10

This is to certify that the
dissertation entitled

MICROBIAL COMMUNITY ANALYSIS ASSESSED BY
PYROSEQUENCING OF rRNA GENE: COMMUNITY
COMPARISONS, ORGANISM IDENTIFICATION, AND ITS
ENHANCEMENT

presented by

WOO JUN SUL

has been accepted towards fulfillment
of the requirements for the

Doctor of Philosophy degree in Crop and Soil Sciences – Environmental Toxicology

Major Professor's Signature

Dec. 16, 2009

Date

MSU is an Affirmative Action/Equal Opportunity Employer

MICROBIAL COMMUNITY ANALYSIS ASSESSED BY PYROSEQUENCING OF rRNA GENE: COMMUNITY COMPARISONS, ORGANISM IDENTIFICATION, AND ITS ENHANCEMENT

By

Woo Jun Sul

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Crop and Soil Sciences – Environmental Toxicology

2009

ABSTRACT

MICROBIAL COMMUNITY ANALYSIS ASSESSED BY PYROSEQUENCING OF rRNA GENE: COMMUNITY COMPARISONS, ORGANISM IDENTIFICATION, AND ITS ENHANCEMENT

By

Woo Jun Sul

There are more than 10^{30} bacteria on Earth, with their members embedding 3.8 billion years evolutionary history and having evolved to take advantage of virtually every energy-yielding niche hospitable to life. This makes the microbial world extremely diverse, ubiquitous and essential to Earth's habitability. Hence, determining which microbes make up these communities is an initial goal for understanding microbial communities. Recently, pyrosequencing of ribosomal RNA genes has become a popular tool for in-depth analyses of microbial communities. I used pyrosequencing of rRNA's hypervariable V4-region to characterize a wide variety of microbial communities.

Soil microbial communities in the tropics are potentially more dynamic than temperate ones due to longer and more favorable temperature, moisture and energy resources from primary productivity. I studied the effect on soil Bacteria of different soil-crop management systems in Eastern Ghana, one of which lost 50% of its stored soil organic carbon (SOC) within 4 years. Canonical correspondence analysis and stepwise multiple regression of the 290,000 V4-rRNA sequences showed that SOC was the most important factor that explained differences in microbial community structure among managements. The data indicate that the use of a pigeon-pea crop (a legume) during the winter season (normally fallow) promotes a higher microbial diversity and sequesters more soil organic carbon, which is important for soil structure, nutrient retention and

recycling, and general soil health. I also evaluated analysis methods for 211 rRNA-determined bacterial assemblages, comprising 1.3 million rRNA sequences from seven habitat types. A taxonomy-supervised method, using taxonomy-bins, was advantageous in its ability to compare non-overlapping sequences, and requiring minimal computation capacity compared to the non-taxonomy-supervised (clustering-determined) method. The taxonomy supervised method produced results that were significantly correlated to the clustering method, which is the current standard, and as taxonomy improves should provide even better resolution. Because of the much greater depth and replication provided by pyrosequencing, more robust determination of microbial species distribution, diversity, organism identification, community comparisons and dynamics is possible.

As a result of microbes' long history, they harbor considerable genetic diversity and some of their genes likely have more desirable properties that those known. I used stable isotope probing (SIP) with [^{13}C]-biphenyl as substrate to retrieve novel biphenyl dioxygenase subunits *bphAE* which showed PCB oxidative activity in the 31.8 kb cosmid clone made from the [^{13}C]-DNA. The discrepancy of G+C content near the *bphAE* genes implies their recent acquisition, possibly by horizontal transfer, and suggests dispersed dioxygenase gene organization in nature. I also used V4-16S rRNA gene pyrosequencing of the [^{13}C]-biphenyl-derived DNA from three PCB-contaminated environmental matrices: rhizosphere, industrial soil, and river sediment to more specifically identify the PCB- and biphenyl-utilizing populations of the three sites. I found little commonality in the abundant members of three sites but new candidate groups that may be involved in PCB degradation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Images in this dissertation are presented in color.

CHAPTER I

WHAT WE LEARN FROM MICROBIAL COMMUNITY PROFILING BY rRNA GENE PYROSEQUENCING: AN INTRODUCTION

BLOSSOMING OF 16S rRNA GENE SEQUENCES

Since the complete 16S ribosomal RNA gene in *Escherichia coli* was sequenced in 1978 (Brosius *et al.*, 1978), the demands and usages of 16S rRNA gene sequencing have increased in the fields of microbiology and microbial ecology. Ribosomal RNA genes sequences have been used not only as marker genes to shape bacterial phylogeny but also as surrogates to reveal microbial community composition. Over past the 30 years, the numbers of rRNA gene sequences per study have greatly increased (Figure 1) because of lower sequencing costs and, recently, because of massively parallel capacity, such as by 454's pyrosequencing technology. This technology has been successfully used as a rapid and efficient tool for in-depth analysis of microbial communities including comparisons of microbial communities and the pre-diagnosis of microbial communities prior to metagenomic analysis (Tringe and Hugenholtz, 2008).

DEEP SEQUENCING

Cost-effective microbial community fingerprinting methods such as denaturing gradient gel electrophoresis (DGGE), temperature gradient gel electrophoresis (TGGE), single-strand conformation polymorphism (SSCP), terminal restriction fragment length polymorphism (T-RFLP), amplified rDNA restriction analysis (ARDRA), amplified ribosomal intergenic spacer analysis (ARISA), were widely applied in the previous decade due to their reasonable costs and convenience of performing and interpreting data

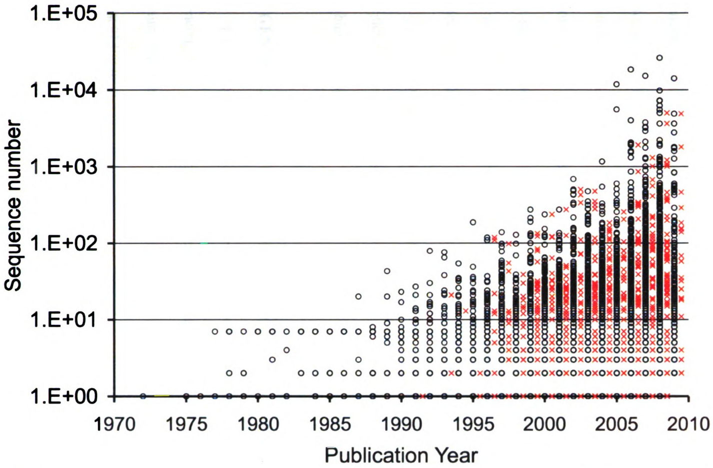


Figure 1.1 The numbers of 16S rRNA gene sequences per study. Black open circles indicated the numbers of sequences of studies deposited in RDP database and published. Red crosses include studies from searching keywords "soil" or "sediment" in literature titles.

(Anderson and Cairney, 2004). However, these methods detected only the most dominant community members, and thus they have limited resolution for describing microbial community structure. Conventional 16S rRNA clone libraries determined by Sanger sequencing are more informative about those members sampled but the cost is too great to reach the numbers of sequences to provide sufficient community resolution, although a few studies attained more than ten thousand sequences with incredible sequencing investments (Ley *et al.*, 2008). Hence, the large numbers of sequences produced by 16S rRNA gene pyrosequencing is major breakthrough in overcoming the obstructions in cost and low resolution of the previous methods. For a similar cost, pyrosequencing can produce hundreds of times more sequences than the 16S rRNA clone conventional libraries, thus providing better community descriptions (Figure 2). These huge numbers of sequences along with phylogenetic information, provide more precise signatures of microbial communities.

GENERAL PROCEDURE FOR 16S rRNA GENE PYROSEQUENCING

The procedures of 16S rRNA gene pyrosequencing consist of four parts: 1) pre-sequencing steps include sampling design, DNA extraction, 16S rRNA primer design or selection, barcodes selection, and PCR to produce the amplicons for pyrosequencing, and mixing of the amplicons for the sequencing plate or section, 2) pyrosequencing itself, 3) initial processing of sequences data, including trimming of barcodes, filtering out bad sequences, alignment of sequences, sequences clustering to generate OTUs, assignment of sequences' taxonomy, and 4) data analysis of processed sequences, including calculation of microbial community richness, evenness, and diversity as well as

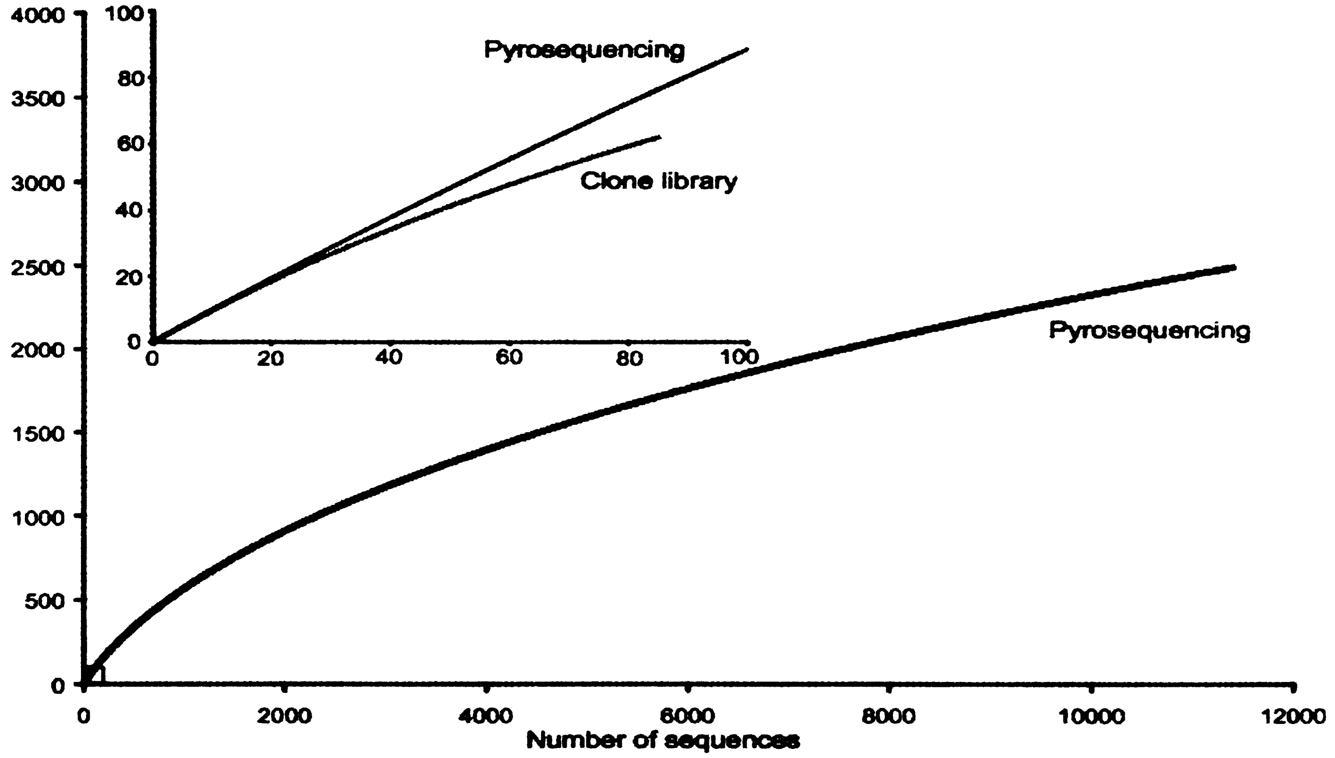
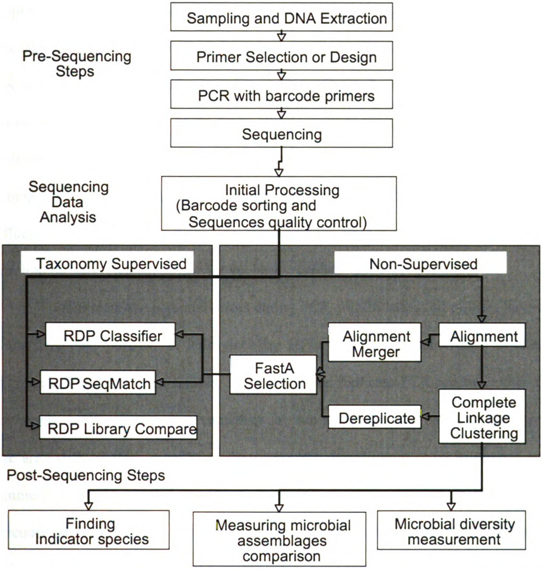


Figure 1.2. Comparison of rarefaction curves with the outcome from pyrosequencing and Sanger sequencing of conventional clone libraries from a PCB-contaminated rhizosphere sample. Small box in the lower graph is expanded in the upper graph.

community comparison index to facilitate the interpretation of sequence data, and the necessary analyses, e.g. statistical, to derive scientific conclusions (Figure 3).

CONSIDERATIONS IN PROCEDURES FOR 16S rRNA PYROSEQUENCING

When selecting or designing 16S rRNA primer sets, several aspects need to be considered: 1) an appropriate PCR product length for currently available pyrosequencing reads (~ 240 bps for GS FLX and ~400 bps for GS Titanium), 2) adequacy of 16S primers for coverage of Bacterial or Archael groups, 3) high resolution and accuracy of selected regions for organism identification, and 4) low frequency of insertions and deletions to simplify sequence alignment and to retain more comparable sequences. The choice of 16S primers strongly influences the coverage of 16S rRNA genes in microbial communities and, therefore, can lead to a biased representation of microbial communities. 16S primers preferentially selects or rules out specific taxa (reviewed in Hamady and Knight, 2009) and over- / underestimates their microbial richness (Youssef *et al.*, 2009). Different primer selection may potentially lead the different research conclusions (Hamp *et al.*, 2009). Due to the inability to cover the entire 16S rRNA gene sequence by current high-throughput sequencing methods, selection of a "good" region is important for good taxonomy assignment (Liu *et al.*, 2007; Wang *et al.*, 2007). Furthermore, the low frequency of insertions and deletions in the sequencing region is also important to simplify sequence alignment and to retain more comparable sequences (Cole *et al.*, 2009). Currently the more popular regions used in 16S rRNA pyrosequencing include the regions surrounding V2, V4, and V6 (Sogin *et al.*, 2006; Huse *et al.*, 2007; Roesch *et al.*, 2007; Andersson *et al.*, 2008; Chapter 2, 4, and 5).



Pre-Sequencing Steps
Sampling and DNA Extraction
Primer Selection or Design
PCR with barcode primers
Sequencing
Sequencing Data Analysis
Initial Processing (Barcode sorting and Sequences quality control)
Taxonomy Supervised
Non-Supervised
RDP Classifier
RDP SeqMatch
RDP Library Compare
FastA Selection
Alignment Merger
Alignment
Dereplicate
Complete Linkage Clustering
Post-Sequencing Steps
Finding Indicator species
Measuring microbial assemblages comparison
Microbial diversity measurement

Figure 1.3. Suggested procedure for 16S rRNA gene pyrosequencing. RDP Pyrosequencing Pipeline provides a trained **aligner** on a small hand-curated set of high-quality, full-length rRNA gene sequences. These aligned sequences can be clustered by **Complete Linkage Clustering**, a method of calculating distance between clusters in hierarchical cluster analysis. For identifying clusters' bacterial taxonomy, **Dereplicate Request** allows users to select a representative sequence from each cluster. The sequence with the minimum sum of the square of distances between sequences within a cluster is assigned as the representative sequence for that cluster. Representative sequences can easily be retrieved from original sample's sequences using **FASTA Sequence Selection**. **Alignment Merger** helps sequence retrieval from multiple alignment files.

Positioning barcode (key or tag) nucleotides, such as those calculated by Parameswaran *et al.* (2007) and Hamady *et al.* (2009), by positioning them between adaptor sequences to pyrosequencing beads and 16S rRNA gene primers allows one to mix multiple samples in one pyrosequencing run (Figure 4). Also, RDP's Pyrosequencing Pipeline lists 72 barcodes of 8-base length (V4-adaptor A primer specific) that have a minimum difference of 2 bases from all other barcodes (Cole *et al.*, 2009), avoid problematic order of nucleotide addition in pyrosequencing flow, and do not include homopolymers that increase possibility of sequencing error (Quinlan *et al.*, 2008). The influence of the barcode sequence on biasing the sequences amplified by extending the match to the target sequence has not yet been established.

To minimize the potential errors during PCR amplification, all primers have to be synthesized and purified at least once by HPLC to remove incorrectly synthesized oligonucleotides. For each sample, more than three replicate PCR reactions with DNA polymerases with proofreading capability are run in parallel and bands of the expected size are extracted from a gel after electrophoretic separation in order to remove primer dimmers and primer residues. When analyzing multiple samples in the same run, barcoded primers are used for amplification and the amplicons are carefully quantified and mixed together in equimolar amounts before applying to the sequencing plate.

Processing raw sequence data includes filtering out low-quality reads, although error rates for pyrosequencing was only 0.4% with the GS 20 instrument (Huse *et al.* 2007). The suggested procedure is to discard reads with any errors in the 16S primers and barcodes or below the average quality score of 25 (Huse *et al.* 2007). In addition, checking the error in reverse primers (3'end, opposite primer from sequencing start) is a

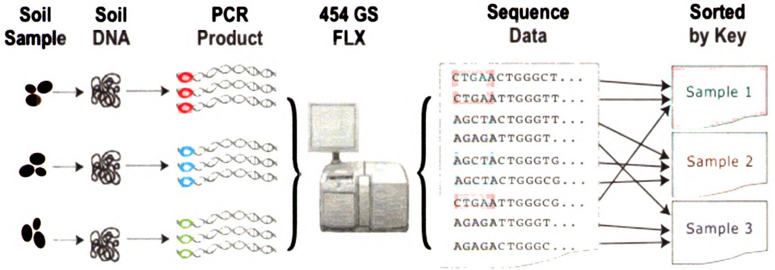


Figure 1.4. Schematic diagram of 16S rRNA gene pyrosequencing with barcode (tag, key) primers. Multiple samples can be accommodated in one 454 run. Sequences are sorted by barcode using RDP's Pyrosequencing Pipeline.

further option (allowing maximum reverse primer edit distance) to filter out low-quality reads because of the greater tendency for errors to occur when sequence reads reach the 3′-end.

Filtered sequences can be assigned to bacterial taxonomy by several applications: RDP Classifier, a naïve Bayesian rRNA classifier (Wang *et al.*, 2007), searching for nearest neighbor SeqMatch (Wang *et al.*, 2007), SILVA (Pruesse et al., 2007), Greengene (DeSantis *et al.*, 2006). In order to cluster the sequences to generate OTUs (operational taxonomic units), sequence alignment can be performed with the Infernal aligner, a SCFG-based, secondary-structure aware aligner, (Nawrocki & Eddy, 2009) adapted in RDP Pyrosequencing Pipeline, and NAST; Nearest Alignment Space Termination (DeSantis *et al.*, 2006b).

QUANTIFICATION OF COMMUNITY STRUCTURE BY 16S rRNA GENE PYROSEQUENCING

The quantification of bacterial species abundance by rRNA gene pyrosequencing has been compared with other abundance measures such as FISH, quantitative PCR and 16S rRNA gene clone libraries. For example, relative abundances of *Exiquobacterium* and *Psychrobacter* measured by V4-rRNA gene pyrosequencing were correlated to relative abundances from Q-PCR of the same organism's 16S rRNA gene, after correction for copy number (Figure 5). The relative abundances of certain bacterial groups using V4-rRNA pyrosequencing in wastewater treatment systems was found to be correlated to results from other methods although there were some exceptions: *Chloroflexi* and *Nitrospira* abundances were overestimated by FISH (or underestimated by rRNA pyrosequencing), while the reverse was true for *Betaproteobacteria* by clone libraries (Table 1). The uncertainty, due to potential primer bias, in PCR-based measurements makes that the reliability of quantification by 16S rRNA pyrosequencing uncertain (Figure 6). The measurements of *Acetobacterium* abundances using V6-rRNA pyrosequencing did not correspond with FISH-measurements using *Acetobacterium*-specific probe. The rRNA pyrosequencing result might be over-represented due to a combination of DNA extraction and PCR bias (Gaidos *et al.*, 2009).

Detection of bacterial species by rRNA pyrosequencing can also be compared to culture-based methods. When these methods were compared the "culture-negative/pyrosequencing-positive discordant pairs" (found only in pyrosequencing data set) were found, but "culture-positive/pyrosequencing-negative discordant pairs" (only by culturing) were rarely found (Price *et al.*, 2009). The genus *Rhodococcus* was dominant by isolation, but ruled not detected in a clone library from a Czech PCB-contaminated soil (Leigh *et al.*, 2006; Leigh *et al.*, 2007).

However, results of rRNA pyrosequencing showed that *Rhodococcus* was present in low abundance but preferentially cultured in this case (Chapter 4).

PHYLOGENY BY 16S rRNA GENE PYROSEQUENCING

Phylogenetic analyses using pyrosequencing data has proven useful (Andersson *et al.*, 2008; Chapter 2). However, studying the bacterial phylogeny with pyrosequencing sequences is strictly limited by the degree of polymorphism of bacterial groups within the sequenced 16S rRNA gene region. Short read lengths makes the phylogenetic analysis less robust due to decreased resolution, certainly the case at the species level for FLX

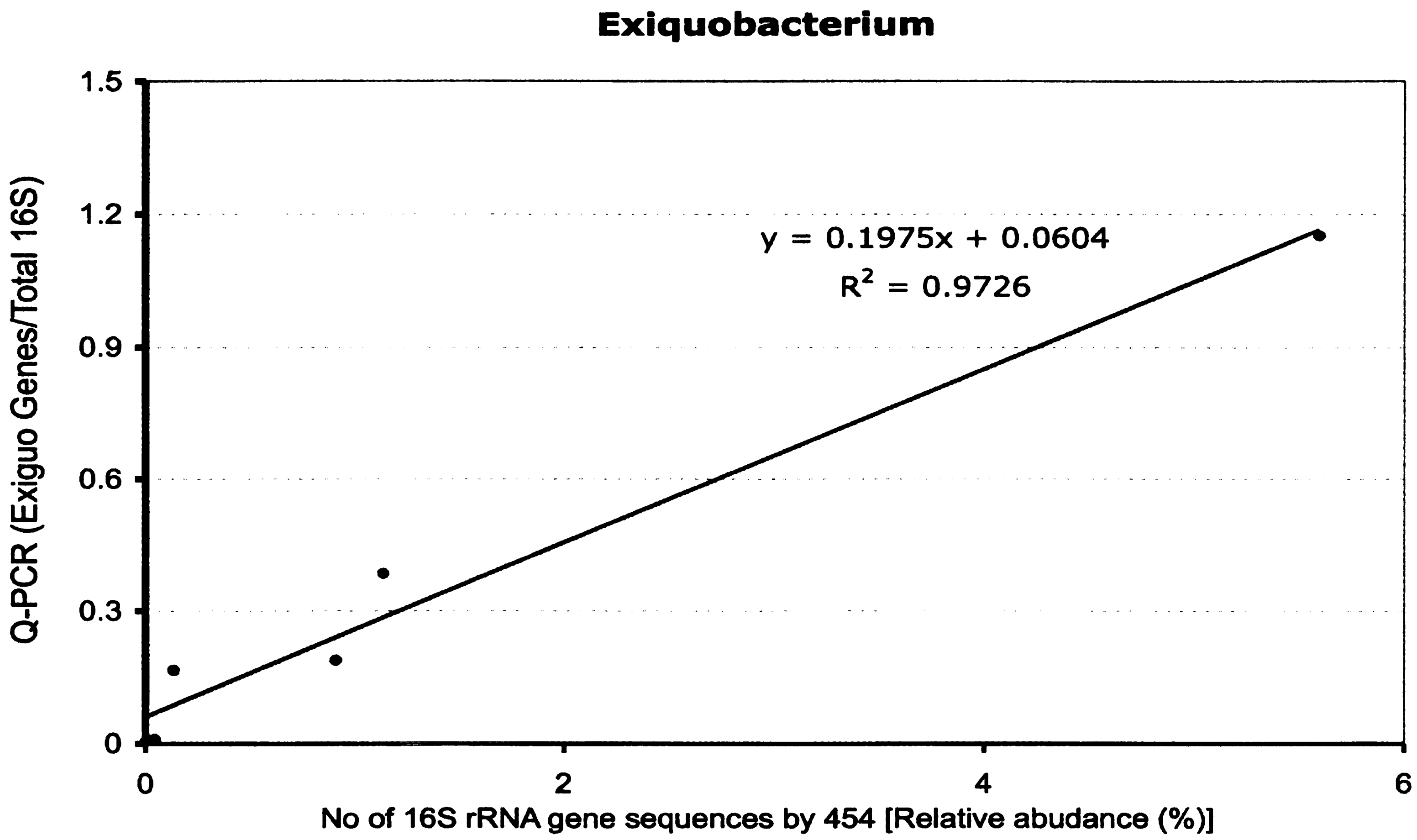


Figure 1.5. Comparison of relative abundance by quantitative PCR and 16S pyrosequencing. *Exiquobacterium* sp. specific 16S rRNA primer and universal rRNA primers were used for quantitative PCR (Rodrigues *et al*, 2009). V4-16S rRNA pyrosequencing was performed on the same 6 permafrost samples and *Exiquobacterium* sequences were determined by RDP Classifier at 50% confidence level.

	Pyrosequencing	FISH	FISH/Bacteria	Clone library
Alphaproteobacteria	16.2	18.5	12.58	9
Betaproteobacteria	13.1	27	18.36	50
Hydrogenophaga	1.2	NA	NA	11
Comamonas	5.8	NA	NA	1
Deltaproteobacteria	1.3	NA	NA	1
Epsilonproteobacteria	0.8	NA	NA	3
Gammaproteobacteria	9.6	NA	NA	9
Planctomycetes	4.8	NA	NA	7
Bacteroidetes	0.6	NA	NA	5
Chloroflexi	3.3	32	21.76	1
Nitrosomonas	1.6	5	3.4	ND
Nitrospira	0.05	18	12.24	ND
Nitrobacter	ND	ND	ND	ND
Nitrosospira	ND	ND	ND	ND

ND: not detetected
NA: not available

Table 1. Comparisons of relative abundances in a high nitrate wastewater treatment system in Uruguay, measured by rRNA gene pyrosequencing, FISH, and conventional clone library

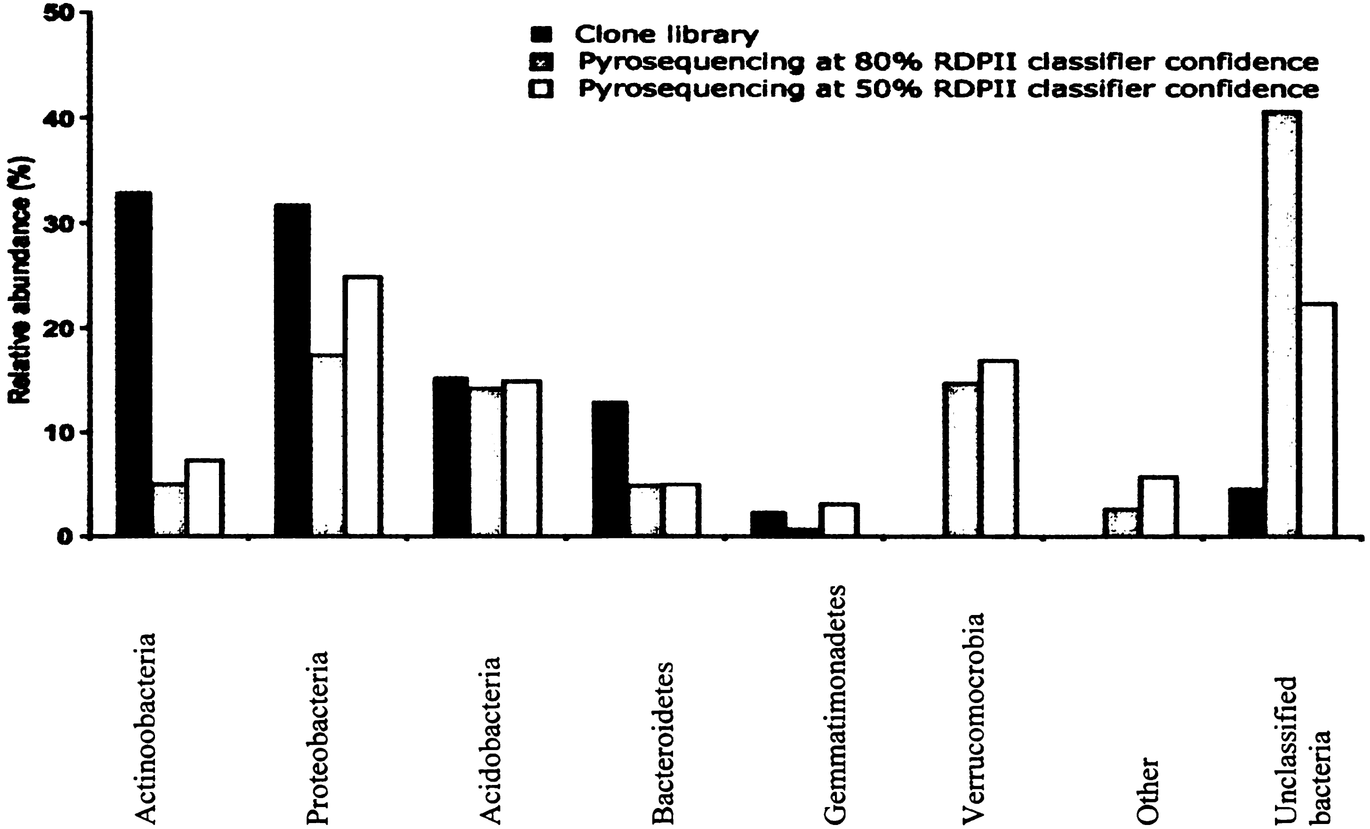


Figure 1.6. Relative abundance (%) at the Phylum level by V4-16S pyrosequencing and by Sanger sequencing of the clone library of a PCB-contaminated rhizosphere soil. The differences may be due to differences numbers of sampling (sequences), primer coverage, and classification accuracy.

sequencing (Armougom *et al.*, 2009). If two bacteria have the same 16S rRNA gene sequences within the sequenced region, it is impossible to differentiate their phylogeny. In addition, there is the discrepancy in phylogenetic relationship between full-length 16S rRNA gene and short pyrosequencing reads (Figure 7). The phylogenetic trees with short pyrosequencing read or full-length 16S rRNA gene sequences sometimes conflict with each other by altering the position of major phyla in the tree. Moreover, the actual phylogeny could possibly be overwhelmed by inherited pyrosequencing error rates. The phylogeny with pyrosequencing reads should be carefully done.

DIVERSITY AND SPECIES DISTRIBUTIONS IN BACTERIAL COMMUNITIES

16S rRNA gene pyrosequencing reveals a "rare biosphere" in that thousands of low-abundance populations are now detected (Sogin *et al.*, 2006). This large sampling by this exhaustive sequencing can make more valid richness estimates by assuming a species/taxa-abundance distribution (TAD). Previously, diversity has been estimated by fitting data of 16S rRNA gene clone libraries (Hong *et al.*, 2006) or T-RFLP (Doroghazi and Buckley, 2008) to taxa-abundance curves and extrapolating from this to estimate richness (Curtis *et al.*, 2006). Having large numbers of sequences circumvents the limitations of previous sampling methods makes it possible to apply rigorous statistical methods to fit TADs to rRNA pyrosequencing data, resulting in better prediction of microbial diversity (Quince *et al.*, 2008). Although the true bacterial taxa-abundance distribution is unclear as the ultimate statistical model is to fit TADs, estimation of richness can be used for pre-metagenomic analysis to decide the depth of sequencing

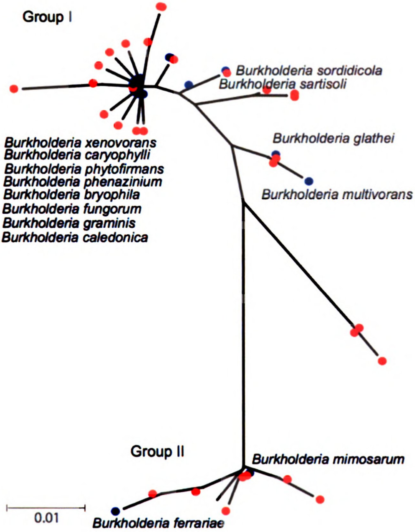


Figure 1.7. Examples of the phylogenetic analysis: Distribution of *Burkholderia* species in California grassland

efforts required to cover the microbial genetic component and is the basis for the systematic exploration of microbial diversity on the planet.

