The ISME Journal (2009) 3, 992–1000

www.nature.com/ismej

ORIGINAL ARTICLE

Abundance, composition, diversity and novelty of soil *Proteobacteria*

Anne M Spain[1], Lee R Krumholz[1] and Mostafa S Elshahed[2]
[1]*Department of Botany and Microbiology and Institute for Energy and the Environment, University of Oklahoma, Norman, OK, USA and* [2]*Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, OK, USA*

Small subunit (16S) rRNA gene surveys generating near full-length 16S rRNA clones offer a unique opportunity for in-depth phylogenetic analysis to highlight the breadth of diversity within various major bacterial phyla encountered in soil. This study offers a detailed phylogenetic analysis of the *Proteobacteria*-affiliated clones identified from 13 001 nearly full-length 16S rRNA gene clones derived from Oklahoma tall-grass prairie soil. *Proteobacteria* was the most abundant phylum in the community, and comprised 25% of the total clones. The most abundant and diverse class within the *Proteobacteria* was *Alphaproteobacteria*, followed by the *Delta-*, *Beta-* and *Gammaproteobacteria*. Members of the *Epsilon-* and *Zetaproteobacteria* were not detected in the dataset. Our analysis identified 15 novel order-level and 48 novel family-level *Proteobacteria* lineages. In addition, we show that the majority of *Proteobacteria* clones in the dataset belong to orders and families containing no described cultivated representatives (50% and 65%, respectively). An examination of the ecological distribution of the six most abundant *Proteobacteria* lineages in this dataset with no characterized pure culture representatives provided important information regarding their global distribution and environmental preferences. This level of novel phylogenetic diversity indicates that our understanding of the functions of soil microorganisms, even those belonging to phyla with numerous and diverse well-characterized cultured representatives such as the *Proteobacteria*, remains far from adequate.
The ISME Journal (2009) **3,** 992–1000; doi:10.1038/ismej.2009.43; published online 30 April 2009
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: microbial communities; culture-independent studies; *Proteobacteria*; soil

Introduction

Small subunit (16S rRNA) gene-based surveys have clearly shown that the scope of phylogenetic diversity in soil is much broader than that implied using culture-based approaches (Ovreas and Torsvik, 1998; Dunbar *et al.*, 1999; Smit *et al.*, 2001; Lipson and Schmidt, 2004). Although having a remarkably stable phylum level diversity, soil is an extremely diverse ecosystem at the order, family, genus and species levels (Fulthorpe *et al.*, 2008), with multiple yet-uncultured lineages within virtually each of the major bacterial phyla in soil (for example, *Proteobacteria*, *Acidobacteria* and *Actinobacteria*) (Janssen, 2006). The detailed phylogenetic analysis and taxonomic placements of 16S rRNA gene sequences has traditionally been the main focus of soil diversity studies. However, with the availability of newer sequencing technology and curated databases and the subsequent creation of large (>1000) datasets, the focus of the data analysis process has recently shifted more towards computing more accurate estimates of species richness and evenness (Schloss and Handelsman, 2006; Roesch *et al.*, 2007; Quince *et al.*, 2008; Youssef and Elshahed, 2008a), identification of novel bacteria phyla (Elshahed *et al.*, 2008), accessing members of the rare soil biosphere (Elshahed *et al.*, 2008) and computational comparisons of communities between different soils (Roesch *et al.*, 2007; Fulthorpe *et al.*, 2008). Detailed phylogenetic analysis of these datasets has often been overlooked, either because of the short amplicon size created, or the sheer number of clone sequences analyzed. This is unfortunate, as such datasets, especially those with near full-length 16S rRNA gene sequence, offer a unique opportunity for an in-depth evaluation of the phylogenetic diversities within each of the major bacterial phyla in soil.

In a recent study, a near full-length 16S rRNA gene clone library was constructed from Oklahoma tall-grass prairie soil and 13 001 clones were sequenced (Elshahed *et al.*, 2008). The most abundant phylum was shown to be the *Proteobacteria* as is typically observed in soil libraries (for a review, see Janssen, 2006). The *Proteobacteria* encompass an enormous

Correspondence: MS Elshahed, Department of Microbiology and Molecular Genetics, Oklahoma State University, 1110 S. Innovation way, Stillwater, OK 74074, USA.
E-mail: mostafa@okstate.edu
Received 5 February 2009; revised 23 March 2009; accepted 27 March 2009; published online 30 April 2009

level of morphological, physiological and metabolic diversity, and are of great importance to global carbon, nitrogen and sulfur cycling (Kersters *et al.*, 2006). Despite this phylum containing more validly described isolates than any other phylum (Kersters *et al.*, 2006), the vast majority of soil *Proteobacteria* are yet to be cultivated. In this study, we describe the composition of *Proteobacteria* clones from Oklahoma tall-grass prairie soil, in which the majority of clones belong to family- and order-level lineages containing no characterized cultivated isolates, and compare the ecological distribution of some of the dominant uncharacterized orders whose functions in soil remain unknown.

