The ISME Journal (2008) 2, 901–910

www.nature.com/ismej

ORIGINAL ARTICLE

Distantly sampled soils carry few species in common

Roberta R Fulthorpe[1], Luiz FW Roesch[2], Alberto Riva[3] and Eric W Triplett[2]
[1]*Department of Physical and Environmental Sciences, University of Toronto at Scarborough, Ontario, Canada;* [2]*Department of Microbiology and Cell Science, University of Florida, Gainesville, FL, USA and* [3]*Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL, USA*

The bacterial phylogenetic structure of soils from four distinctly different sites in South and North America was analyzed. One hundred and thirty-nine thousand sequences of the V9 region of the small subunit of the bacterial ribosomal RNA gene generated for a previous study were used for this work. Whereas the previous work estimated levels of species richness, this study details the degree of bacterial community overlap between the four soils. Sequences from the four soils were classified and grouped into different phyla and then assigned to operational taxonomic units (OTUs) as defined by 97 or 100% sequence similarity. Pairwise Jaccard and θ similarity indices averaged over all phyla equalled 6 and 12% respectively at the 97% similarity level, and 15% for both at the 100% similarity level. At 100 and 97% sequence similarity, 1.5 and 4.1% of OTUs were found in all four soils respectively, and 87.9 and 74.4%, respectively were a unique particular soil. These analyses, based on the largest soil bacterial sequence retrieval to date, establish the high degree of community structure difference for randomly sampled dissimilar soils and support the idea that wide sampling is important for bioprospecting. The 10 most abundant cultured genera were determined in each soil. These 10 genera comprised a significant proportion of the reads obtained from each soil (31.3–37.4%). *Chitinophaga* was the most abundant or the second most abundant genus in all four soils with 7.5–13.8% of the total bacterial sequences in these soils. The striking result is that several culturable genera, whose roles in soil are virtually unknown, were found among these dominant sequences.
The ISME Journal (2008) **2,** 901–910; doi:10.1038/ismej.2008.55; published online 5 June 2008
Subject Category: microbial populations and community ecology
Keywords: biogeography; community structure; pyrosequencing

Introduction

Roesch *et al*. (2007) used 16S rRNA sequences from four soils across the western hemisphere to estimate the number of operational taxonomic units (OTUs) in a gram of soil. Rarefaction curve extrapolations estimated this number to be between 11 000 and 21 000 per gram, which was considerably below an estimate by Gans *et al*. (2005). In Roesch *et al*. (2007) pyrosequencing was used to determine the depth of microbial diversity and to estimate the abundance of phyla in each soil. In this study, using the same data set, we look at the degree of similarity of the soils using the data from this deep pyrosequencing effort.

The availability of molecular tools has dramatically increased our ability to determine microbial community structures in natural environments. Many studies involving the cloning and sequencing of whole small subunit RNA genes from bacteria and archaea from various habitats have revealed vast diversity and altered our understanding of who lives where. Soil community structures differ markedly between systems, and even across landscapes that appear macroscopically homogeneous. (for example, Axelrood *et al*., 2002; Franklin and Mills, 2003; Hackl *et al*., 2004; Sliwinski and Goodman, 2004; Oline, 2006; Herrera *et al*., 2007; Lamarche *et al*., 2007). Whole community studies and studies on the genotypes of individual taxa have revealed that, at high levels of taxonomic resolution, there is a degree of prokaryote endemism apparent across broad geographic scales (Fulthorpe *et al*., 1998; Cho and Tiedje, 2000; Wawrik *et al*., 2007).

There are two mechanisms for the development of prokaryotic endemism. One is the strong selective effect of local conditions on a ubiquitous base of current diversity. The other is the evolution and extinction of endemic populations at rates that outstrip the pace of global mixing (Hughes-Martiny *et al*., 2006). Distinguishing between these mechanisms is difficult without more thorough sampling of comparative sites that would definitively show if all

Correspondence: RR Fulthorpe, Physical and Environmental Science, University of Toronto at Scarborough, 1265 Military Trail, Toronto, Canada M1C 1A4.
E-mail: fulthorpe@utsc.utoronto.ca
Received 21 February 2008; revised 28 April 2008; accepted 28 April 2008; published online 5 June 2008

species are in fact present in all samples, albeit at very low populations.

In a recent study Roesch *et al*. (2007) used pyrosequencing of 16S RNA gene fragments from four different soils to empirically estimate their bacterial species richness. Even with a data set of 139 819 sequences, the authors did not attain complete sampling. However, if the extrapolations in Roesch *et al*. (2007) are correct, 34–48% of the estimated number of OTUs present in the soils were identified in these samples. Those results represent an unprecedented, if still imperfect, opportunity to test the assumption that soil bacteria are ubiquitous.

In this work, all of the sequences from four distantly sampled soils were classified into OTUs defined at two different taxonomic levels. This allowed us to determine the degree of similarity between soils and the number of OTUs that were found in all four of our samples. Very few OTUs were present in all soils. In addition, the numerically dominant identifiable bacterial genera found within these soils were culturable organisms about which very little is known.

Materials and methods

Data sets

The 16S rRNA data sets were obtained from a previously described study by Roesch *et al*. (2007). In that work, a hypervariable region of the highly conserved 16S rRNA gene was amplified from soil DNA derived from three agricultural fields and one forest site. A total of 139 819 bacterial sequences from the V9 region of the 16S gene with an average read length of 103 bases were obtained through pyrosequencing (454 Life Sciences Branford, CT, USA). A small portion of those sequences were shorter than 50 bases. In this study, those sequences were deleted from the data sets. After removing the smaller sequences the data set consisted of 26 115 bacterial sequences from a Distrophic oxisol sampled in Rio Grande do Sul, Brazil; 28 247 sequences from a Lauderhill euic hyperthermic lithic sampled in Florida; 31 745 bacterial sequences from a Mesic aquic argiudoll sampled in Illinois and 53 245 bacterial sequences from a Dystric brunisol sampled in northern Ontario, Canada. The sequences are available in GenBank with the accession numbers: EF222481–EF361836.