Non-parametric estimations are used to measure bacterial diversity for practical reasons although the estimation tends to be underestimated when sampling sizes are small. Pyrosequencing overcomes the limit of the small sample size, and has been used to measure and compare diversity. For instance, non-parametric estimate of diversity of commensal human oral microflora was at least one order of magitude higher (>19,000 species) using pyrosequencing than previous estimates based on (Keijser *et al.*, 2008). It is worthy to note that short fragment sequences of pyrosequencing, gives various species richness estimates depending on which variable regions the sequence fragments span. By comparing to richness estimates from complete 16S rRNA gene fragments, richness values were overestimated by the V1+V2, and V6 regions, underestimated by V3, V7, and V7+V8 regions, and nearly comparable by V4, V5+V6, and V6+V7 regions (Youssef *et al.*, 2009).

In bacterial communities with less taxa at the phyla level but high numbers at the species and strain level, taxonomic richness would likely be underestimated because short variable regions of the 16S rRNA gene would have insufficient resolution. An example is the architecture of highly speciated, but phyla impoverished human gut microbiota. This is not the case for the soil environment, which has more uniform distributions of its phylogenetic architecture.

COMMUNITY PROFILING AND COMPARISONS

The main purpose of rRNA gene pyrosequencing is the profiling of various bacterial communities, for instance, the deep marine biosphere (Sogin *et al.*, 2006; Huber *et al.*, 2007), soils (Roesch *et a*l., 2007), oral microflora (Keiser *et al.*, 2008), oligarchic microbial assemblages in anoxic bottom waters of a volcanic lake (Gaidos *et al.*, 2009), bacterial and archaeal communities in tidal flat sediments (Kim *et al.*, 2008), active PCB-degrading populations (Chapter 4), airborne microbial community (Bowers *et al.*, 2009), rhizosphere soils (Figure 8).

Barcoding of 16S rRNA gene pyrosequencing also provide for analyzing a larger number of replicates that previously possible by the clone library approach. Hence, comparisons of microbial communities can be reliably achieved along with changes due to ecological raison d'être. We used this strategy to compare different soil management systems, one of which rapidly altered stored soil carbon, in agricultural plots in Africa. Soil organic carbon (SOC) was the most important factor that explained differences in microbial community structure among treatments. Most notably, members of the *Acidobacteria* subdivisions GP4, GP6, and *Alphaproteobacteria* were more abundant in soils with relatively high SOC whereas *Acidobacteria* subdivisions GP7 and GP1, *Actinobacteria*, and *Gemmatimonadetes* were more prevalent in soil with lower SOC (Chapter 2).

Bacterial communities in stools from bio-breeding diabetes-prone, and bio-breeding diabetes-resistant rats were compared and different species were found to be dominant. However, the relatedness of these species to diabetes could not be determined (Roesch *et al.*, 2009a). V6-rRNA pyrosequencing was applied to human microbiomes in throat, stomach and fecal samples in study focused on effects of the presence of *Helicobactor pylori* in stomach. Hierarchical clustering based on Unifrac distance showed that *H. pylori* positive stomach samples have a different signature in its bacterial

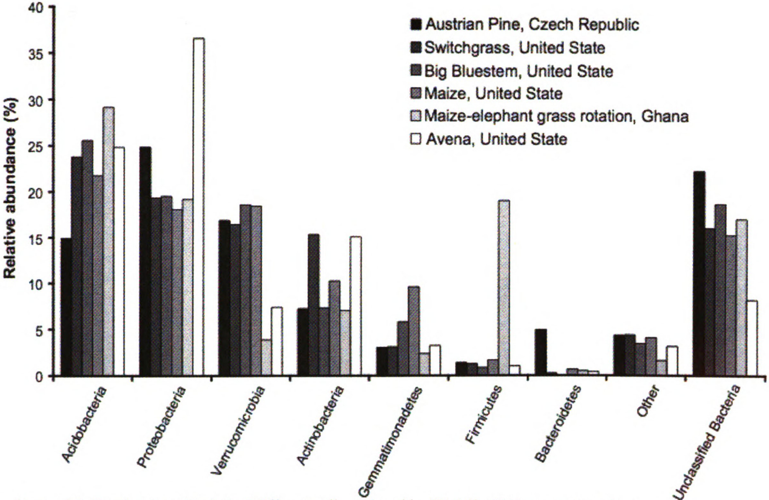


Figure 1.8. The dominant Phyla in 6 different soils analyzed by V4-16S rRNA pyrosequencing

community compared to negative *H. pylori* samples (Andersson *et al.*, 2008). The impact of diabetes and antibiotics on chronic wound microbiota characterized by V3-16S rRNA pyrosequencing showed that wound microbiota from antibiotic treated patients was significantly different from untreated patients. Also, antibiotic use among diabetics decreased *Streptococcaceae* abundance, which was more abundant among diabetics as compared to non-diabetics. The authors conclude that some bacteria might be involved in the non-healing state of some chronic wounds (Price *et al.*, 2009). Hamsters' fecal bacterial populations determined by pyrosequencing of 16S rRNA tags were analyzed to understand the influence of grain sorghum lipid extract (GSL) through feeding the hamsters GSL. Pyrosequencing results revealed that families *Coriobacteriaceae* and *Erysipelotrichaceae* were negatively correlated to GSL intake, and *Allobaculum* was positively correlated with GSL while phylum level composition had no differences. Hence, alterations of taxa occurred a deeper levels (small groups) were linked to diet (Martínez *et al.*, 2009). These findings suggest that rRNA gene pyrosequencing can used to detect and quantify community differences and to analyze disease-associated microbial gut ecology.

MEASURING BACTERIAL COMMUNITY DYNAMICS

Bacterial community dynamics also can be measured by 16S rRNA gene pyrosequencing. Population dynamics in fermented foods, e.g. pearl millet slurries, revealed that *Firmicutes* and lactic acid bacteria were detected throughout 24 h of fermentation whereas other bacteria were only detected at beginning of fermentation (Humblot and Guyot, 2009). Dethlefsen and colleagues (2008) analyzed the antibiotic (Ciprofloxacin)-associated disturbance of the human gut microbiota. Ciprofloxacin treatment influenced the abundance of about a third of the bacterial taxa in the gut, and decreased the taxonomic richness, diversity, and evenness of the community, however, the bacterial community returned to the pretreatment state indicating this community's resilience. Also, rRNA pyrosequencing may be used to measure the outcome of management of microbial community composition to aid functional stability in bioreactors (unpublished) and wastewater treatment systems (Appendix B3).

BACTERIAL GROUPS THAT CORRELATE TO HABITAT CHARACTERISTICS

Several studies have tried to find correlations between characteristics of habitats and the presence or relative abundance of certain bacterial groups. Bacterial community composition from 87 different soils, was significantly correlated with differences in soil pH, largely driven by changes in the relative abundances of *Acidobacteria*, *Actinobacteria* and *Bacteroidetes* across the range of soil pHs. Phylogenetic diversity of the bacterial communities was also correlated with soil pH (Lauber *et al.*, 2009). Relative abundance, diversity, and composition of the Phylum *Acidobacteria* were correlated strongly with soil pH (Jones *et al.*, 2009), suggesting the ecological relevance of this poorly-cultivated, less-known group — *Acidobacteria.* Also, a comparison of four geographically distant microbial communities showed few shared members, indicating environmental characteristics are strong features determining microbial community composition (Fulthorpe *et al.*, 2008).

METHOD VALIDATION

The bias that can be caused by sample handling and experimental procedures such as sample storage and DNA extraction also can be investigated by rRNA pyrosequencing. The changes in bacterial community composition and diversity was studied in samples of healthy children's feces analyzed immediately at sampling and after storing at room temperature up to 72h. In the latter samples, members of *Bacteroides* and *Clostridium* decreased and the members of the Enterobacteriaceae increased (Roesch *et al.*, 2009b). Understanding of the bias of DNA extraction was studied by comparing the bacterial composition in the DNA recovered after first extraction and 6th serial extraction (Feinstein *et al.*, 2009). Rarely-cultivated groups such as *Acidobacteria*, *Gemmatimonades*, and *Verrucomicrobia* were extracted more efficiently in the first extraction, while proportionally more *Proteobacteria* and *Actinobacteria* were recovered in DNA from the 6th extraction.

AMPLICON PYROSEQUENCING OF PROTEIN ENCODING GENES

Describe earlier, short read length offers a limited phylogenetic information for more conserved genes, like the ribosomal genes, which may be addressed by targeting genes other, faster-evolving, phylogenetically-informative genes. Pyrosequencing of a protein-encoding gene, e.g. Chaperonin-60 universal target (*cpn*60 UT), provided better resolution at the species level than 16S rRNA genes when describing the vaginal microbial community (Schellenberg *et al.*, 2009).

CONCLUSIONS AND FUTURE DIRECTIONS

Pyrosequencing of rRNA genes has been opening a new path to assess microbial communities, in respect to species distribution, diversity, the organism identification, community comparisons and dynamics. Although rRNA pyrosequencing is currently (arguably) the most effective bacterial community analysis method, we have often faced the problem in linking these 16S rRNA sequences to biological functions in the microbial community, especially when sequences reflected dominant species whose functions are unknown (Fulthorpe *et al.*, 2008). Also, rare members, which usually comprise more than half the species of natural environments, are the outcome of evolutionary history and have a seemingly the infinite source of genomic inventory (Sogin *et al.*, 2006). Gathering genomic information and physiology of unknown groups and rare members is beginning to be addressed by the GEBA Project (the Genomic Encyclopedia of Bacteria and Archaea) which aims to systematically fill the gaps in genome sequence of major branches in Bacterial and Archaeal of the Tree of Life.

Microbial ecologists would benefit from consensus in a standard operating procedure for rRNA pyrosequencing. Mostly because the short read length of current pyrosequencing technique, has led to use of different universal primers and targeting of different regions in SSU rRNA resulting non-comparable datasets generated by numerous laboratories (discussed in Chapter 5). Even though rRNA pyrosequencing is powerful, it still provides a rather the sketchy vies of microbial communities since the resolution of an already conserved gene is much, much less that for whole metagenomic analysis. Community level-MLST/A (Multi Locus Sequence Typing/Analysis) may become possible in near future and if so, should provide better insight into microbial community diversity and perhaps membership, and a good bridge to metagenomic data.

REFERENCES

Anderson IC, Cairney JW (2004) Diversity and ecology of soil fungal communities: increased understanding through the application of molecular techniques. *Environ Microbiol* **6**:769-779

Andersson AF, Lindberg M, Jakobsson H, Bäckhed F, Nyrén P, Engstrand L (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One* **3**:e2836

Armougom F, Bittar F, Stremler N, Rolain JM, Robert C, Dubus JC, Sarles J, Raoult D, La Scola, B (2009) Microbial diversity in the sputum of a cystic fibrosis patient studied with 16S rDNA pyrosequencing. *Eur J Clin Microbiol Infect Dis* (in process)

Bowers RM, Lauber CL, Wiedinmyer C, Hamady M, Hallar AG, Fall R, Knight R, Fierer N (2009) Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei. *Appl Environ Microbiol* **75**:5121-5130

Brosius J, Palmer ML, Kennedy PJ, Noller HF (1978) Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli. Proc Natl Acad Sci U S A* **75**:4801-4805

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis *Nucleic Acids Res* **37**:D141-145

Curtis TP, Head IM, Lunn M, Woodcock S, Schloss PD, Sloan WT (2006) What is the extent of prokaryotic diversity. *Philos Trans R Soc Lond B Biol Sci* **361**:2023-2037

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

Dethlefsen L, Huse S, Sogin ML, Relman DA (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* **6**:e280

Doroghazi JR, Buckley DH (2008) Evidence from GC-TRFLP that bacterial communities in soil are lognormally distributed. *PLoS One* **3**:e2910

Feinstein LM, Sul WJ, Blackwood CB (2009) Assessment of bias associated with incomplete extraction of microbial DNA from soil. *Appl Environ Microbiol* (in print)

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

Gaidos E, Marteinsson V, Thorsteinsson T, Jóhannesson T, Rúnarsson AR, Stefansson A, Glazer B, Lanoil B, Skidmore M, Han S, Miller M, Rusch A, Foo W (2009) An oligarchic microbial assemblage in the anoxic bottom waters of a volcanic subglacial lake. *ISME J* **3**:486-497

Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* **19**:1141-1152

Hamp TJ, Jones WJ, Fodor AA (2009) Effects of experimental choices and analysis noise on surveys of the "rare biosphere". *Appl Environ Microbiol* **75**:3263-3270

Hong SH, Bunge J, Jeon SO, Epstein SS (2006) Predicting microbial species richness. *Proc Natl Acad Sci U S A* **103**:117-122

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML (2007) Microbial population structures in the deep marine biosphere. *Science* **318**:97-100

Jones RT, Robeson MS, Lauber CL, Hamady M, Knight R, Fierer N (2009) A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *ISME J* **3**:442-453

Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale *Appl Environ Microbiol* **75**:5111-5120

Leigh MB, Pellizari VH, Uhlík O, Sutka R, Rodrigues J, Ostrom NE, Zhou J, Tiedje JM (2007) Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs). *ISME J* **1**:134-148

Leigh MB, Prouzová P, Macková M, Macek T, Nagle DP, Fletcher JS (2006) Polychlorinated biphenyl (PCB)-degrading bacteria associated with trees in a PCB-contaminated site. *Appl Environ Microbiol* **72**:2331-2342

Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, Gordon JI (2008) Evolution of mammals and their gut microbes. *Science* **320**:1647-1651

Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* **35**:e120

Martínez I, Wallace G, Zhang C, Legge R, Benson AK, Carr TP, Moriyama EN, Walter J (2009) Diet-induced metabolic improvements in a hamster model of hypercholesterolemia are strongly linked to alterations of the gut microbiota. *Appl Environ Microbiol* **75**:4175-4184

Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 10: inference of RNA alignments. *Bioinformatics* **25**:1335-1337

Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M, Fire AZ (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* **35**:e130

Price LB, Liu CM, Melendez JH, Frankel YM, Engelthaler D, Aziz M, Bowers J, Rattray R, Ravel J, Kingsley C, Keim PS, Lazarus GS, Zenilman JM (2009) Community analysis of chronic wound bacteria using 16S rRNA gene-based pyrosequencing: impact of diabetes and antibiotics on chronic wound microbiota. *PLoS One* **4**:e6462

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**:7188-7196

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

Quinlan AR, Stewart DA, Strömberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* **5**:179-181

Rodrigues DF, da C Jesus E, Ayala-Del-Río HL, Pellizari VH, Gilichinsky D, Sepulveda-Torres L, Tiedje JM (2009) Biogeography of two cold-adapted genera: *Psychrobacter* and *Exiguobacterium*. *ISME J* **3**:658-665

Roesch LF, Casella G, Simell O, Krischer J, Wasserfall CH, Schatz D, Atkinson MA, Neu J, Triplett EW (2009) Influence of fecal sample storage on bacterial community diversity. *Open Microbiol J* **3**:40-46

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**:283-290

Roesch LF, Lorca GL, Casella G, Giongo A, Naranjo A, Pionzio AM, Li N, Mai V, Wasserfall CH, Schatz D, Atkinson MA, Neu J, Triplett EW (2009) Culture-independent identification of gut bacteria correlated with the onset of diabetes in a rat model. *ISME J* **3**:536-548

Schellenberg J, Links MG, Hill JE, Dumonceaux TJ, Peters GA, Tyler S, Ball TB, Severini A, Plummer FA (2009) Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. *Appl Environ Microbiol* **75**:2889-2898

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* **103**:12115-12120

Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W, Farmerie W (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* **37**:e76

Tringe SG, Hugenholtz P (2008) A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11**:442-446

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**:5261-5267

Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA, Elshahed MS (2009) A Comparative study of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Appl Environ Microbiol* (in process)

CHAPTER II

COMMUNITY RESPONSES TO AGRICULTURAL PRACTICES IN TROPICAL AFRICA ANALYZED BY PYROSEQUENCING

ABSTRACT

We analyzed the microbial community that developed after four years of testing different soil-crop management systems in the savannah-forest transition zone of Eastern Ghana where management systems can rapidly alter stored soil carbon as well as soil fertility and structure. The treatments were: (i) the native practice of winter regrowth of native elephant grass (*Pennistum purpureum*) followed by burning that biomass before planting maize in the spring, (ii) the same practice but without burning and the maize received mineral nitrogen fertilizer, (iii) a winter crop of a legume, pigeon pea (*Cajanus cajan*), followed by maize, (iv) a treatment kept vegetation free in the winter (bare fallow) followed by maize and (v) and unmanaged elephant grass-shrub vegetation. The mean soil carbon contents of the sampled soils were: 1.29, 1.67, 1.54, 0.80 and 1.34, respectively, differences that could be expected to affect the microbial communities.

From the more than 290,000 sequences obtained by pyrosequencing of the SSU rRNA gene, 80% belonged to seven bacterial phyla common to most soils; Acidobacteria, *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Verrucomicrobia*, *Gemmatimonadetes*, and *Bacteroidetes*. Less than 5% of all sequences were identical to SSU rRNA gene sequences previously recovered from cultivated bacteria, most were 90% or more similar to previous sequences in pubic databases, but 1.2% (2330 sequences) had lower than 85% similarity to any environmental or isolated sequences suggesting potentially novel phyla. Canonical correspondence analysis and stepwise multiple regression showed that soil organic carbon (SOC) was the most important factor that explained differences in microbial community structure among treatments. Most notably, members of the *Acidobacteria* subdivisions GP4, GP6, and *Alphaproteobacteria* were more abundant in soils with relatively high SOC whereas *Acidobacteria* subdivisions GP7 and GP1, *Actinobacteria*, and *Gemmatimonadetes* were more prevalent in soil with lower SOC. While community structure was most affected by SOC, diversity appeared to be influenced by a combination of factors. The data suggest that the use of a pigeon-pea fallow in tropical agriculture promotes a higher microbial diversity and sequesters more soil organic carbon, thus improving soil structure, function, and resiliency.

ABBREVIATIONS

ID: Nuleotide identity
SOC: Soil organic carbon
MD: Mean uncorrected nucleotide distance
EbM: Maize-elephant grass (*Pennisetum* sp) rotation with fallow residue burning
EfM: Fertilized maize-elephant grass rotation with minimum tillage of fallow residue by hand slashing
PM: Maize-pigeon pea (*Cajanus cajan*) rotation with minimum tillage of fallow residue by hand slashing
BF: Maize-bare fallow rotation with complete residue removal during fallow period
Eu: Unmanaged elephant grass

INTRODUCTION

Conversion of natural ecosystems to agriculture results in soil organic carbon (SOC) losses due to increased organic matter oxidation, leaching, and erosion (1). Globally, deforestation rates are greater in the tropics than rates of current or historical changes in any other region (2). SOC retention increases soil cation exchange capacity, improves structure, and conserves nitrogen, phosphorus, potassium and sulfur. Cultivation, in concert with fertilizer application, tillage, and residue removal results in rapid SOC depletion followed by a slower decrease, typically spanning several decades, before a new steady-state is reached (3). These losses can range between 20% to 70% of the original SOC content (4), but can be remediated with the use of cover crops and minimum tillage, when the residue is not removed (5,6). Reduced tillage increases SOC retention through macroaggregate preservation (7), and has been proposed as a primary method for optimizing SOC in fine textured soils (8). Current agricultural practices in tropical regions typically involve fallow residue removal, either by grazing or burning. This practice has in recent years been re-evaluated with the goal of gaining benefits from developing a winter crop that would provide food, sequester more carbon in the soil, improve soil fertility and structure and provide the potential for earning cash from the developing carbon markets (9).

While much study has focused on the chemical and physical changes to soil from different cropping systems, the associated shifts in soil microbial community structure and function remain largely unknown. Soil harbors the largest reservoir of microbial diversity due to an enormous number of niches, small-scale spatial isolation (10, 11), and 3.8 billion years of evolution. Soil microbial communities are responsible for carbon and nutrient cycling and are thus an integral component of the soil productivity and the global element cycles. Therefore, their response needs to be understood when developing new agricultural practices. Recent studies assessing soil microbial community changes due to cropping systems used methods such as clone libraries (12) and denaturing gradient gel electrophoresis (DGGE) (13), which often lack the coverage and resolution necessary to reveal changes among treatments. Pyrosequencing (14) now allows us to define diversity and complexity by targeted SSU rRNA gene sequencing (15-17) at such depth that community responses may be quantified in contrasting soil management schemes.

In this study, we utilized SSU rRNA gene pyrosequencing to determine the effect of different maize-fallow rotations on soil microbial communities in the savannah-forest transition zone of Ghana (18). Soils were sampled from four replicated plots after maize harvest and after 4 years of the following annual rotations: 1) EbM: Growth of elephant grass (*Pennistum* sp) in the winter with its residue burned followed by maize cultivation (native practice), 2) PM: winter Pigeon pea (*Cajanus cajan*) crop, minimal tillage of fallow residues followed by maize cultivation, 3) EfM: Growth of elephant grass with no burn and followed by fertilized maize, 4) BF: bare fallow, i.e. no fallow season plant, followed by maize cultivation and 5) Eu: re-growth of the native elephant grass-shrub vegetation left unmanaged for 4 years (native condition).

RESULTS

Characterization of Microbial Communities and Phylogenetic Structures. After trimming of the forward and reverse primers and passage of trimmed reads through quality filters to minimize the effects of embedded pyrosequencing errors, more than 290,000 sequencing reads with an average length of 207 bp were obtained. The number of high quality sequences per sample was evenly distributed between 7519 to 12204 (Table 1). When operational taxonomic units (OTUs) were defined at 95% identity (ID), rarefaction curves indicated that sampling was, as expected, not fully exhaustive. The phylogenetic architecture (19, 20) of these soil communities showed an extensive deep-lineage variation, with a phyla rich pattern typical for soil habitats. In contrast, the microbial community from a carbon amended aquifer exhibited shallow-lineage variation with a lower taxa level (species) rich pattern, which drastically increased at 98% ID (Fig.

Agricultural plots	SOC (%)	pH	Available P (mg/kg)	Total Nitrogen (%)	Microbial Biomass	Maize Grain Yield	Maize Biomass	Sequencing samples ID	No. of sequences	No. of OTUs at 95% ID	Chao 1	H'
PM	1.31	6.3	6.9	0.10	493	1.48	8.12	PM1	9882	2963	7262	6.99
	1.88	6.6	19.0	0.14	686			PM2	9902	3187	7693	7.13
	1.60	6.6	3.6	0.12	699			PM3	12204	2940	6604	6.71
	1.37	6.2	5.9	0.10	669			PM4	11487	3711	9260	7.23
EbM	1.04	6.7	3.4	0.08	464	1.02	6.18	EbM1	9557	2411	6201	6.48
	1.56	6.8	7.9	0.12	370			EbM2	10080	2853	6307	6.86
	1.47	6.7	5.2	0.11	487			EbM3	8536	2408	5221	6.65
	1.09	6.5	6.0	0.08	619			EbM4	9614	3134	7720	7.07
EfM	1.87	6.9	20.5	0.14	812	2.36	9.76	EfM1	10328	2466	5117	6.69
	1.55	6.6	18.0	0.11	481			EfM2	9885	2172	4044	6.73
	1.42	6.0	12.1	0.11	824			EfM3	9201	2403	5247	6.54
	1.82	6.2	9.7	0.13	913			EfM4	7598	2038	4581	6.38
BF	0.76	6.4	5.9	0.06	392	0.05	0.872	BF1	9995	2819	6023	6.82
	0.61	6.4	6.8	0.05	452			BF2	11040	2665	5408	6.65
	0.91	6.6	3.5	0.07	455			BF3	8985	2404	5451	6.7
	0.91	6.3	4.1	0.07	311			BF4	7519	1527	2980	5.52
Eu	1.34	6.68	ND	0.10	858	NA	NA	Eu1	9557	1553	2676	5.86
								Eu2	8481	1188	1963	5.85
IM	2.7	5.50	ND	0.20	ND	ND	ND	IM	10564	1511	2258	6.15

Table 2.1. Summary of soil characteristics of agricultural plots and pyrosequencing results.

1A). Based on the non-parametric estimator Chao 1, the fallow plant rotations PM and EbM led to higher bacterial richness compared to the BF and EfM (P<0.007)(Table 1).

Microbial Members in Ghana Soils. Taxonomic classification of the sequences was assessed using the RDP Classifier trained on species type-strain sequences from the Taxonomic Outline of the Bacteria (TOBA; http://www.taxonomicoutline.org/) along with additional sequences for regions of bacterial diversity not well-covered by TOBA. The classifier was set to a bootstrap confidence threshold of 50%. Sequences covered by our newly designed primer set were assigned to 23 phyla, 57 orders, 149 families, and 490 genera. Irrespective of the rotation practice, seven phyla accounted for 73% to 86 % of total sequences in a given sample: *Acidobacteria*, *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Verrucomicrobia*, *Gemmatimonadetes*, and *Bacteroidetes* (Fig. 1B). These phyla appear highly ubiquitous as been observed by SSU rRNA clone libraries in most soil environments (21, 22). Notably, BF significantly contained more *Actinobacteria* (15.4, SD 9.7%) sequences than the other treatment samples (ANOVA, P=0.037) (Fig. 1B).

Structural Differences in Microbial Communities among Agricultural Plots. In order to identify differences among the microbial communities, all 192,835 sequences were clustered at 95% ID, yielding 26,287 clusters. All nineteen microbial communities (including IM), were compared by calculating the pair-wise abundance-based adjusted Sørensen similarity index (23). This index was used in Principle Coordinate Analysis (PCoA) which showed that the EfM, EbM and PM generally grouped together whereas BF was unique. The IM was clearly distinct from from the Ghanaian soils, most likely due to a different soil origin and history (Fig. 2A).

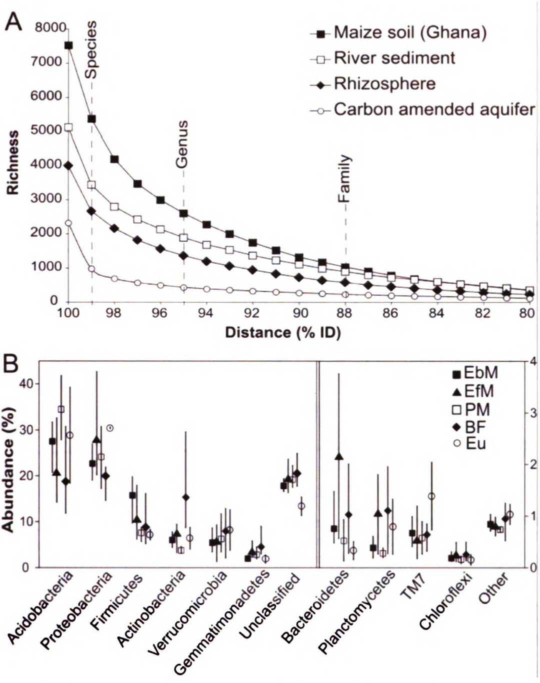


Figure 2.1. Microbial community structure and composition. (A) Phylogenetic architecture of microbial communities among habitats. Richness was estimated by rarefaction curves with randomly selected 10,000 sequences from each habitat. Sequence data sets for the rhizosphere, river sediment, and carbon-amended aquifer samples (unpublished) were obtained as described. (B) Phylum-level composition of the microbial communities in Ghanaian soil.

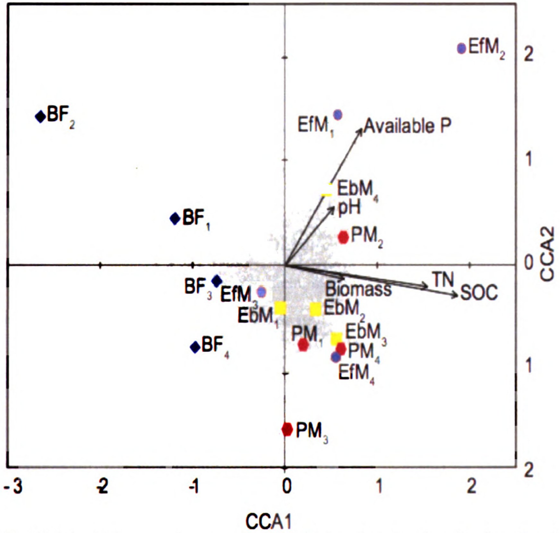


Fig. 2.2 Microbial community comparison (A) PCoA analysis based on abundance-based adjusted Sorenson similarity indices. (B) Two-dimensional CCA ordination plot. The magnitude of the environmental vectors; microbial biomass (biomass), total nitrogen (TN), available phosphorus (available P), and soluble organic carbon (SOC), is represented by arrows. Cluster positions are indicated by grey symbols.

Canonical Correspondence Analysis (CCA) was implemented in order to establish the linkage between cluster abundance and the environment by implicitly embedding the environmental soil data (Table 1) with the cluster abundance. This method explained 36% of the cluster variability at the whole community level. Model significance was confirmed using anovasim (number of permutations=10000, *Pseudo-P*<0.005) and permutation (number of permutations =10000, *Pseudo-P*<0.02) tests. The first ordination axis was positively correlated with both TN and SOC while the second ordination axis correlated with Available P. The BF was negatively correlated with all environmental variables and is clearly distinguishable from the others. Utilizing two independent methods, both PCoA and CCA served to illustrate the distinct structure of the BF.

To identify taxonomic groups that were most responsive to fallow practice, clusters were selected that exhibited at least a three-fold abundance difference in BF compared to the other agricultural treatment. Using this approach [Supporting information (SI) text 1], 620 clusters were identified that accounted for approximately 25% of total sequences. Clusters more predominant under fallow rotation were classified as *Acidobacteria* GP6 and GP4 (EbM and PM), class Bacilli (EbM), and *Alphaproteobateria* (EfM). In contrast, clusters more abundant in BF were mostly affiliated to *Actinobacteria*, *Acidobacteria* GP1, and *Gemmatimonadetes* (Fig. 3).

In order to investigate the environmental factors that influenced cluster abundance, stepwise multiple regressions of the 620 clusters was performed. Significant stepwise multiple regressions were identified with 287 clusters (46%) (adjusted P<0.05). SOC was the most consistent significant predictor of relative cluster abundance among the sites, followed by TN and available P. Among those clusters, 182 (63%) included

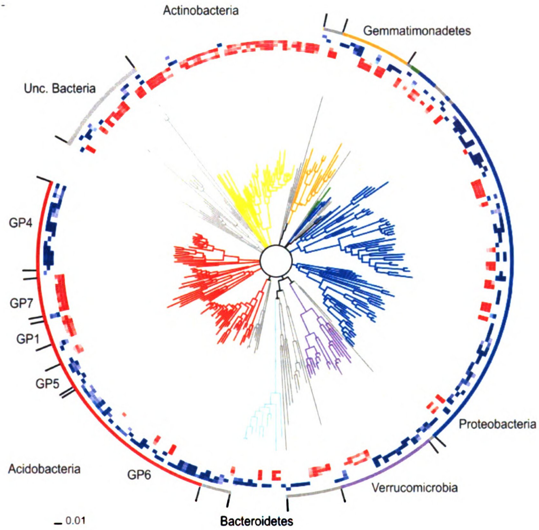


Figure 2.3. Neighbor-joining phylogenetic tree displaying 287 clusters with a significant stepwise multiple regressions to any of the environmental parameters. From inner to outer rings, red-colored rings represent the fold-difference in relative cluster abundance in BF compared to EbM, EfM, and PM. Blue-colored rings represent the fold-difference in cluster relative in EbM, EfM, and PM compared to BF. The outer ring is color-coded according to the taxonomic placement of the clusters.