Materials and methods

Phylogenetic analysis of Kessler Farm soil Proteobacteria *16S rRNA gene sequences*

The dataset used in this study initially consisted of 13 001 16S rRNA clone sequences from soil, described in an earlier study (Elshahed *et al.*, 2008). Briefly, a clone library ($n = 13\,001$ clones) was constructed from 16S rRNA genes (PCR-amplified using primers 27F and 1391R) from community DNA extracted from Kessler Farm Soil (KFS), which was collected from an undisturbed tall-grass prairie preserve in Central Oklahoma. Sequences were binned into operational taxonomic units (OTUs) using a 97% similarity cutoff using DOTUR (Schloss and Handelsman, 2005). Soil characteristics, and details of sampling, DNA extraction, PCR amplification, 16S rRNA clone library construction and sequencing, and initial phylogenetic classification of 16S rRNA sequences can be found in the original manuscript (Elshahed *et al.*, 2008).

Sequences representative of each OTU identified as *Proteobacteria* in the original manuscript were aligned using Greengenes' NAST alignment tool (DeSantis *et al.*, 2006a, b). Aligned KFS and closely related 16S rRNA sequences were imported into Greengenes May 2007 ARB database (DeSantis *et al.*, 2006a) using ARB software package, available online at http://www.arb-home.de/ (Ludwig *et al.*, 2004). We used the on-line program Pintail (Ashelford *et al.*, 2005) to screen individual sequences within the *Proteobacteria* dataset using suspicious sequences (those identified by Bellerophon (Huber *et al.*, 2004) or those with unclear phylogenetic affiliation or that formed unusually long branches in neighbor-joining dendrograms) as the query sequence, and the closest cultured relative or a reliable closely-related abundant KFS OTU sequence ($n > 50$) as the reference sequence. After removal of chimera, 2675 *Proteobacteria* clones belonging to 479 OTUs were classified to the family taxonomic level using phylogenetic tree-building methods. Initial placement of OTUs in already-named families according to the Hugenholtz taxonomic framework (DeSantis *et al.*, 2006a) was determined by parsimony placement of KFS clone sequences into the ARB universal dendrogram. Distance trees of each class within *Proteobacteria* were constructed using the neighbor-joining algorithm and Jukes–Cantor corrections using ARB software package (Ludwig *et al.*, 2004) with filters available for each class of *Proteobacteria*. Branching of distance trees was also verified by constructing trees through the same methods using PAUP 4.0b10 software (Sinauer Associates, Sunderland, MA, USA) and generating bootstrap percentages based on 1000 replicates. Final classifications of KFS OTUs into families, according to the Hugenholtz taxonomic outline (DeSantis *et al.*, 2006a), were determined by placement of each OTU into a bootstrap-supported ($>50\%$) already-named or novel family in constructed trees. In general, novel families were defined as a bootstrap-supported group of at least two clone sequences sharing approximately >92–93% sequence similarity with each other but <92–93% sequence similarity to sequences from an already-named family. Novel orders were defined similarly, using 90% as a general cutoff, though these values varied between each class of *Proteobacteria* (for example, *Deltaproteobacteria* is more divergent than *Alpha* and *Betaproteobacteria*).

Ecological distribution of abundant KFS uncharacterized lineages

We chose the six most abundant *Proteobacteria* order-level lineages containing no characterized, cultivated representatives (*Deltaproteobacteria*-KFS-6, EB1021, Ellin314, MND1, A21b and Ellin339), and recorded the isolation source of all available environmental clone sequences belonging to each order. To determine what environmental clone sequences belonged in an order, we created distance trees in ARB using all sequences belonging to the order based on the universal parsimony tree, using the May 2007 Greengenes database. Second, we used the BLAST algorithm on the NCBI website (in November, 2008) to search for more recently deposited sequences belonging to each order, using the 'type sequence' (the environmental clone sequence after which the order was named, for example, MND1) as the query and 90% similarity as a general cutoff.

Results and discussion

Abundance and composition of Proteobacteria *in KFS and other soils*

The *Proteobacteria*-affiliated clones in KFS represented 25% of the total 16S rRNA clone sequences (Elshahed *et al.*, 2008) compared with an average of 40% abundance in all published soil studies analyzing >1000 16S rRNA sequences, including eight individual soil samples in addition to a composite collection of soil libraries compiled by Janssen (2006) (Table 1). From the clone library studies, including those generating >1000 near full-

Table 1 Comparison of the composition and abundance of *Proteobacteria* in Kessler Farm soil to other soils among published studies analyzing >1000 PCR-amplified 16S rRNA sequences

