of group I represent 83.6 to 92.1% and members of the second group represent 7.9 to 16.4% of the total number of rare KFS clones (at $n = 1$ and $n \leq 5$, respectively) (Fig. 4B). Similar to more abundant members of the community, members of group I belong to common, well-described, and well-sampled soil lineages. On the other hand, members of the unique group II usually belong to novel bacterial phyla, novel lineages within previously described phyla and candidate phyla, or are members of lineages that are ubiquitous in specific environments but rarely encountered in soils. We reason that these novelty and uniqueness patterns provide clues regarding the origins and potential ecological roles of members of the soil's rare biosphere.

The close sequence similarity between nonunique members of the rare biosphere (group I) and dominant OTUs within the community argues against an old, evolutionary distinct origin for this fraction of the rare biosphere, as previously suggested (32). Various lines of ecological evidence suggest that this group of nonunique, nonnovel members of the rare biosphere acts as a backup system and readily responds to seasonal variations encountered in soil temperature, pH, light exposure, and nutrient levels. The constant seasonal promotion of some members of group I rare species to be members of the dominant (and hence readily identifiable) taxa in soil, together with the seasonal demotion of some of the more abundant taxa in soil, is probably responsible for the observation that seasonal variations often result in significant changes in the phylogenetic affiliations of the most numerous members of the community, leading to statistically detectable differences between seasonal clone libraries from the same soil (12, 18, 30, 31). However, these seasonal cyclic changes rarely affect the fundamental soil community structure, and in all seasons, the sampled soils will still have their distinctive community composition (16). We also reason that this fine-tuning function of group I of the rare biosphere is responsible for the fact that within the thousands of soil studies conducted so far (see reference 16 for a review), no two clone libraries have had exactly the same community composition, and exact (100%) sequence matches between the most abundant species and database-deposited sequences (that broadly represent a global repository of more abundant members of soil and other communities) are very rarely encountered. The variations in physical and geochemical characteristics between different soils always select for different species as the most dominant members of the community. Therefore, in all soil surveys reported so far, dominant species identified almost always belong to a previously unencountered strain, species, or genus within well-recognized soil lineages (and hence the tail end of the curve in Fig. 2B never reaches zero).

Within group II of the rare biosphere in the KFS data set (rare bacterial taxa with no close relatives within the dominant species), a fraction belongs to well-described phylogenetic lineages that are prevalent in other ecosystems but are rarely encountered in soil (phyla *Chlorobia*, *Caldithrix*, *Elusimicrobia* candidate phylum BRC-1, *Salinibacter*, and *Clostridiales*-affiliated clones, and clones belonging to the Sup-05 lineage within the *Gammaproteobacteria*) (see Files S4 to S6 in the supplemental material). In addition, we speculate that since members of this group are present in a far less than ideal habitat, the majority will be present in an extremely low number and escaped detection in this study. We suggest that taxa belonging to this fraction of group II of the rare biosphere (together with other species recruited via immigration) respond to more drastic disturbances that could occur in the ecosystem. For example, desertification has been shown to consistently result in an increase in the numbers of organisms from the *Deinococcus-Thermus* group (26), which is otherwise rarely detected in other soil ecosystems. A change in redox potential could regenerate (among other changes) *Clostridiales*-affiliated cells (or spores) present in KFS in low abundance. More drastic and sustained disturbances (e.g., the occurrence of a major hydrocarbon spill and the development of anaerobic conditions in soil) initiate more radical promotion, demotion, and recruitment processes, resulting in a completely different community composition.

Finally, a fraction of group II of the rare biosphere belongs to novel bacterial lineages (phyla and subphyla) with no close relatives in the entire global 16S rRNA gene data set currently available (members of candidate phyla KFS1 to KFS5 and novel lineages within different bacterial phyla and candidate phyla) (see Files S3, S5, and S6 in the supplemental material). The ecological role of members of these novel, unique lineages is not yet known. It has been suggested that members of this group fulfill specific crucial, yet unknown functions within soil ecosystems (14, 32). Alternatively, this fraction of the rare biosphere might represent remnants of microbial evolution that, although currently out-competed in all global ecosystems, possess an exceptional ability to survive and escape extinction.