SOC in the regression, 132 included total nitrogen, 130 included available P, 90 included pH, and 35 clusters included microbial biomass. Regression slopes of SOC were positively correlated to clusters affiliated to mostly *Proteobacteria* and *Acidobacteria* GP4, GP5, and GP6, reflecting increasing cluster abundance with larger SOC values. In contrast, *Actinobacteria* and *Verrucomicrobia* clusters were negatively correlated to SOC.

Characterization of New Clades of Sequences Unaffiliated to Known Sequences. In order to determine the relatedness between our sequences and those in the public database, the uncorrected nucleotide distance to the closest public isolate and environmental sequence was calculated. The mean uncorrected nucleotide distance (MD) when sequences were compared to the environmental plus isolate database (MDENV) or isolate database (MDISO) were 96.8% and 88.6% ID, respectively. Interestingly, each bacterial phyla or class exhibited a distinct MDISO. For *Bacilli* and *Alphaproteobacteria* the MDISO were 98.2% and 95.4%, respectively, whereas for *Acidobacteria*, *Gemmatimonadetes*, *Verrucomicrobia*, *Chloroflexi*, *Planctomycetes*, and *Nitrospira* the MDISO were below 90% ID (Fig 4A).

Interestingly, 2330 sequences had a similarity below 85% ID to any environmental or isolate sequence in public databases. When clustered at 95% ID, 286 OTUs (941 sequences) contained at least two sequences, whereas 1389 OTUs contained a single sequence. Notably, 144 OTUs (including a large OTU containing 27 sequences) originated from multiple samples. This suggests that novel, yet to be sequenced bacterial

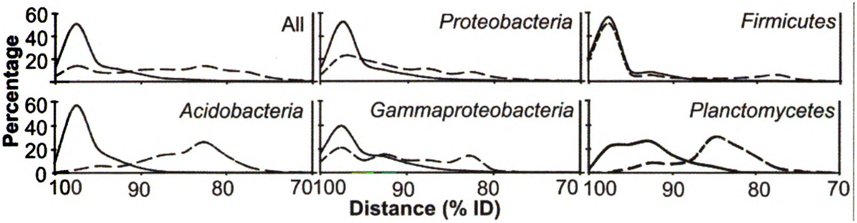


Figure 2.4. (A) Frequency distributions of the uncorrected distances of sequences affiliated to selected taxonomic groups. Solid lines were based upon comparison with the environmental plus isolates database, and dashed lines upon comparison with the isolates database.

clades, exist that are poorly classified by both the modified TOBA (24) and Hugenholtz schemes (http://greengenes.lbl.gov) (Fig 4B).

DISCUSSION

Pyrosequencing of the SSU rRNA gene was used to contrast microbial community structures at a greater depth and with more replication than typically attainable through previous methods. These analyses promote when properly selecting sequencing region in SSU rRNA genes (SI text 2 and Fig. S1)(25) and providing highthroughput analysis tools (26).

We identified changes in bacterial community diversity responsive to agricultural fallow treatments. Bacterial richness was higher in all agricultural plots when compared to an elephant grass-shrub dominated, unmanaged plot (Eu). This illustrated the influence of agricultural management on microbial community structure. After four years, bacterial groups responsive to particular treatments were opportunistic additions to the endogenous community represented by Eu. These groups are likely part of the "rare biosphere" in the Eu community and serve as a genetic or functional reservoir. Physical disturbance of the soil due to plowing, planting and burning of fallow plants may increase spatial resource competition. Among the treatments, low diversity occurred in the bare fallow treatment and was likely due to overall resource limitation from low exogenous organic matter input. However, fertilizer application also restricted diversity, perhaps due to the higher nutrient availability driving a less metabolically diverse r-selected community. While fertilizer application exhibited the highest organic matter deposition, the PM treatment served to sequester carbon in woody biomass. Microbial diversity was highest in this

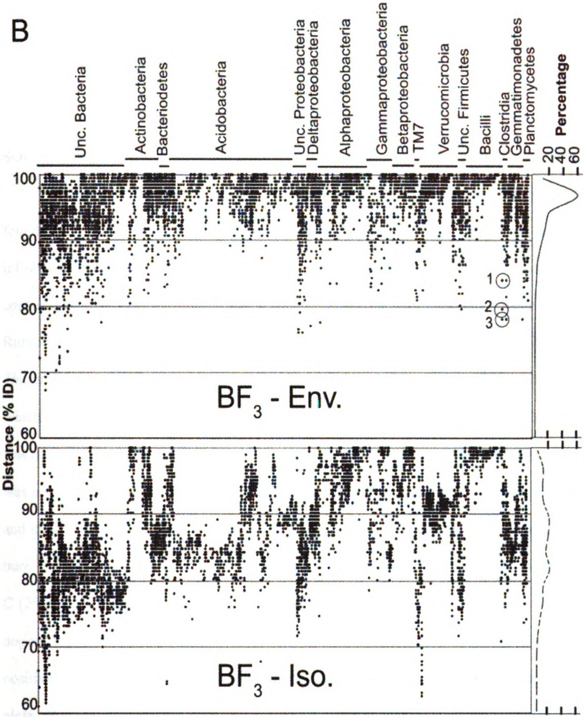


Figure 2.4. (B) NEO plots for a representative microbial community under bare fallow treatment (BF3). Open circles 1, 2, and 3 are example of OTUs with a similarity below 85% ID.

treatment, possibly due to slower nutrient release from the more recalcitrant pigeon pea organic matter structure, steady N addition from N2 fixation, and P solubilization. Based on these results, pigeon pea appears to be the most appropriate cover crop in a tropical ecosystem such as this, by sustaining a diverse bacterial community while sequestering SOC, thus improving overall soil health.

Overall soil microbial community structure and specific taxa distribution were found to be most affected by SOC abundance. Sequestered carbon appears to largely influence Actinobacteria abundance in soil. The lowest-SOC, (bare fallow) treatment consistently exhibited the highest abundance of Actinobacteria, largely of the subclass Rubrobacteridae. Previously isolated bacteria within this subclass, *Rubrobacter* (27) and *Thermoleophilum* (28), are resistant to radiation and are found primarily in arid soil, which consistent with the more harsh condition of this soil since the summer maize crop was meager in years 3 and 4 (Table 1). Though not selected by regression as temperature was not a variable, the high Bacilli abundance in the burned treatment (EbM) was notable and may be due to the heat resistance of these spore-forming bacteria. The traditional burn of the fallow season vegetation has resulted in measured soil temperatures as high as C (29), which could influence survivors in surface soil communities. In contrast to the dogma that all Acidobacteria are oligotrophs (20), we found that certain groups were positively correlated to SOC and were present in high abundance in the nutrient enriched plots. However, overall it does not appear that Acidobacteria uniformly respond to environmental variables, which could be expected for this very large and diverse phylum.

Our observations support that SSU rRNA gene pyrosequencing can be used to assess microbial abundances in soils among different environments, and can be used to test widely held inferences that were perhaps based on insufficient data. First, our data show that 4.1% of sequences were identical to SSU rRNA gene sequences previously recovered from cultivated bacteria, in comparison to the common notion that "less than 1% of bacteria are cultivated". However, since our reads covered only a small portion of the total SSU rRNA gene and our sampling was not exhuastive, this estimation may be artificially inflated. In order to extrapolate to the full diversity coverage of the samples, we calculated the ratio of Chao 1 with sequences identical to isolates at 100% ID against Chao 1 with all sequences at 100% ID. This estimate was 0.13%, the adjusted value expected with exhaustive sequencing. Secondly, our data indicate that most members of the *Acidobacteria*, *Verrucomicrobia*, *Gemmatinomadetes*, *Nitrospira*, and *Planctomycetes* are poorly cultivated, whereas many *Proteobacteria* and *Firmicutes*, and most of the *Bacilli*, have been isolated (30, 31). Within the Proteobacteria, however, the *Gammaproteobacteria* have a large number of highly divergent, uncultivated members (Fig. 4A). This is particularly interesting since it has been generally assumed that the *Gammaproteobacteria* are easily cultured and most of their diversity is known. . Thirdly, the massive compilation of SSU rRNA gene sequences yielded highly divergent sequences from groups that were not previously sequenced. Based on our evidence, we suggest that the 2330 sequences with less than 85% ID threshold against SSU rRNA genes sequences in public database, are deeply divergent taxa that have yet to be isolated or characterized. As such, this method is useful in discovering novel bacterial clades and in providing potential probes to aid in their recovery of for studying their ecology.

In conclusion, this study illustrates the usefulness of pyrosequencing for the comparison of microbial community structures. Land use change, including the expansion of agriculture in the tropics is having major effects on ecosystems and on our climate. These changes will most likely change the supporting microbial communities and perhaps the soil processes and ecosystem services they provide. The new sequencing methodologies now provide the depth and replication needed to assess microbial change as a part of evaluating management and land use impacts. In this case, our data suggest that the use of a pigeon-pea winter crop in tropical agriculture not only promotes a higher microbial diversity but also serves to sequester soil organic carbon, thus improving soil structure, function, and resiliency.

MATERIALS AND METHODS

Experimental Design and Sampling. The research site was located at the Kpeve Agricultural Experimental Station (KAES) in Volta Region, Ghana (coordinates 6o 43.15'N, 0o 20.45'E). Classified as a savanna to forest transitional zone, the area is dominated by Haplic Lixisols (sandy clay loams), Haplic Acrisols and Leptic Haplic Acrisols. Soil samples were taken from each of four replicate plots (50 m by 80 m) in a randomized complete block design with a 2.5 cm x 18.5 cm corer on September 10, 2006 after the maize harvest and after 4 years of the same annual rotations (32). Each replicate sample was a homogenized composite of ten random sub-samples (18), with the exception of Eu, composites of two sub-samples, separated by 0.7 m. The soils were immediately place on ice and then stored at -20C until DNA extraction.. The soil was cultivated at the time of plot establishment but not after. The Iowa soil (IM), classified as Tama silty clay loam, was collected on Dec. 1, 2006 following a maize crop which was preceeded by soybean and was under no-till management for over 5 years.

SSU rRNA Gene Amplicon Pyrosequencing. Soil DNA was extracted with the Mobio PowerSoil DNA Isolation Kit (Mobio, Carlsbad, CA) according to the manufacturer's instructions. Primers were designed with barcodes for pyrosequencing to accommodate multiple samples in a single PicoTiterPlate (Roche Applied Science, Indianapolis, IN). The forward key-tagged primers were composed of sequencing adaptor A, sample-specific 4 or 6-bp keys, and a eubacterial 563F primer (bold in sequences below). The reverse fusion primer consisted of sequencing adaptor B, and a eubacterial 802R primer. All primers were passed through dual HPLC-purification (Integrated DNA Technologies, Coralville, IA) in order to increase specificity of primers and minimize the miss-sorting of samples by primer synthesis error. The forward primer sequence is 5'-GCCTCCCTCGCGCCATCAG(keys)AYTGGGYDTAAAGVG-3' and the reverse primer is 5'-GCCTTGCCAGCCCGCTCAGTACNVGGGTATCTAATCC-3'. PCR mixtures contained 1 μM of each primer (IDT, Coralville, IA), 1.8 mM MgC12, 0.2 M dNTPs, 1.5 X BSA (New England Biolabs, Beverly, MA), 1 unit of FastStart High Fidelity PCR System enzyme blend (Roche Applied Science, Indianapolis, IN), and 10 ng of DNA template. Amplification conditions were as follows: initial incubation for 3 min at 95oC; 30 cycles of 95oC for 45 sec, 57oC for 45 sec, and 72oC for 1 min; and a final 4 min incubation at 72oC. For each sample, three replicate PCR reactions were run in parallel, PCR products were purified by agarose gel electrophoresis, and excised bands of 270-300 bps were combined. Amplicon recovery was performed with Qiagen Gel extraction (Qiagen, Valencia, CA) followed by an extra Qiagen PCR Purification step. DNA was quantified spectrophotometrically using the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) and equimolar amounts of each sample were subsequently combined and subjected to pyrosequencing using the Genome Sequencer FLX System (454 Life Sciences, a Roche company, Bradford, CT)

Pyrosequencing Data. Raw reads were processed, filtered, aligned, clustered, and bias-corrected Chao1 species richness estimates obtained using programs from the RDP Pyrosequencing Pipeline (26). Sequences were assigned to bacterial taxa using the RDP Classifier version 2 using the RDP release 9.53 training set (25). Chao's abundance-based adjusted Sørensen similarity (23) were calculated for each pair of samples using EstimateS (purl.oclc.org/estimates) after first clustering each sample pair together.

For the phylogenetic tree, aligned representative sequences of 287 selected clusters were exported and a neighbor-joining phylogenetic tree was constructed (35). Tree and fold-difference color codes were visualized by iTOL (36) and resorted based on phylum-level classification. For the NEO plots, sequences were first ordered by classification results either at the phylum or class level. Each symbol indicates the uncorrected distance of a given sequence read to its closest match within the isolates database (ISO) and its closest match within the environmental plus isolates (ENV) SSU rRNA database. Each sequence was used as a query to the RDP's SequenceMatch tool (37, 38) to identify the sequence in the RDP's database with the largest number of matching words. The uncorrected pairwise distance was calculated between the aligned query and the SequenceMatch result sequence.

Statistical Analyses and Implementation. ANOVA and canonical correspondence analysis (CCA) was performed using the R statistical program (R Development Core Team) running the vegan package. Clusters were assigned at 5% ID using the complete linkage clustering method and soil environmental data from each replicate was used. The joint effect or "significance" of constraints in the CCA model was tested using both an anova permutation test (anova.cca, α=0.05, n=10000) and a CCA permutation test (permutest.cca, n=10000). Except where otherwise indicated, processing software was written in Java (API v1.5.0) and executed on the Macintosh (OS 10.4) or Linux (2.4.23) operating systems running Java virtual machines from Apple or Sun, respectively.

ACKNOWLEDGEMENTS.

This work was supported by grants from The Office of Science (BER), U.S. Department of Energy(DE-FG02-99ER62848, DE-FG02-04ER63933); National Science Foundation (DBI-0328255); and the U.S. Department of Agriculture (NRI).

AUTHOR CONTRIBUTIONS

Stella Asuming-Brempong, Samuel Adiku, and James Jones designed and managed agricultural plots in Ghana for four years. Jorge Rodrigues managed DNA samples. Jim Cole, and Qiong Wang set up computational sequences analysis: quality controls, alignment, clustering, Neo's plot, etc. Dieter Tourlousse and Ryan Penton performed stastical analyses: CCA and multiple regression.

Designed research: S.A-B, S.G.K.A., and J.W.J.
Performed research: W.J.S. S.A-B, and J.L.M.R.
Analyzed data: W.J.S, Q.W., D.M.T., C.R.P., and J.R.C.
Wrote the paper: W.J.S., C.R.P., D.M.T., Q.W., J.R.C., and J.M.T.

SUPPORTING INFORMATION 1 TEXT

Selection of Clusters Contrast to BFs. These clusters were identified by pairwise comparison of each practice to BF, with the filtering criteria that: 1) the number of sequences in each cluster were found in all replicates, and 2) clusters exhibited at least a three-fold prevalence as a replicate average in either each of the agricultural plots or BF. For example, when identifying clusters that are more prevalent in PM compared to BF, only those clusters with non-zero sequence counts in all PM replicates irrespective of BF, EbM, and EfM were included. The average number of those clusters among the four replicates was required to be 3x higher than BF.

SUPPORTING INFORMATION 2 TEXT

Bacterial Primer Design for Pyrosequencing of SSU rRNA Genes. Regions in the SSU rRNA gene suitable for pyrosequencing were identified that exhibited: 1) an appropriate amplicon length for pyrosequencing reads, 2) high coverage by bacterial universal primers, 3) high resolution and accuracy for bacterial classification and identification, and 4) a low frequency of insertions and deletions to simplify sequence alignment. A new set of bacterial universal primer, designed that encompassed the hypervariable V4 region (corresponding to *Escherichia coli* SSU rRNA gene positions 563 to 802), allows for accurate bacterial taxa identification with the RDP Classifier (1). Its applicability for pyrosequencing was further supported by *in-silico* Unifrac analysis (2). The universality of the primers was determined by internal alignment of perfect matches against SSU rRNA gene sequences in the Ribosomal Database Project II (RDP) (94.6% coverage) and from the metagenomic database of the Sorcerer II Global Ocean Sampling Expedition (94.7% coverage) (3). Specifically, the primers designed in this study targeted an overwhelming majority of known SSU rRNA gene sequences throughout all phyla while providing deep taxa classification useful for community comparisons (SI Figure 5).

SI MATERIALS AND METHOD

Initial Processing and Filtering. Raw reads were sorted into individual samples using the assigned tag sequence. Forward and reverse primers were then removed from the sequences. Trimmed sequences less than 150 bases in length were discarded. Also discarded were sequences with a simple edit distance of greater than two in the forward primer sequence. The read length was not always sufficient to cover the entire reverse primer. Depending on the end point in the reverse primer, a maximum edit distance 0 to 2 to the covered portion of the reverse primer was allowed. After this work was completed, additional control experiments indicated that sequences with incomplete reverse primer sequences or imperfect reverse primer sequences had an above average sequence error rate (not shown). We would suggest that a perfect reverse primer sequence filter be included in future work.

Sequence alignment. Sequences were aligned using the INFERNAL version 8.1, a stochastic context-free grammar based aligner (http://infernal.janelia.org/). The rRNA gene region corresponding to the region between primers (E.coli position 578 to 784) was extracted from the RDP version 9 alignment for the 5341 representative sequences used to train the RDP Classifier (1). The INFERNAL aligner was trained using this subalignment along with the Bacterial 16S rRNA secondary-structure model of Gutell and colleagues (4). The 205 residues estimated to be present in greater than 95% of all bacterial 16S rRNA sequences were selected as model positions for training. Sequences

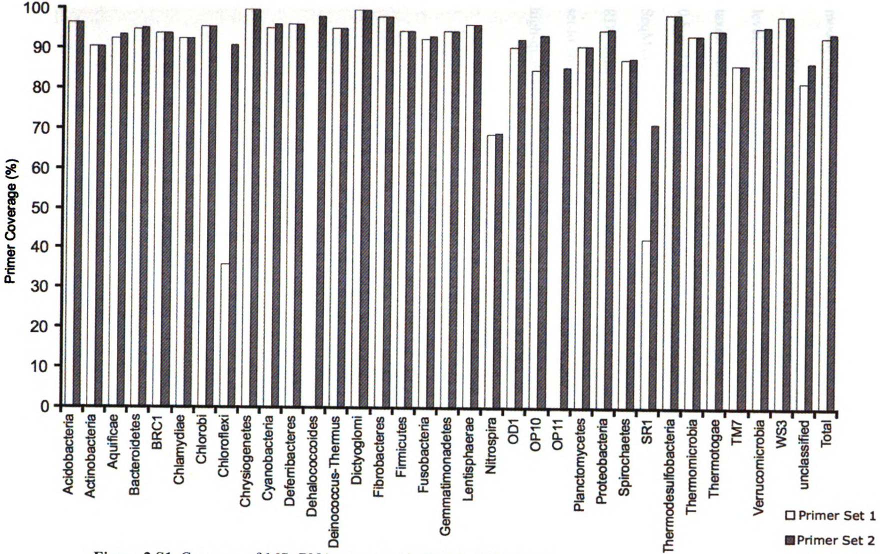


Figure 2.S1. Coverage of 16S rRNA sequences in RDP by V4 primers.

were aligned using this model and the options "--hbanded" and "--full". With this short model, Infernal aligns approximate 2200 reads per minute.

NEO plots. Sequences were first ordered by classification results at the phylum level, and for Firmicutes and Proteobacteria at the class level. Sequences assigned to each taxon were then ordered by successive complete linkage clustering at distances between 0.5 and 0.0 with a step size of 0.01. Each sequence was used as query to the RDP's SeqMatch tool trained on the RDP release 9.56 data set (6, 7) to find the sequence in the RDP's database with the largest number of matching words. The program options were set to search among all high-quality sequences greater than 1200 bases in length or only high-quality sequences from cultured isolates of length greater than 1200 bases.

REFERENCES

1. Lal R (2007) Carbon sequestration. Phil Trans R Soc B doi:10.1098/rstb.2007.2185.

2. Houghton RA (1994) The worldwide extent of land use change. Bioscience 44:305-313.

3. Scholes MC, Powlson D, Tian G (1997) Input control of organic matter dynamics. Geoderma 79:25-47.

4. Mann LK (1986) Changes in soil carbon storage after cultivation. Soil Sci 142:279-288.

5. Mann L, Tolbert V, Cushman J (2002) Potential environmental effects of corn (*Zea mays* L.) stover removal with emphasis on soil organic matter and erosion. Agric Ecosyst Environ 89:149-166.

6. Lal R et al. (2004) Managing Soil Carbon. Science 304:393.

7. Grandy AS, Robertson GP (2007) Land use intensity effects on soil organic carbon accumulation rates and mechanisms. Ecosystems 10:58-73.

8. Chivenge PP, Murwira HK, Giller KE, Mapfumo P, Six J (2007) Long-term impact of reduced tillage and residue management on soil carbon stabilization: Implications for conservation agriculture on contrasting soils. Soil Till Res 94:328-337.

9. Sandor R, Walsh M, Marques R (2002) Greenhouse-gas-trading markets. Philos Transact A Math Phys Eng Sci 360:1889-1900.

10. Zhou J et al. (2002) Spatial and resource factors influencing high microbial diversity in soil. Appl Environ Microbiol 68:326-334.

11. Treves DS, Xia B, Zhou J, Teidje JM (2003) A two-species test of the hypothesis that spatial isolation influences microbial diversity in soil. Microbial Ecol 45:20-28.

12. Ndour NYB et al. (2008) Characteristics of microbial habitats in a tropical soil subject to different fallow management. Appl Soil Ecol 38:51-61.

13. Muyzer G, De Wall BC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified gene coding for 16S rRNA. Appl Environ Microbiol 59:695-700.

14. Margulies M et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

15. Sogin ML et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103:12115-12120.

16. Huber JA et al. (2007) Microbial population structures in the deep marine biosphere. Science 318:97-100.

17. Roesch LF et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1:283-290.

18. Asuming-Brempong S et al. (2008) Changes in the biodiversity of microbial populations in tropical soils under different fallow treatments. Soil Biol Biochem 40:2811-2818 .

19. Acinas SG et al. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430:551-554.

20. Ley R, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces that shape microbial diversity and genome content in the human intestine. Cell 124:837-848.

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

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

23. Chao A, Chazdon RL, Colwell RK, Shen TJ (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62:361-371.

24. Garrity GM, Bell JA, Lilburn TG (2004) Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology. second edition. Release 5.0. Springer-Verlag New York.

25. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261-5267.

26. Cole JR et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research 37:D141-D145.

27. Chen MY (2004) Rubrobacter taiwanensis sp. nov., a novel thermophilic, radiation-resistant species isolated from hot springs. Int J Syst Evol Microbiol 54:1849-1855.

28. Yakimov MM, Lünsdorf H, Golyshin PN (2003) Thermoleophilum album and Thermoleophilum minutum are culturable representatives of group 2 of the Rubrobacteridae (Actinobacteria). Int J Syst Evol Microbiol 53:377-380.

29. Giardina CP, Sandford RL, Døkersmith IC, Jaramillo VJ (2000) The effects of slash burning on ecosystem nutrients during the land preparation phase of shifting cultivation. Plant Soil 220: 247-260.

30. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57:369-394.

31. Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 180:4765-4774.

32. Adiku SGK, Narh S, Jones JW, Laryea KB, Dowuona GN (2008) Short-term effects of crop rotation, residue management, and soil water on carbon mineralization in a tropical cropping system. Plant Soil 311, 29-38.

33. Nawrocki EP, Eddy SR (2007) Query-dependent banding (QDB) for faster RNA similarity searches. PLoS Comput Biol 3:e56.

34. Cannone JJ et al. (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3:2.

35. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406-425.

36. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127-128.

37. Cole JR et al. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res 33:D294-296.

38. Cole JR et al. (2007) The Ribosomal Database Project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res 35:D169-172.

SUPPORTING INFORMATION REFERENCE

1. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261-5267.

2. Liu Z *et al.* (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* 35:e120.

3. Rusch DB *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5:e77.

4. Cannone JJ *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* 3:2.

5. Cole J R *et al.* (2009) The Ribosomal Database Project: improved alignments and new

tools for rRNA analysis. *Nucleic Acids Res.* 37:D141-D145.

6. Cole JR *et al.* (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* 33:D294-296.

7. Cole JR *et al.* (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* 35:D169-D172.

8. Chao A, Chazdon RL, Colwell RK, Shen TJ (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* 62:361-371.

CHAPTER III

DNA-STABLE ISOTOPE PROBING INTEGRATED WITH METAGENOMICS: RETRIEVAL OF BIPHENYL DIOXYGENASE GENES FROM PCB-CONTAMINATED RIVER SEDIMENT

The work presented in this chapter has been published:

Woo Jun Sul, Joonhong Park', John F. Quensen III, Jorge L. M. Rodrigues, Laurie Seliger, Tamara V. Tsoi, Gerben J. Zylstra, and James M. Tiedje (2009) *Appl Environ Microbiol* 75:5501-5506

Author contributions:

Laurie Seliger and Gerben Zylstra performed sequencing and analysis of two cosmid clones. John Quensen measured PCB-transformation and biphenyl disappearance, Joonhong Park, and Tamara Tsoi were involved in experimental design and project development. Jorge Rodrigues helped with the phylogenetic analysis of 16S gene.

ABSTRACT

Stable isotope probing with [^{13}C]-biphenyl was used to explore the genetic properties of indigenous bacteria able to grow on biphenyl in PCB-contaminated River Raisin sediment. A bacterial 16S rRNA gene clone library generated from [^{13}C]-DNA after a 14-day incubation with [^{13}C]-biphenyl revealed the dominant organisms to be *Achromobacter* and *Pseudomonas*. A library from PCR amplification of genes for aromatic ring hydroxylating dioxygenases from the [^{13}C]-DNA fraction revealed two sequence groups similar to *bphA* (encoding biphenyl dioxygenase) of *Comamonas testosteroni* strain B-356 and of *Rhodococcus* sp. RHA1. A library of 1,568 cosmid clones was produced from the [^{13}C]-DNA fraction. A 31.8 kb cosmid clone, detected by aromatic dioxygenase primers, contained genes of biphenyl dioxygenase subunits *bphAE*, while the rest of the clone's sequence was similar to an unknown γ-Proteobacteria. The discrepancy of G+C content near the *bphAE* genes implies their recent acquisition possibly by horizontal transfer. The biphenyl dioxygenase from the cosmid clone oxidized biphenyl and unsubstituted and para-only substituted rings of polychlorinated biphenyl (PCB) congeners. DNA-stable isotope probing based cosmid libraries enabled the retrieval of functional genes from an uncultivated organism capable of PCB metabolism and suggests dispersed dioxygenase gene organization in nature.

INTRODUCTION

Commercially used polychlorinated biphenyls (PCBs), which are mixtures of more than 60 individual chlorinated biphenyl congeners, are among the most persistent anthropogenic chemical pollutants that threaten natural ecosystems and human health (1). Numerous biphenyl-degrading microorganisms have been isolated and studied, especially for the range of PCB congeners degraded. Research has been primarily focused on the biodegradative pathways and the biphenyl dioxygenases responsible for initial PCB oxidation by isolated bacteria (14, 27). Knowledge, however, is limited concerning the indigenous microbial populations that metabolize PCBs in the environment. Stable isotope probing (SIP) coupled with metagenomics is one approach to more directly explore which organisms and genetic information may be involved PCB degradation in PCB contaminated sites.

SIP was developed to separate and concentrate nucleic acids or fatty acids of microbial populations that metabolize and hence assimilate the isotopically labeled substrates into new cell material (4, 5, 28). Recently, the active PCB degraders in a biofilm community on PCB droplets were revealed as *Burkholderia* species using DNA-SIP (32). In another DNA-SIP study, 75 different genera that acquired carbon from [^{13}C]-biphenyl were found in the PCB contaminated root zone of a pine tree (22). In addition, that heavy [^{13}C]-DNA fraction revealed new dioxygenase sequences and possible PCB degradation pathways from GeoChip (16) results and from PCR amplified sequences using primers targeting aromatic ring hydroxylating dioxygenase (ARHD) genes (22).

A major hurdle in using DNA-SIP for metagenomic analyses (9) is the very small amount of heavy DNA that is produced and hence recovered making library construction difficult. Two studies have shown the feasibility of DNA-SIP for metagenomic analyses for C-1 compound utilizing communities but they first increased the amount of the heavy

DNA fraction by multiple displacement amplification (6, 10) or enriched the community by growth in sediment slurries. (18).

In this study, we used [^{13}C]-biphenyl to probe for potential PCB-degrading populations in a PCB-contaminated river sediment and to recover genes potentially involved in the critical first step of PCB degradation, the dioxygenase attack. We found a 31.8 kb cosmid clone that contained a biphenyl dioxygenase sequence (*bphAE*) and demonstrated its activity on PCBs.

MATERIALS AND METHODS

Sample description and SIP microcosms. Sediment historically contaminated with Aroclor 1248 at concentrations of 0.2 to 4.6 mg kg^{-1} was collected in October 2000 from River Raisin at Monroe, Michigan, USA. Sediment samples were stored at 4°C under river water until use.

Five replicate microcosms, each containing 5 g of sediment amended with 10 mg of uniformly labeled [^{13}C]-biphenyl (99 atom % ^{13}C)(Sigma-Aldrich) and 10 ml K1 minimal medium (34) was placed in 160-ml serum bottles. Sample bottles were sealed with Teflon stoppers and aluminum crimp-caps and incubated at room temperature in the dark on a horizontal shaker at 150 rpm. The microcosms were aerated by opening flasks in sterile conditions for 10 min every 3-4 days, and after 14 d, DNA was extracted from all microcosms.

To monitor biphenyl metabolism, nine microcosms amended with 10 mg unlabeled-biphenyl, and three sterile microcosms with twice-autoclaved sediment and unlabeled-biphenyl were established in parallel and incubated as described above. After 0, 7 and 14 d of incubation, triplicate microcosms were sacrificed for biphenyl extraction by the addition of 10 ml saturated KCl and 10 ml dichloromethane. Biphenyl concentrations were determined by gas chromatography with flame ionization detection. Split injections (50:1) were made on a J&K Scientific ICB-PAH capillary column (15 m, 0.25 mm ID, 0.15 μm film thickness). Temperature conditions were: inlet at 220°C; oven at 80°C for 1 min and then ramped 40°C min^{-1} to 220°C; detector at 325°C. Colony counts at each time point were obtained using R2A (29) agar plates and counted after 3 weeks of incubation.