The data set was analyzed on an HP DL585 Proliant Server with 64 Gigabytes of RAM and two dual core Opteron processors at 1.8GHz, running the RedHat Enterprise Linux OS. The amount of information (sequences) contained in the data sets exceeded the capacity of the server to analyze all sequences together. To compare the sequences, the four libraries were merged and the sequences were classified and grouped into phyla. Twenty-one files (one for each phyla found and one for all those unclassifiable), were generated and analyzed separately.

Phylogenetic classification of 16S rRNA gene fragments

The 16S rRNA gene sequences were phylogenetically assigned according to their best matches to sequences in the NAST (Nearest Alignment Space Termination) database (DeSantis *et al*., 2006). The sequences were submitted to a web-interface tool (http://greengenes.lbl.gov/NAST) and aligned against the 16S green genes rRNA reference database (188 073 aligned 16S rDNA records). Once aligned, the sequences were taxonomically classified according to the best match with the reference database using the NCBI taxonomic nomenclature. Each sequence was grouped with its own phyla generating one file for each phylum comprising sequences from the four libraries tested. Sequences that could not be classified were grouped together under the designation 'Noclass' and analyzed as one phylum.

Shared OTUs and similarity

For each phylum, sequences from all four soils were aligned using NAST. Each query sequence in the uploaded file was searched for 16S rRNA gene sequences and aligned according to a core set of alignment templates. Based on the alignment, a distance matrix was constructed using DNAdist from the PHYLIP suite of programs version 3.6 with default parameters (Felsenstein, 1989, 2005). These pairwise distances served as input to DOTUR (Schloss and Handelsman, 2005) for clustering the sequences into OTUs defined by 100% sequence similarity or 97% sequence similarity. Community similarities were determined using SONS (Schloss and Handelsman, 2006) which uses the OTU data output files from DOTUR. SONS calculates the similarities between pairs of communities using various indices. We present the Jaccard similarities—determined as the quotient of the number of OTUs shared and the total number of OTUs in both samples (equation 3 in Schloss and Handelsman, 2006)—as it is the simplest measure of shared species that does not involve abundance data. We also present θ indices (Yue and Clayton 2005, equation 9 in Schloss and Handelsman, 2006) as the θ index does take abundance data, that is, community structure, into account. Data from the OTU files were reduced to binary data (OTUs were counted as present or absent) in Excel to calculate the number of OTUs shared between the soils.

Results

Phylogenetic classification of numerically abundant sequences

The taxonomic classifications of the 10 most numerically abundant genera from each of the four soils are shown in Table 1. Their abundance is

Table 1 The ten most abundant genera found in each of the four soils sampled as percent of total sequences from each soil

Genus (citations)	*Florida (%)*	*Brazil (%)*	*Canada (%)*	*Illinois (%)*	*Phylum*
Chitinophaga (11)	7.5	8.3	7.5	13.8	Bacteroidetes
Acidobacterium (77)	2.4	9.3	5.3	4.0	Acidobacteria
Acidovorax (195)	7.5	1.8	—	2.0	β-proteobacteria
Thiobacter (1)	1.7	—	2.2	2.0	β-proteobacteria
Gemmatimonas (2)	—	—	3.7	1.2	Gemmatimonadetes
Nitrospira (202)	2.8	2.0	—		α-proteobacteria
Sphingomonas (1232)	—	1.8	—	3.0	α-Proteobacteria
Chondromyces (104)	—	—	3.0	1.5	δ-proteobacteria
Pedobacter (34)	2.2	—	1.0	—	Bacteroidetes
Bradyrhizobium (3285)	—	2.3	2.5	—	α-proteobacteria
Haliangium (5)	—	—	2.2	1.7	δ-proteobacteria
Nevskia (13)	—	4.0	—	—	γ-proteobacteria
Hydrocarboniphaga (2)	—	3.1	—	—	γ-proteobacteria
Flavobacterium (2523)	3.1	—	—	—	Bacteroidetes
Aquabacterium (18)	—	—	2.7	—	β-proteobacteria
Dyadobacter (10)	—	2.5	—	—	Bacteroidetes
Pseudomonas (69 906)	2.3	—	—	—	γ-proteobacteria
Methylophilus (239)	—	2.3	—	—	β-proteobacteria
Stenotrophomonas (2)	—	—	—	1.3	γ-proteobacteria
Bacillus (70 165)	—	—	—	1.2	Firmicutes
Nitrobacter (565)	—	—	1.2	—	α-proteobacteria
Sporocytophaga (40)	1.1	—	—	—	Bacteroidetes
Prevotella (1973)	0.9	—	—	—	Bacteroidetes
Total as percentage of sequences	31.5	37.4	31.3	31.7	

No data indicates less than 1%.
The number of citations in Web of Science (as of 11 February 2008) for each genus is listed in parentheses after the name of each genus. These data suggest that the abundance of genera in soil is not well correlated with our level of knowledge on these genera.

indicated as a percentage of the total number of sequences for that soil. When combined this list includes 22 genera, of which six are from the Bacteroidetes, 13 from the Proteobacteria (four β, four α, four γ and one δ) and one genus from each of the Acidobacteria, Gemmatimonadetes and Firmicutes. The genus *Chitinophaga* and *Acidovorax* were found to be dominant in the soil sampled in Florida. Both genera comprised 7.5% of 28 247 sequences analyzed from this soil. In the soil sampled in South Brazil, the most abundant genus found was *Acidobacterium* (9.3% of 26 115 sequences analyzed). The genus *Chitinophaga* was the most abundant in both soils sampled in Canada and Illinois representing 7.5% of 53 245 bacterial sequences and 13.8% of 31 745 bacterial sequences respectively. The genera *Chitinophaga* and *Acidobacterium* were found in large numbers in all of the four soils tested.

We have also compiled a list of all genera found in these soils and the number of sequences found for each genus. (Supplementary Table S1).