Soil sample	*Type of community analysis*	*No. sequences analyzed*	*Average length (bp)*	*% Abundance of Proteobacteria*	*% Of each class within Proteobacteria*					*Reference*
					α	β	γ	δ	ε	
Oklahoma tall-grass prairie (KFS) soil	Clone library	13 001	>1300	25	39	16	8	37	0	Elshahed *et al.* (2008)
Wisconsin trembling aspen rhizosphere soil	Clone library									Lesaulnier *et al.* (2008)
Ambient (CO_2)		1155	>1300	37	55	27	9	9	0	
Elevated (CO_2)		1132	>1300	40	58	25	11	7	0	
Minnesota farm soil	Clone library	1700	>1300	32	35	23	17	25	0	Tringe *et al.* (2005); von Mering *et al.* (2007)
Review of 21 different soils from various locations	Clone library	2290	>300	39	48	25	6	21	0.1	Janssen (2006)
Soil samples from different regions	Pyrosequencing									
Brazil		26 140	103	50	24	30	10	36	0	
Florida		28 328	102	48	17	42	25	16	0.2	Roesch *et al.* (2007)
Illinois		31 818	104	42	24	40	17	19	0	
Canada		53 533	102	47	26	39	9	27	0.2	
			Average	40	36	30	12	22	0.06	
			s.d.	8	15	9	6	11	0.09	

length 16S rRNA genes and the Janssen compilation study (analyzing 16S rRNA gene sequences >300 bp), *Proteobacteria* comprised 25–40% abundance (relative to total sequences) and 42–50% abundance from shorter (~100 bp) fragments generated by pyrosequencing (Table 1). Although such larger proportion of *Proteobacteria* in pyrosequencing-based studies might be a true reflection of the communities analyzed, it might also indicate the existence of a cloning bias or that classification based on small 16S rRNA gene fragments could lead to different taxonomic assignments than classification based on near to full-length sequences, as suggested earlier (Elshahed *et al.*, 2008). Nevertheless, *Proteobacteria* remains the most abundant soil phylum, regardless of the utilized approach, which aside from PCR-based clone libraries and pyrosequencing has included metagenomics (Liles *et al.*, 2003; Tringe *et al.*, 2005), fluorescent *in situ* hybridization (Zarda *et al.*, 1997) and microarray analysis (Yergeau *et al.*, 2009).

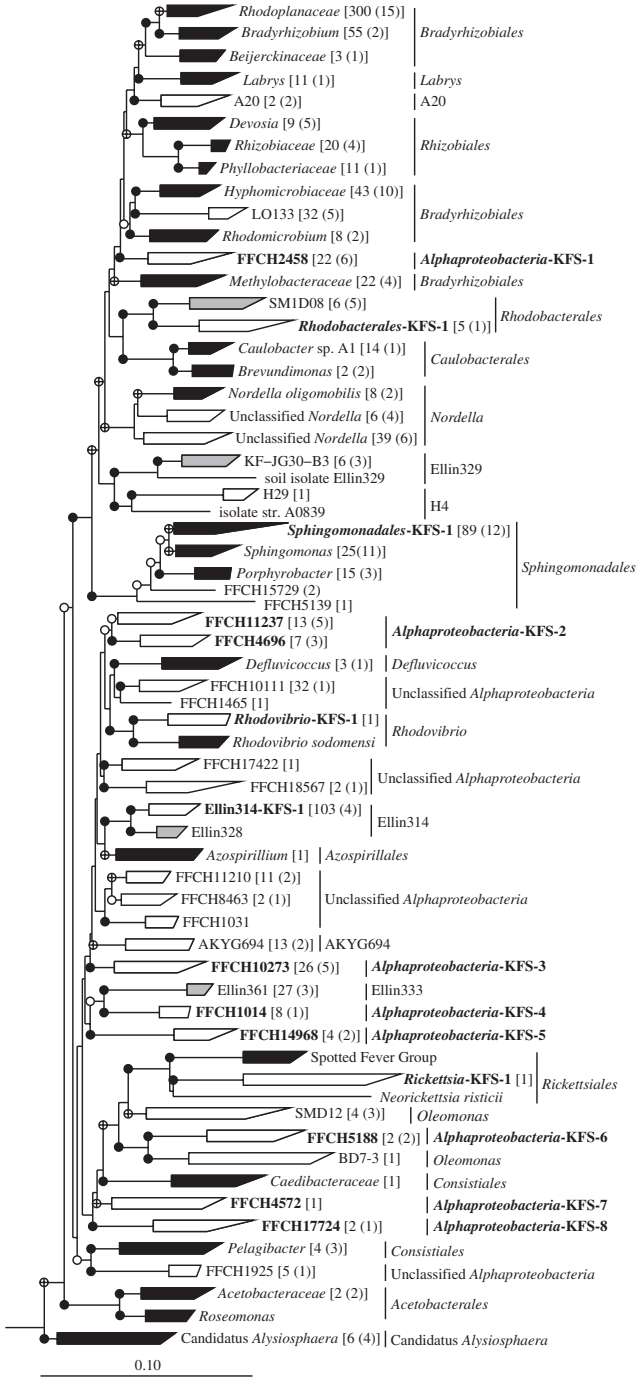
The most abundant class (39% of total *Proteobacteria* clones) in KFS was *Alphaproteobacteria*, followed by *Delta-* (37%), *Beta-* (16%) and *Gammaproteobacteria* (7.6%). Among all clone library datasets (>1000 sequences) of PCR-amplified 16S rRNA genes from soil (Table 1), *Alphaproteobacteria* is the most abundant class, relative to total sequences, comprising 35–58% of *Proteobacteria* clones, whereas *Gammaproteobacteria* is typically, though not always the least abundant (5.9–17%). *Deltaproteobacteria* was overrepresented in KFS compared with other large soil datasets, whereas *Betaproteobacteria* was underrepresented (Table 1). *Epsilonproteobacteria*, which has not been detected in many of the large 16S rRNA soil libraries (Table 1) was not detected in KFS, suggesting that this class is either extremely rare in soil or is not ubiquitous as are the other classes within *Proteobacteria*. Likewise, the recently discovered class *Zetaproteobacteria*, which seems to have a limited ecological distribution and metabolic abilities (Emerson *et al.*, 2007), was undetected in KFS and other large soil clone libraries (Table 1).