DNA extraction and [^{13}C]-DNA separation. DNA was extracted following a previous protocol (35) but modified as follows to recover high molecular weight DNA. All sediment slurries were centrifuged at 3500 × *g* and 4 g of sediment pellet was transferred to a disposable 50-ml polypropylene centrifuge tube where 13.5 ml extraction buffer containing 0.1 M PIPES (pH 6.4), 100 mM EDTA, 1.5 M NaCl and 1% CTAB was added. Tubes were amended with 1.5 ml 20% SDS (w/v) and incubated in a 65°C water bath for 2 h with gentle inversion every 10 min. Supernatant without whitish material was collected after centrifugation at 3000 × *g* for 5 min and transferred into another 50-ml polypropylene tube and extracted with an equal volume of chloroform. DNA was precipitated with isopropanol, washed with ethanol, and dissolved in water at 50°C. For removing humic substances, the DNA solution was adjusted to 0.3 M NaCl by adding 1 M NaCl in TE (10 mM TrisCl, pH 8.0) and placed into 1 ml DEAE Sephacel (Sigma-Aldrich) columns pre-equilibrated with 0.3 M NaCl in TE. The columns were washed with 4 ml of 0.3 M NaCl in TE, and DNA was eluted with 4 ml of 0.5 M NaCl in

TE. DNA was again precipitated with isopropanol, washed with ethanol and dissolved in water at 50°C.

A total of 70 μg DNA at 0 d (D_0) and 14 d (D_{14}) was loaded in 18.5 ml cesium trifluoroacetate (CsTFA) (Amersham, Piscataway, New Jersey) solution without the addition of ethidium bromide and with a starting buoyant density of 1.60 g ml^{-1}. The CsTFA solution with DNA was transferred to 18.5 ml-Ultracrimp tubes (Sorvall, Waltham, Mass.). The tubes were centrifuged in a TV-865B vertical rotor (Sorvall) at 179,000 × *g* (43,500 rpm) for 40 h at 20°C. The gradients were fractionated into 500 μl fractions (up to 37 fractions) by displacement with water using a syringe pump at a flow rate of 1 ml min^{-1}. The buoyant density of each fraction was measured at 25°C by a refractometer. DNA fractions were precipitated with 1/10 volume of 3 M sodium acetate (pH 5.2) and isopropanol. The DNA pellets were then washed and re-suspended in EB elution buffer (Qiagen, Valencia, Calif.) and incubated at 50°C for 1 h. Fractionated DNA was quantified with a ND-1000 spectrophotometer (NanoDrop, Wilmington, Delaware). Secondary isopycnic density gradient centrifugation of combined DNA and quantitative PCR (Q-PCR) were conducted as described (22).

16S rRNA and Aromatic Ring Hydroxylating Dioxygenase (ARHD) gene clone libraries. Amplifications of 16S rRNA genes for clone libraries were conducted using primers 27F (17) and 529R (33), on D_0, and 27F and 1392R (17) on D_{14}H (H=heavy DNA fraction). Cycling conditions were as follows: denaturation for 5 min at 94°C, then 25 cycles of 1 min at 94°C, 1 min at 55°C, and 1 min (D_0) or 1 min 40 s (D_{14}H) at 72°C, and an additional 7 min extension at 72°C. PCR amplification of ARHD genes was performed using primers ARHD1F (5'-TTYRYNTGYANNTAYCAYGGNTGGG-3') and ARHD2R (5'-AANTKYTCNGCNGSNRMYTTCCA-3') with $D_{14}H$ as previously described (22). PCR amplicons of both 16S rRNA and ARHD genes were gel-purified using a QIAquick Gel Extraction Kit (Qiagen) and cloned using a TOPO TA Cloning Kit for Sequencing (Invitrogen, Carlsbad, Calif.). Clone libraries were sequenced using primers T7 or T3 at the Michigan State University, Research Technology Support Facility with an ABI 3730 Genetic Analyzer (Applied Biosystems Inc., Foster City, Calif.). The phylogenetic identification of 16S rRNA gene consensus sequences was performed using the RDP-II Classifier (7).

Cosmid library construction and screening library with ARHDs primers. Size-selected $D_{14}H$ (25-40 kb) was obtained by electrophoresis on 1% (w/v) low melting point agarose TAE gel, and the desired size DNA was recovered using Gelase (Epicentre Inc., Madison, Wisc.) without UV irradiation, end-repaired with T4 DNA polymerase, and then inserted into pWEB™ cosmid (Epicentre Inc.) at *SmaI* site. A cosmid library was constructed by using pWEB™ cosmid cloning kit. All cosmid clones were stored at -80°C. PCR amplification with ARHD primers was used for cosmid library screening as described above. Every 96 cosmid clones were pooled as templates for PCR screening.

Sequencing cosmid clone and genomic analysis. The cosmid clone L11E10 was sheared into approximately 4 kb fragments using a GeneMachines HydroShear device (Genomic Solutions, Ann Arbor, Mich.). The fragments were end repaired with T4 DNA polymerase and phosphorylated with T4 polynucleotide kinase (Epicentre). The DNA fragments were then ligated into the vector pCR-Blunt (Invitrogen) and transformed into E. coli TOP10. A total of 192 colonies were picked and then grown in LB plus 50 μg ml^{-1} kanamycin in deep well microtitre plates. Plasmid DNA was isolated using the Invitrogen PureLink 96 well lysis technique. The two ends of the inserted DNA fragment were sequenced using either the primer BL (5'-TCGGATCCACTAGTAACGGC-3') or BR (5'-CCAGTGTGATGGATATCTGC-3'). Sequences were trimmed and assembled using the Lasergene software (DNAStar, Madison, Wisc.).

PCB transformation by expression in *E. coli.* The *bphAE* of *Burkholderia xenovorans* LB400 was amplified from genomic DNA using primers (5'-CACCATGAGTTCAGCAATCAAGAA-3') (Underlined sequences were for directional cloning described below) for the forward sequence of *bphA* and (5'-CTAGAAGAACATGCTCAGGTT-3') for reverse sequence of *bphE*. PCR for LB400-*bphAE* was performed with Platinum® *Pfx* polymerase (Invitrogen) and 30 pmol of each primer for 25 cycles of 1 min at 94°C, 1 min at 55°C, and 4 min at 72°C. The *bphAE* genes of L11E10 were amplified from the cosmid clone DNA using (5'-CACCATGAATACTTTGATCAAAGAA-3') for forward sequence of *bphA* with modification of start codon GTG to ATG and (5'-TTAGAAGAACATGCTCAGGTT-3') for reverse sequence of *bphE*. PCR for L11E10 was performed for 25 cycles of 1 min at 94°C, 1 min at 55°C, and 6 min at 68°C. Both pET101[LB400-*bphAE*] and pET101[L11E10-*bphAE*] were generated using Champion™ pET101 Directional TOPO Expression Kit (Invitrogen). pET101[LB400-*bphAE*] or pET101[L11E10-*bphAE*] and pDB31[LB400-*bphFGBC*](2) were co-transformed into *Escherichia coli* BL21 Star(DE3).

PCB degradation capabilities of transformants were assessed using a resting cell assay. *E. coli* BL21 containing pET101[LB400-*bphAE*] or pET101[L11E10-*bphAE*], plus pDB31[LB400-*bphFGBC*] was grown in LB medium containing 100 μg ml^{-1} ampicillin and 25 μg ml^{-1} kanamycin in addition to 0.8 mM IPTG at 37°C. Log phase cells were washed and resuspended to an optical density of 1.75 at 600 nm in M9 medium containing 0.8 mM IPTG and 0.1% (w/v) sodium acetate. Portions (2 ml) were pipetted into glass vials, amended separately with one of two PCB mixtures in 10 μl of acetone, and sealed with Teflon-lined stoppers and aluminum crimp caps. The PCB mixtures were identical to mixtures 1B and 2B (3) except that 2,2′,4,4′,6,6′-CB (chlorinated biphenyl) was used as the internal standard instead of 2,2′,4,4′,6-CB; the final concentration of each congener was 1 μg ml^{-1}. The tubes were then incubated at 37°C with shaking at 200 rpm for 18 h. Following incubation, the contents of the tubes were acidified with three to four drops of concentrated HCl, and the PCBs were extracted three times with 1 ml of hexane:acetone (1:1, v:v). The extracts from each sample were combined and analyzed for PCBs using a gas chromatograph fitted with an electron capture detector and a DB-5 capillary column (30 m length, 0.32 mm ID, 0.25 μm film thickness). The oven temperature program was 140 °C for 1 min, then increased 2°C min^{-1} to 260 °C. The inlet and detector temperatures were 220 °C and 325 °C, respectively. PCBs were quantified using a four-point calibration curve and the internal standard method. In a separate experiment, accumulation of 2-hydroxy-6-oxo-6-

phenylhexa-2,4-dienoate (HOPDA) by transformants was determined at 434 nm (19) with a UV-Vis spectrophotometer (Varian Inc., Palo Alto, Calif.) after addition of biphenyl.

Nucleotide sequence accession numbers. The GenBank accession numbers are: ARHD of $D_{14}H$ (accession no. GQ231323-GQ231332), 16S rRNA clone libraries of $D_{14}H$ (accession no. GQ231333-GQ231378), and D_0 (accession no. GQ231379-GQ231433), and cosmid clone L11E10 (accession no. GQ231434).

RESULTS

Disappearance of biphenyl during the incubation. To confirm the feasibility of this sediment for the SIP experiment, biphenyl disappearance was measured in microcosms incubated with unlabelled biphenyl. Only 0.6% of the biphenyl remained after a 14 d aerobic incubation, whereas none of the biphenyl disappeared in the sterile microcosms. During the period, total culturable bacteria increased from 4.6×10^5 to 1.79×10^8 CFU's g^{-1} dry sediment as determined by plate counts.

DNA extraction and isopycnic centrifugation. DNA (D_0: DNA from sediment at 0 time, D_{14}: DNA from sediment in microcosms incubated with [^{13}C]-biphenyl for 14 d) was extracted by our high molecular weight DNA extraction method. Both D_0 and D_{14} were separately loaded, approximately 70 μg each, to 18.5 ml-scaled isopycnic centrifugation. [^{13}C]-DNA fractions of D_{14} were collected for buoyant densities from

1.634 to 1.656 g ml^{-1}, where DNA was detected in D_{14} but not in D_0. For confirmation that this fraction had [^{13}C]-DNA, the collected DNA from the heavy fraction (D_{14}H), from the unlabeled biphenyl incubated microcosms at 14 d (unlabeled D_{14}), and from D_0, were applied to 2 ml-scaled isopycnic centrifugation, followed by quantitative PCR of 16S rRNA genes on the separated fractions (Fig. 1). These results confirmed D_{14}H consisted of only [^{13}C]-DNA, clearly separated from either D_0 or unlabeled D_{14}. The approximately 3 μg of D_{14}H, was enough to construct a 16S rRNA gene clone library, a metagenomic library, and a PCR-based ARHD library.

Analysis of 16S rRNA and ARHDs genes in clone libraries. Fifty-five 16S rRNA gene clones from D_0 and 46 clones from D_{14}H were sequenced. The two libraries exhibited distinct microbial community composition and diversity (∫-LIBSHUFF *P* values for both ΔCxy and ΔCyx were <0.001) (30). The D_{14}H clone library, which should include active biphenyl degrading microorganisms, contained members of genera *Achromobacter*, *Pseudomonas*, *Acidovorax*, *Ramlibacter*, *Azoarcus*, and *Hydrogenophaga*, which were not found in the D_0 clone library (Table 1).

A library of ARHDs gene sequences in D_{14}H yielded five unique ARHD sequences from 10 clones, which could be divided in two groups, based on the translated amino acid sequences (99-106 aa). Clones 8, 13 (numbers of identical sequences=3), and 17 (n=2) exhibited 92%, 94%, and 94%, respectively, amino acid identities to a biphenyl dioxygenase large subunit of *Comamonas testosteroni* strain B-356 (31) (now *Pandoraea*

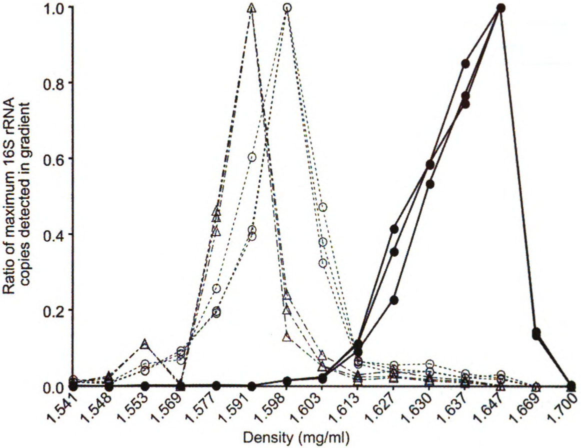


Figure 3.1. Separation of [^{12}C]- and [^{13}C]-DNA by small-scaled secondary isopycnic centrifugation and quantified by Q-PCR of 16S rRNA genes on triplicate samples. Solid circles and lines $D_{14}H$; open circles and dashed lines D_0; and open triangles and dashed lines D_{14}.

Phylogenetic group[a]	Number of clones		Genera[b] (Number of clones)
	D_0	$D_{14}H$	
Actinobacteria			
Intrasporangiaceae (c)	2		
Propionibacteriaceae (c)		1	
Unclassified Actinobacteria	1		
Acidobacteria	4		
Bacteroidetes	3		
Chloroflexi			
Caldilineacea(c)	7		*Levilinea* (1*), *Leptolinea* (1*),
Unclassified Anaerolineae	10		
Firmicutes	1		
Planctomycetes	1		*Pirellula* (1*)
Proteobacteria			
α-Proteobacteria			
Rhodobacteraceae(c)	2		*Rhodobacter* (1*)
Unclassified α-Proteobacteria	1		
β-Proteobacteria			
Rhodocyclaceae(c)		1	*Azoarcus* (1)
Gallionellaceae(c)	1		*Gallionella* (1*)
Comamonadaceae(c)	1	9	*Acidovorax* (6), *Ramlibacter* (2) *Hydrogenophaga* (1), Rhodoferax (1*)
Alcaligenaceae(c)		22	*Achromobacter* (22)
Hydrogenophilaceae(c)	2		*Thiobacillus* (2*)
Unclassified β-Proteobacteria	5	1	
γ-Proteobacteria			
Pseudomonadaceae (c)		9	*Pseudomonas* (9)
Xanthomonadaceae(c)		1	
δ-Proteobacteria	2		*Smithella* (1*), *Pelobacter* (1*)
Unclassified Proteobacteria		1	
OP10	1		
Unclassified bacteria	11	1	
Total	55	46	

a. The taxonomic assignment was based on the lowest taxonomic level that gave a > than 80% confidence level for assignment by the RDP-II Classifier release 9.50 (7).

b. Genera is indicated when more than 80% confidence.

c. Indicated taxonomy unit family.

*. Genera found in D_0

Table 3.1. Phylogenetic classification of 16S rRNA genes in clone libraries at zero (D_0) and 14 ($D_{14}H$) days.

pnomenusa (15)). Another group including clone 11 (n=2) and 12 (n=2) were similar to a dioxygenase large subunit of the gram-positive *Rhodococcus* sp. strain RHA1 (24) with amino acid identities of 82% and 77%, respectively.

Screening for and analysis of biphenyl dioxygenases. A library of 1568 cosmid clones, which contained DNA inserts averaging 30 to 40 kb (data not shown), from $D_{14}H$ was constructed and screened for genes encoding large subunits of biphenyl dioxygenases (*bphAs*) using primers to detect ARHD-encoding DNA. Five of the clones yielded ARHD amplicons of 300-330 bps, but sequencing of the amplicons showed that only one clone, L11E10, actually contained a *bphA* sequence. The *bphA* sequence from L11E10 was not an exact match with any of the PCR amplified ARHD sequences found in $D_{14}H$.

The clone L11E10 contained an insert of 31,850 bps with 67.38% G+C content. Seventeen of 22 open reading frames (ORFs) in L11E10 gave top BlastX hits against ORFs in the genera of *Xanthomonas* and *Stenotrophomonas*. Genes for subunits of the biphenyl dioxygenase (*bphA and E*) were found in L11E10. L11E10 contained no other genes directly relevant to the known biphenyl degradation pathway (Fig 2A). The *bphA* was highly similar to *bphA* in *Pseudomonas* sp. strain Cam-1 (90%) and *bphA1* in *Pseudomonas pseudoalcaligenes* KF707 (89.5%)(13). The *bphA* also encoded the motif Cys-X-His-X17-Cys-X2-His that forms the Rieske-type [2Fe-2S] cluster of iron-sulfur proteins. The *bphE* in L11E10 was 93% identical to *bphE* (a small subunit of biphenyl dioxygenase) in *B. xenovorans* LB400 and *bphA2* in *P. pseudoalcaligenes* KF707.

Functional analysis of biphenyl dioxygenases. To determine the activity of bphAE encoded in L11E10 (*bphAE*-L11E10) toward biphenyl and PCBs, *bphAE*-L11E10

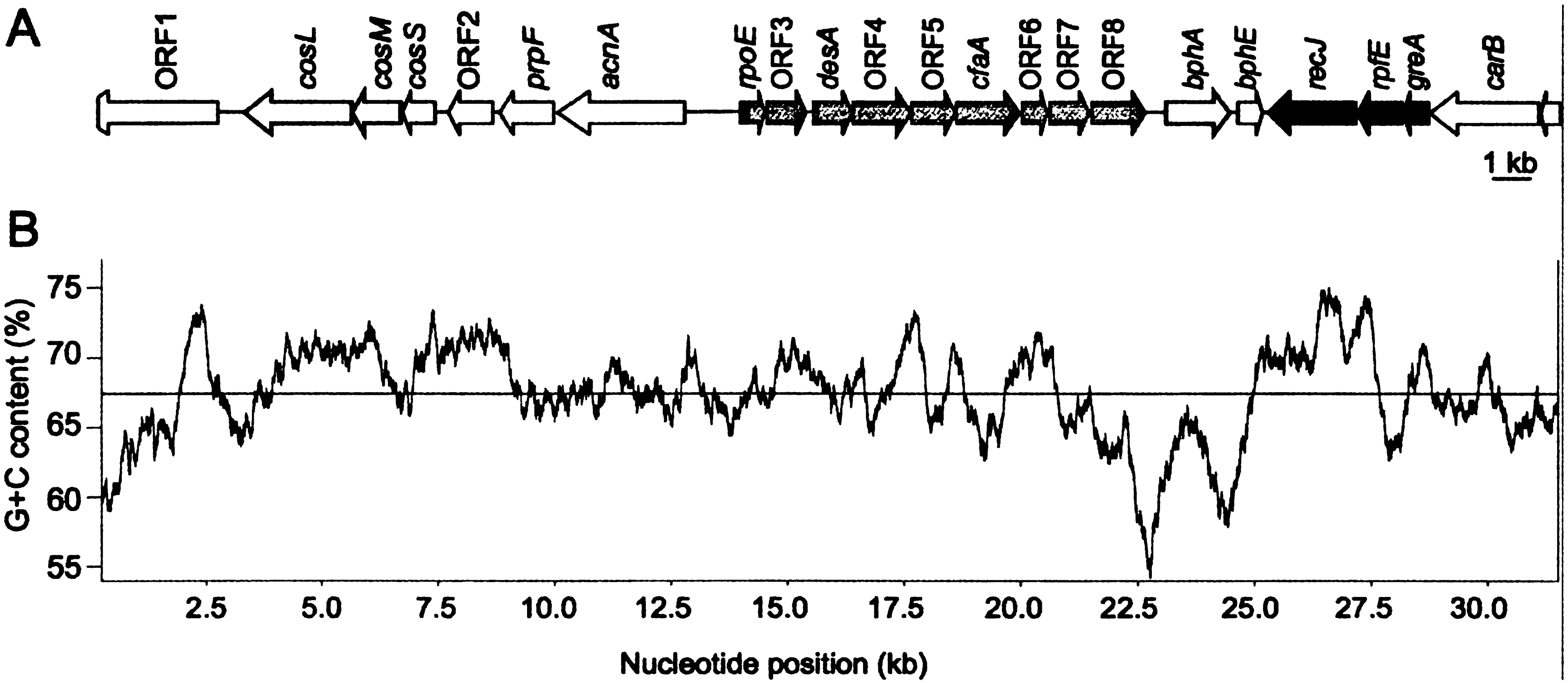


Figure 3.2. A. Schematic diagram of gene order in clone L11E10. Following are gene description: ORF, open reading frame; *cosL*, carbon monoxide dehydrogenase large subunit; *coxM*, carbon monoxide dehydrogenase middle subunit; *coxS*, carbon monoxide dehydrogenase small subunit; *prpF*, *AcnD*(the homolog of *acnA)*-accessory protein; *acnA*, aconitate hydratase; *rpoE*, RNA polymerase sigma 70 factor; *desA*, fatty acid desaturase; *cfaA*, cyclopropane fatty acyl phospholipid synthase; *bphA*, biphenyl dioxygenase, large subunit; *bphE*, biphenyl dioxygenase, small subunit; *recJ*, single stranded DNA specific exonuclease; *rpfE*, regulatory protein; *greA*, transcription elongation factor; and *carB*, carbamoyl phosphate synthase, large subunit. **B**. G+C contents of window size 500. Line in the middle refers to average G+C content of insert DNA in L11E10.

was expressed in *E. coli* BL21 along with *bphFGBC* from *B. xenovorans* LB400 (*bphFGBC*-LB400). The *bphFGBC*-LB400 encodes ferredoxin (BphF), ferredoxin reductase (BphG), biphenyl-2,3-dihydrodiol 2,3-dehydrogenase (BphB), and 2,3-dihydroxybiphenyl 1,2-dioxygenase (BphC), involved in the upper pathway of biphenyl catabolism. In this pathway, biphenyl is transformed to HOPDA producing a yellow color (23). When *E. coli* BL21 transformants containing *bphAE*-L11E10 were induced with IPTG and incubated with biphenyl, they produced the yellow color indicative of HOPDA within 2 h. In resting cell assays with PCB mixtures, the same transformants metabolized 2,3-CB, 2,4′-CB, 4,4′-CB, 2,4,4′-CB, and 2,4′,5-CB; the 4,4′-CB, 2,4,4′-CB to a greater extent than similar transformants containing *bphABFG* genes from LB400. These results are consistent with activities of resting cell assays of *P. pseudoalcaligenes* KF707 (11), with the exceptions that KF707 also exhibited some transformation of 2,2′,3,3′-CB and 2,3′,4,4′-CB (Table 2).

DISCUSSION

A major hurdle in DNA-SIP based metagenomics is the recovery of [^{13}C]-DNA in sufficient quantity for cosmid library construction and the production of a target number of clones. Due to these constraints, we used sediment slurries that were able to increase biphenyl consumption compared to our SIP study using [^{13}C]-biphenyl in rhizosphere soil (22), thus enhancing the incorporation of labeled carbon into cell material and obtaining sufficient [^{13}C]-DNA to produce a cosmid library. The resulting community, $D_{14}H$, seems to have less bacterial diversity than the heavy fraction from

Congener	% Depletion			
	L11E10	LB400	LB400[a]	KF707[a]
2,2′	<10	100	100	5
2,3	100	100	100	100
2,4′	100	100	100	100
4,4′	100	<10	15	100
2,2′,5	0	100	100	0
2,4,4′	92	22	45	93
2,5,4′	89	99	94	83
2,2′,3,3′	<10	96	94	60
2,2′,3,5′	0	96	96	0
2,2′,4,4′	0	16	38	0
2,2′,5,5′	0	99	95	0
2,3′,4,4′	0	0	16	24
2,3′,4′,5	<10	94	83	0
3,3′,4,4′	0	0	0	0
2,2′,3′,4,5	0	<10	38	0
2,2′,3,4,5′	0	29	58	0
2,2′,4,5,5′	0	64	73	0

a. Resting-cell assay data were obtained from previous study (11).

Table 3.2. Depletion of PCB congeners by the biphenyl dioxygenases of L11E10 and LB400.

using [^{13}C]-biphenyl in rhizosphere soil (22), as would be expected from the addition of the larger amount of biphenyl. This approach is useful for recovering functional genes from potentially unculturable populations and for analyzing their natural genetic context, but would not be useful for recovering genes from populations that might be specialists for low substrate concentrations. The dioxygenase clone we recovered did not overlap with the sequences amplified by the ARDH primers. The most likely explanation is that PCR bias favored genes not recovered in the cosmid.

The $D_{14}H$ community analysis showed that the dominant bacterial groups were closely related to previously known PCB and biphenyl-utilizing bacteria. The most dominant group, genera *Achromobacter*, includes *Achromobacter xylosoxidans* KF701, which can grow on biphenyl, 4-methylbiphenyl, 2-hydroxybiphenyl, benzoate and salicylate (12). Seven sequences in family *Comamonadaceae*, classified as *Acidovorax* and *Hydrogenphaga* by the RDP classifier, are most similar to PCB and biphenyl-degrading *Acidovorax* sp. (formerly *Pseudomonas* sp.) strain KKS102 (20, 26), and biphenyl-utilizing and PCB-cometabolizing psychrotrophic *Hydrogenophaga taeniospiralis* IA3-A (21). Also, genera *Pseudomonas* includes *P. pseudoalcaligenes* KF707, a well-known biphenyl and PCB-degrading microorganism.

It is interesting that L11E10 had only the *bphAE* genes of the biphenyl pathway and that the genetic organization differs from the upper *bph* operons of known biphenyl degrading microorganisms (27). In addition, the G+C content around *bphAE* was lower than average for the clone (Fig. 2B). Furthermore, the gene order of *rpoE*-ORF3-*desA*-ORF4-ORF5-*cfaA*-ORF6-ORF7-ORF8 (Fig. 2A, grey arrows) and *recJ*-*rpfE*-*greA* (black arrows) in L11E10 were identical to six sequenced *Xanthomonas* genomes, none of which have the upper *bph* operons. Therefore, *bphAE* in L11E10 could have been recently acquired from another microorganism, perhaps an outcome of the at least 40-year exposure to Aroclor 1248 in these sediments. It is possible that the gene organization of *bph* operons in nature is dispersed while the *bph* operons found in biphenyl-degrading microorganisms typically isolated by enrichment culture are less common, but better arranged for rapid growth and hence isolation.

Analysis of the origin of L11E10 suggests that the insert DNA came from a γ-Proteobacterium because the homology in L11E10 of *recJ,* a single stranded DNA specific exonuclease required for efficient recovery of DNA synthesis (8), was highly similar to those in γ-Proteobacteria.

BphAE-L11E10 showed a PCB congener transformation spectrum similar to but narrower than the KF707 biphenyl dioxygenase. It appeared to transform only PCB congeners without chlorines at the 2,3 positions. This is consistent with BphA protein sequences in which regions I, II, III and IV of L11E10, responsible for substrate specificity (25), are identical to KF707 biphenyl dioxygenase except Val-337 (L11E10) instead of Ile-335 (KF707) at LB400 position 336 (Fig. 3). As such, Val-337 (L11E10) may effect a narrow specificity toward 2,2′,3,3′-CB and 2,3′,4,4′-CB. Even though the difference in the N-terminus (31 amino acid differences before position 196) and C-terminus (11 amino acid differences after position 395) between BphA-L11E10 and KF707 or LB400 is greater than between LB400 and KF707 (only one amino acid difference), this does not appear to affect PCB substrate specificity (14).

Combining DNA-SIP and metagenomic analyses should increase our understanding of genomic features of microbial populations in nature since it avoids

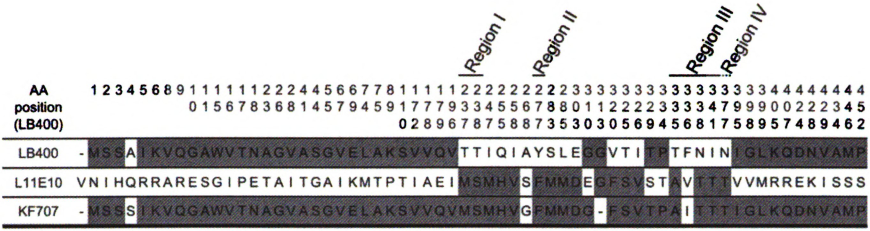

Region I
Region II
Region III
Region IV
AA position (LB400)
1 2 3 4 5 6 8 9 1 1 1 1 1 1 2 2 2 4 4 5 6 6 7 7 8 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4
0 1 5 6 7 8 3 6 8 1 4 5 7 9 4 5 9 1 7 7 7 9 3 3 4 5 5 6 7 8 8 0 1 1 2 2 2 2 3 3 3 3 4 7 9 9 9 0 0 2 2 2 3 4 5
0 2 8 9 6 7 8 7 5 8 8 7 3 5 3 0 3 0 5 6 9 4 5 6 8 1 7 5 8 9 5 7 4 8 9 4 6 2
LB400 - M S S A I K V Q G A W V T N A G V A S G V E L A K S V V Q V T T I Q I A Y S L E G G V T I T P T F N I N I G L K Q D N V A M P
L11E10 V N I H Q R R A R E S G I P E T A I T G A I K M T P T I A E I M S M H V S F M M D E G F S V S T A V T T T V V M R R E K I S S S
KF707 - M S S S I K V Q G A W V T N A G V A S G V E L A K S V V Q V M S M H V G F M M D G - F S V T P A I T T T I G L K Q D N V A M P

Figure. 3.3. Amino acid sequence alignment of large subunit of LB400, L11E10, and KF707 biphenyl dioxygenases. Only the positions that are not identical among three BphAs are shown in the alignment. The numbers of amino acid position are for LB400 BphA. The shaded area represents conserved amino acid sequences.

cultivation bias and minimizes interference from nonfunctional genes. The efficiency of the methods, particularly the sufficient recovery of labeled nucleic acids of high molecular weight, and its use under conditions that typify the natural environment, e.g. little disturbance and natural substrate concentrations, need further development.

ACKNOWLEDGEMENTS

WJS thanks Vivian Pellizari and Stephan Gantner for providing primers for library screening, and Jong-Chan Chae for technical assistance of cosmid library construction. We acknowledge Michel Sylvestre for advice and for providing plasmids containing *bphFGBC*-LB400 for PCB transformation. This work was supported by NIEHS grant P42-ES004911 under the Superfund Basic Science Program, and the Center for Microbial Ecology.

REFERENCES

1. **ATSDR.** 2000. Toxicological profile for Polychlorinated Biphenyls (PCBs). *In* Agency for Toxic Substances and Disease Registry. Public Health Service.

2. **Barriault, D., and M. Sylvestre.** 1999. A ColE1-compatible expression vector for the production of His-tagged fusion proteins. Antonie Van Leeuwenhoek **75:**293-7.

3. **Bedard, D. L., R. Unterman, L. H. Bopp, M. J. Brennan, M. L. Haberl, and C. Johnson.** 1986. Rapid assay for screening and characterizing microorganisms for the ability to degrade polychlorinated biphenyls. Appl Environ Microbiol **51:**761-8.

4. **Boschker, H. T. S., S. C. Nold, P. Wellsbury, D. Bos, W. de Graaf, R. Pel, R. J. Parkes, and T. E. Cappenberg.** 1998. Direct linking of microbial populations to specific biogeochemical processes by 13C-labelling of biomarkers. Nature **392:**801-805.

5. **Buckley, D. H., V. Huangyutitham, S. F. Hsu, and T. A. Nelson.** 2007. Stable isotope probing with 15N achieved by disentangling the effects of genome G+C content and isotope enrichment on DNA density. Appl Environ Microbiol **73:**3189-95.

6. **Chen, Y., M. G. Dumont, J. D. Neufeld, L. Bodrossy, N. Stralis-Pavese, N. P. McNamara, N. Ostle, M. J. Briones, and J. C. Murrell.** 2008. Revealing the uncultivated majority: combining DNA stable-isotope probing, multiple displacement amplification and metagenomic analyses of uncultivated Methylocystis in acidic peatlands. Environ Microbiol **10:**2609-22.

7. **Cole, J. R., B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje.** 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res **35:**D169-72.

8. **Courcelle, C. T., K. H. Chow, A. Casey, and J. Courcelle.** 2006. Nascent DNA processing by RecJ favors lesion repair over translesion synthesis at arrested replication forks in Escherichia coli. Proc Natl Acad Sci U S A **103:**9154-9.

9. **Dumont, M. G., and J. C. Murrell.** 2005. Stable isotope probing - linking microbial identity to function. Nat Rev Microbiol **3:**499-504.