Similarities between pairs of soils

The overall degree of similarity between the soils was examined at two phylogenetic levels (100 and 97% sequence similarity—referred to below as fine and coarse, respectively). For the region of 16S gene analyzed, 100% similarity best correlates to species level (Roesch *et al.*, 2007), but 97% similarity was analyzed here as it is the accepted level for species definition when the entire 16S sequence is taken into account. The Jaccard and θ similarity indices for each pair of soils, averaged over all phyla, were determined (Figure 1). Overall Canada and Florida exhibited the least similarity between them, with Jaccard and θ indices of 0.04 and 0.11, respectively at the fine level; and 0.11 and 0.24 respectively, at the coarse level. Canada/Brazil and Brazil/Illinois exhibited the highest similarities and these were statistically greater than those between Florida/Canada (*t*-tests on Jaccard indices, $P<0.05$ or less). Jaccard and θ similarity indices for Brazil/Illinois were 0.09 and 0.24, respectively at the fine level, and 0.24 and 0.38, respectively at the coarse level.

Phylum level differences in average soil similarities

The necessity of subdividing the data set for analysis allowed us to observe differences between phyla. The similarity indices between all pairs of soils were averaged for each phylum (Table 2). Average Jaccard similarities for individual phyla ranged from 0.02–0.10 at the fine level, and from 0.05–0.28 at the coarse level. θ indices ranged between 0.01–0.29 and 0.07–0.28, respectively. Significantly higher levels of similarity are seen in the Bacteroidetes, α-proteobacteria, Acinetobacteria, β-proteobacteria and the Candidatus division OP10. Most of these phyla were represented by high numbers of total sequences, so the phyla differences could be attributed merely to sample size effects. However, the amount of variance in the similarity indices

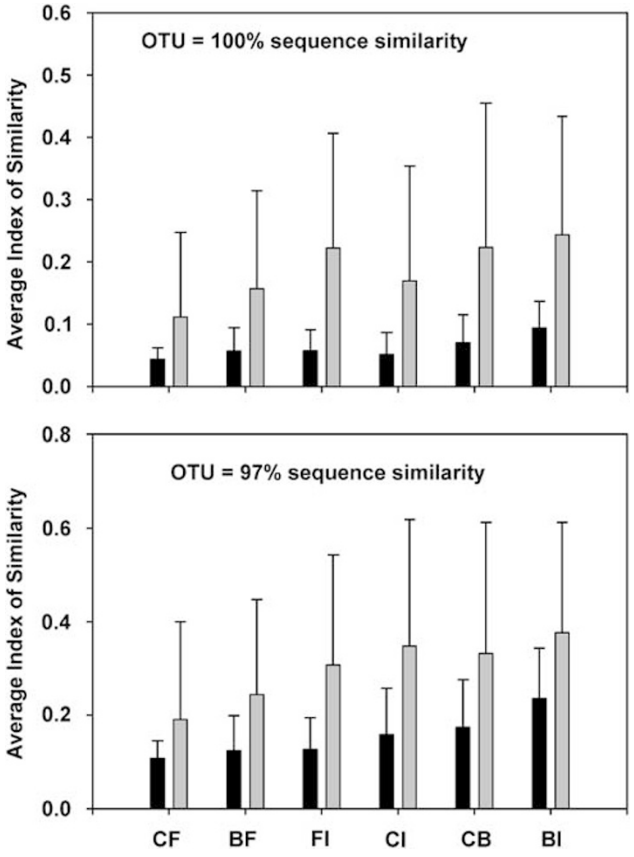

Figure 1 Jaccard and θ indices of similarity between pairs of soils, averaged over all phyla. Bars indicate means with s.d. error bars. First bar (black) = Jaccard Index, second bar (light gray) = Theta index. *x* axis indicates soils compared—B = Brazil, C = Canada, F = Florida, I = Illinois.

explained by the number of sequences in each phylum is only 12% at the fine level and 21% at the coarse level (Figures 2a and b).

OTU occurrences in all four soils

Determining the average level of similarity seen between pairs of soils is not the same as determining how many OTUs were found in all four soils. We carried out this analysis for each phyla (Table 3). The percentage of OTUs found in all four soils is correlated to the number of sequences analyzed (Figure 3), since in the rarer phyla there were too few sequences per OTU to detect them in all soils (Table 3). However, a linear regression analysis of percent OTUs found in all four soils versus number of sequences in the phyla provides linear slopes that are not significantly different from zero, indicating that sequence number alone is insufficient to predict the number of OTUs that are widespread. This suggests that a greater sampling would not lead to significantly higher estimates of shared OTUs. The phyla with the most shared OTUs are the Bacteroidetes, Actinobacteria, β-proteobacteria and the Candidatus division OP10. Figures 4 and 5 summarize the data for all phyla together in four component Venn diagrams. When using 97% similarity as the definition of an OTU (coarse level) −4.1% of OTUs could be detected in all four soils and 74.4% of OTUs were unique to the soils they were found in. When using 100% similarity as the definition of an OTU (fine level), 1.5% of the OTUs' were found in all four soils and 87.9% were unique to one soil. Therefore the majority of OTUs, regardless of how they were defined, were found to be unique to the soil they were sampled in and only a small portion of the OTUs were found to be common to all of the soils tested.

Discussion

These data illustrate the amount of similarity found between randomly chosen soils, separated by up to 9000 km, with the largest soil 16S rRNA gene sampling effort to date. Although, we analyzed short pyrosequencing reads (103 bases in average), the choice of a hypervariable region of the 16S gene can yield the same clustering as full length sequences (Liu *et al.*, 2007) and allow for substantial resolution of similarities and differences between two bacterial communities. The main finding of this study is that less than 5% of the bacterial OTUs, as defined at the 97% similarity level, were detected with pyrosequencing in all four soils, If we use 100% sequence similarity in this 16S region as definition of species, then only 1.5% of the bacterial species were found in all soils.

The taxonomic correlates to 97 and 100% sequence similarity in this region of the 16S gene are not certain. An analysis of more than 2702 sequences from known species/genera suggests that OTUs defined by 100% similarities underestimates the number of species (as determined by simple rarefaction, see Table 2 in Roesch *et al.*, 2007). OTUs defined by 97% similarities overestimates the number of genera. Therefore, our fine level of resolution is not quite as fine as species, our coarse level somewhat finer than genera. For example, different species of *E. coli* and even some *Klebsiella* have identical sequences in this region. Different known genera can have more than 3% difference in sequence. We conclude that we cannot reliably detect species using this region of the 16S gene, but that both our definitions of OTUs lie between species and genera.