Family and order-level diversities within KFS Proteobacteria

The use of classifier programs, available from Greengenes and the Ribosomal Database Project (Cole *et al.*, 2005; DeSantis *et al.*, 2006a), provide useful tools for initial classification of 16S rRNA gene sequences; however, inaccurate taxonomic assignments may be made without tree-building phylogenetic analyses, especially at the subphylum levels. In addition, uncertain placements of clones with low-sequence similarity to their closest relative has been observed with both classification programs, resulting in outputs with multiple placement suggestions (Greengenes), or low confidence in order and family-level affiliation outputs (Ribosomal Database Project). In addition, satisfactory identification and documentation of novel lineages requires detailed phylogenetic analysis and

Table 2 Composition and novel and uncharacterized lineages within the different classes of *Proteobacteria*

Lineage	*No. of clones*	*No. of OTUs*	*No. of orders*	*No. of novel orders*	*% Clones in uncharacterized orders*	*No. of families*	*No. of novel families*	*% Clones in uncharacterized families*
Proteobacteria (total)	2675	479	60	15	50	120	48	65
Alpha-	1043	168	29	8	28	45	14	37
Delta-	998	165	15	6	64	33	15	85
Beta-	432	78	5	1	69	23	14	85
Gamma-	202	68	11	0	58	19	5	66



tree-building approaches. In this study, phylogenic associations at the class, order and family levels were initially determined using both Greengenes and Ribosomal Database Project classification programs, and were verified by parsimony analysis using the ARB software package and neighbor-joining analysis using PAUP 4.01b10. Using this combined approach, 120 family-level lineages were identified belonging to 60 orders (Table 2). *Alphaproteobacteria* had the highest number of families and orders, consisting of 45 families within 29 orders, and was followed by *Deltaproteobacteria* (33 families within 15 orders) (Table 2, Figures 1 and 2). *Beta-* and *Gammaproteobacteria* were less diverse, containing 23 and 19 families within five and 11 orders, respectively (Table 2, Figures 3 and 4). This pattern of order and family level diversity rankings between various *Proteobacteria* classes is in agreement with the diversity ranking estimated from the same datasets based on rarefaction curve analysis and diversity ordering approaches of KFS $OTU_{0.03}$ (Youssef and Elshahed, 2008b).

Figure 1 Distance phylogram of *Alphaproteobacteria* KFS OTU sequences based on aligned near full-length 16S rRNA gene sequences (approximately 1350 bp) from KFS clone library as well as representative sequences from each family-level lineage downloaded from GenBank, totaling 329 sequences, with each clade shown representing a family-level lineage (unless otherwise noted), consisting of at least two sequences. The tree was rooted with the 16S rRNA gene sequence from *Chloroflexus aurantiacus* (GenBack accession no. AJ308501). Bootstrap values are based on 1000 replicates and are shown to the left of each branch with bootstrap support >90% (●), 70–89% **(⊗) and 50–69% (○). Black clades represent families with characterized, described cultivated representatives. Gray and unfilled clades represent uncharacterized families, consisting of clone sequences and sequences from unpublished or uncharacterized isolates (gray) or only environmental clone sequences (unfilled). Numbers aside each clade denote the number of clone sequences and OTUs detected from each family in the KFS clone library. Orders, according to Hugenholtz taxonomy and the Greengenes ARB May, 2007 database, are shown to the right of families. Novel lineages are shown in bold, with novel orders labeled as *Proteobacteria* class-KFS-# (for example, *Alphaproteobacteria*-KFS-1). Novel families within novel orders are labeled according to clone names (for example, FFCH2458), and novel families within characterized orders are labeled as order name-KFS-# (for example, *Sphingomonadales*-KFS-1).