The comprehensive data set obtained in this study should prove extremely valuable in future research aimed at understanding community dynamics in response to environmental fluctuations, as well as a starting point for elucidating the physiological capabilities and metabolic potential of the numerous novel, as-yet-uncultured lineages in the rare biosphere. We are currently evaluating the effect of simulated global warming on the dynamics of the KFS bacterial community at different levels of phylogenetic resolution, ranging from the phylum level (using quantitative PCR) to the species level (using Phylochip, a comprehensive 16S rRNA gene microarray [4]). Further, targeted metagenomic approaches such as fluorescent in situ hybridization coupled to fluorescence-activated cell sorting and multiple displacement amplification, or microfluidic separation of individual cells coupled to multiple displacement amplification, are two promising approaches that could help in elucidating the metabolic potential of novel yet-uncultured groups with low abundance in soil and other complex environments (23, 25).

This work provides an overall assessment of the phylogenetic diversity and evolutionary relationships between rare and more abundant members of the soil biosphere. The data demonstrate that even in extensively studied habitats, the rare biosphere harbors novel lineages (with no representatives in

the database) and unique lineages (that are evolutionary distinct, dissimilar to more abundant members of the community). We anticipate that similar efforts in different soils will greatly expand our understanding of the nature of the soil rare biosphere. Similarly, future efforts examining the rare biosphere in ecosystems currently estimated to have a higher level of yet-unexplored bacterial diversity (see reference 20 for a list of these environments) will greatly expand our understanding of the phylum-level global bacterial diversity. The identification of multiple novel bacterial phyla and subphyla within one of the most intensively studied and sampled habitats on earth clearly indicates that while the probability of identifying novel bacterial groups within numerically abundant members of various microbial communities appears to be nearing saturation, the potential of identifying novel lineages, genes, and genomes, as well as potentially novel metabolic abilities and microbial secondary metabolites within the rare biosphere, is just starting to be realized.

ACKNOWLEDGMENTS

This work has been supported by the National Science Foundation Microbial Observatories program (grant no. MCB_0240683), the DOE Small Laboratory Science Program (V.L.B.), and the Oklahoma State University Start-up fund (M.S.E.).