The geographic distribution of bacterial OTUs becomes more restricted as one defines an OTU by greater degrees of 16S rRNA similarity or by finer methods of genotyping. In all molecular studies of community composition in soils, the same major phyla are always discovered regardless of soil location. The community structure, at the phylum level, is nonetheless highly site specific (for example, Dunbar *et al.*, 1999; Hackl *et al.*, 2004; Oline,

Table 2 Average similarity indices for all pairwise comparisons. Data averaged over all pairwise comparisons are shown for Jaccard and θ indices at two levels of OTU definition, 100% sequence similarity and 97% similarity

	Jaccard similarity		*θ similarity*		*Number of sequences*
	Mean	*Std*	*Mean*	*Std*	
OTU = 97% similarity					
Proteobacteria					
α-proteobacteria	0.24	0.03	0.13	0.14	8865
β-proteobacteria	0.23	0.04	0.21	0.22	13 862
δ-proteobacteria	0.13	0.05	0.07	0.08	2857
γ-proteobacteria	0.17	0.04	0.04	0.06	2611
Acidobacteria	0.13	0.04	0.08	0.16	2459
Actinobacteria	0.24	0.05	0.19	0.11	5233
Aquificae	0.05	0.08	0.07	0.08	47
Bacteroidetes	0.28	0.04	0.28	0.11	24 951
Chlamydiae	0.08	0.07	0.04	0.08	106
Chloroflexi	0.10	0.06	0.07	0.13	259
Cyanobacteria	0.07	0.05	0.03	0.03	144
Firmicutes	0.13	0.04	0.04	0.05	3337
Gemmatimonadetes	0.19	0.08	0.15	0.16	483
Nitrospirae	0.14	0.05	0.22	0.26	1633
Planctomycetes	0.05	0.08	0.07	0.16	1042
Spirochaete	0.10	0.07	0.28	0.29	295
Thermus	0.17	0.12	0.25	0.35	77
Verrumicrobia	0.11	0.12	0.20	0.23	404
Candidate division OP10	0.26	0.12	0.23	0.17	369
Candidate division TM7	0.18	0.06	0.34	0.25	2073
Noclass	0.05	0.08	0.07	0.16	3963
OTU = 100% similarity					
Proteobacteria					
α-proteobacteria	0.10	0.02	0.12	0.13	8865
β-proteobacteria	0.09	0.02	0.21	0.22	13 862
δ-proteobacteria	0.05	0.02	0.06	0.06	2857
γ-proteobacteria	0.07	0.02	0.04	0.05	2611
Acidobacteria	0.06	0.02	0.08	0.15	2459
Actinobacteria	0.09	0.03	0.16	0.11	5233
Aquificae	0.02	0.05	0.02	0.04	47
Bacteridetes	0.09	0.01	0.29	0.11	24 951
Chlamydiae	0.02	0.02	0.01	0.02	106
Chloroflexi	0.03	0.02	0.01	0.02	259
Cyanobacteria	0.04	0.04	0.01	0.01	144
Firmicutes	0.05	0.02	0.03	0.05	3337
Gemmatimonadetes	0.06	0.01	0.11	0.13	483
Nitrospirae	0.06	0.02	0.14	0.19	1633
Planctomycetes	0.02	0.03	0.03	0.08	1042
Spirochaete	0.06	0.06	0.26	0.30	295
Thermus	0.07	0.06	0.12	0.19	77
Verrumicrobia	0.06	0.06	0.04	0.07	404
Candidate division OP10	0.09	0.03	0.17	0.23	369
Candidate division TM7	0.07	0.02	0.34	0.27	2073
Noclass	0.09	0.03	0.17	0.23	3963

2006; Herrera *et al.*, 2007; Labbe *et al.*, 2007; Lamarche *et al.*, 2007). In detailed studies community phyla structures also prove to be seasonally dynamic and highly heterogenous even in homogeneous landscapes (Axelrood *et al.*, 2002; Franklin and Mills, 2003; Sliwinski and Goodman, 2004). A few studies have attempted to correlate community structure to ecological parameters (Sessitsch *et al.*, 2001; Fierer and Jackson 2006; Fierer *et al.*, 2007).

Few studies have looked at finer phylogenetic levels in soils, but those that do show that at phylogenetic levels of resolution approaching that of species, distributions become more restricted. The more definitive examples of bacterial endemism have involved the analysis of isolates at the subspecies levels. Fulthorpe *et al.* (1998) found regional endemisim of 3-chlorobenzoate degrading soil bacteria sampled from six geographic regions at the level of REP genotype (whole genome fingerprinting), but not at the ARDRA (16S rRNA) level. Cho and Tiedje (2000) used the same soil collection to isolate widely dispersed fluorescent Pseudomonads, and also found no geographic pattern at the ARDRA level, some at the ITS level, but strong endemicity at BOX genotype level. More recently, Wawrik *et al.* (2007) demonstrated the distinctive-