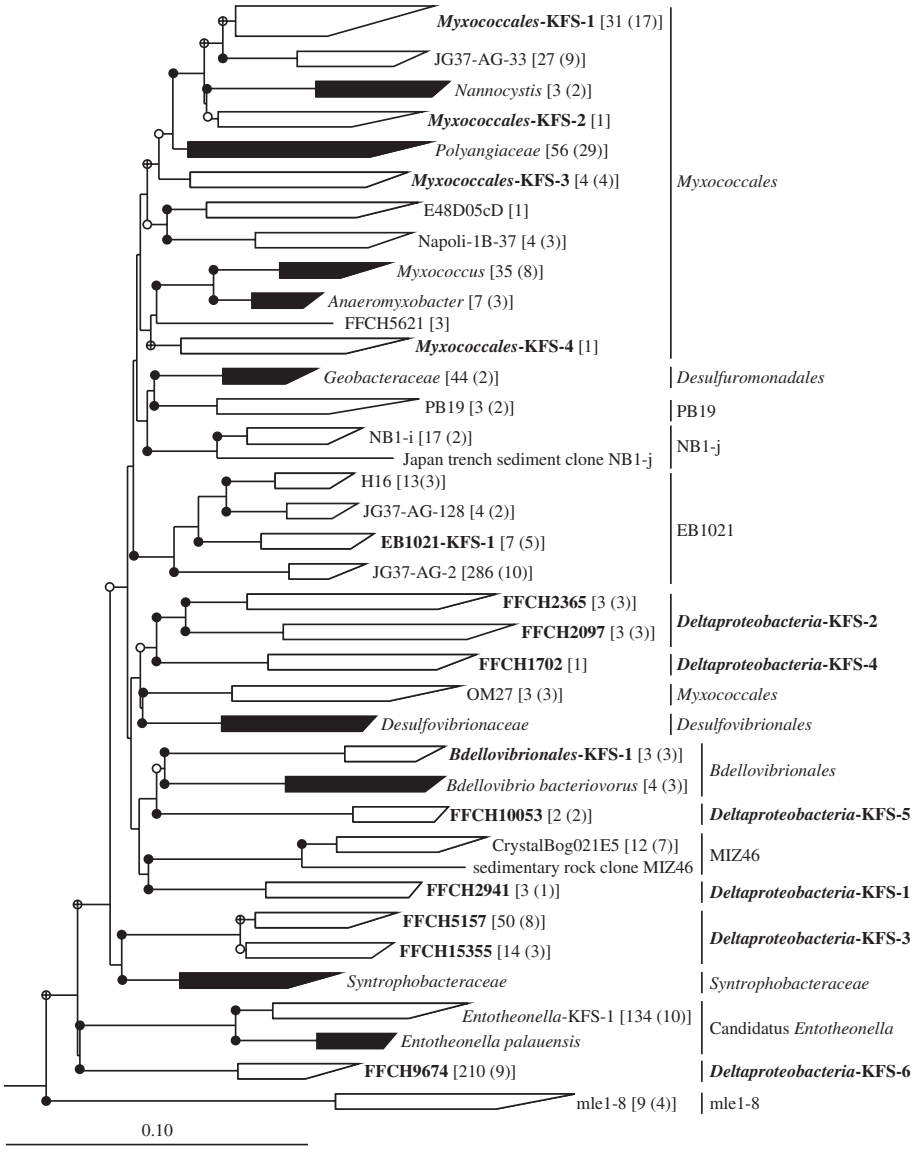


Figure 2 Distance phylogram of *Deltaproteobacteria* KFS OTU sequences based on aligned near full-length 16S rRNA gene sequences from KFS clone library as well as representative sequences from GenBank, totaling 241 sequences. Tree construction and notations are the same as described in Figure 1.

Prevalence of uncharacterized and novel lineages within KFS Proteobacteria

The vast majority of KFS *Proteobacteria* clones belonged to uncharacterized lineages (families or orders containing no validly described species); in total, 50% and 65% of KFS *Proteobacteria* clones belonged to uncharacterized orders and families, respectively (Table 2). It is important to note; however, that among the *Alpha-*, *Beta-* and *Gammaproteobacteria*, some microorganisms have been cultivated among these uncharacterized lineages, but have not been characterized nor validly described (Figures 1–3). Indeed, within all *Proteobacteria* classes in KFS with the exception of *Alphaproteobacteria*, the most abundant orders contained no cultivated or characterized pure cultures. The most abundant order in *Alphaproteobacteria* was *Bradyrhizobiales* (Figure 1), which consisted of 463 clones (39 OTUs) and contained the most abundant OTU in the KFS dataset ($n = 204$). The most abundant orders in *Deltaproteobacteria* were EB1021 (310 clones, 20 OTUs) and novel order *Deltaproteobacteria*-KFS-6 (210 clones, nine OTUs) (Figure 2), neither of which contain any cultivated microorganisms. The dominant orders in *Beta-* and *Gammaproteobacteria* in KFS were MND1 and Ellin339, respectively (Figures 3 and 4), which are also uncharacterized lineages. *Deltaproteobacteria* contained the highest number of clones belonging to undescribed lineages, with 637 clones (64%) belonging to uncharacterized orders and 848 clones (85%) belonging to uncharacterized families. These *Deltaproteobacteria* lineages were comprised solely of environmental clone sequences, none