10. **Dumont, M. G., S. M. Radajewski, C. B. Miguez, I. R. McDonald, and J. C. Murrell.** 2006. Identification of a complete methane monooxygenase operon from soil by combining stable isotope probing and metagenomic analysis. Environ Microbiol **8:**1240-50.

11. **Erickson, B. D., and F. J. Mondello.** 1993. Enhanced biodegradation of polychlorinated biphenyls after site-directed mutagenesis of a biphenyl dioxygenase gene. Appl Environ Microbiol **59:**3858-62.

12. **Furukawa, K., N. Hayase, K. Taira, and N. Tomizuka.** 1989. Molecular relationship of chromosomal genes encoding biphenyl/polychlorinated biphenyl catabolism: some soil bacteria possess a highly conserved bph operon. J Bacteriol **171:**5467-72.

13. **Furukawa, K., and T. Miyazaki.** 1986. Cloning of a gene cluster encoding biphenyl and chlorobiphenyl degradation in Pseudomonas pseudoalcaligenes. J Bacteriol **166:**392-8.

14. **Furukawa, K., H. Suenaga, and M. Goto.** 2004. Biphenyl dioxygenases: functional versatilities and directed evolution. J Bacteriol **186:**5189-96.

15. **Gomez-Gil, L., P. Kumar, D. Barriault, J. T. Bolin, M. Sylvestre, and L. D. Eltis.** 2007. Characterization of biphenyl dioxygenase of Pandoraea pnomenusa B-356 as a potent polychlorinated biphenyl-degrading enzyme. J Bacteriol **189:**5705-15.

16. **He, Z., T. J. Gentry, C. W. Schadt, L. Wu, J. Liebich, S. C. Chong, Z. Huang, W. Wu, B. Gu, P. Jardine, C. Criddle, and J. Zhou.** 2007. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME J **1:**67-77.

17. **Johnson, J. L.** 1994. Similarity analyses of rRNAs. American Society for Microbiology, Washington, DC.

18. **Kalyuzhnaya, M. G., A. Lapidus, N. Ivanova, A. C. Copeland, A. C. McHardy, E. Szeto, A. Salamov, I. V. Grigoriev, D. Suciu, S. R. Levine, V. M. Markowitz, I. Rigoutsos, S. G. Tringe, D. C. Bruce, P. M. Richardson, M. E. Lidstrom, and L. Chistoserdova.** 2008. High-resolution metagenomics targets specific functional types in complex microbial communities. Nat Biotechnol **26:**1029-34.

19. **Khan, A., and S. Walia.** 1989. Cloning of bacterial genes specifying degradation of 4-chlorobiphenyl from Pseudomonas putida OU83. Appl Environ Microbiol **55:**798-805.

20. **Kimbara, K., T. Hashimoto, M. Fukuda, T. Koana, M. Takagi, M. Oishi, and K. Yano.** 1989. Cloning and sequencing of two tandem genes involved in degradation of 2,3-dihydroxybiphenyl to benzoic acid in the polychlorinated biphenyl-degrading soil bacterium Pseudomonas sp. strain KKS102. J Bacteriol **171:**2740-7.

21. **Lambo, A. J., and T. R. Patel.** 2006. Isolation and characterization of a biphenyl-utilizing psychrotrophic bacterium, Hydrogenophaga taeniospiralis IA3-

A, that cometabolize dichlorobiphenyls and polychlorinated biphenyl congeners in Aroclor 1221. J Basic Microbiol **46:**94-107.

22. **Leigh, M. B., V. H. Pellizari, O. Uhlik, R. Sutka, J. Rodrigues, N. E. Ostrom, J. Zhou, and J. M. Tiedje.** 2007. Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs). ISME J **1:**134-48.

23. **Maltseva, O. V., T. V. Tsoi, J. F. Quensen, 3rd, M. Fukuda, and J. M. Tiedje.** 1999. Degradation of anaerobic reductive dechlorination products of Aroclor 1242 by four aerobic bacteria. Biodegradation **10:**363-71.

24. **Masai, E., A. Yamada, J. M. Healy, T. Hatta, K. Kimbara, M. Fukuda, and K. Yano.** 1995. Characterization of biphenyl catabolic genes of gram-positive polychlorinated biphenyl degrader Rhodococcus sp. strain RHA1. Appl Environ Microbiol **61:**2079-85.

25. **Mondello, F. J., M. P. Turcich, J. H. Lobos, and B. D. Erickson.** 1997. Identification and modification of biphenyl dioxygenase sequences that determine the specificity of polychlorinated biphenyl degradation. Appl Environ Microbiol **63:**3096-103.

26. **Ohtsubo, Y., H. Goto, Y. Nagata, T. Kudo, and M. Tsuda.** 2006. Identification of a response regulator gene for catabolite control from a PCB-degrading beta-proteobacteria, Acidovorax sp. KKS102. Mol Microbiol **60:**1563-75.

27. **Pieper, D. H.** 2005. Aerobic degradation of polychlorinated biphenyls. Appl Microbiol Biotechnol **67:**170-91.

28. **Radajewski, S., P. Ineson, N. R. Parekh, and J. C. Murrell.** 2000. Stable-isotope probing as a tool in microbial ecology. Nature **403:**646-9.

29. **Reasoner, D. J., and E. E. Geldreich.** 1985. A new medium for the enumeration and subculture of bacteria from potable water. Appl Environ Microbiol **49:**1-7.

30. **Schloss, P. D., B. R. Larget, and J. Handelsman.** 2004. Integration of microbial ecology and statistics: a test to compare gene libraries. Appl Environ Microbiol **70:**5485-92.

31. **Sylvestre, M., M. Sirois, Y. Hurtubise, J. Bergeron, D. Ahmad, F. Shareck, D. Barriault, I. Guillemette, and J. M. Juteau.** 1996. Sequencing of Comamonas testosteroni strain B-356-biphenyl/chlorobiphenyl dioxygenase genes: evolutionary relationships among Gram-negative bacterial biphenyl dioxygenases. Gene **174:**195-202.

32. **Tillmann, S., C. Strompl, K. N. Timmis, and W. R. Abraham.** 2005. Stable isotope probing reveals the dominant role of Burkholderia species in aerobic degradation of PCBs. FEMS Microbiol Ecol **52:**207-17.

33. **Weisburg, W. G., S. M. Barns, D. A. Pelletier, and D. J. Lane.** 1991. 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol **173:**697-703.

34. **Zaitsev, G. M., and Y. N. Karasevich.** 1985. Preparatory metabolism of 4-chlorobenzoic and 2,4-dichlorobenzoic acids in Corynebacterium sepedonicum. Mikrobiologiya **54:**356–359.

35. **Zhou, J., M. A. Bruns, and J. M. Tiedje.** 1996. DNA recovery from soils of diverse composition. Appl Environ Microbiol **62:**316-22.

CHAPTER IV

UNIQUE PCB- AND BIPHENYL-UTILIZNG POPULATIONS IN THREE DIFFERENT ENVIRONMENTAL MATRICES

ABSTRACT

PCB- and biphenyl-utilizing populations in three PCB-contaminated environmental matrices: plant rhizosphere, sandy industrial soil, and river sediment were characterized using stable isotope probing with ^{13}C-biphenyl substrate and subsequent V4-16S rRNA gene pyrosequencing. Among the sites, PCB- and biphenyl-utilizing populations were mostly affiliated with Phyla *Proteobacteria*, *Actinobacteria* and *Acidobacteria* as well as *Firmicutes* particularly in the sediment. However, there is less phylogenetic redundancy among these PCB- and biphenyl-utilizing populations. Abundant members of PCB- and biphenyl-utilizing population were suggested to possess aromatic degradation genes or to have activity on aromatic compounds from previous studies. Phylum *Acidobacteria* and Genus *Escherichia* are new candidate groups that may be involved in PCB degradation in the environment. Ratios of richness (biphenyl-utilizing population / original community) suggested that 10-40% of total bacteria might utilize biphenyl carbon. Information attained by profiling populations active in PCB degradation in different environments might provide the clues for bioaugmentation of PCB.

INTRODUCTION

Polychlorinated biphenyls (PCBs) are widely distributed, persistent, anthropogenic pollutants (ATSDR, 2000). Removal of PCB from the environment occurs mostly by the way of bacterial oxidative degradation, anaerobic dechlorination or a combination of both, an important mechanism for ecosystem sustainability. Laboratory-based research shows that there were successes in the introduction of bacteria, known as bioaugmentation, can result inresponsible for extensive PCB degradation from extensive laboratory-based research to contaminated site materials (Hickey *et al.*, 1993; Focht and Brunner, 1985). However, *in-situ* studies with by introductedion of PCB-degrading strains to PCB contaminated environments often find that PCB degradation is minimal. This is thought to be due to several factors including failure of introduced strains to survive and/or grow, insufficient distribution and poor bioavailability and propagation failure in natural conditions. It is, therefore, necessary to investigate the composition of natural PCB-degrading populations in concert with thorough analysis of the chemical and physical properties of contaminated matrices (Rysavy *et al.*, 2005; Yan *et al.*, 2006). This will serve as a guide for improving the successful bioaugmentation strategies by selected indigenous PCB-degrading organisms.

As of February 2009, there were hundreds of 16S rRNA gene sequences that were "tagged" to "PCB/biphenyl" isolated bacteria deposited in Ribosomal Database Project (http://rdp.cme.msu.edu/index.jsp). These sequences were mostly affiliated with known aerobic PCB degraders: *Burkholderia*, *Pseudomonas*, and *Rhodococcus* as well as the anaerobic *Dehalococcoides* known for its dechlorination abilities. Although isolation of bacteria is necessary for evaluation of bioaugmentation strains, there is often a limited range of bacterial taxa that are cultivated from PCB-contaminated environments (Leigh *et al.*, 2006). Isolated bacteria likely do not represent actual PCB degrading community (Leigh *et al.*, 2007).

Thus, culture independent methods, such as 16S rRNA gene clone libraries, have been employed to study indigenous bacterial communities in PCB-contaminated environments. Sequences similar to *Burkholderia* and *Sphingomonas*, well-known PCB-degrading bacteria, were retrieved from PCB-contaminated soil. In addition, there were numbers of sequences affiliated with the phylum *Acidobacteria,* which is one of the most abundant phyla in soil, but are not known PCB-degraders (Nogales *et al.*, 1999; Nogales *et al.*, 2001). Another study identified increased abundance of *Rhizobiales* and *Acidobacteria* in rhizomediated PCB-contaminated sites (de Cárcer et al., 2007). These authors speculated that the identified bacteria were involved in either direct or indirect PCB utilization since PCB was a major carbon source.

Bacterial members responsive to PCB addition have been determined by assessing community structure before and after exposure to PCB droplets. Members of the active PCB-degrading population were found to be closely related to the genera *Aquabacterium*, *Caulobacter*, *Imtechiu*, *Nevskia*, *Parvibaculum,* and *Burkholderia* (Macedo *et al.*, 2007).

Alternatively, stable isotope probing (SIP) (Radajewski *et al.*, 2000) has been used to directly trace active bacteria involved in aerobic PCB degradation. This method takes advantage of the incorporation of labeled substrate into DNA and RNA of cells growing on the labeled substrate, which allows for taxonomic classification of the organisms and identification of functional genes that have become labeled. This has been used to target PCB-degrading bacteria in the rhizosphere of Austrian pine (*Pinus nigra*) growing in a PCB-contaminated industrial site. The most frequently identified members from the ^{13}C-DNA fraction were *Pseudonocardia, Kribella, Nocardiodes* and *Sphingomonas* (Leigh *et al.*, 2007).

In this study, we investigated active PCB-degrading communities in three PCB-contaminated environments using a combination of SIP and 16S rRNA gene pyrosequencing of the hypervariable V4 region. This study focuses on whether common PCB populations are selected from different soil or sediment communities.

MATERIALS AND METHODS

Site Description. Rhizosphere (Cz) soil (15 mg/kg of PCB, pH 7.7) was collected in the root zone of an *Austrian pine* (*Pinus nigra*) in the Czech Republic (Leigh *et al.*, 2006). Sandy soil (Pi) (120 mg/kg of PCB, pH 7) was collected at Picatinny Arsenal, NJ, USA. Sediment (4.8 mg/kg of PCB, pH 7.6) was collected from River Raisin, Monroe, MI. DNA of Cz 0d, Cz 4d, Cz 14d, Rr 0d, and Rr 14d (d=days) was obtained from previous studies (Leigh *et al.*, 2007; Sul *et al.*, 2009). Other DNA was collected by SIP following incubation with ^{13}C-biphenyl as follows. Microcosms for SIP were established following previous studies (Leigh *et al.*, 2007). Briefly, uniformly 1 mg ^{13}C-labeled biphenyl was added per 5 g environmental material. Isopycnic density gradient centrifugation and fractionation protocols were conducted following DNA extraction as previously described (Leigh *et al.*, 2007). ^{13}C-DNA fractions were determined by real-time PCR using 16S rRNA genes (Leigh *et al.*, 2007).

V4-16S rRNA Gene Pyrosequencing. PCR for amplicon pyrosequencing was performed with barcode primers, which targeted the 16S rRNA gene V4 region as previously described (Chapter 2). Pyrosequencing was performed using the Genome Sequencer FLX System (454 Life Sciences, Bradford, CT). Raw reads were processed, filtered, aligned, and clustered through the RDP Pyrosequencing Pipeline (Cole *et al.*, 2009). All 122,651 sequences were assigned to bacterial taxa with the RDP Classifier version 2, using the Taxonomic Outline of the Bacteria and Archaea (TOBA), release 7.8 (Cole et al., 2007). Bacterial assemblages were compared with *Chao* abundance-based adjusted *Sørensen* similarity calculated using *EstimateS* (purl.oclc.org/estimates) and then performed Principle Coordinate Analysis (PCoA) using the R statistical program (R Development Core Team) running the vegan package.

Estimates of Bacterial Richness. We implemented 7 parametric models: single point mss, gamma, lognormal, Inverse Gaussin, Pareto, mixture of two exponentials, and mixture of three exponentials to rank-frequency matrix of each sample. Model selection followed empirical procedures (Bunge and Barger, 2008). Briefly, we require that both GOF5 and GOF0 > 0.01 and then sort the results first by decreasing "tau" (right truncation point) and second by increasing AICc. Then the minimum-AIC model within each *tau* block (models evaluated at the same tau) is examined, and the one with the largest *tau* such that SE<= est/2 is selected. This may result in competing models, in which case we have to use expert judgment. Also, eleven nonparametric estimators were calculated using the software SPADE.

RESULTS

Bacterial communities in PCB-contaminated sites and their biphenyl-utilizing populations. The bacterial composition at the phyla level of three PCB-contaminated sites (rhizosphere, river sediment, and sandy soil), differed by soil type and PCB concentration, was determined by V4 16S rRNA gene pyrosequencing. The rhizosphere soil (Cz 0d) was dominated by three phyla: *Proteobacteria*, *Acidobacteria*, and *Verrucomicrobia* (Figure 1A). River sediment (Rr 0d) exhibited a high *Proteobacterial* abundance and contained more sequences affiliated to *Bacteroidetes, Firmicutes*, and *Chloroflexi* than rhizosphere and sandy soil (Figure 1B). *Actinobacteria* dominated the

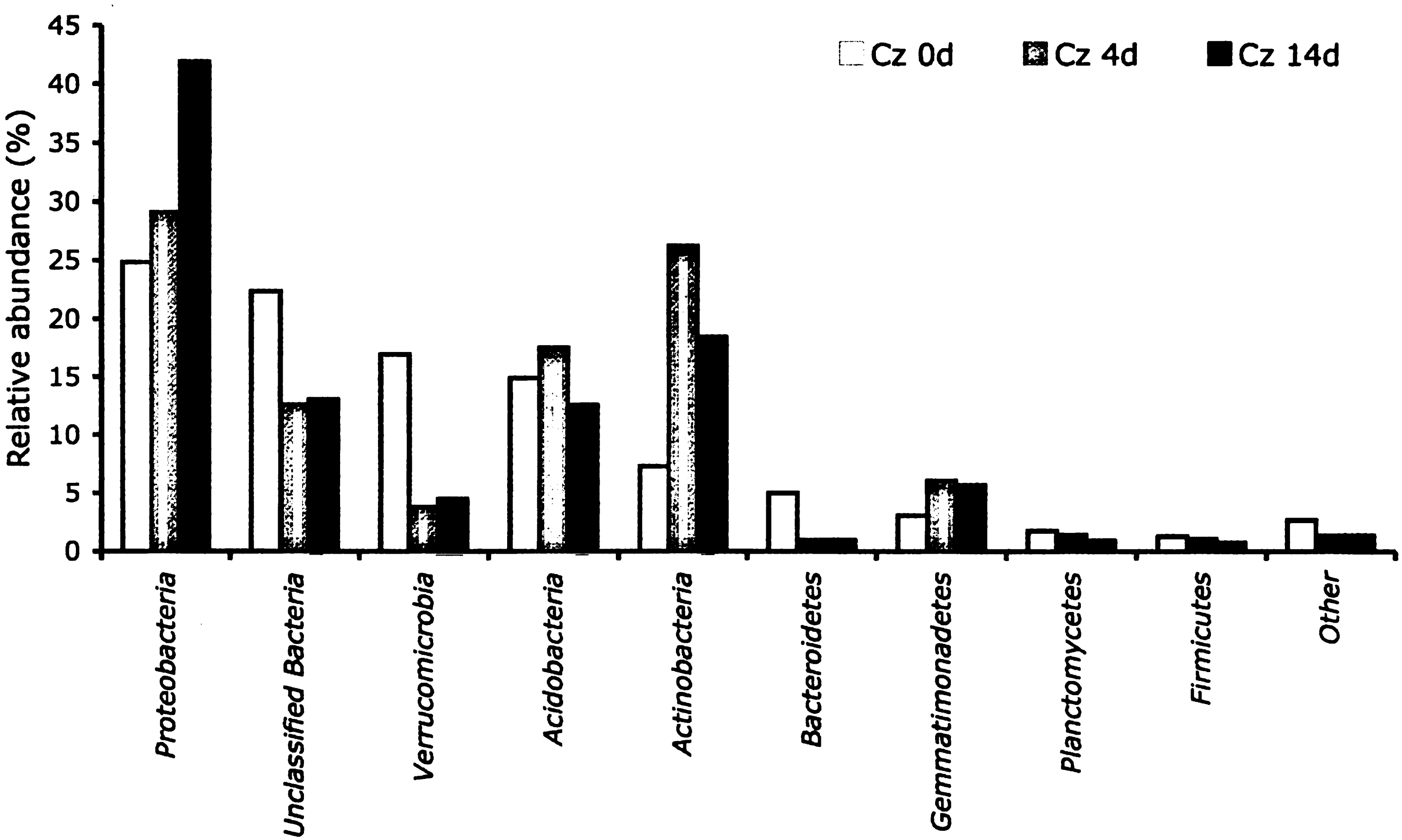


Figure 4.1. Bacterial phylum composition in three PCB-contaminated sites initially (0d) and after 4 and 14 days of incubation with biphenyl. A. Czech rhizosphere soil. B. River Raisin sediment, C. Sandy soil All sequences were classified by RDP-Classifier-II at a 50% confidence level.

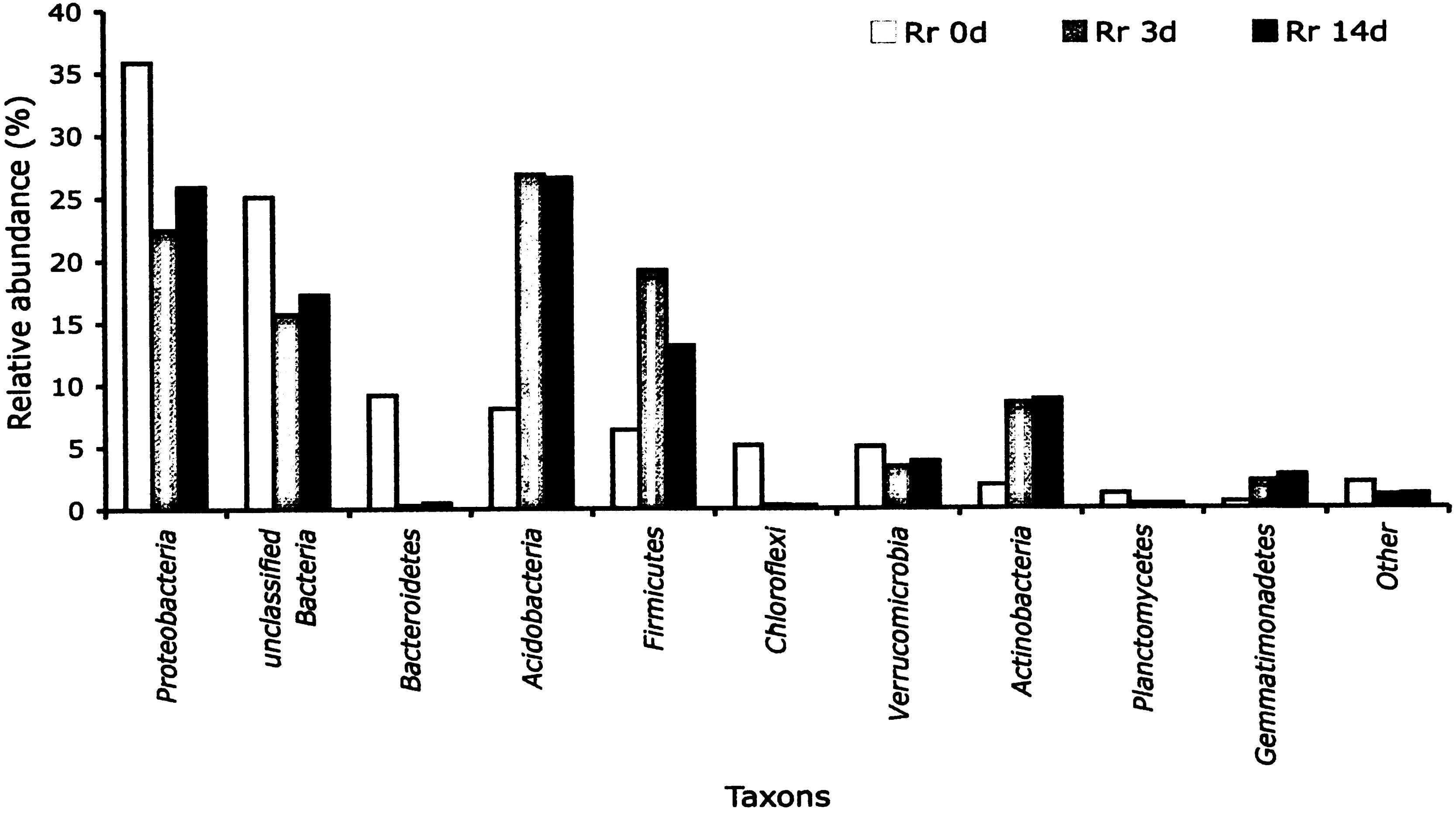


Figure 4.1B. River Raisin sediment.

sandy soil (Pi 0d) and included the genera *Streptomyces* (5.2%), *Nocardioides* (2.8%), and *Solirubrobacter* (2.6%) (Figure 1C).

Biphenyl-utilizing populations were analyzed using the collected heavy DNA derived from ^{13}C-biphenyl-SIP after 4 d and 14 d incubations. Both rhizosphere time points (Cz 4d and Cz 14d) contained sequences most closely classified as *Proteobacteria*, *Actinobacteria*, and *Acidobacteria* (Figure 1A). Notably, these samples were dominated by genera affiliated with *Actinobacteria*: *Nocardioides*, *Pseudonocardia*, *Kribbella*, and *Sphingomonas*, and with *Proteobacteria*: *Escherichia*, and *Bradyrhizobium*, and lastly to *Acidobacteria* Gp6 (Appendix A). In river sediment (Rr 4d & Rr 14d), *Firmicutes* were higher in relative abundance to other soils and were marked by a high abundance of *Proteobacteria* and *Acidobacteria* (Figure 1B). The most dominant genera were *Bacillus*, *Arthrobacter*, *Burkholderia,* and *Escherichia* (Appendix A). There was a lower abundance of sequences affiliated with *Bacteroidetes and Chloroflexi*, which were more than 5% of the relative abundance in the original matrix (Rr 0d) (Figure 1B). In the sandy industrial area soil (Pi), *Proteobacteria* had grown to 80% at 14d in relative abundance (20% at 0d) with less *Actinobacteria* compared to its 45% at 0d (Figure 1C). High abundances of *Phenylobacterium*, *Azospirillum*, *Lysobacter*, *Wautersia*, *Pseudoxanthomonas*, *Escherichia*, *Sphingomonas*, (ordered by relative abundance) as well as Acidobacteria Gp6 were identified in Pi at 14d (Appendix A). Among all three PCB contaminated sites, the ^{13}C-biphenyl utilizing populations were mostly *Proteobacteria*, *Actinobacteria*, *Acidobacteria* as well as *Firmicutes*, the later particularly in the sediment.

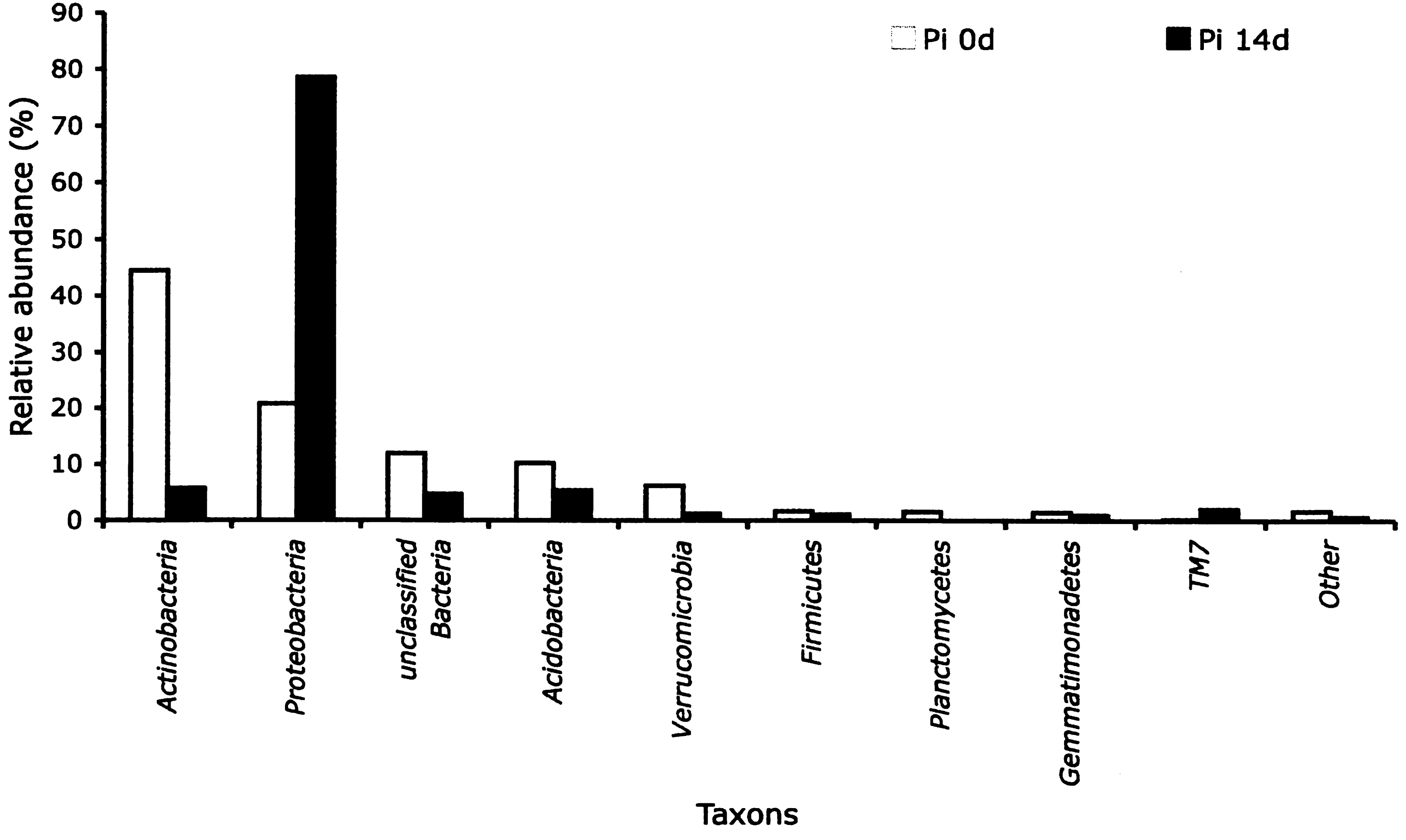


Figure 4.1C. Sandy soil.

PCB- and Biphenyl- Population Shifts During Incubation. A distance-based (*Chao's abundance based Sørenson Similarity*) principal coordinate analysis (PCoA) at a 97% OTU clustering illustrates the shift in bacterial community structure between that of the original total community and the biphenyl-utilizing populations over the 14 day incubation for the three PCB-contaminated sites (Figure 2). Shared OTUs between Cz 4d and Cz 14d contain 85% the sequences while shared OTUs between Rr 4d and Rr 14d) contain 75% of those sequences Most of the lower abundance OTUs in Cz4d were *Actinobacteria* whereas *Proteobacteria* increased at Cz 14d (Figure 5A). This increase was also found in the Rr incubation at 14d, but was accompanied by a decrease in *Bacillus* (Figure 5B).

Richness of both the total bacterial and biphenyl-utilizing communities was estimated by both parametric and non-parametric methods (supplemental materials). Regardless of sample origin, an estimation carried out at lower OTUs (90%) selected an inverse Gaussian as the appropriate abundance model. In contrast, 2-mixed or 3-mixed exponential models were better fits at higher OTU clustering levels. The proportions of the biphenyl-utilizing populations relative to total bacteria can be calculated from the ratio of richness estimations (biphenyl-utilizing population / total bacteria). Ratios at 97% OTUs are 27% (Cz 4d/Cz 0d with parametric), 27% (Cz 4d/Cz 0d with nonparametric), 43% (Cz 14d/Cz 0d with parametric) and 36% (Cz 14d/Cz 0d with nonparametric). The sandy soil has a lower proportion of biphenyl-utilizing populations: 16% (Pi 14d/Pi 0d with parametric), and 10% (Pi 14d/Pi 0d with nonparametric), while richness estimations of biphenyl-utilizing populations in the sediment were larger than the total bacteria population estimates: 218%, 153% (Rr 3d/Rr 0d with parametric, nonparametric,

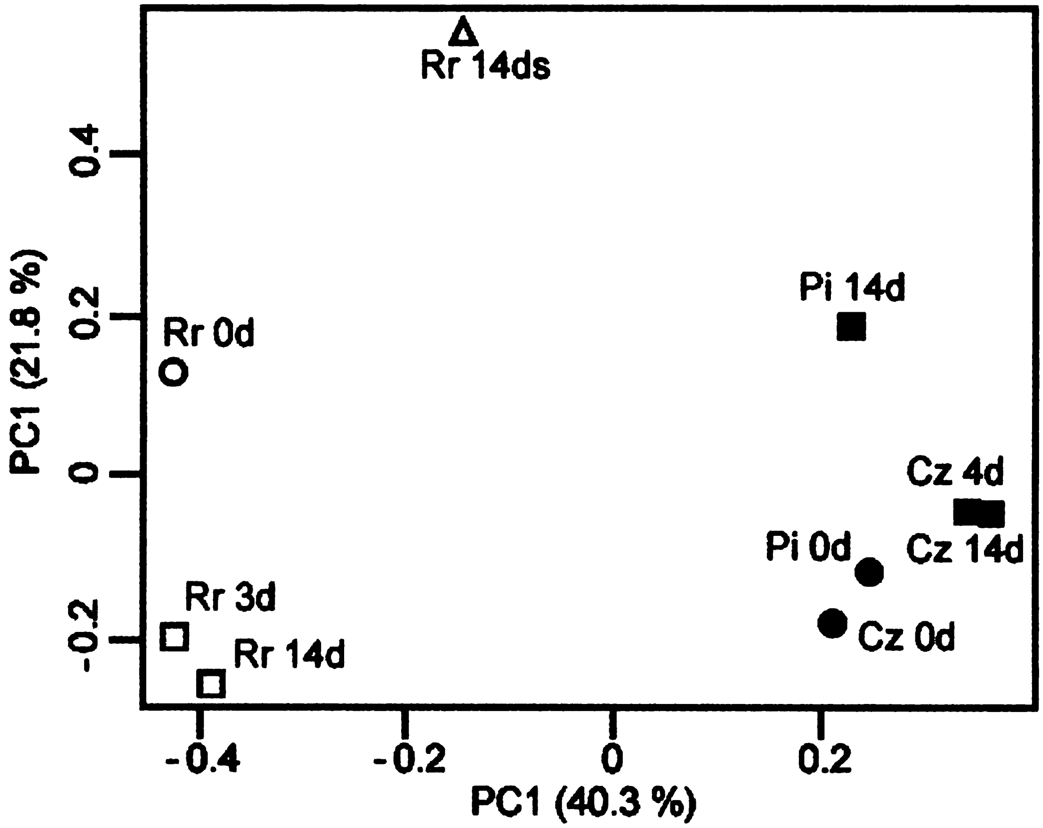


Figure 4.2. Principal Coordinate Analysis (PCoA) plot. Circles represents original PCB-contaminated matrix, square represent PCB- and biphenyl-utilizing community.

respectively), 128% and 109% (Rr 14d/Rr 0d with parametric, nonparametric, respectively) (Table 1, 2, and 3).