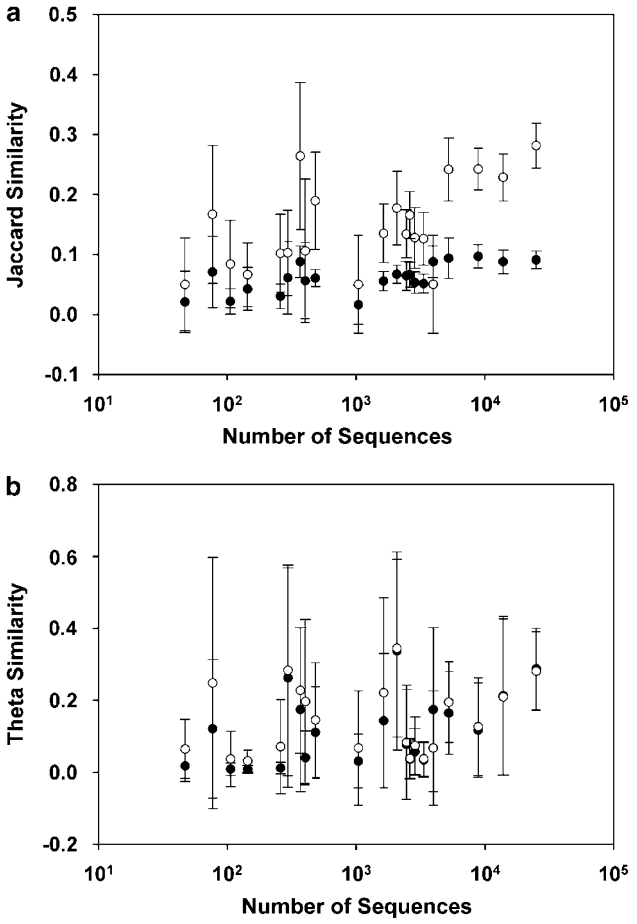

Figure 2 (**a**) Phyla average pairwise Jaccard similarity values versus number of sequences analyzed. Open circles—data derived from OTUs defined at 97% sequence similarity, $r^2 = 0.0.218$. Black circles—data derived from OTUs defined at 100% sequence similarity, $r^2 = 0.121$. (**b**) Phyla average pairwise θ similarity values versus number of sequences analyzed. Open circles—data derived from OTUs defined at 97% sequence similarity, $r^2 = 0.216$. Black circles—data derived from OTUs defined at 100% sequence similarity, $r^2 = 0.125$.

ness of New Jersey versus Uzbekistan actinomycete populations by looking at their polyketide synthase (PKS) genes. Greater than 30% of the isolates had PKS genes endemic to their sites, while TRFLP profiles of 16S rRNA genes demonstrated less isolation.

These studies demonstrate that geographic location and environmental conditions exert strong selection pressure on species composition. The question remains, are all species present although not dominant? This study provides the strongest evidence so far that even at the 16S rRNA sequence level; all taxa are not extant and ubiquitous. However, this statement must be tempered by what we know of bacterial community structure, and by the limitations of our methods. In general, bacterial species abundance curves generally exhibit log-normal distributions with very long tails (Hughes-Martiny *et al.*, 2006) In other words, a large percentage of species are present in extremely low numbers. The species that dominate the community can change rapidly with environmental conditions such as moisture, temperature or energy sources. It is entirely possible that the long tail of the rare species, those individuals that we could not detect, harbors a very large number of species that are truly ubiquitous. The species composition of the tails remain hidden to us because of (a) the number of sequence reads we obtained, (b) biases in the data due to incomplete cell lysis and other DNA extraction artifacts, (c) the tendency of PCR to preferentially amplify more numerically dominant sequences. Although we used a low number of cycles (20) to amplify the 16S fragments, and although we did so with 96 separate PCRs for each sample, we still could have under sampled rare taxa.

Therefore, we might conclude that not until metagenomic studies can provide enough sequence reads, without prior PCR amplification, from complete DNA extractions, can we be sure of the ubiquity of bacterial species. We must also acknowledge that not enough is known yet about the degree of variation of community structure within one soil 'sample'. Does the absence of species from a half-gram sample mean that those same species are absent from a different adjacent half-gram sample?

For now, we know that using current methods and according to the calculations in Roesch *et al.* (2007), it would require from 400 000 to 1.8 million more sequences per soil sample to obtain a complete picture of the species present. However, the correlations between the similarity estimates and sequence samples sizes were weak, especially at the fine level of resolution. The regression of OTUs shared by all four soils on sequence sample size gave nonsignificant slopes (that is, not statistically different from zero).

For practical purposes such as bioprospecting, these data support the idea that the sampling and study of different soils will uncover novel genotypes, species and even genera. For biogeochemical studies, the activities of dominant species are more important than the taxonomic makeup of the long tails.

The taxonomic classification of the most numerically dominant identifiable (and presumably culturable) genera in the samples tested was somewhat surprising (Table 1). In 1977, Martin Alexander included a list of the nine most significant genera in his textbook on soil microbiology (Alexander, 1977). These genera were *Agrobacterium, Alcaligenes, Arthrobacter, Bacillus, Flavobacterium, Micromonospira, Nocardia, Pseudomonas* and *Streptomyces.* Only three of these were among the 10 most dominant in our soils—*Bacillus, Flavobacterium* and *Pseudomonas* (and *Acidovorax* previously named *Pseudomonas*). The rest conform to the more recent wisdom on the abundance of phyla in soils, with members of Proteobacteria, Acidobacteria, Bacterioidetes, Firmicutes and Gemmatimonadetes

Table 3 Number and percentage of OTUs found in all four soil samples for each of the phylum analyzed, percentage of OTUs restricted to only one soil and total number of sequences analyzed per phylum

Phylum	*100% similarity*			*97% similarity*			*Number of sequences*		
	OTUs	*All (%)*	*Restricted*[a] *(%)*	*OTUs*	*All (%)*	*Restricted*[a] *(%)*	*Total*	*per OTU*[b]	*per OTU*[c]
Proteobacteria									
α-proteobacteria	1738	2.42	85.9	1738	2.42	85.8	8865	5.1	5.1
β-proteobacteria	1832	1.47	85.8	682	6.01	66.6	13 862	13.6	36.6
δ-proteobacteria	853	0.82	90.4	471	1.49	75.6	2857	3.3	6.1
γ-proteobacteria	723	1.52	90.5	375	4.53	76.8	2611	3.6	7.0
Acidobacteria	428	1.17	89.7	212	2.83	75.5	2459	5.7	11.6
Actinobacteria	1871	1.87	86.1	724	7.60	60.4	5233	2.8	7.2
Aquificae	43	0.00	97.7	23	0.00	87.0	47	1.1	2.0
Bacteriodetes	3658	1.94	85.5	1083	9.60	58.4	24 951	6.8	23.0
Chlamydiae	68	0.00	92.6	41	0.00	70.7	106	1.6	2.6
Chloroflexi	112	0.00	92.0	79	1.27	81.0	259	2.3	3.3
Cyanobacteria	79	0.00	92.4	60	0.00	86.7	144	1.8	2.4
Firmicutes	1134	0.79	91.0	660	1.82	77.0	3337	2.9	5.1
Gemmatimonadetes	149	0.67	89.9	76	3.95	71.1	483	3.2	6.4
Nitrospirae	310	0.97	91.6	162	3.09	82.7	1633	5.3	10.1
Planctomycetes	246	0.00	95.1	135	0.00	87.4	1042	4.2	7.7
Spirochaete	59	0.00	91.5	41	0.00	85.4	295	5.0	7.2
Thermus	32	0.00	84.4	18	0.00	66.7	77	2.4	4.3
Verrumicrobia	122	0.00	88.5	65	0.00	81.5	404	3.3	6.2
Candidatus division OP10	109	2.75	89.0	46	8.70	67.4	369	3.4	8.0
Candidatus division TM7	451	1.33	89.4	191	3.66	69.6	2073	4.6	10.9
Noclass	1366	0.66	92.1	809	1.85	81.7	3963	2.9	4.9
Overall	15 383	1.50	87.9	7691	4.10	74.4	86 160	5.6	11.2

[a]The sum of those that are restricted to one of the four soils analyzed.
[b]Number of sequences per OTU at 100% similarity.
[c]Number of sequences per OTU at 97% similarity.