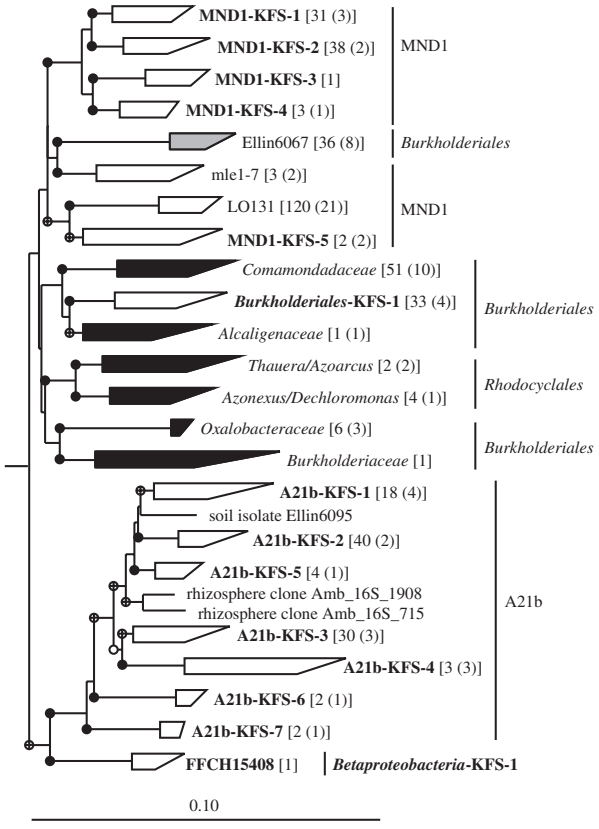


Figure 3 Distance phylogram of *Betaproteobacteria* KFS OTU sequences based on aligned near full-length 16S rRNA gene sequences from KFS clone library as well as representative sequences from GenBank, totaling 128 sequences. Tree construction and notations are the same as described in Figure 1.

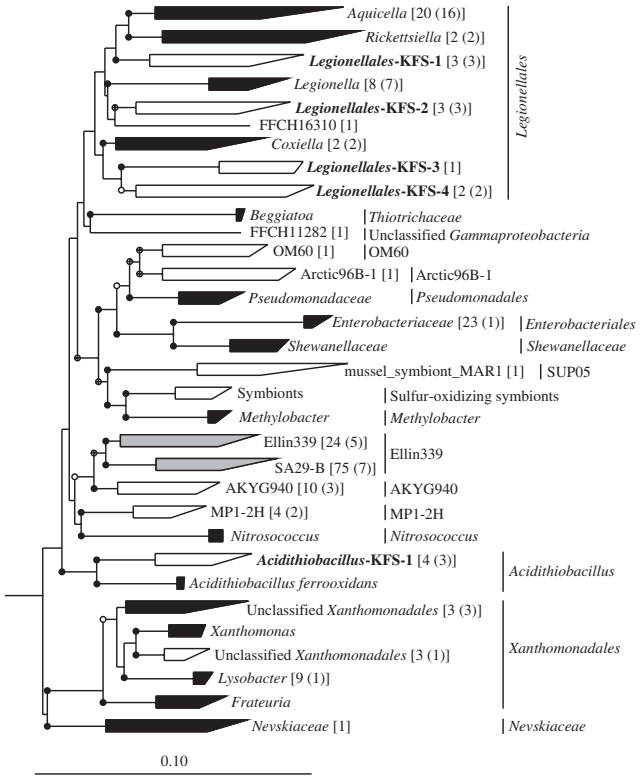


Figure 4 Distance phylogram of *Gammaproteobacteria* KFS OTU sequences based on aligned near full-length 16S rRNA gene sequences from KFS clone library as well as representative sequences from GenBank, totaling 183 sequences. Tree construction and notations are the same as described in Figure 1.

containing any cultivated representatives, suggesting that soil *Deltaproteobacteria* may be extremely difficult to cultivate in pure culture in the laboratory using standard heterotrophic growth media.

In addition, KFS contained numerous novel lineages within the *Proteobacteria* dataset (Table 2). In total, 15 novel orders and 48 novel families among the four classes were named in this study (Figures 1–4; for detailed descriptions of *Proteobacteria* KFS OTU phylogenetic affiliations, including all novel lineages, see Supplementary Table 1). The large number of novel family and orders identified from a single clone library clearly suggests that global soil *Proteobacteria* diversity is far broader than our current database collection suggests. Likewise, despite *Proteobacteria* being the most abundant soil phylum, containing more validly described species than any other phylum, the functions of the majority of *Proteobacteria* in soil remain to be shown.

Ecological distribution of abundant uncharacterized order-level lineages

As the majority of KFS *Proteobacteria* clones belong to family- and order-level lineages with no characterized representatives, the functions of these groups of microorganisms in soils is completely unknown. To gain insight into the rarity of and ecological distribution of uncharacterized lineages within *Proteobacteria*, we chose the six most abundant KFS uncharacterized orders, *Deltaproteobacteria*-KFS-6 (*Deltaproteobacteria*, $n=210$), EB1021 (*Deltaproteobacteria*, $n=310$), Ellin314 (*Alphaproteobacteria*, $n=103$), MND1 (*Betaproteobacteria*, $n=198$), A21b (*Betaproteobacteria*, $n=99$) and Ellin339 (*Gammaproteobacteria*, $n=99$) and mapped their distribution among different environmental categories using data available from 16S rRNA sequences deposited into GenBank. We found that these six lineages, collectively, have been identified in 174 different sampling sites that fall into 30 general environmental categories, the most abundant of which was soil, whereas many samples also came from aquatic and subsurface ecosystems (Table 3; for details and references for each study, see Supplementary Table 2).