Shared OTUs of Three Biphenyl-Utilizing Populations After 14 Days Incubation. Over the same incubation period (Cz 14d, Rr 14d, Pi 14d), only 46 of 11,951 OTUs of biphenyl-utilizing bacterial populations were shared among all three samples. Representative sequences of each shared OTU, defined as those with the lowest sum distance to others within OTU's, were mostly *Acidobacteria*, *Actinobacteria*, and *Proteobacteria*. Two OTUs assigned to the genera *Escherichia* and unclassified Enterobacteriaceae were present at a relatively high abundance in all three samples (Figure 4). Most of the remaining OTUs were identified at high abundances in only one or two samples.

Different Incubation Methods Altered Biphenyl-Utilizing Populations. Different biphenyl-utilizing populations were detected depending on the SIP incubation conditions. A previously studied incubation on River Raisin sediments at 14 days (Rr 14d) used a slurry incubation instead of the static one as used in the experiments presented so far. The dominant genera in the slurry were *Pseudomonas* (47.8%), *Acidovorax* (6.9%), *Chitinophaga* (4.7%), and *Achromobacter* (3.6%). Using the static method these genera comprised less than 0.3% of the community in either Rr 3d or 14d. The top ten high abundance 97% OTUs of the current Rr 14d are rare members in Rr 14d slurry: <0.15% of relative abundance (Figure 2). The ten most abundant Rr 14d slurry OTUs accounted for only 0.46% of the sequences in Rr 14d static.

at 90% OTUs	No. of sequences	Obseved OTUs	Parametric estimate	Abundance Model	non-Parametric estimate	Estimator
Cz 0d	11400	1390	3171±249	Inverse Gaussian	2530±54	ACE-1
Cz 4d	4089	586	1006±73	Inverse Gaussian	824±16	ACE
Cz 14d	12338	898	1270±57	Inverse Gaussian	1138±19	ACE
Rr 0d	12697	1547	3368±234	Inverse Gaussian	2737±63	ACE-1
Rr 3d	22716	2274	3535±73	2-Mixed Exponential	3006±34	ACE
Rr 14d	24217	2167	2856±39	2-Mixed Exponential	2551±25	ACE
Rr 14ds	21449	551	1249±191	2-Mixed Exponential	830±40	ACE
Pi 0d	10609	1113	2973±338	Inverse Gaussian	2108±194	ACE-1
Pi 14d	3136	255	397±37	Inverse Gaussian	338±19	ACE

Table 4.1. Bacterial richness estimations at 90% OTUs. Abundance model of parametric estimates and estimator of nonparametric estimates were selected by empirical procedures to calculate "best" estimation.

97% OTUs	No. of sequences	Obseved OTUs	Parametric estimate	Abundance Model	non-Parametric estimate	Estimator
Cz 0d	11400	2846	9060±726	3-Mixed Exponential	7451±119	ACE-1
Cz 4d	4089	1075	2456±180	Inverse Gaussian	2045±192	ACE-1
Cz 14d	12338	1871	3938±406	3-Mixed Exponential	2647±34	ACE
Rr 0d	12697	2923	6994±241	2-Mixed Exponential	6827±114	ACE-1
Rr 3d	22716	6162	15225±1447	3-Mixed Exponential	10429±90	ACE
Rr 14d	24217	5493	8952±113	3-Mixed Exponential	7449±54	ACE
Rr 14ds	21449	926	2151±155	3-Mixed Exponential	2081±250	ACE-1
Pi 0d	10609	2324	6440±358	3-Mixed Exponential	6375±130	ACE-1
Pi 14d	3136	402	1030±159	Inverse Gaussian	646±40	ACE

Table 4.2. Bacterial richness estimations with 97% OTUs.

99% OTUs	No. of sequences	Obseved OTUs	Parametric estimate	Abundance Model	non-Parametric estimate	Estimator
Cz 0d	11400	3931	18527±2527	3-Mixed Exponential	13556±193	ACE-1
Cz 4d	4089	1432	3866±283	3-Mixed Exponential	3433±81	ACE-1
Cz 14d	12338	2824	7306±681	3-Mixed Exponential	5573±96	ACE-1
Rr 0d	12697	4132	14734±950	3-Mixed Exponential	12977±182	ACE-1
Rr 3d	22716	10095	25161±374	2-Mixed Exponential	19373±152	ACE
Rr 14d	24217	9016	16864±194	3-Mixed Exponential	13425±84	ACE
Rr 14ds	21449	1428	3833±273	3-Mixed Exponential	3583±90	ACE-1
Pi 0d	10609	3224	12463±846	3-Mixed Exponential	11848±176	ACE-1
Pi 14d	3136	542	1326.2±123	2-Mixed Exponential	1272.±223	ACE-1

Table 3. Bacterial richness estimations with 99% OTUs.

Figure 4.3A. Increase and decrease in relative abundance of shared OTUs in Cz 4d and Cz 14d. Solid line in the middle represents mean ratio of OTUs' relative abundance between two samples. OTUs indicated by lower case characters have at least two fold higher abundance than Cz 14d and more than 0.5% in relative abundance in Cz 4d. OTUs representative sequences were classified as: a, *Nocardioides*; b, unclassified bacteria; c, unclassified Nocardioidaceae; d, unclassified Micromonosporaceae; e, *Nocardioides*; f, *Nocardioides*; g, *Promicromonospora*; h, *Kribbella*; I, Acidobacteria Gp16; j, Acidobacteria Gp6. OTUs indicated by italic characters have consistent abundance both samples less than two fold difference to either side. OTUs indicated by numbers have at least two fold higher abundance than Cz 4d and more than 0.5% in relative abundance in Cz 14d. OTUs representative sequences were classified as: 1, *Pedomicrobium*; 2, *Escherichia*; 3, unclassified Rhizobiales; 4, unclassified Comamonadaceae; 5, unclassified Comamonadaceae; 6, *Sphingomonas*; 7, unclassified bacteria; 8, Verrucomicrobia Subdivision 3; 9, unclassified Rhizobiales; 10, unclassified Sphingomonadaceae.

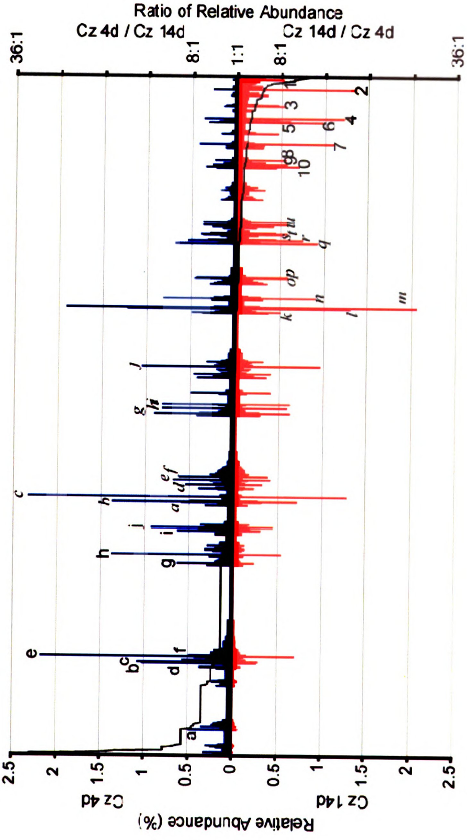
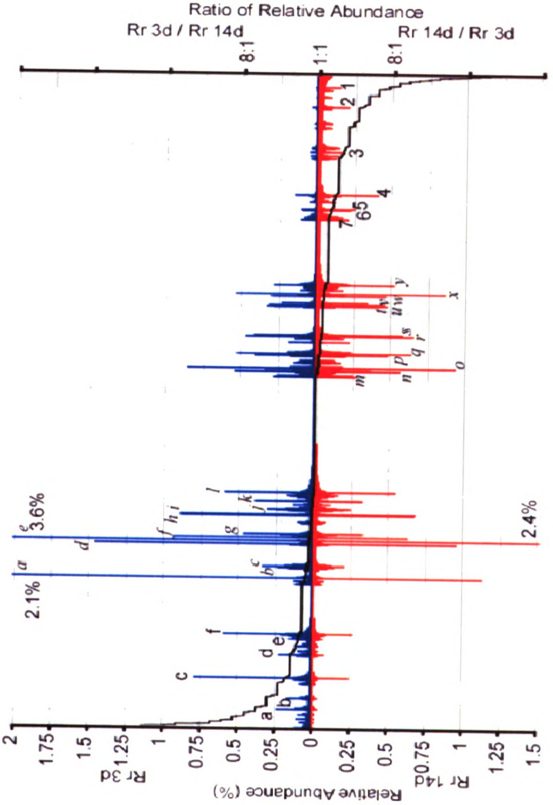

Ratio of Relative Abundance
Cz 4d / Cz 14d
Cz 14d / Cz 4d
36:1
8:1
1:1
8:1
36:1
1
2
3
4
5
6
7
8
9
10
u
t
s
r
q
o
p
n
m
l
k
j
g
h
i
f
e
d
c
b
a
e
c
b
d
f
h
g
i
j
a
2.5
2
1.5
1
0.5
0
0.5
1
1.5
2
2.5
Cz 4d
Cz 14d
Relative Abundance (%)

Figure 4.3B. Increase and decrease in relative abundance of shared OTUs in Rr 3d and Rr 14d. Solid line in the middle represents ratio of OTUs' relative abundance between two samples. OTUs indicated by small cap characters have at least two fold higher abundance than Rr 14d. Notable OTUs' representative sequences were classified as: a, *Acidobacteria* Gp7; b, *Acidobacteria* Gp4; c, *Burkholderia*; d, *Bacillus*;e, *Bradyrhizobium*; f, *Sporosarcina*; g, *Acidobacteria* Gp5. OTUs indicated by italic characters have consistent abundance both samples less than two fold difference to either side. Notable OTUs are: *a*, *b*, and *c*, *Bacillus*; *d*, *Arthrobacter*; *e* and *f*, *Bacillus*; g, *Acidobacteria* Gp4; *h*, unclassified *Proteobacteria*; *i*, *Bacillus*; *j*, *Acidobacteria* Gp4; *k*, *Methylobacterium*; *l*, unclassified bacteria; *m* and *n*, *Acidobacteria* Gp6; *o*, *Verrucomicrobia*; *p, Acidobacteria* Gp4; *q*, *Blastochloris*; *r*, *Acidobacteria* Gp6; *s, Escherichia*; *t*, *Acidobacteria* Gp6; *u*, *Acidobacteria* Gp4; *v*, unclassified bacteria; *w*, Rhodoplanes; *x*, *Gemmatimonas*; y, *Verrucomicrobia.* OTUs indicated by numbers have at least two fold higher abundance than Rr 4d. Notable OTUs' representative sequences were classified as: 1, *Clostridium*; 2, *Pseudomonas*; 3, unclassified Rhizobiales; 4, unclassified Sphingomonadaceae; 5, unclassified Beijerinckiaceae; 6, unclassified Bacteria; 7, *Gemmatimonas.*

Rr 3d
dance
fold
ified
us;e,
italic
ither
s; g.
4; k,
o,
); s,
; w,
have
nces
; 4,
fied



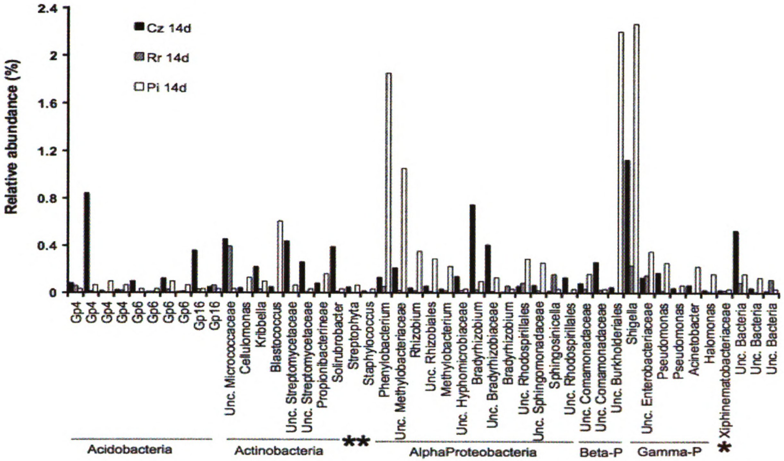


Figure 4.4. Shared OTUs among three PCB- and biphenyl-utilizing populations after 14 days incubation with ^{13}C-biphenyl (Pi). P is abbreviation of *Proteobacteria.*

DISCUSSION

We focused on the characterization of indigenous bacterial communities in three different PCB-contaminated sites and their PCB- and biphenyl-utilizing populations. Bacterial communities in these PCB-contaminated sites had very low phylogenetic commonality. These trends were also found in a previous study that showed four randomly chosen soils shared just a few common species, <5% at 97% OTUs (Fulthorpe *et al.*, 2008). Since the presence of PCBs is the only apparent common attribute in our soils, the differences in geographical distances, soil characteristics, plant interactions, and PCB concentrations can explain the taxonomic differences.

PCB- and biphenyl-degrading populations in PCB-contaminated sites differed by sample origin. The dominant genera in these sites are either known as PCB- and biphenyl-degrading bacteria, possess aromatic compound degradative genes, or were previously found in PCB-contaminated sites. Among PCB- and biphenyl-degrading populations of rhizosphere soil, were members of *Nocardioides*, *Pseudonocardia*, *Kribbella*, and *Sphingomonas*, which were previously identified in the 16S rRNA clone library from thee soils (Leigh *et al.*, 2007). In addition, *Bradyrhizobium* was found, which has members known to degrade 4-chlorobenezoate (Gentry *et al*, 2004) was also found in PCB-contaminated soil (Nogales *et al.*, 1999; Nogales *et al.*, 2001) and in PCB-biofilms (Tillmann *et al.*, 2005; Macedo *et al.*, 2007). Among PCB-and biphenyl-degrading populations in river sediment, *Bacillus* is known a thermophilic PCB-degrader isolated from compost (Shimura *et al.*, 1999). *Arthrobacter* can transform PCB congeners (Kohler *et al.*, 1988), induce PCB degradation by plant compounds (Gilbert and Crowley, 1997) and was also found in a chlorobenzene-contaminated aquifer (Abraham *et al.*,

2005) and Antarctica (Michaud *et al.*, 2007). *Burkholderia* are well-known PCB-degraders (reviewed in Pieper, 2008). Among PCB- and biphenyl-degrading populations in sandy soil, *Phenylobacterium* spp. possessed (herbicide) Chloridazon catechol dixoygenase (Blecher *et al.*, 1981), *Azospirillum* species showed chemotaxis to aromatic compounds such as protocatechuate, catechol, and 4-hydroxybenzoate (Lopez-de-Victoria and Lovell, 1993), *Lysobactor* species can degrade naphthalene and phenanthrene (Maeda *et al.*, 2009), *and Pseudoxanthomonas* species were able to degrade BTEX compounds (Kim *et al.*, 2008).

Most of the abundant genera have a relevancy to PCB or its intermediates degradation, while several dominant bacterial groups in biphenyl-degrading populations were not previously identified as known PCB- and biphenyl-degraders. The presence of *Acidobacteria* in the biphenyl-degrading populations in all three samples is of particular interest. *Acidobacteria*, especially of subdivision 4 and 6, may be members of an initial biphenyl-degrading consortium. However, there is no proof their biphenyl degradation due to difficulty in cultivation of members of this Phylum. *Acidobacteria* dominated in a highly PCB-contaminated soil (Nogales *et al.*, 1999) and the presence of aromatic ring dioxygenases such as protocatechuate 3,4-dioxygenase, albeit a more common aromatic metabolism pathway, was found in complete *Acidobacteria* genomes (Ward *et al.*, 2009). Surprisingly, sequences of the genera *Escherichia* was also consistently found in three biphenyl-degrading populations (Figure 4 and appendix A). *Escherichia* can be found outside of animal intestinal tracts, and environmental strains may harbor more metabolic diversity (Whitman and Nevers, 2003). The biphenyl-selected OTU, whose median (representative) sequence was classified as *Shigella,* seems most like clade V of

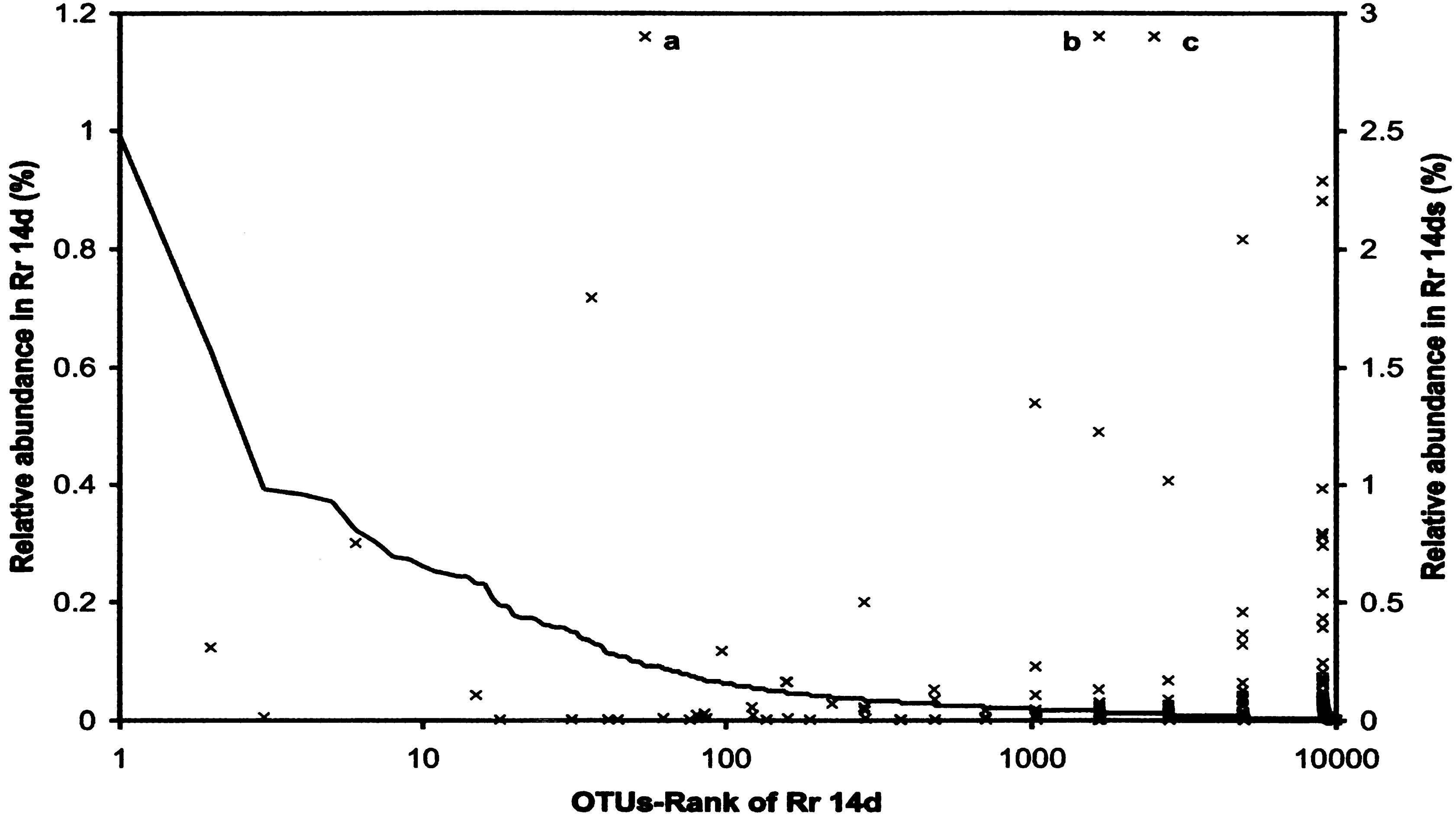


Figure 4.5. Relative abundances of Rr 14ds OTUs ordered by the rank of Rr 14d OTUs. Solid line is rank-abundance curves of Rr 14d and cross symbols indicated relative abundance in Rr 14d slurry following Rr 14d'slurry OTUs-rank. Relative abundances of three dominant OTUs in Rr 14ds slurry are 39.8% (indicated b), 11.3% (c), and 6.5% (a) classified as *Pseudomonas*, *Achromobacter*, and *Acidovorax*, respectively.

environmental *E. coli* based on sequence identity although there are no polymorphisms within the V4 region among clade V environmental, pathogenic *E. coli*, and *Shigella*. Regardless of whether this group is environmental *E. coli* or not, this group of bacteria hasn't yet been reported contain any biphenyl degradation related genes, although little is known about the metabolic capacity of the understudied environmental *Escherichia*. It is known, however, The *E. coli* possess enzymes for downstream steps of the biphenyl pathway. The consistently higher abundance of the *Escherichia* OTU in 14 d rather than 3d in sediment and rhizosphere soil is consistent with utilization of PCB intermediates.

A caveat of using SIP incubations is that primary biphenyl-degraders initially metabolize biphenyl but also produce secondary and intermediate metabolites that can be utilized by cross-feeders or non-specific carbon substrate scavengers. Hence, it is impossible to distinguish between primary or secondary biphenyl-C utilizing populations. This complexity is illustrated by the difference in biphenyl degrading populations among our sites (Figure 6). Although there was a general lack of common biphenyl degrading populations among our PCB-contaminated sites, 46 OTUs were common and may represent cosmopolitan bacteria able to degrade biphenyl or consume intermediate biphenyl substrates regardless of environmental barriers.

The application of deep sequencing to SIP (heavy DNA) samples has advantages in searching for and identifying less abundant possible PCB-degraders. For instance, in both Cz 4d and Cz 14d, we found 0.1% of sequences to be of *Rhodococcus*, which were previously the dominant isolates from the same sample (Leigh *et al.*, 2006), although not detected in the previous clone library. Another benefit is more reliable bacterial richness estimations that enables calculation of the portion of the community that can derive

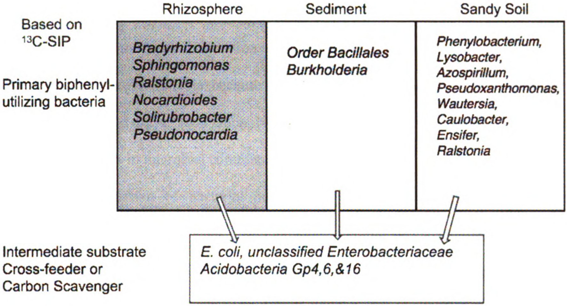


Figure 4.6. Schematic summary of biphenyl-utilizing bacteria and cross-feeders in three PCB-contaminated sites.

carbon from the single source. Based on our calculation, biphenyl can be utilized by 10-45% of the total community. Estimation ratios between Rr 0d, and RR 3d and RR 14d in river sediment are not reliable because we altered the environmental condition form anaerobic to aerobic during incubation. Nonetheless, this might be the first estimation of single carbon effect in microbial community.

Our comparison of bacterial populations between two different enrichment methods (Rr 14d slurry and Rr14d static) indicated that the slurry addition caused rapid growth of specific *r*-strategy bacterial groups. The slurry condition had greater substrate availability due to a 10x higher biphenyl concentration and resulted in an even carbon source distribution. The static conditions probably favored populations like those that would naturally encounter PCBs while the slurry favored the fast-growing soil consortium.

Overall, these findings indicate that ^{13}C-biphenyl utilizing population change as a function of the inherent site characteristics, incubation time, and incubation method. The lack of a common biphenyl degrading population among sites illustrates that soil heterogeneity plays a large role in promoting and maintaining these populations. This suggests that successful bioaugmentation of PCB contaminated soils requires that the capability of the native soil to sustain an augmented population is known. An appropriate augmented population can then be chosen to increase success rates in the remediation of PCB contaminated soils.

ACKNOWLEDGMENTS

W.J.S thanks to Seth Walk for advice on environmental *Escherichia* and to Ribosomal Database Project group for incredible support in analyses of rRNA sequence data.

AUTHOR CONTRIBUTIONS

Mary Beth Leigh provided the ^{13}[C]-DNA. John Bunge performed parametric and nonparametric estimates calculation. Ryan penton involved in statistical analysis and project improvement.

REFERENCES

Abraham WR, Wenderoth DF, Glässer W (2005) Diversity of biphenyl degraders in a chlorobenzene polluted aquifer. *Chemosphere* **58**:529-533

Blecher H, Blecher R, Wegst W, Eberspaecher J, Lingens F (1981) Bacterial degradation of aminopyrine. *Xenobiotica* **11**:749-754

Bunge J, Barger K (2008) Parametric models for estimating the number of classes. *Biom J* **50**:971-982.

Chao A, Chazdon RL, Colwell RK, Shen TJ (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* **62**:361-371

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**:D141-145

Focht DD, Brunner W (1985) Kinetics of Biphenyl and Polychlorinated Biphenyl Metabolism in Soil. *Appl Environ Microbiol* **50**:1058-1063

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

Gentry TJ, Wang G, Rensing C, Pepper IL (2004) Chlorobenzoate-degrading bacteria in similar pristine soils exhibit different community structures and population dynamics in response to anthropogenic 2-, 3-, and 4-chlorobenzoate levels. *Microb Ecol* **48**:90-10

Gilbert ES, Crowley DE (1997) Plant compounds that induce polychlorinated biphenyl biodegradation by Arthrobacter sp. strain B1B. *Appl Environ Microbiol* **63**:1933-1938

Hickey WJ, Searles DB, Focht DD (1993) Enhanced mineralization of polychlorinated biphenyls in soil inoculated with chlorobenzoate-degrading bacteria. *Appl Environ Microbiol* **59**:1194-1200

Kim JM, Le NT, Chung BS, Park JH, Bae JW, Madsen EL, Jeon CO (2008) Influence of soil components on the biodegradation of benzene, toluene, ethylbenzene, and *o*-, *m*-, and *p*-xylenes by the newly isolated bacterium *Pseudoxanthomonas spadix* BD-a59. *Appl Environ Microbiol* **74**:7313-7320

Kohler HP, Kohler-Staub D, Focht DD (1988) Cometabolism of polychlorinated biphenyls: enhanced transformation of Aroclor 1254 by growing bacterial cells. *Appl Environ Microbiol* **54**:1940-1945

Leigh MB, Pellizari VH, Uhlík O, Sutka R, Rodrigues J, Ostrom NE, Zhou J, Tiedje JM. (2007) Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs). *ISME J* **1**:134-148

Lopez-de-Victoria G, Lovell CR (1993) Chemotaxis of *Azospirillum* Species to Aromatic Compounds. *Appl Environ Microbiol* **59**:2951-2955

Macedo AJ, Kuhlicke U, Neu TR, Timmis KN, Abraham WR (2005) Three stages of a biofilm community developing at the liquid-liquid interface between polychlorinated biphenyls and water. *Appl Environ Microbiol* **71**:7301-7309

Maeda R, Nagashima H, Zulkharnain AB, Iwata K, Omori T (2009) Isolation and characterization of a car gene cluster from the naphthalene, phenanthrene, and carbazole-degrading marine isolate *Lysobacter* sp. strain OC7. *Curr Microbiol* **59**:154-159

Michaud L, Di Marco G, Bruni V, Lo Giudice A. (2007) Biodegradative potential and characterization of psychrotolerant polychlorinated biphenyl-degrading marine bacteria isolated from a coastal station in the Terra Nova Bay (Ross Sea, Antarctica). *Mar Pollut Bull* **54**:1754-1761

Nogales B, Moore ER, Abraham WR, Timmis KN (1999) Identification of the metabolically active members of a bacterial community in a polychlorinated biphenyl-polluted moorland soil. *Environ Microbiol* **1**:199-212

Nogales B, Moore ER, Llobet-Brossa E, Rossello-Mora R, Amann R, Timmis KN (2001) Combined use of 16S ribosomal DNA and 16S rRNA to study the bacterial community of polychlorinated biphenyl-polluted soil. *Appl Environ Microbiol* **67**:1874-1884

Pieper DH, Seeger M (2008) Bacterial metabolism of polychlorinated biphenyls. *J Mol Microbiol Biotechnol* **15**:121-138

Radajewski S, Ineson P, Parekh NR, Murrell JC (2000) Stable-isotope probing as a tool in microbial ecology. *Nature* **403**:646-649

Rysavy JP Yan T, Novak PJ (2005) Enrichment of anaerobic polychlorinated biphenyl dechlorinators from sediment with iron as a hydrogen source. *Water Res* **39**:569-578

Shimura M, Mukerjee-Dhar G, Kimbara K, Nagato H, Kiyohara H, Hatta T (1999) Isolation and characterization of a thermophilic *Bacillus* sp. JF8 capable of degrading polychlorinated biphenyls and naphthalene. *FEMS Microbiol Lett* **178**:87-93

Sul WJ, Park J, Quensen JF III, Rodrigues JLM., Seliger L, Tsoi TV, Zylstra, GJ, Tiedje JM (2009) DNA-Stable Isotope Probing Integrated with Metagenomics: Retrieval of Biphenyl Dioxygenase Genes from PCB-Contaminated River Sediment. *Appl Environ Microbiol* (in process)

Tillmann S, Strömpl C, Timmis KN, Abraham WR (2005) Stable isotope probing reveals the dominant role of *Burkholderia* species in aerobic degradation of PCBs. *FEMS Microbiol Ecol* **52**:207-217

Ward NL, Challacombe JF, Janssen PH, Henrissat B, Coutinho PM, Wu M, Xie G, Haft DH, Sait M, Badger J, Barabote RD, Bradley B, Brettin TS, Brinkac LM, Bruce D, Creasy T, Daugherty SC, Davidsen TM, DeBoy RT, Detter JC, Dodson RJ, Durkin AS, Ganapathy A, Gwinn-Giglio M, Han CS, Khouri H, Kiss H, Kothari SP, Madupu R, Nelson KE, Nelson WC, Paulsen I, Penn K, Ren Q, Rosovitz MJ, Selengut JD, Shrivastava S, Sullivan SA, Tapia R, Thompson LS, Watkins KL, Yang Q, Yu C, Zafar N, Zhou L, Kuske CR (2009) Three genomes from the phylum *Acidobacteria* provide insight into the lifestyles of these microorganisms in soils. *Appl Environ Microbiol* **75**:2046-2056

Yan T, Lapara TM, Novak PJ (2006) The Impact of Sediment Characteristics on PCB-dechlorinating Cultures: Implications for Bioaugmentation. *Bioremediat J* **10**:143-151

CHAPTER V

MICROBIAL COMMUNITY (ASSEMBLAGES) COMPARISONS BY BACTERIAL TAXONOMY-SUPERVISED METHOD BYPASSING SEQUENCE ALIGNMENT AND CLUSTERING

Author contributions:

Ryan Farris provide computational support and bacteria classifications. Ederson Jesus helped with statistical analysis, project improvement and provided some sets of sequences. Other providers of DNA samples, environmental matrix or pyrosequences were: Mary Beth Leigh, David Emerson, Chris Blackwood, Ederson Jesus, Erick Cardenas, Stres Blaz, Stephan Gantner, Claudia Etchebehere, Thad Stanton, Debora Rodrigues, Aviaja Hansen, Mathew Marshall, Alexandre Soares Rosado, and Dan Fisher.