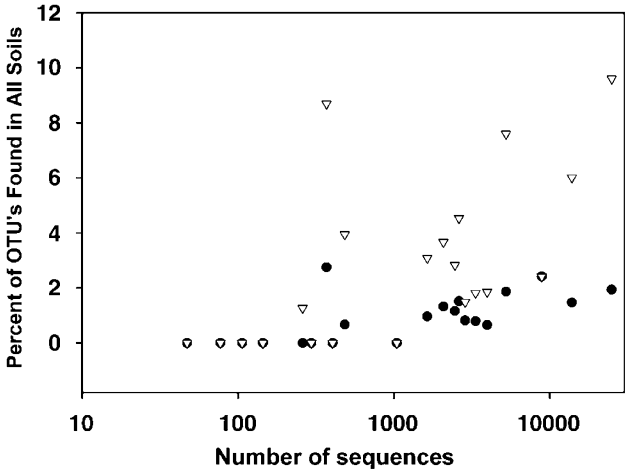

Figure 3 Percent of OTUs found in all four soils versus number of sequences analyzed. Open triangles—data derived from OTUs defined at 97% sequence similarity, $r^2 = 0.41$. Black circles—data derived from OTUs defined at 100% sequence similarity, $r^2 = 0.27$. Note *x* axis is log scale. Slope of linear regression is not significantly different from zero for both data sets.

being well represented. This list of identifiable OTUs is heavy with members of the Bacteroidetes. This phylum seems to include a large number of culturable strains, relative to other 'new' phyla such as *Acidobacteria*. In a cross-study survey of soil bacteria, Janssen (2006) noted that certain genera of the Bacteroidetes that is, *Chitinophaga, Flavobacterium* and *Pedobacter* were within the top 34–62% of Bacteroidetes sequences, although the phyla as a whole averaged about 5% of soil communities.

Table 1 includes genera that have barely been studied or even isolated. *Acidobacterium* species were found in relative abundance in all the soils as expected. In Janssen's survey of soil libraries, *Acidobacteria* are the second largest phylum after *Proteobacteria* and make up an average of 20% of the sequences found, despite being first described in 1991 (Kishimoto *et al.*, 1991; Janssen 2006).

An unexpected result from this work was the prevalence of *Chitinophaga* in all four soils—it was either the most abundant or second most abundant cultured genus in all four soils. This genus was represented by 7.5–13.8% of the reads in each soil. *Chitinophaga* was first described in 1981, for a new group of filamentous, chitinolytic gliding bacteria (Sangkhobol and Skerman 1981). These chitinolytic bacteria may be using fungal hyphae and insects as sources of chitin. Their high numbers may be the result of the fact that fungal hyphae and insects comprise a significant proportion of soil biomass. A related but more well-known genus, *Sporocytophaga* made up 1.1% of sequences in the soil sampled in Florida. *Sporocytophaga* are spore forming gliding bacteria that do not form fruiting bodies, and are considered to be active cellulose degraders in soils.

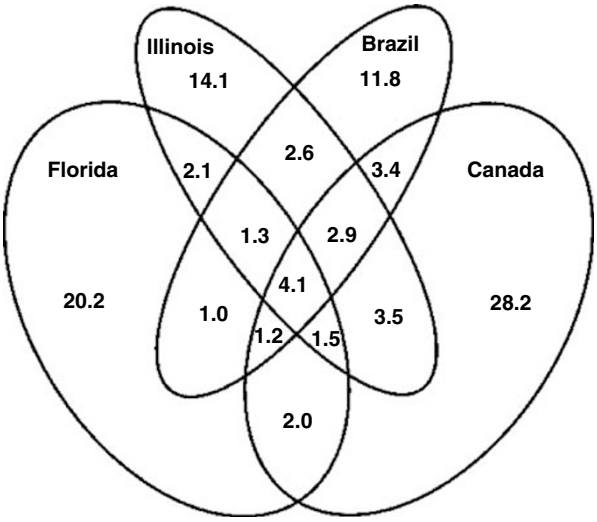

Figure 4 Venn diagram showing overall overlap of OTUs between soils. OTUs defined at 97% sequence similarity or more.

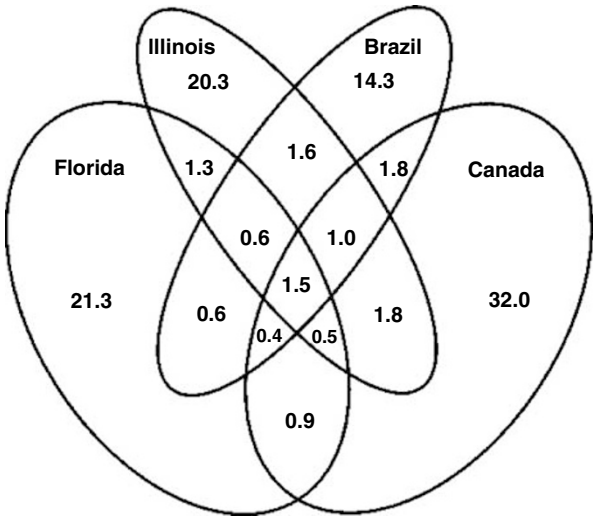

Figure 5 Venn diagram showing overall overlap of OTUs between soils. OTUs defined at 100% sequence similarity.