The two *Deltaproteobacteria* orders were the most abundant of the uncharacterized orders; however, novel order *Deltaproteobacteria*-KFS-6 was detected in only four sites, all from soil. EB1021 contained the most clones out of any of the KFS uncharacterized orders, and was detected in 52 total samples from 15 different ecosystem types. This order was

Table 3 Distribution of six abundant uncharacterized order-level lineages from Kessler Farm soil among different types of ecosystems

Ecosystem type	*No. of sampling sites detected from each ecosystem type*						
	Total	*Delta proteobacteria-KFS-6 (δ)*	*EB1021 (δ)*	*Ellin314 (α)*	*MND1 (β)*	*A21b (β)*	*Ellin339 (γ)*
Soil ecosystems:	61	4	25	25	32	23	27
Soil/rhizosphere (uncontaminated)	34	3	16	11	15	16	18
Organic-contaminated soil	11	1	2	5	6	1	5
Soil from cold/Polar regions	7		1	3	5	3	1
Metal/radionuclide mill tailings	4		4	3	3		2
Wetland soils	3		2	3	2	2	1
Peat soils	2				1	1	
Aquatic ecosystems:	54	0	18	16	25	12	14
Freshwater	11			1	5	1	7
Freshwater sediments	9		4	4	5	4	3
Coastal/estuarine sediments	10		4	1	5	2	3
Marine water	2				2		
Deep sea sediment	10		9	6			
Wastewater/activated sludge	9		1	2	6	4	1
Drinking water	3			2	2	1	
Subsurface ecosystems:	25	0	4	6	18	1	2
Subsurface groundwater/sediment	10		3	2	7		1
Contaminated subsurface	15		1	4	11	1	1
Extreme ecosystems:	14	0	2	5	6	1	6
Caves	4			3	3	1	1
Acid mine drainage	5						5
Volcanic/terrestrial mats	3		1	2	2		
Hot springs	2		1		1		
Animal-associated ecosystems:	10	0	2	3	3	3	2
Animal GI tract	4			1	1	2	1
Marine sponges	2		2				
Oyster shell	1				1		
Deep sea octacoral	1			1	1	1	
Freshwater sponge	1			1			
Human skin	1						1
Other ecosystems:							
Anaerobic organic-degrading enrichments/consortia	4		1	3		1	
Air	3			2			1
Waste-gas biofilter	1					1	
Decayed velvetleaf seed	1						1
Biofilm on O_2-transfer membrane	1			1			
Total sampling sites	174	4	52	61	84	42	53
		No. of ecosystem types in which each order was detected					
Total ecosystem types	30	2	15	21	20	16	17

detected in 25 out of the 61 different soil sample sites, but was detected in 90% of the deep-sea sediment sites (Table 3 and Supplementary Table 2) and both of the marine sponge studies. Interestingly, among aquatic environments, EB1021 was detected in all sediment ecosystems (freshwater, estuarine and marine) but was not detected in any of the overlying water ecosystems, suggesting EB1021 could be preferentially distributed in anoxic ecosystems. Thus, members of EB1021 might be living in anoxic or hypoxic microenvironments within soil aggregates, and the use of anaerobic techniques could prove useful in trying to cultivate members of EB1021.

From the *Alphaproteobacteria*, uncharacterized order, Ellin314 was detected in more ecosystem types than any of the other KFS uncharacterized orders (Table 3, Supplementary Table 2). Most notably, members of this order have been detected in 75% of samples detected from anaerobic enrichments or consortia degrading organic pollutants. Like EB101, Ellin314 was detected in 25 of the 61 soil sites, and was more frequently detected in aquatic sediments rather than overlying water, including 60% of the deep-sea sediment sites. Unlike EB1021, however, organisms belonging to Ellin314 have been cultivated but not characterized (Joseph *et al.*, 2003).

From the *Betaproteobacteria*, MND1 (the dominant order in KFS *Betaproteobacteria*) was detected in 84 different samples sites, more than any of the other KFS uncharacterized orders (Table 3 and Supplementary Table 2), being detected more frequently in soil, aquatic and subsurface ecosystems, which suggests that MND1 may be diverse in function and/or capable of a wide range of environmental conditions. MND1 was detected in 18 of the 25 total subsurface sites, which is triple the number of any other KFS uncharacterized order. Originally, MND1 was first detected in ferromanganous-coated sediment (Stein *et al.*, 2001; Joseph *et al.*, 2003), but it shows no preferential distribution towards either aerobic vs anaerobic environments. A21b (*Betaproteobacteria*) has a similar distribution pattern to MND1, but is detected in fewer samples, and has been rarely documented among subsurface community studies, and has not been detected in any marine environments to date (Table 3 and Supplementary Table 1). Like A21b, Ellin339 (the dominant order in KFS *Gammaproteobacteria*) was rare in subsurface sites and was not detected in any marine samples. However, unlike other KFS uncharacterized orders, Ellin339 was detected among more freshwater sites and was the only order detected in several acid mine drainage sites (Table 3 and Supplementary Table 1). In addition, Ellin339 was detected in an acid-impacted lake (Percent *et al.*, 2008) and an extremely acidic river (Garcia-Moyano *et al.*, 2007), suggesting this uncharacterized order likely contains acid-tolerant or acidophilic bacteria.