ABSTRACT

Two different species-sites matrices, the abundance list of species as rows and sites (bacterial assemblages) as columns, from taxonomy-bins based on existing bacterial taxonomy and non-taxonomy-supervised (clustering-determined) OTUs were compared by classic Q-mode analysis, to describe interrelationships between sites and bacterial assemblages. Similarity index measures and morphology of points in principle coordinate analysis (PCoA) from two matrices based on 1.3 million 16S rRNA gene sequences from pyrosequencing were significantly correlated to each other. The taxonomy-supervised method, using taxonomy-bins, is able to compare non-overlapping sequences, which are often found in various regions within 16S rRNA genes sequences generated by pyrosequencing, and is not limited by the exhaustive computation required for the alignment and clustering required by the non-taxonomy-based method, but it does not resolve as well were the current taxonomy is limited.

INTRODUCTION

Recently, the increasing abundance of 16S rRNA genes sequences has provided new insight into the analysis of microbial communities (Tringe and Hugenholtz, 2008), mostly due to reduced sequencing cost by new sequencing technologies. Although short read lengths make it difficult to assign sequences for the purpose of bacterial taxonomy, deep sequencing with these new formats (e.g. 454 pyrosequencing [Margulies *et al.*, 2005]) is an emerging trend (Sogin *et al.*, 2006; Huber *et al.*, 2007; Roesch *et al.*, 2008; Chapter 2). More comprehensive sequencing provides better opportunities for intensive bacterial community profiling and bacterial community comparisons. When comparing bacterial assemblages with 16S rRNA gene sequences by classic Q-mode analysis to describe interrelationships between sites (bacterial assemblages), each sequence is allocated to species or OTUs (operational taxonomic units) by alignment-based clustering at a specified nucleotide distance, usually at a 97% similarity. This species-site OTU matrix, which is exclusively based on the nucleotide distances among 16S rRNA sequences, is aligned as rows with sites or assemblages as columns. This matrix can be generated and used for measuring site similarities either with presence / absence or abundance data. Site clustering and site ranking can also be performed with this site-site distance based matrix by ordination-based or hierarchical clustering. This process is termed "taxonomy non-supervised analysis", and is based simply on the distribution of sequences to OTUs.

When applying taxonomy non-supervised analysis, the large numbers of sequences ($>10^6$) generated by new sequencing technologies are an issue. Analysis requires a large computational capacity in order to process the sequence data (Hamady and Knight, 2009). The alignment and clustering of sequences that requires calculation of pair-wise nucleotide distances is the bottleneck when this method is used. Taxonomy non-supervised OTU analysis is advantageous in that it includes sequences which are yet unassignable to taxonomy. However, the current computational limitations make pursuing comparisons between among samples difficult.

Thus, we investigated an alternative method which is to allocate sequences into taxonomy-supervised OTUs or 'taxonomy-bins' based on the existing bacterial taxonomy, which rooted in 'polyphasic taxonomy' (Colwell, 1970) reflecting physiological, morphological, and genetic information. We define taxonomy-bins as all taxonomic units (Genus to Phylum) provided by the Taxonomic Outline of the Bacteria and Archaea (TOBA), release 7.8 (Cole *et al.*, 2007) augmented with non-validated taxa to cover sequences unassigned to the current bacterial taxonomy. Currently, several ribosomal RNA databases (i.e. RDP [Wang *et al.*, 2007], Greengenes [DeSantis *et al.*, 2006], and SILVA [Pruesse *et al.*, 2007]) are dedicated to sequence deposition and provide algorithm-based 16S rRNA gene classification tools.

In this study, taxonomy non-supervised OTUs and taxonomy-bins are compared using two similarity measures using 1.3 million sequences from 211 bacterial assemblages (Appendix B3).

MATERIALS AND METHODS

We used approximately 1.3M V4 region-16S rRNA gene sequences collected from 211 samples previously described in Chapter 2. We choose the following priori: The habitat grouping was based on the habitat definitions (Category of priori group G01-G11 were listed in Appendix B2; Group assignment of 211 samples were listed in Appendix B3.) suggested in Habitat-Lite Version 0.4 (Hirschman *et al.*, 2008; definition of terms were listed in Appendix B1).

For the non-supervised analysis, species-site matrices were generated as previously described in appendix B5. Briefly, all sequences were aligned by secondary structure using Infernal (Nawrocki *et al.*, 2009)), clustered by complete-linkage clustering, and then allocated into 97% OTUs through RDP's pyrosequencing pipeline (Cole *et al.*, 2009).

For the taxonomy-supervised analysis, all sequences were allocated into taxonomy bins: 1400 genus and 492 artificial 'unclassified' taxa provided by RDP classifier-II at 80%, 50%, and 0% confidence thresholds. Each of the lowest taxonomy units, i.e. genera and 'unclassified' taxa were considered as taxonomy-bins. The reliability of classification of each sequence was estimated using bootstrapping, and sequences that could not be assigned, as they were below a bootstrap confidence threshold, were located to an artificial 'unclassified' taxon.

Similarity measures of 211 samples (bacterial community assemblages) were calculated by pair-wise Chao's corrected *Sorensen* index (quantitative measures)(Chao *et al.*, 2006) and *Jaccard* index (presence/absence measures)(Jaccard, 1901) using EstimateS (http://viceroy.eeb.uconn.edu/EstimateS). Two site-by-site distance based matrices (1- Chao's corrected *Sorensen* index and 1- *Jaccard* index) from species-sites matrices of OTUs and taxonomy-bins were compared by *Mantel* test (Mantel, 1967) based on *Spearman*'s rank correlation *rho*. Site rank (ranks of bacterial assemblages) based Principal Coordinate Analysis (PCoA) was visualized in two dimensions to represent the greatest variability. The shape of points (assemblages) in PCoA plots was compared by *Procrustes* analysis, a statistical shape analysis that compares the distribution of points' shapes with all 211 points in 210 Principal Coordinates (PC) dimensions.

Three different sets of full-length (>1200bp) 16S rRNA gene sequence collections were used: RDP-II classifiers training set, human gut (Dethlefsen *et al.*, 2008), and soil (Elshahed *et al.*, 2008) were aligned and cut into V3, V4, and V6 hypervariable regions based on the reference positions of the *Escherichia coli* 16S rRNA gene. A query of full-length sequences to RDP-II classifier were compared to the query of the V3, V4, and V6 hypervariable regions.

RESULTS

Allocation of 1.3M sequences to taxonomy-bins or 97% OTUs. Each rRNA query sequence was assigned to a set of bins, 1400 genus and 492 artificial 'unclassified' taxa using a naïve Bayesian rRNA classifier (RDP-II classifier version 10). When the Classifier cutoffs were set at 80%, 50%, and 0% threshold (the latter forced all sequences to genus bins), 48%, 64% and 100% of the sequences were classified up to the genus level (Figure 1), and total number of taxonomy-bins (genera and 'unclassified' taxa) covering the 1.3 M sequences was 903, 1170, and 1259 bins, at 80%, 50%, and 0%, respectively. The mean value of maximum distance among the sequences within each bin was increased when the Classifier threshold was set lower. For taxonomy non-supervised OTUs, all sequences were clustered into 112,233 OTUs at 97% 16S rRNA sequence identity.

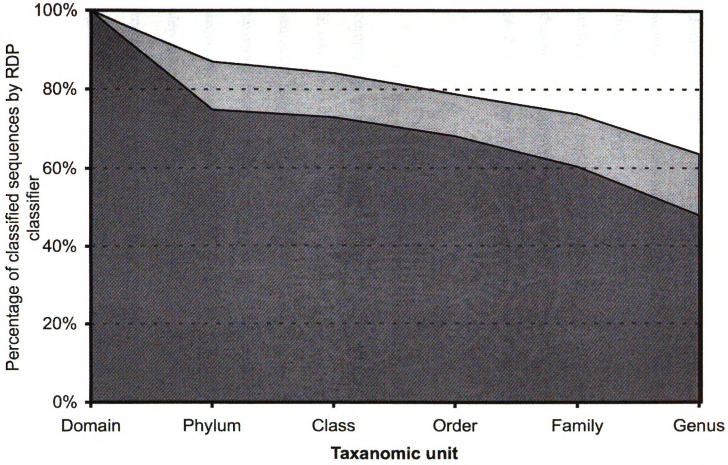


Figure 5.1. Sequence classification percentages at different confidence thresholds determined by RDP-II Classifier for different taxonomic levels.

A total of 22,154 pair-wise similarity index (*Chao's corrected Sorenson* similarity index or *Jaccard* similarity index) calculations of 211 bacterial assemblages were performed with both the taxonomy non-supervised OTUs-sites and the taxonomy-bins-sites matrices. We used *Mantel* matrix correlation test to compare the two site-site distance (1-similarity) based matrices (Table 1). The site-site matrix from taxonomy non-supervised OTUs was significantly and highly correlated with three site-site matrices from taxonomy-bins (Table 1 and Figure 2). All ordinations of principle coordinated analysis (PCoA) from the OTU-based dissimilarity matrix and taxonomy-bins-based dissimilarity matrices were also highly correlated to each other when all ordinations (k=210) of PCoA plots were compared by *Procrustes* rotation (Table 1).

DISCUSSION

The major advantage of the taxonomy-supervised method is the possibility for comparison between any region of the 16S rRNA gene without alignment and clustering, in contrast to the non-taxonomy supervised OTU method. Depending on the 16S rRNA sequence length and the resolution of the bacterial taxonomy classification, the taxonomy-based method can also compare the bacterial assemblages of 16S rRNA sequences spanning other hypervariable regions or bacterial assemblages with previously deposited sequences. For example, the RDP classifier-II returns similar classification results when compared to full-length queries at the genus level, regardless of the hypervariable region (Table 2). Therefore one can obtain compatible data regardless of the sequenced region. However, the coverage of the eubacterial primers used must be

	Similarity Index	Comparisons	Taxonomy bins at 80%	Taxonomy bins at 50%	Taxonomy bins at 0%
97% OTU-Based	1-Chao corrected *Sørensen* index	*Mantel* test *r* statstics	0.7763*	0.8008*	0.8146*
		Procrustes Analysis (r)	0.9396*	0.9406*	0.9404*
	1-*Jaccard* index	*Mantel* test *r* statstics	0.7856*	0.8595*	0.7856*
		Procrustes analysis (r)	0.6853*	0.7007*	0.6853*

Table 5.1. Similarity index measures and morphology of points in principle coordinate analysis (PCoA). *Mantel* statistic based on *Spearman*'s rank correlation rho and *Procrustes* rotation.

a. The significance of the statistic is evaluated by permuting rows and columns of the first dissimilarity matrix, * P value < 0.001

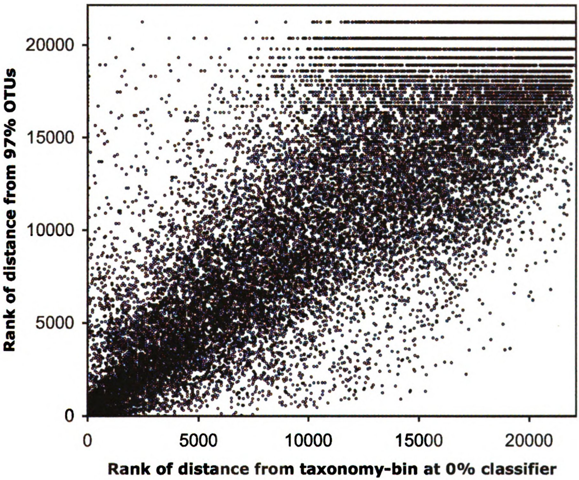


Figure 5.2. Rank comparison of distances (1-Chao's corrected *Sørenson* similarity) calculated using non taxonomy-supervised 97% OTUs and taxonomy-bins at 0% RDP classifier threshold.

	V3[a]			V4[b]			V6[c]		
Bootstrap cutoff	0%	50%	80%	0%	50%	80%	0%	50%	80%
Human Gut									
*% classified	100	92.4	82.3	100	97	87.9	100	73.5	40.4
**% matching	92	95	98.1	92.8	94.5	95.7	79	96.5	98.7
Soil									
*% classified	100	71.3	48.3	100	74.4	56.3	100	32.7	16.7
**% matching	70	85.5	94.6	84.1	93.3	96.8	48	80	84.3

[a] V3 region range: *E. coli* position 358~514 and length 130~160 bps; [b] V4 region: *E. coli* position 578~784 and length 200~210 bps; [c] V6 region: *E. coli* position 986~1045 and length 50~70 bps.

*% classified: fraction of sequences classified to genus level

**% matching: matching of each regions' classifications with full-length's classification.

Table 5.2. Bacterial classification accuracy of partial sequences spanning hypervariable V3, V4, and V6 regions in 16S rRNA gene considered because the different sets of primers preferentially covered or does not cover certain group of bacteria that derives the conflict community compositions.

Another advantage of the taxonomy-based method is that, due to the fixed number of taxonomy-bins, it is simple to add and delete bacterial assemblages from a pre-formulated bacterial assemblage comparison. Using taxonomy non-supervised OTUs, the addition and deletion of bacterial assemblages affects the species-sites matrices because the number and composition of sequences within OTUs are affected by re-alignment and re-clustering causing the addition and deletion of sequences. In addition, taxonomy-bin allocation is faster computationally than taxonomy non-supervised OTUs, which requires significantly longer processing times with the addition of sequences (complete linkage clustering requires increasing memory as the square root of the number of added sequences).

We focused on defining the differences between using taxonomy non-supervised OTUs and taxonomy-bin when comparing bacterial assemblages. Both a distance-based matrix and the morphology of points in PCoA ordinations confirmed that the two methods are significantly correlated such that the conclusions would be comparable. However, the resolution in comparing the bacterial assemblages is more limited with the taxonomy method due to the coarser average distance among taxa. The mean distance among the sequences inside the taxonomy-bins was 5.6%, 7.4%, 14.6% at 80%, 50%, 0% threshold, respectively. For example, there was a decreased resolution of priori G01 (basically soils) in taxonomy-bin based PCoA plots as compared to taxonomy non-supervised OTUs. This is due to the more limited number of taxonomy-bins in the Phylum *Acidobacteria* (26 genera and 4 'unclassified' taxa), *Verrucomicrobia* (10 genera

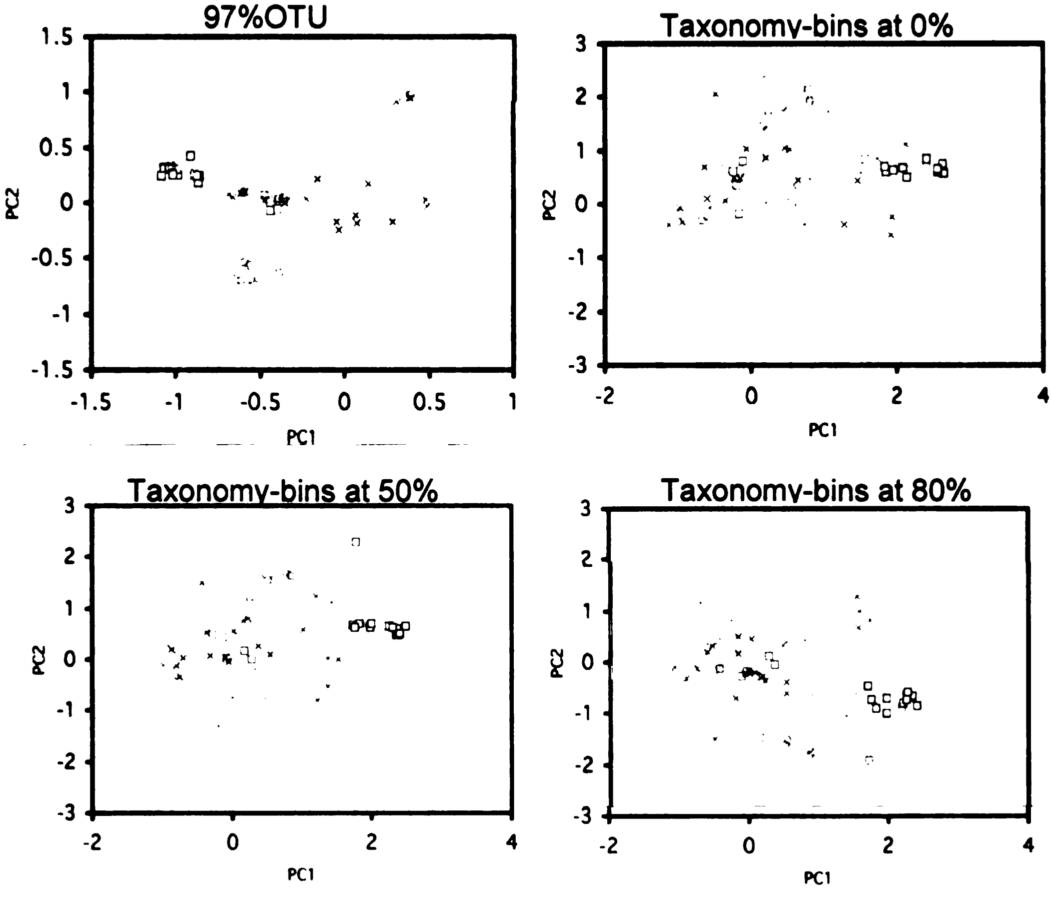


Figure 5.3A. PCoA plot comparisons by abundance based distance

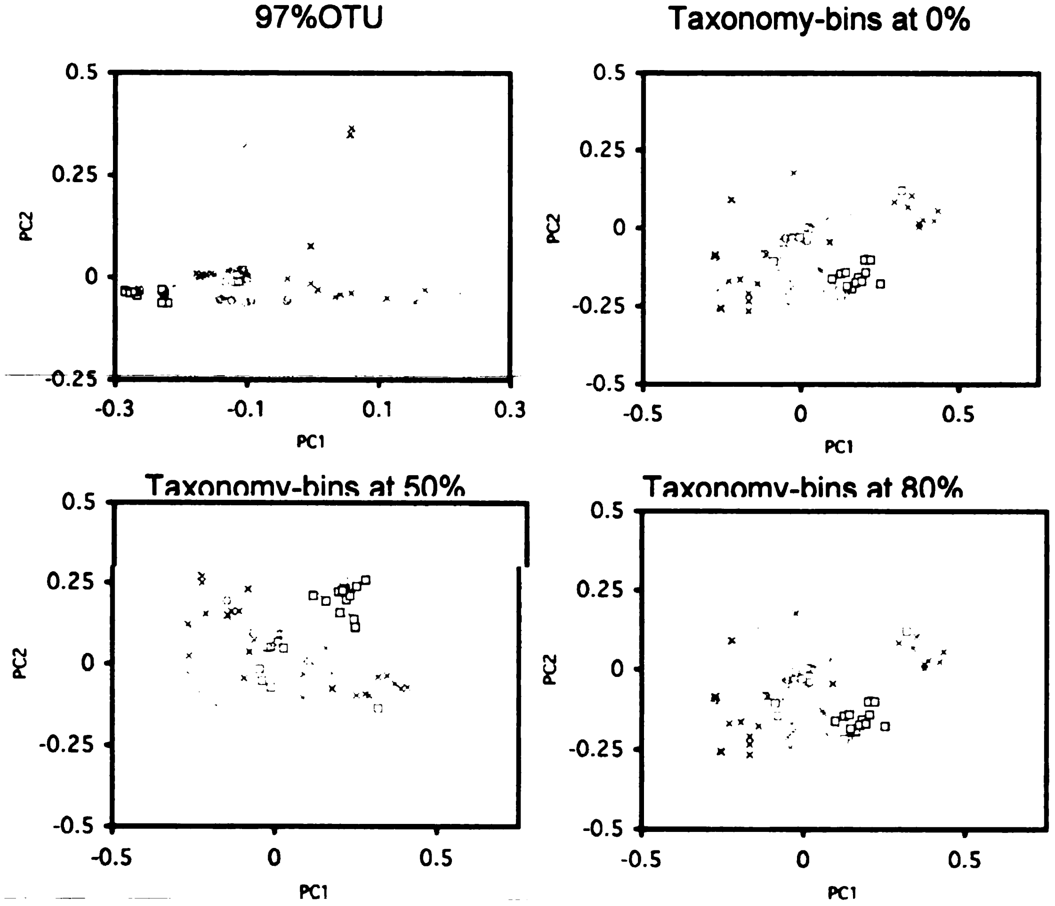


Figure 5.3B. PCoA plot comparisons by occurrence based distances

and 8 'unclassified' taxa), and *Gemmatimonadetes* (2 genera and 5 'unclassified' taxa). These bins have a relatively large number of sequences in priori G01 to the low number of isolated bacteria or described clusters. As such, their taxonomy is currently incomplete. In contrast, the assemblages in priori G04 (animal feces) were mostly composed of well-characterized groups and exhibited better separation to other groups with the taxonomy-bin method rather than the taxonomy non-supervised OTU method. When better classification of the bacterial taxonomy is available for these phyla and the 'unclassified' taxa, the bacterial assemblage comparison result should exhibit a higher resolution and more accurately reflect microbial community composition.

Revolutionary sequencing technologies continue to emerge, generating tremendous numbers of 16S rRNA gene sequences. However, current clustering tools are limited in both their flexibility and computational requirements. The taxonomy-based method has the potential to overcome these limitations as a fast and simple bacterial assemblage comparison method. Its value could be further improved if the microbiologists advanced the taxonomy for the poorly characterized groups.

REFERENCES

Chao A, Chazdon RL, Colwell RK, Shen TJ (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* **62**:361-71

Colwell RR (1970) Polyphasic taxonomy of the genus *vibrio*: numerical taxonomy of *Vibrio cholerae*, *Vibrio parahaemolyticus*, and related *Vibrio* species. *J Bacteriol* **104**:410-433

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

Dethlefsen L, Huse S, Sogin ML, Relman DA (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* *6*:e280

Elshahed MS, Youssef NH, Spain AM, Sheik C, Najar FZ, Sukharnikov LO, Roe BA, Davis JP, Schloss PD, Bailey VL, Krumholz LR (2008) Novelty and uniqueness patterns of rare members of the soil biosphere. *Appl Environ Microbiol* **74**:5422-5428

Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* **19**:1141-1152

Hirschman L, Clark C, Cohen KB, Mardis S, Luciano J, Kottmann R, Cole J, Markowitz V, Kyrpides N, Morrison N, Schriml LM, Field D, Novo Project (2008) Habitat-Lite: a GSC case study based on free text terms for environmental metadata. *OMICS* **12**:129-136

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML (2007) Microbial population structures in the deep marine biosphere. *Science* **318**:97-100

Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* **37**:547–579

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res* **27**:209-220

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer, ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz

SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**:376-380

Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**:1335-1337

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**:283-290

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**:7188-7196

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* **103**:12115-12120

Tringe SG, Hugenholtz P (2008) A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11**:442-446

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**:5261-5267

Appendix A.

	Cz	Cz	Cz	Pi	Pi	Rr	Rr	Rr	Rr
	0d	4d	14d	0d	14d	0d	3d	14d	14ds
no rank Root	100	100	100	100	100	100	100	100	100
domain Bacteria	99.9	100	100	100	100	99.9	99.9	99.9	100
UC) Bacteria	22.3	12.5	13.0	11.9	4.66	25.0	15.7	17.2	2.30
P) Actinobacteria	**7.25**	**26.2**	**18.4**	**44.4**	**5.71**	**1.87**	**8.50**	**8.75**	**0.15**
C) Actinobacteria	7.25	26.2	18.4	44.4	5.71	1.87	8.50	8.75	0.15
UC) Actinobacteria	0.98	1.96	1.66	2.48	0.10	0.20	2.09	2.31	0.00
SC) Actinobacteridae	3.26	19.4	12.6	21.1	4.91	1.18	4.51	3.72	0.11
O) Bifidobacteriales				0.01		0.94			
F) Bifidobacteriaceae				0.01		0.94			
G) Bifidobacterium				0.01		0.94			
O) Actinomycetales	3.22	19.4	12.6	21.1	4.91	0.23	4.44	3.69	0.11
UC) Actinomycetales	1.21	1.10	0.92	1.58	0.10	0.06	0.48	0.49	0.02
SO) Streptosporangineae		0.02	0.18	0.08	0.03		0.04	0.08	
F) Streptosporangiaceae			0.14	0.03			0.00		
G) Streptosporangium			0.11				0.00		
SO) Micrococcineae	0.55	2.13	1.95	2.76	0.54	0.06	2.64	1.87	0.02
UC) Micrococcineae	0.16	0.17	0.15	0.08			0.10	0.10	
F) Cellulomonadaceae	0.16	0.07	0.08	0.10	0.16		0.04	0.05	
G) Cellulomonas	0.16	0.07	0.08	0.10	0.16		0.04	0.05	
F) Promicromonosporaceae	0.03	0.68	0.26	0.06	0.03			0.00	
G) Promicromonospora		0.64	0.24	0.04					
F) Microbacteriaceae	0.07	0.42	0.53	0.17	0.13	0.02	0.02	0.03	0.01
UC) Microbacteriaceae	0.04	0.20	0.19	0.08	0.10	0.01	0.00	0.03	
G) Agromyces	0.01	0.17	0.25	0.04	0.03		0.01		0.01
F) Intrasporangiaceae	0.11	0.29	0.33	0.35	0.10		0.02	0.04	
UC) Intrasporangiaceae	0.03	0.02	0.16	0.12	0.03		0.00	0.01	
G) Janibacter	0.09	0.27	0.11	0.17			0.01		
F) Micrococcaceae	0.03	0.49	0.60	2.01	0.13	0.04	2.45	1.64	0.01
UC) Micrococcaceae							0.29	0.17	
G) Renibacterium				0.03		0.02	0.47	0.31	
G) Arthrobacter	0.03	0.49	0.60	1.98	0.13	0.02	**1.66**	**1.16**	0.01
SO) Frankineae	0.17	0.15	0.21	0.76	1.95	0.01	0.05	0.12	0.03
F) Kineosporiaceae	0.04	0.07	0.01	0.31	0.13		0.02	0.01	
G) Kineosporia	0.04	0.07	0.01	0.25	0.13		0.01	0.00	
F) Nakamurellaceae	0.03		0.06	0.17					
G) Nakamurella	0.03		0.06	0.17					
F) Geodermatophilaceae	0.06	0.07	0.11	0.20	1.82		0.01	0.08	0.02
G) Blastococcus	0.03	0.07	0.06	0.17	1.79			0.05	0.02
SO) Pseudonocardineae	0.08	4.55	4.46	1.87	0.10	0.03	0.09	0.06	0.00
F) Actinosynnemataceae	0.02	0.17	0.05	0.56	0.06	0.01	0.00	0.01	0.00
G) Actinosynnema				0.18	0.03				
G) Lentzea		0.17	0.04	0.35	0.03				
F) Pseudonocardiaceae	0.06	4.30	4.38	1.28	0.03		0.07	0.05	
UC) Pseudonocardiaceae	0.01	0.05	0.37	0.07				0.01	
G) Kutzneria	0.01	0.02	0.09	0.10					
G) Saccharopolyspora			0.04	0.50			0.02		
G) Pseudonocardia	0.04	**4.23**	**3.83**	0.57	0.03		0.04	0.01	
SO) Propionibacterineae	0.77	8.58	2.92	6.20	1.56	0.05	0.39	0.47	0.00
F) Nocardioidaceae	0.76	8.54	2.92	6.20	1.56	0.05	0.38	0.45	0.00

Appendix A cont'd

SO) Propionibacterineae	0.77	8.58	2.92	6.20	1.56	0.05	0.39	0.47	0.00
F) Nocardioidaceae	0.76	8.54	2.92	6.20	1.56	0.05	0.38	0.45	0.00
UC) Nocardioidaceae	0.15	0.22	0.11	0.77		0.03	0.10	0.12	
G) Nocardioides	0.55	6.53	1.95	3.30	1.31	0.02	0.24	0.27	0.00
G) Kribbella	0.01	1.39	0.65	0.89	0.10		0.01	0.04	
G) Aeromicrobium	0.04	0.34	0.17	1.14	0.16			0.00	
SO) Micromonosporineae	0.04	1.17	0.48	0.83	0.03		0.20	0.14	
F) Micromonosporaceae	0.04	1.17	0.48	0.83	0.03		0.20	0.14	
UC) Micromonosporaceae	0.01	0.32	0.06	0.34			0.13	0.10	
G) Micromonospora		0.34	0.15	0.32			0.00	0.01	
G) Actinoplanes	0.02	0.32	0.12	0.08				0.02	
SO) Streptomycineae	0.06	0.78	0.91	6.08	0.35	0.02	0.18	0.19	
F) Streptomycetaceae	0.06	0.78	0.91	6.08	0.35	0.02	0.18	0.19	
G) Streptomyces	0.03	0.68	0.83	5.95	0.29	0.02	0.11	0.16	
SO) Glycomycineae		0.32	0.18						
F) Glycomycetaceae		0.32	0.18						
G) Stackebrandtia		0.05	0.17						
G) Glycomyces		0.27	0.01						
SO) Corynebacterineae	0.33	0.68	0.35	0.96	0.26	0.01	0.36	0.26	0.02
F) Nocardiaceae	0.04	0.15	0.11	0.45	0.13		0.02	0.00	0.00
G) Rhodococcus	0.02	**0.10**	**0.10**	0.42	0.06		0.00		0.00
F) Mycobacteriaceae	0.25	0.39	0.21	0.49	0.06	0.01	0.30	0.21	
G) Mycobacterium	0.25	0.39	0.21	0.49	0.06	0.01	0.30	0.21	
SC) Rubrobacteridae	2.97	4.72	4.03	20.5	0.70	0.47	1.88	2.70	0.03
O) Rubrobacterales	2.97	4.72	4.03	20.5	0.70	0.47	1.88	2.70	0.03
SO) Rubrobacterineae	2.97	4.72	4.03	20.5	0.70	0.47	1.88	2.70	0.03
UC) Rubrobacterineae	0.37	0.44	0.30	0.57	0.06	0.09	0.44	0.58	
F) Rubrobacteraceae	2.60	4.21	3.67	19.9	0.64	0.38	1.38	2.10	0.03
UC) Rubrobacteraceae	1.32	1.66	1.65	7.11	0.13	0.14	0.86	1.24	
G) Solirubrobacter	0.57	1.12	0.85	4.68	0.26	0.02	0.16	0.23	0.00
G) Conexibacter	0.61	1.08	0.95	5.14	0.06	0.17	0.20	0.38	0.02
G) Thermoleophilum	0.08	0.34	0.23	2.97		0.03	0.09	0.24	
G) Rubrobacter	0.02			0.08	0.19	0.01	0.07	0.00	
SC) Acidimicrobidae	0.03	0.05	0.12	0.23		0.01	0.01	0.01	
O) Acidimicrobiales	0.03	0.05	0.12	0.23		0.01	0.01	0.01	
SO) Acidimicrobineae	0.03	0.05	0.12	0.23		0.01	0.01	0.01	
F) Acidimicrobiaceae	0.03	0.05	0.12	0.23		0.01	0.01	0.01	
G) Acidimicrobium	0.03	0.05	0.12	0.23		0.01	0.01	0.01	
P) Bacteroidetes	4.99	0.95	0.96	0.44	0.29	**9.17**	0.29	0.44	6.53
UC) Bacteroidetes	0.18	0.07	0.03			2.06	0.01	0.03	0.25
C) Flavobacteria	2.40	0.05	0.07	0.01		5.42	0.00	0.00	1.05
O) Flavobacteriales	2.40	0.05	0.07	0.01		5.42	0.00	0.00	1.05
UC) Flavobacteriales	0.03	0.02	0.03			2.52			0.83
F) Flavobacteriaceae	2.25	0.02	0.01	0.01		0.58	0.00		0.13
G) Flavobacterium	2.25	0.02	0.01	0.01		0.28			0.10
G) Lutibacter						0.29			0.02
F) Cryomorphaceae	0.13		0.03			2.32		0.00	0.10
UC) Cryomorphaceae	0.04		0.02			1.31			0.01
G) Brumimicrobium						0.61			0.01
G) Crocinitomix						0.27		0.00	0.07
G) Fluviicola	0.09		0.01			0.13			0.01
C) Sphingobacteria	2.38	0.83	0.85	0.43	0.29	1.22	0.26	0.40	5.23