Notable among the most abundant identifiable/culturable taxa is *Thiobacter.* It represented about 2% of the community in three soils and yet the genus was only described in 2005. *Thiobacter subterraneus* was isolated from a subsurface aquifer with an ambient temperature of 70 °C (Hirayama *et al.*, 2005). Clearly much more ubiquitous members of this exist and are important in mesophilic soils.

Gemmatimonas was described as new genera within a new phylum of the BD group by Zhang *et al.* (2003). The type strain is a polyphosphate accumulating aerobe. Taxa of this genus were common in the soil sampled in Canada and Illinois.

Chondromyces and *Haliangium* representatives were abundant in two of the soils tested (Canada and Illinois). Both of these genera are myxobacteria and the global distribution of this interesting group of 'social bacteria' in soils has been detailed by Dawis (2000). We are unaware of any reports that have shown them to be numerically important before (that is the two genera together rival the abundance of the *Acidobacteria* in the soil samples from Canada and Illinois).

The genus *Pedobacter* was named in 1998 (Steyn *et al.*, 1998). It is another gliding bacterium of the *Bacteroidetes*, in the novel family *Sphingobacteriaceae*. A strict aerobe, the type strain is best known for heparinase production, and is frequently isolated from soils or activated sludges and freshwater. Four new species have been described in 2007 alone, many in Korea (that is, Yoon *et al.*, 2007a, b).

Nevskia members are commonly found at water surfaces—a habitat termed the 'neuston' (Pladdies *et al.*, 2004), but they have not been described as being important in the soil habitat. Juteau *et al.* (1999) found significant numbers of *Nevskia* in a biofilter treating toluene. Related *Hydrocarboniphaga* species were named in 2004 for their ability to degrade alkanes (Palleroni *et al.*, 2004).

Aquabacterium is another genus seemingly out of place. It has been found to make up 2–53% of the cells found in drinking water system biofilms in Germany (Kalmbach *et al.*, 2000), and Loy *et al.* (2005) noted that they commonly make up a large portion of the microbial communities found in bottled natural mineral water. It made up 2.7% of the sequences from Canada.

Dyadobacter is a novel genus described in 2000 after isolation of a corn endophyte (*Dyadobacter fermentans*, Chelius and Triplett, 2000). Since then, numerous other members of this genus have been isolated from diverse terrestrial habitats (Chaturvedi *et al.*, 2005; Reddy and Garcia-Pichel 2005; Liu *et al.*, 2006; Baik *et al.*, 2007; Dong *et al.*, 2007; Suihko *et al.*, 2007).

Stentrophomonas, previously classified as a *Pseudomonas* and then *Xanthomonas*, is recognized as being ubiquitous in waters and soils. It is best known by the opportunistic pathogen *S. maltophilia* that is infamous for possessing broad-spectrum antibiotic resistance.

Along side each genus in Table 1 is the number of literature citations on that genus in Web of Science as of 11 February 2008. It is clear that opportunities exist for greatly expanding our understanding of a number of important soil bacteria using classical approaches—most of the top 10 genera are not well studied. Perhaps the most striking example of this is *Chitinophaga*, which is either the most or second-most abundant genus in every soil but only 11 papers discuss this organism in the literature. Although the recent emphasis on culture-independent analyses is warranted, much will be gained by studying many of the soil organisms that are already available in culture.

References

Alexander M. (1977). *Introduction to Soil Microbiology*. Wiley: New York.

Axelrood PR, Chjow ML, Radomski CC, McDermott JM, Davis J. (2002). Molecular characterization of bacterial diversity from British Columbia forest soils subjected to disturbance. *Can J Microbiol* **48**: 655–674.

Baik KS, Kim MS, Kim EM, Kim HR, Seong CN. (2007). *Dyadobacter koreensis* sp nov., isolated from fresh water. *Int J Syst Evol Microbiol* **57**: 1227–1231.

Chaturvedi P, Reddy GSN, Shivaji S. (2005). *Dyadobacter hamtensis* sp. nov., from Hamta glacier, located in the Himalayas, India. *Int J Syst Evol Microbiol* **55**: 2113–2117.

Chelius MK, Triplett EW. (2000). *Dyadobacter fermentans* gen. nov. sp. nov. a novel Gram-negative bacterium isolated from surface sterilized Zea mays stems. *Int J Syst Evol Microbiol* **50**: 751–758.

Cho JC, Tiedje JM. (2000). Biogeography and degree of endemicity of fluorescent Pseudomonas strains in soil. *Appl Environ Microbiol* **66**: 5448–5456.

Dawis W. (2000). Biology and global distribution of myxobacteria in soils. *FEMS Microbiol Rev* **24**: 403–427.

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

Dong Z, Guo X, Shang X, Qiu F, Sun L, Gong H *et al.* (2007). *Dyadobacter beijingensis* sp nov., isolated from the rhizosphere of turf grasses in China. *Int J Syst Evol Microbiol* **57**: 862–865.

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

Felsenstein J. (1989). PHYLIP-Phylogeny inference package (version 3.2). *Cladistics* **5**: 164–166.

Felsenstein J. (2005). Using the quantitative genetic threshold model for inferences between and within species. *Philos T R Soc B* **360**: 1427–1434.

Fierer N, Bradford MA, Jackson RB. (2007). Toward an ecological classification of soil bacteria. *Ecology* **88**: 1354–1364.

Fierer N, Jackson RB. (2006). The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103**: 626–631.

Franklin RB, Mills AL. (2003). Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia agricultural field. *FEMS Microb Ecol* **44**: 335–346.

Fulthorpe RR, Rhodes AN, Tiedje JM. (1998). High levels of endemicity apparent in 3-chlorobenzoate degrading soil bacteria. *Appl Environ Microbiol* **64**: 1620–1627.

Gans J, Wolinsky M, Durbar J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**: 1387–1390.

Hackl E, Zechmeister-Boltenstern S, Bodrossy L, Sessitsch A. (2004). Comparison of diversities and compositions of bacterial populations inhabiting natural forest soils. *Appl Environ Microbiol* **70**: 5057–5065.