This study highlights the importance of detailed subphylum level phylogenetic analysis of large 16S rRNA datasets, a process that is increasingly overlooked in favor of automated phylum-level assignment. The discovery and documentation of 15 novel orders and 46 novel families within the *Proteobacteria* in a single dataset indicates that even in phyla with multiple cultured representatives, the breadth of the subphylum level diversity is not completely understood. Finally, our survey of the ecological distribution of six abundant, yet-uncultured *Proteobacteria* orders suggests that most of these uncharacterized lineages may be ecologically important in not only soil but many ecosystems globally, and that specific enrichment and isolation approaches that have rarely been tested (for example, acidic, hypoxic or anoxic conditions) might prove useful in obtaining these lineages in pure cultures.

Acknowledgements

This work has been supported by the National Science Foundation microbial observatories program (Grant no. MCB_0240683 and EF0801858), DOE small laboratory science program and OSU start-up funds to MSE.

References

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol* **71**: 7724–7736.

Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM *et al.* (2005). The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* **33**: D294–D296.

DeSantis Jr TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM *et al.* (2006b). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* **34**: W394–W399.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006a). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Dunbar J, Takala S, Barns SM, Davis JA, Kuske CR. (1999). Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Appl Environ Microbiol* **65**: 1662–1669.

Elshahed MS, Youssef NH, Spain AM, Sheik C, Najar FZ, Sukharnikov LO *et al.* (2008). Novelty and uniqueness patterns of rare members of the soil biosphere. *Appl Environ Microbiol* **74**: 5422–5428.

Emerson D, Rentz JA, Lilburn TG, Davis RE, Aldrich H, Chan C *et al.* (2007). A novel lineage of proteobacteria involved in formation of marine Fe-oxidizing microbial mat communities. *PLoS ONE* **2**: e667.

Fulthorpe RR, Roesch LF, Riva A, Triplett EW. (2008). Distantly sampled soils carry few species in common. *ISME J* **2**: 901–910.

Garcia-Moyano A, Gonzalez-Toril E, Aguilera A, Amils R. (2007). Prokaryotic community composition and ecology of floating macroscopic filaments from an extreme acidic environment, Rio Tinto (SW, Spain). *Syst Appl Microbiol* **30**: 601–614.

Huber T, Faulkner G, Hugenholtz P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**: 2317–2319.

Janssen PH. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol* **72**: 1719–1728.

Joseph SJ, Hugenholtz P, Sangwan P, Osborne CA, Janssen PH. (2003). Laboratory cultivation of widespread and previously uncultured soil bacteria. *Appl Environ Microbiol* **69**: 7210–7215.

Kersters K, De Vos P, Gillis M, Swings J, Vandamme P, Stackebrandt E. (2006). Introduction to the Proteobacteria. In: Dwarkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E (eds). *The Prokaryotes*, 3rd edn, vol. 5. Springer: New York, pp 3–37.

Lesaulnier C, Papamichail D, McCorkle S, Ollivier B, Skiena S, Taghavi S *et al.* (2008). Elevated atmospheric CO_2 affects soil microbial diversity associated with trembling aspen. *Environ Microbiol* **10**: 926–941.

Liles MR, Manske BF, Bintrim SB, Handelsman J, Goodman RM. (2003). A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol* **69**: 2684–2691.

Lipson DA, Schmidt SK. (2004). Seasonal changes in an alpine soil bacterial community in the colorado rocky mountains. *Appl Environ Microbiol* **70**: 2867–2879.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar A *et al.* (2004). ARB: a software

environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

Ovreas L, Torsvik VV. (1998). Microbial diversity and community structure in two different agricultural soil communities. *Microb Ecol* **36**: 303–315.

Percent SF, Frischer ME, Vescio PA, Duffy EB, Milano V, McLellan M *et al.* (2008). Bacterial community structure of acid-impacted lakes: what controls diversity? *Appl Environ Microbiol* **74**: 1856–1868.

Quince C, Curtis TP, Sloan WT. (2008). The rational exploration of microbial diversity. *ISME J* **2**: 997–1006.

Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**: 1501–1506.

Schloss PD, Handelsman J. (2006). Toward a census of bacteria in soil. *PLoS Comput Biol* **2**: e92.

Smit E, Leeflang P, Gommans S, van den Broek J, van Mil S, Wernars K. (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.

Stein LY, La Duc MT, Grundl TJ, Nealson KH. (2001). Bacterial and archaeal populations associated with freshwater ferromanganous micronodules and sediments. *Environ Microbiol* **3**: 10–18.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al.* (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554–557.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ *et al.* (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126–1130.

Yergeau E, Schoondermark-Stolk SA, Brodie EL, Dejean S, Desantis TZ, Goncalves O *et al.* (2009). Environmental microarray analyses of Antarctic soil microbial communities. *ISME J* **3**: 340–351.

Youssef NH, Elshahed MS. (2008a). Species richness in soil bacterial communities: a proposed approach to overcome sample size bias. *J Microbiol Methods* **75**: 86–91.

Youssef NH, Elshahed MS. (2008b). Diversity rankings among bacterial lineages in soil. *ISME J* (in press).

Zarda B, Hahn D, Chatzinotas A, Schonhuber W, Neef A, Amann RI *et al.* (1997). Analysis of bacterial community structure in bulk soil by *in situ* hybridization. *Arch Microbiol* **168**: 185–192.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)