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Ashby, M. N., J. Rine, E. Mongodin, K. E. Nelson, and D. Dimster-Denk.** 2007. Serial analysis of rRNA genes and the unexpected dominance of rare members of microbial communities. Appl. Environ. Microbiol. **73:**4532–4542.
3. **Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman.** 2006. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Appl. Environ. Microbiol. **72:**5734–5741.
4. **Brodie, E. L., T. Z. DeSantis, D. C. Joyner, S. M. Baek, J. T. Larsen, G. L. Andersen, T. C. Hazen, P. M. Richardson, D. J. Herman, T. K. Tokunaga, J. M. Wan, and M. K. Firestone.** 2006. Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl. Environ. Microbiol. **72:**6288–6298.
5. **Colwell, R. K.** 2006. EstimateS: statistical estimation of species richness and shared species from samples, version 8. http://viceroy.eeb.uconn.edu/estimateS.
6. **DeSantis, T. Z., E. L. Brodie, J. P. Moberg, I. X. Zobieta, Y. M. Piceno, and G. L. Anderson.** 2007. High-density universal 16S rRNA microarray analysis reveals broader diversity that typical clone libraries when sampling the environment. Microb. Ecol. **53:**371–383.
7. **DeSantis, T. Z., P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen.** 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acid Res. **34:**D394–D399.
8. **DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Anderson.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. **72:**5069–5072.
9. **Dunbar, J., S. M. Barns, L. O. Ticknor, and C. R. Kuske.** 2002. Empirical and theoretical bacterial diversity in four Arizona soils. Appl. Environ. Microbiol. **6:**3035–3045.
10. **Fierer, N., M. A. Bradford, and R. B. Jackson.** 2007. Toward an ecological classification of soil bacteria. Ecology **88:**1354–1364.
11. **Fierer, N., J. A. Jackson, R. Vilgalys, and R. B. Jackson.** 2005. Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. Appl. Environ. Microbiol. **71:**4117–4120.
12. **Griffiths, R. I., A. S. Whiteley, A. G. O'Donnell, and M. J. Bailey.** 2003. Influence of depth and sampling time on bacterial community structure in an upland grassland soil. FEMS Microbiol. Ecol. **43:**35–43.
13. **Hong, S.-H., J. Bunge, S.-O. Jeon, and S. S. Epstein.** 2006. Predicting microbial species richness. Proc. Natl. Acad. Sci. USA **103:**117–122.
14. **Huber, J. A., D. B. M. Welch, H. G. Morrison, S. M. Huse, P. R. Neal, D. A. Butterfield, and M. L. Sogin.** 2007. Microbial population structures in the deep marine biosphere. Science **318:**97–100.
15. **Hugenholtz, P., B. M. Goebel, and N. R. Pace.** 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. **180:**4765–4774.
16. **Janssen, P. H.** 2006. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Appl. Environ. Microbiol. **72:**1719–1728.
17. **Lane, D. J.** 1991. 16S/23S rRNA sequencing, p. 115–174. *In* E. Stackebrandt and M. Goodfellow (ed.), Nucleic acid techniques in bacterial systematics. John Wiley & Sons, Chichester, United Kingdom.
18. **Lipson, D. A., and S. K. Schmidt.** 2004. Seasonal changes in an alpine soil bacterial community in the Colorado Rocky Mountains. Appl. Environ. Microbiol. **70:**2867–2879.
19. **Liu, Z., C. Lozupone, M. Hamady, F. D. Bushman, and R. Knight.** 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. **35:**120–130.
20. **Lozupone, C. A., and R. Knight.** 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. USA **104:**11436–11440.
21. **Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Y. kumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Förster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lüßmann, 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. **Luo, Y., S. Wan, D. Hui, and L. L. Wallace.** 2001. Acclimatization of soil respiration to warming in a tall grass prairie. Nature **413:**622–625.
23. **Marcy, Y., C. Ouverney, E. M. Bik, T. Losekann, N. Ivanova, H. G. Martin, E. Szeto, D. Platt, P. Hugenholtz, and S. R. Quake.** 2007. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc. Natl. Acad. Sci. USA **140:**11889–11894.
24. **Pedros-Alio, C.** 2006. Marine microbial diversity: can it be determined. Trends Microbiol. **14:**257–263.
25. **Podar, M., C. B. Abulencia, M. Walcher, D. Hutchison, K. Zengler, J. A. Garcia, T. Holland, D. Cotton, L. Hauser, and M. Keller.** 2007. Targeted access to the genomes of low abundance organisms in complex microbial communities. Appl. Environ. Microbiol. **73:**3205–3214.
26. **Rainey, F. A., K. Ray, M. Ferreira, B. Z. Gatz, M. F. Nobre, D. Bagaley, B. A. Rash, M.-J. Park, A. M. Earl, N. C. Shank, A. M. Small, M. C. Henk, J. R. Battista, P. Kämpfer, and M. S. da Costa.** 2005. Extensive diversity of ionizing-radiation-resistant bacteria recovered from Sonoran Desert soil and description of nine new species of the genus *Deinococcus* obtained from a single soil sample. Appl. Environ. Microbiol. **71:**5225–5235.
27. **Roesch, L. F. W., R. R. Fulthorps, A. Riva, G. Casella, A. K. M. Hadwin, A. D. Kent, S. M. Daroub, F. A. O. Camargo, W. G. Farmerie, and E. W. Triplett.** 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. **1:**283–290.
28. **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.
29. **Schloss, P. D., and J. Handelsman.** 2006. Toward a census of bacteria in soil. PLoS Comput. Biol. **2:**786–793.
30. **Smalla, K., G. Wieland, A. Buchner, A. Zock, J. Parzy, S. Kaiser, N. Roskot, H. Heuer, and G. Berg.** 2001. Bulk and rhizosphere soil bacterial communities studied by denaturing gradient gel electrophoresis: plant-dependent enrichment and seasonal shifts revealed. Appl. Environ. Microbiol. **67:**4742–4751.
31. **Smit, E., P. Leeflang, S. Gommans, J. V. d. Broek, S. van Mil, and K. Wernars.** 2001. Diversity and seasonal fluctuations of the dominant members of the bacterial soil community in a wheat field as determined by cultivation and molecular methods. Appl. Environ. Microbiol. **67:**2284–2291.
32. **Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. A. Arrieta, and G. H. Herndl.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. USA **103:**12115–12120.
33. **Taylor, A. F. S.** 2002. Fungal diversity in ectomycorrhizal communities: sampling effort and species detection. Plant Soil **244:**19–28.
34. **Tourova, T. P.** 2003. Copy number of ribosomal operons in prokaryotes and its effect on phylogenetic analyses. Microbiology **72:**437–452.
35. **Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. M. 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.
36. **Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole.** 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. **73:**5261–5267.
37. **Zhou, X., and Y. Luo.** 2007. Source components and interannual variability of soil CO_2 efflux under experimental warming and clipping in a grassland ecosystem. Global Change Biol. **13:**761–775.