Herrera A, Hery M, Stach JEM, Jaffre T, Norman P, Navarro E. (2007). Species richness and phylogenetic diversity comparison of soil microbial communities affected by nickel-mining and revegetation efforts in New Caledonia. *Eur J Soil Biol* **43**: 130–139.

Hirayama Hdirayama H, Takai K, Inagaki F, Nealson KH, Horikoshi K. (2005). *Thiobacter subterraneus* gen. nov., sp. nov., and obligately chemolithoautotrophic, thermophilic, sulfur-oxidizing bacterium from a subsurface hot aquifer. *Int J Sys Evol Microbiol* **55**: 467–472.

Hughes-Martiny JB, Bohannan BHM, Brown JH, Colwell RK, Fuhrman JA, Green JL *et al.* (2006). Microbial biogeography: putting microorganisms on the map. *Nature Rev Microb* **4**: 102–112.

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

Juteau P, Rho D, Larocque R, LeDuy A. (1999). Analysis of the relative abundance of different types of bacteria capable of toluene degradation in a compost biofilter. *Appl Microb Biotech* **52**: 863–868.

Kalmbach S, Manz W, Bendinger B, Szewzyk U. (2000). *In situ* probing reveals *Aquabacterium commune* as a widespread and highly abundant bacterial species in drinking water biofilms. *Wat Res* **34**: 575–581.

Kishimoto N, Kosako Y, Tano T. (1991). *Acidobacterium capsulatum* gen. nov., sp. nov.: an acidphilic chemoorganiotrophic bacterium containing menaquinone from an acidic mineral environment. *Curr Microbiol* **22**: 1–7.

Labbe D, Margesin R, Schinner F, Whyte LG, Greer CW. (2007). Comparative phylogenetic analysis of microbial communities in pristine and hydrocarbon-contaminated alpine soils. *FEMS Microb Ecol* **59**: 466–475.

Lamarche J, Bradley RL, Hooper E, Shipley B, Beaunoir A_MS, Beaulieu C. (2007). Forest floor bacterial community composition and catabolic profiles in relation to landscape features in Quebec's southern boreal forest. *Microbial Ecol* **54**: 10–20.

Liu QM, Im WT, Lee M, Yang DC, Lee ST. (2006). *Dyadobacter ginsengisoli* sp nov., isolated from soil of a ginseng field. *Int J Syst Evol Microbiol* **56**: 1939–1944.

Liu Z, Lozupone C, Hamady M, Bushman D, Knight R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nuc Acid Res* **35**: 1–10.

Loy A, Beisker W, Meier H. (2005). Diversity of bacteria growing in natural mineral water after bottling. *Appl Environ Microbiol* **71**: 3624–3632.

Oline DK. (2006). Phylogenetic comparisons of bacterial communities from serpentine and nonserpentine soils. *Appl Environ Microbiol* **72**: 6965–6971.

Palleroni NJ, Port AM, Chang HK, Zylstra GJ. (2004). *Hydrocaroniphage effusa* gen. nov., sp. nov., a novel member of the gamma-Proteobacteria active in alkane and aromatic hydrocarbon degradation. *Int J Syst Evol Microbiol* **54**: 1203–1207.

Pladdies T, Babenzien HD, Cypionka H. (2004). Distribution of *Nevskia ramosa* and other rosette-forming neustonis bacteria. *Microb Ecol* **47**: 218–223.

Reddy G, Garcia-Pichel F. (2005). *Dyadobacter crusticola* sp nov., from biological soil crusts in the Colorado Plateau, US, and an emended description of the genus *Dyadobacter* Chelius and Triplett 2000. *Int J Syst Evol Microbiol* **55**: 1295–1299.

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

Sangkhobol V, Skerman VBD. (1981). *Chitinophaga*, a new genus of chitinolytic myxobacteria. *Int J Syst Bacteriol* **31**: 285–293.

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). Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol* **72**: 6773–6779.

Sessitsch A, Weilharter A, Gerzabek MH, Kirchmann H, Kandeler E. (2001). Microbial population structures in soil particle size fractions of a long-term fertilizer field experiment. *Appl Environ Microb* **67**: 4215–4224.

Sliwinski MK, Goodman RM. (2004). Spatial heterogeneity of crenarchaeal assemblages within mesophilic soil ecosystems as revealed by PCR-single stranded conformation polymorphism profiling. *Appl Environ Microb* **70**: 1811–1820.

Steyn PL, Segers P, Vancanneyt M, Sandra P, Kersters K, Joubert JJ. (1998). Classification of heparinolytic bacteria in to a new genus, *Pedobacter*, comprising four species: *Pedobacter* heparainus comb. nov., *Pedobacter* piscium comb. nov., Pedobacter africanus sp. nov. and *Pedobacter* saltans sp. nov. Proposal of the family Sphingobacteriaceae fam. nov. *Int J Syst Bact* **48**: 165–177.

Suihko ML, Alakomi HL, Gorbushina A, Fortune I, Marquardt E, Saarela M. (2007). Characterization of aerobic bacterial and fungal microbiotia on surfaces of historic Scottish monuments. *Syst Appl Microbiol* **30**: 494–508.

Wawrik B, Kutliev D, Abdivasievna UA, Kukor JJ, Zylstra GJ, Kerkhof L. (2007). Biogeography of actinomycete communities and type II polyketide synthase genes in soils collected in New Jersey and Central Asia. *Appl Environ Microbiol* **73**: 2982–2989.

Yoon J-H, Kang S-J, Oh H-W, Oh T-K. (2007a). *Pedobacter insulae* sp. nov. isolated from soil. *Int J Syst Evol Microb* **57**: 1999–2003.

Yoon J-H, Kang S-J, Park S, Oh T-K. (2007b). *Pedobacter lentus* sp. nov. and Pedobacter terricola sp. nov. isolated from soil. *Int J Syst Evol Microb* **57**: 2089–2095.

Yue JC, Clayton MK. (2005). A similarity measure based on species proportions. *Commun Stat Theor Methods* **34**: 2123–2131.

Zhang H, Sekiguchi Y, Hanada S, Hugenholtz P, Kim H, Kamagata Y *et al.* (2003). *Gemmatimonas aurantiaca* gen. nov., sp. nov., a Gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. *nov* **53**: 1155–1163.

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