Appendix A cont'd

O) Sphingobacteriales	2.38	0.83	0.85	0.43	0.29	1.22	0.26	0.40	5.23
UC) Sphingobacteriales	0.18		0.05	0.01		0.20	0.03	0.03	0.15
F) Crenotrichaceae	1.52	0.49	0.59	0.30	0.10	0.29	0.17	0.31	4.79
UC) Crenotrichaceae	0.31	0.10	0.04	0.01		0.04	0.04	0.02	0.05
G) Terrimonas	1.01	0.32	0.43	0.23	0.03	0.20	0.12	0.24	0.05
G) Chitinophaga	0.19	0.07	0.11	0.07	0.06	0.01	0.01	0.04	**4.69**
F) Sphingobacteriaceae		0.02			0.19	0.01			
G) Pedobacter		0.02			0.19	0.01			
F) Saprospiraceae	0.11	0.02		0.06		0.25			0.01
G) Lewinella	0.05			0.04		0.19			0.01
F) Flexibacteraceae	0.58	0.29	0.21	0.07		0.47	0.06	0.07	0.28
UC) Flexibacteraceae	0.09	0.02	0.03	0.01		0.24	0.01		0.01
G) Niastella	0.46	0.22	0.18	0.06		0.16	0.05	0.06	0.26
C) Bacteroidetes	0.03					0.47	0.01		0.01
O) Bacteroidales	0.03					0.47	0.01		0.01
UC) Bacteroidales	0.03					0.37			
F) Porphyromonadaceae						0.10			0.00
G) Paludibacter						0.10			0.00
P) Nitrospira	0.07	0.15		0.06		0.32	0.07	0.10	
C) Nitrospira	0.07	0.15		0.06		0.32	0.07	0.10	
O) Nitrospirales	0.07	0.15		0.06		0.32	0.07	0.10	
F) Nitrospiraceae	0.07	0.15		0.06		0.32	0.07	0.10	
UC) Nitrospiraceae						0.13			
G) Nitrospira	0.07	0.15		0.06		0.01	0.07	0.10	
G) Magnetobacterium						0.18			
P) Acidobacteria	14.8	17.5	12.5	10.1	5.36	8.06	26.7	26.5	1.59
C) Acidobacteria	14.8	17.5	12.5	10.1	5.36	8.06	26.7	26.5	1.59
O) Acidobacteriales	14.8	17.5	12.5	10.1	5.36	8.06	26.7	26.5	1.59
F) Acidobacteriaceae	14.8	17.5	12.5	10.1	5.36	8.06	26.7	26.5	1.59
UC) Acidobacteriaceae	0.32	0.05	0.04	0.06	0.06	0.06	0.39	0.36	0.02
G) Gp4	2.66	**2.76**	**2.33**	3.38	**1.24**	0.85	**7.46**	**7.09**	0.09
G) Gp22	0.59		0.06	0.02		0.02	0.07	0.05	0.08
G) Gp16	0.46	2.13	1.51	2.18	0.22	0.58	0.61	0.64	0.13
G) Gp10	0.41	0.05	0.02	0.05		0.01	0.08	0.02	0.05
G) Gp5	0.32	0.42	0.24	0.12	0.03	0.08	1.35	0.87	
G) Gp18	0.02		0.01	0.03		0.59	0.08	0.06	0.18
G) Gp6	7.88	**9.98**	**6.57**	3.15	**2.42**	4.48	**11.4**	**11.2**	0.81
G) Gp23						0.65	0.01	0.01	0.08
G) Gp11	0.34	0.12	0.12	0.04		0.01	0.08	0.06	
G) Gp3	0.43	0.24	0.31	0.34	0.64	0.17	0.56	0.79	0.02
G) Gp1	0.02	0.15	0.04	0.16	0.54	0.05	1.69	2.44	
G) Gp2	0.04	0.02		0.01		0.03	0.11	0.10	
G) Gp25	0.12		0.03	0.07		0.07	1.35	1.48	
G) Gp17	0.92	0.93	0.81	0.20	0.10	0.18	0.26	0.21	0.03
G) Gp7	0.22	0.61	0.45	0.35	0.10	0.14	1.12	0.94	0.08
P) Proteobacteria	24.8 4	29.1 0	41.8 4	20.7 8	78.5 4	35.8 4	22.4 6	25.8 8	83.9 8
UC) Proteobacteria	5.41	3.77	5.61	1.88	0.32	5.76	2.35	2.21	0.73
C) Epsilonproteobacteria	0.03			0.18		0.25	0.07	0.00	
O) Campylobacterales	0.03			0.18		0.25	0.07	0.00	
F) Campylobacteraceae	0.03			0.18		0.24			
G) Campylobacter	0.03			0.18		0.23			

Appendix A cont'd

C) Deltaproteobacteria	3.51	2.64	3.85	1.00	0.29	4.93	1.81	1.88	1.93
UC) Deltaproteobacteria	1.90	1.12	1.36	0.45		1.82	0.95	0.98	0.48
O) Syntrophobacterales	0.22	0.34	0.09	0.01		0.98	0.06	0.07	0.32
F) Syntrophaceae	0.07		0.02			0.49	0.02	0.03	0.17
G) Smithella						0.37	0.01	0.03	0.13
F) Syntrophobacteraceae	0.15	0.34	0.06	0.01		0.48	0.04	0.03	0.13
UC) Syntrophobacteraceae	0.11	0.29	0.05	0.01		0.22	0.02	0.02	0.07
G) Syntrophobacter						0.18			0.01
O) Desulfuromonales	0.19		0.02	0.17		0.28	0.11	0.11	0.03
F) Geobacteraceae	0.13		0.02	0.17		0.16	0.09	0.09	0.03
G) Geobacter	0.11		0.02	0.17		0.13	0.01		0.03
O) Desulfobacterales	0.02			0.01		1.21	0.08	0.03	0.71
F) Desulfobacteraceae						0.88	0.07	0.02	0.62
UC) Desulfobacteraceae						0.37	0.01	0.00	0.12
G) Desulfobacterium						0.39	0.00	0.01	0.14
G) Desulfonema						0.11	0.05	0.00	0.35
F) Desulfobulbaceae	0.02			0.01		0.32	0.01	0.00	0.09
G) Desulfobulbus	0.02					0.14		0.00	0.07
G) Desulfocapsa						0.11			0.01
O) Desulfovibrionales						0.19	0.03		0.00
F) Desulfovibrionaceae						0.17	0.03		0.00
G) Desulfovibrio						0.17	0.03		0.00
O) Myxococcales	0.96	1.05	2.32	0.33	0.22	0.43	0.54	0.67	0.23
UC) Myxococcales	0.55	0.42	1.15	0.19	0.03	0.18	0.16	0.19	0.00
SO) Cystobacterineae	0.05	0.27	0.31	0.03	0.10	0.16	0.24	0.28	0.23
UC) Cystobacterineae		0.10	0.15			0.02	0.09	0.06	
F) Cystobacteraceae	0.04	0.07	0.10	0.01		0.11	0.13	0.14	0.04
G) Anaeromyxobacter	0.03	0.02	0.07	0.01		0.10	0.04	0.05	0.03
F) Myxococcaceae	0.01	0.10	0.06	0.02	0.10	0.02	0.02	0.08	0.19
SO) Nannocystineae	0.04	0.20	0.36				0.02	0.03	
F) Nannocystaceae	0.03	0.10	0.17					0.02	
UC) Nannocystaceae			0.12						
F) Haliangiaceae	0.01	0.05	0.19				0.01	0.00	
G) Haliangium	0.01	0.05	0.19				0.01	0.00	
SO) Sorangineae	0.32	0.17	0.49	0.11	0.10	0.09	0.12	0.17	
F) Polyangiaceae	0.32	0.17	0.49	0.11	0.10	0.09	0.12	0.17	
UC) Polyangiaceae	0.17	0.02	0.26	0.08	0.10	0.02	0.03	0.09	
G) Byssovorax	0.08	0.12	0.10	0.03		0.04	0.07	0.04	
O) Bdellovibrionales	0.21	0.12	0.06	0.03	0.06	0.04	0.04	0.02	0.14
F) Bdellovibrionaceae	0.16	0.07	0.02	0.03		0.04	0.04	0.02	0.05
G) Bdellovibrio	0.16	0.07	0.02	0.03		0.04	0.04	0.02	0.05
C) Alphaproteobacteria	11.0	16.7	20.6	14.3	53.4	3.43	10.3	11.7	4.90
UC) Alphaproteobacteria	0.42	0.66	0.39	0.26	1.21	0.21	0.29	0.52	0.04
O) Caulobacterales	0.87	1.20	0.74	0.21	18.4	0.18	0.33	0.35	2.49
F) Caulobacteraceae	0.87	1.20	0.74	0.21	18.4	0.18	0.33	0.35	2.49
G) Caulobacter	0.18	0.29	0.06	0.01	1.79	0.03	0.00	0.00	0.45
G) Phenylobacterium	0.55	0.71	0.61	0.18	**16.65**	0.15	0.29	0.26	0.19
G) Brevundimonas	0.11	0.20	0.06		0.03			0.05	1.85
O) Sphingomonadales	0.74	2.64	5.39	0.92	13.8	0.64	1.80	2.42	1.02
F) Sphingomonadaceae	0.74	2.64	5.39	0.92	13.8	0.64	1.80	2.42	1.02
UC) Sphingomonadaceae	0.25	0.61	0.96	0.12	11.3	0.15	0.45	0.82	0.05

Appendix A cont'd

G) Novosphingobium	0.08	0.12	0.13	0.08		0.28	0.08	0.04	0.02
G) Sphingosinicella	0.09	0.32	0.71	0.09	0.19	0.10	0.97	1.07	0.00
G) Sphingomonas	0.24	**1.47**	**3.36**	0.54	**2.23**	0.04	0.22	0.43	0.91
O) Rhodobacterales	0.31	0.24	0.39	0.55	0.16	0.49	0.01	0.01	0.03
F) Rhodobacteraceae	0.31	0.24	0.39	0.55	0.16	0.49	0.01	0.01	0.03
UC) Rhodobacteraceae	0.18	0.02	0.09	0.33	0.06	0.14	0.01	0.01	0.00
G) Amaricoccus	0.11	0.22	0.27	0.15	0.10				
G) Rhodobacter	0.01		0.02			0.16			0.02
O) Rhodospirillales	1.07	1.98	2.11	2.24	12.5	0.50	1.28	1.43	0.93
UC) Rhodospirillales	0.17	0.29	0.42	0.20	0.67	0.07	0.26	0.28	0.02
F) Acetobacteraceae	0.71	1.35	1.45	1.44	5.42	0.40	0.81	0.91	0.33
UC) Acetobacteraceae	0.49	0.95	1.13	1.18	4.11	0.24	0.68	0.80	0.09
G) Belnapia		0.05	0.01	0.04	1.08				
G) Roseomonas	0.06	0.02		0.01		0.13	0.01	0.02	0.24
G) Stella	0.12	0.15	0.23	0.08	0.19		0.07	0.06	
G) Rhodopila	0.03	0.17	0.01	0.10	0.03		0.01	0.00	
F) Rhodospirillaceae	0.19	0.34	0.24	0.60	6.47	0.02	0.21	0.25	0.58
UC) Rhodospirillaceae	0.11	0.05	0.15	0.22	0.16	0.02	0.18	0.21	0.00
G) Skermanella	0.08	0.24	0.05	0.36	0.51		0.01	0.03	
G) Inquilinus	0.01	0.02	0.04	0.03	0.89		0.01	0.00	
G) Azospirillum		0.02			4.91		0.01	0.00	0.57
O) Rhizobiales	7.67	9.98	11.6	10.1	7.17	1.41	6.59	7.02	0.39
UC) Rhizobiales	1.04	0.98	1.90	1.23	0.73	0.19	1.18	1.37	0.02
F) Phyllobacteriaceae	0.26	0.44	0.84	0.26	0.45	0.06	0.07	0.08	
G) Mesorhizobium	0.24	0.22	0.52	0.12	0.32	0.02	0.07	0.05	
G) Phyllobacterium		0.20	0.24	0.07	0.13	0.01		0.02	
F) Rhizobiaceae	0.18	0.34	0.17	0.24	2.01	0.02	0.18	0.11	0.09
G) Ensifer	0.04	0.02	0.03	0.03	0.89		0.00		
G) Rhizobium	0.12	0.32	0.11	0.20	1.08	0.01	0.16	0.08	0.09
F) Bradyrhizobiaceae	4.34	5.33	4.48	3.44	0.96	0.29	1.41	1.43	0.20
UC) Bradyrhizobiaceae	0.91	1.20	1.65	1.08		0.13	0.59	0.81	0.00
G) Bosea	0.18	0.42	0.18	0.12	0.64	0.01	0.01	0.05	0.18
G) Afipia	0.10	0.17	0.17	0.03		0.02	0.03		0.02
G) Rhodopseudomonas	0.43	0.49	0.30	0.04	0.03	0.01			
G) Nitrobacter	0.14	0.20	0.25	0.03		0.02	0.05	0.05	
G) Bradyrhizobium	**2.56**	**2.81**	**1.94**	2.14	0.29	0.10	0.68	0.51	
F) Hyphomicrobiaceae	1.06	2.08	3.57	2.83	0.26	0.63	2.49	3.03	0.02
UC) Hyphomicrobiaceae	0.35	0.64	1.12	0.63	0.06	0.24	1.26	1.40	0.00
G) Rhodoplanes	0.11		0.13	0.08		0.10	0.50	0.72	0.01
G) Pedomicrobium	0.10	0.17	0.80	0.65	0.10	0.01	0.07	0.14	
G) Hyphomicrobium	0.25	0.71	0.83	0.94		0.13	0.05	0.05	0.00
G) Devosia	0.04	0.49	0.36	0.08	0.10	0.02	0.05	0.12	
G) Blastochloris	0.13	0.07	0.18	0.16		0.12	0.50	0.52	
G) Labrys			0.02	0.18			0.04	0.02	
F) Beijerinckiaceae	0.14	0.10	0.10	0.25	0.06	0.06	0.10	0.12	0.03
G) Chelatococcus	0.05	0.02	0.02	0.16	0.03	0.01	0.08	0.04	0.00
F) Methylocystaceae			0.15	0.02	0.03	0.02			
G) Methylopila			0.15	0.02	0.03				
F) Methylobacteriaceae	0.61	0.66	0.41	1.79	2.61	0.06	1.09	0.85	0.02
UC) Methylobacteriaceae	0.10	0.02	0.06	0.41	0.61	0.02	0.25	0.20	0.00
G) Methylobacterium	0.22	0.32	0.17	0.28	0.89	0.02	0.40	0.39	0.00
G) Microvirga	0.29	0.32	0.17	1.10	1.12	0.01	0.44	0.26	0.01

Appendix A cont'd

C) Gammaproteobacteria	3.11	2.96	6.75	1.74	11.2	11.3	3.99	4.62	50.5
UC) Gammaproteobacteria	1.56	1.25	3.24	0.70	0.19	8.12	1.77	1.59	0.70
O) Alteromonadales	0.01		0.16		0.19	0.02	0.00	0.02	
O) Pseudomonadales	0.41	0.39	0.72	0.28	0.99	0.06	0.28	0.34	48.5
F) Moraxellaceae	0.02	0.22	0.24	0.02	0.26	0.02	0.05	0.02	0.56
G) Acinetobacter		0.12	0.15	0.02	0.26	0.01	0.04	0.02	0.55
F) Pseudomonadaceae	0.39	0.17	0.48	0.26	0.73	0.04	0.23	0.32	48.0
G) Pseudomonas	0.17	0.17	0.48	0.25	0.73	0.04	0.19	0.29	**47.9**
G) Cellvibrio	0.22			0.01			0.00	0.02	
O) Enterobacteriales	0.04	0.27	1.35	0.17	3.06	0.16	0.88	1.16	0.11
F) Enterobacteriaceae	0.04	0.27	1.35	0.17	3.06	0.16	0.88	1.16	0.11
UC) Enterobacteriaceae	0.01		0.02		0.10	0.01	0.19	0.21	
G) Klebsiella	0.01			0.02	0.06		0.09	0.13	
G) Shigella	0.01	**0.27**	**1.31**	0.14	**2.81**	0.13	**0.55**	**0.77**	0.11
O) Chromatiales	0.18	0.22	0.40	0.20	0.03	1.17	0.07	0.15	0.01
UC) Chromatiales	0.05	0.02	0.06	0.05		0.19	0.03	0.05	0.01
F) Ectothiorhodospiraceae	0.11	0.20	0.32	0.15	0.03	0.31	0.04	0.05	
UC) Ectothiorhodospiraceae	0.10	0.20	0.32	0.15	0.03	0.28	0.04	0.03	
F) Chromatiaceae	0.02		0.02			0.67		0.04	0.00
UC) Chromatiaceae	0.02		0.01			0.47		0.03	0.00
G) Marichromatium						0.17			
O) Methylococcales						0.86	0.00		
F) Methylococcaceae						0.86	0.00		
UC) Methylococcaceae						0.13	0.00		
G) Methylobacter						0.65			
O) Xanthomonadales	0.52	0.37	0.63	0.33	6.60	0.65	0.42	1.00	1.20
F) Xanthomonadaceae	0.52	0.37	0.63	0.33	6.60	0.65	0.42	1.00	1.20
UC) Xanthomonadaceae	0.22	0.02	0.11	0.10		0.34	0.12	0.19	0.06
G) Luteimonas	0.20		0.05	0.03	0.06	0.13	0.03	0.02	0.02
G) Stenotrophomonas		0.05	0.07	0.03	0.03		0.15	0.57	0.87
G) Lysobacter	0.07	0.24	0.28	0.11	**3.86**	0.13	0.05	0.06	0.23
G) Pseudoxanthomonas	0.01		0.01	0.01	**2.65**		0.01		0.01
O) Legionellales	0.34	0.46	0.21	0.06	0.03	0.29	0.50	0.35	
F) Legionellaceae	0.09	0.17	0.10	0.01		0.01	0.01	0.03	
F) Coxiellaceae	0.23	0.29	0.10	0.02	0.03	0.28	0.45	0.32	
G) Rickettsiella	0.11		0.02			0.20	0.01		
G) Aquicella	0.09	0.29	0.04	0.01	0.03	0.04	0.39	0.27	
O) Oceanospirillales	0.02		0.03	0.01	0.19	0.02	0.01	0.01	0.00
F) Halomonadaceae			0.03		0.16			0.01	
G) Halomonas			0.03		0.16			0.01	
C) Betaproteobacteria	1.70	3.03	4.97	1.67	13.2	10.1	3.94	5.41	25.8
UC) Betaproteobacteria	0.40	0.10	0.22	0.39	0.03	3.17	0.95	1.51	0.14
O) Neisseriales	0.01	0.02	0.03		0.03	0.24	0.14	0.19	0.00
F) Neisseriaceae	0.01	0.02	0.03		0.03	0.24	0.14	0.19	0.00
UC) Neisseriaceae		0.02			0.03	0.14	0.09	0.13	0.00
G) Formivibrio	0.01		0.03			0.10	0.05	0.06	
O) Nitrosomonadales	0.05		0.12	0.07		0.17	0.00	0.01	0.19
F) Nitrosomonadaceae	0.05		0.12	0.07		0.15	0.00	0.01	0.19
G) Nitrosomonas				0.01		0.06		0.00	0.18
O) Methylophilales	0.01	0.02				0.24			0.02
F) Methylophilaceae	0.01	0.02				0.24			0.02

Appendix A cont'd

G) Methylophilus	0.01	0.02				0.24			0.02
O) Rhodocyclales	0.23	0.05	0.74	0.11	0.03	1.17	0.12	0.17	2.61
F) Rhodocyclaceae	0.23	0.05	0.74	0.11	0.03	1.17	0.12	0.17	2.61
UC) Rhodocyclaceae	0.11		0.62	0.07	0.03	0.79	0.06	0.08	0.03
G) Dechloromonas	0.03					0.17	0.02	0.03	
G) Azoarcus	0.06	0.05	0.04	0.04		0.05	0.02	0.00	2.57
O) Hydrogenophilales						0.55	0.03	0.10	0.17
F) Hydrogenophilaceae						0.55	0.03	0.10	0.17
G) Thiobacillus						0.55	0.03	0.10	0.17
O) Burkholderiales	1.00	2.84	3.84	1.10	13.1	4.59	2.70	3.43	22.7
UC) Burkholderiales	0.25	0.15	0.20	0.32		2.28	0.53	0.66	0.28
F) Oxalobacteraceae	0.11	0.83	0.25	0.13	2.65	0.08	0.53	0.50	1.27
UC) Oxalobacteraceae	0.02	0.07	0.06	0.05	0.19	0.03	0.15	0.16	0.87
G) Herbaspirillum			0.01	0.01	0.51	0.01	0.02	0.04	0.25
G) Duganella	0.03	0.15		0.02		0.01	0.12	0.07	
G) Massilia	0.04	0.15	0.02	0.02	0.35		0.14	0.17	0.03
G) Herminiimonas		0.29	0.11		0.03	0.01	0.05	0.02	0.01
G) Janthinobacterium		0.15	0.03	0.02	0.54	0.02	0.02	0.03	0.06
G) Naxibacter	0.02	0.02			1.02				0.05
F) Comamonadaceae	0.47	1.17	2.39	0.28	1.18	1.43	0.26	1.37	9.05
UC) Comamonadaceae	0.25	0.17	0.59	0.08	0.22	0.24	0.12	0.45	0.67
G) Comamonas		0.02	0.01	0.01		0.17	0.01	0.03	
G) Hydrogenophaga			0.02		0.03	0.02	0.00	0.49	0.22
G) Polaromonas	0.06	0.37	0.22	0.03			0.00	0.02	0.03
G) Acidovorax	0.13	0.05	0.06	0.03		0.81	0.07	0.22	**7.39**
G) Variovorax	0.01	0.49	1.36	0.08	0.03		0.01	0.04	
G) Rhodoferax	0.01					0.11		0.02	
G) Ottowia	0.01		0.11	0.04			0.00	0.02	
G) Ramlibacter		0.07	0.02	0.03	0.86	0.06	0.02	0.08	0.74
F) Burkholderiaceae		0.54	0.69	0.04	4.59	0.15	1.21	0.68	0.14
G) Cupriavidus			0.05	0.04	0.83	0.01	0.02	0.02	0.13
G) Wautersia		0.02			**2.71**		0.04	0.01	0.01
G) Burkholderia		0.12			0.03	0.02	**1.04**	**0.45**	
G) Ralstonia		0.37	0.58		0.83	0.01	0.09	0.14	
F) Incertae sedis 5	0.17	0.07	0.28	0.31	3.48	0.46	0.15	0.19	0.05
UC) Incertae sedis 5	0.11	0.05	0.22	0.17	2.65	0.23	0.07	0.11	0.02
G) Azohydromonas	0.01	0.02	0.02	0.11	0.61	0.03	0.05	0.02	
G) Aquabacterium	0.04		0.02		0.22	0.19			0.01
F) Alcaligenaceae		0.07	0.03	0.02	1.21	0.18	0.02	0.03	11.9
G) Tetrathiobacter		0.02	0.02		0.45				
G) Bordetella		0.02		0.02		0.17			0.07
G) Achromobacter		0.02			0.67		0.02	0.02	11.8
P) Chloroflexi	0.32	0.39	0.35	0.54	0.22	5.09	0.26	0.25	0.17
C) Anaerolineae	0.28	0.27	0.27	0.27		5.04	0.11	0.11	0.17
O) Anaerolinaeles						0.67			
F) Anaerolinaeceea						0.67			
G) Anaerolinea						0.67			
SC) Caldilineae	0.28	0.27	0.27	0.27		4.30	0.11	0.10	0.14
O) Caldilineales	0.28	0.27	0.27	0.27		4.30	0.11	0.10	0.14
F) Caldilineacea	0.28	0.27	0.27	0.27		4.30	0.11	0.10	0.14
UC) Caldilineacea	0.12	0.05	0.12	0.10		1.07	0.04	0.05	0.09
G) Levilinea	0.10	0.15	0.14	0.08		1.10	0.04	0.01	0.02

Appendix A cont'd

G) Leptolinea	0.01	0.07		0.01		1.61	0.00		0.01
G) Caldilinea	0.05		0.01	0.08		0.51	0.03	0.04	0.01
C) Chloroflexi	0.02	0.12	0.06	0.22	0.22	0.01	0.10	0.10	
O) Chloroflexales	0.02	0.12	0.06	0.22	0.22	0.01	0.08	0.10	
UC) Chloroflexales	0.01	0.12	0.04	0.14		0.01	0.03	0.03	
F) Oscillochloridaceae	0.01		0.02	0.03	0.19			0.00	
G) Oscillochloris	0.01		0.02	0.03	0.19			0.00	
P) TM7	0.18	0.20	0.41	0.23	2.01	0.02	0.39	0.59	0.00
G) TM7_genera_IS	0.18	0.20	0.41	0.23	2.01	0.02	0.39	0.59	0.00
P) Spirochaetes	0.08			0.05		0.31	0.01		0.01
C) Spirochaetes	0.08			0.05		0.31	0.01		0.01
O) Spirochaetales	0.08			0.05		0.31	0.01		0.01
F) Spirochaetaceae	0.04			0.05		0.28	0.01		0.01
UC) Spirochaetaceae	0.02			0.04		0.19			0.00
P) WS3	0.81	0.05	0.02	0.11		0.49	0.32	0.17	0.14
G) WS3_genera_IS	0.81	0.05	0.02	0.11		0.49	0.32	0.17	0.14
P) OD1	0.46	0.32	0.32	0.18		0.05		0.00	
G) OD1_genera_IS	0.46	0.32	0.32	0.18		0.05		0.00	
P) OP 10	0.11		0.03	0.02		0.27	0.01	0.03	0.02
G) OP10_genera_IS	0.11		0.03	0.02		0.27	0.01	0.03	0.02
P) Verrucomicrobia	16.9	3.74	4.47	6.17	1.18	4.95	3.35	3.79	2.53
C) Verrucomicrobiae	16.9	3.74	4.47	6.17	1.18	4.95	3.35	3.79	2.53
O) Verrucomicrobiales	16.9	3.74	4.47	6.17	1.18	4.95	3.35	3.79	2.53
UC) Verrucomicrobiales	1.00	0.20	0.25	0.17	0.10	0.06	0.05	0.02	0.06
F) Sub 3	5.32	1.20	2.42	1.98	0.38	2.57	0.40	0.41	1.40
G) Sub 3_genera_IS	5.32	1.20	2.42	1.98	0.38	2.57	0.40	0.41	1.40
F) Xiphinematobacteriaceae	7.09	1.52	1.22	3.81	0.67	0.59	2.84	3.20	0.24
UC) Xiphinematobacteriaceae	0.22	0.02	0.06	0.03			0.00	0.01	
G) Xiphinematobacteriaceae	6.82	1.49	1.16	3.78	0.67	0.59	2.83	3.20	0.24
F) Sub 5	0.01					0.21		0.02	
G) Sub 5_genera_IS	0.01					0.21		0.02	
F) Opitutaceae	1.52	0.64	0.58	0.17		0.09	0.03	0.09	0.38
G) Opitutus	1.52	0.64	0.58	0.17		0.09	0.03	0.09	0.38
F) Verrucomicrobiaceae	1.98	0.20	0.02	0.05	0.03	1.43	0.04	0.05	0.44
UC) Verrucomicrobiaceae	0.27	0.02	0.01			0.09		0.01	0.01
G) Verrucomicrobiaceae	1.11	0.07		0.05	0.03	1.22	0.02	0.02	0.33
G) Prosthecobacter	0.12	0.07	0.01			0.01	0.00	0.01	
G) Verrucomicrobium	0.48	0.02				0.09	0.01		0.10
P) BRC1	0.02		0.02			0.16	0.04	0.02	0.01
G) BRC1_genera_IS	0.02		0.02			0.16	0.04	0.02	0.01
P) Cyanobacteria	0.36		0.06	0.03	0.06	0.02	0.10	0.11	0.00
C) Cyanobacteria	0.36		0.06	0.03	0.06	0.02	0.10	0.11	0.00
F) Chloroplast	0.36		0.06	0.02	0.06		0.04	0.08	0.00
G) Streptophyta	0.36		0.05	0.01	0.06		0.00	0.05	
P) Firmicutes	1.26	1.05	0.75	1.63	1.02	6.33	19.1	13.0	2.16
UC) Firmicutes	0.47	0.34	0.46	0.20	0.16	1.17	0.92	0.94	0.50
C) Bacilli	0.15	0.17	0.07	1.03	0.77	**0.95**	**17.6**	**11.2**	**1.37**
UC) "Bacilli"			0.01	0.01		0.02	0.18	0.17	0.00
O) Lactobacillales		0.05	0.01		0.13	0.07	0.01	0.02	

Appendix A cont'd

O) Bacillales	0.15	0.12	0.06	1.02	0.64	0.86	17.4	11.0	1.36
UC) Bacillales	0.03			0.07		0.07	1.27	0.98	0.13
F) Bacillaceae	0.11	0.05	0.01	0.80	0.03	0.69	14.1	8.70	0.37
UC) Bacillaceae				0.02		0.02	0.52	0.41	
SF) Bacillaceae 1	0.11	0.05	0.01	0.77	0.03	0.67	13.5	8.23	0.37
UC) "Bacillaceae 1"	0.02	0.02		0.19		0.24	4.06	2.90	
superG) Bacillus	0.10	0.02	0.01	0.58	0.03	0.35	**9.31**	**5.22**	0.37
UC) Bacillus	0.04	0.02		0.20	0.03	0.11	2.62	1.64	0.36
G) Bacillus d				0.14		0.06	**3.50**	**1.73**	
G) Bacillus h						0.01	0.47	0.19	
G) Bacillus c	0.04		0.01	0.22		0.15	**2.44**	**1.42**	0.00
G) Bacillus k						0.02	0.19	0.21	
G) Anoxybacillus						0.07	0.21	0.12	
F) Listeriaceae						0.02	0.12	0.21	0.00
SF) Paenibacillaceae 2						0.02	0.12	0.21	0.00
G) Oxalophagus						0.02	0.08	0.19	
F) Paenibacillaceae		0.02	0.03	0.08	0.51	0.02	0.49	0.33	0.86
SF) Paenibacillaceae 1		0.02	0.03	0.08	0.51	0.02	0.49	0.33	0.86
G) Brevibacillus						0.01	0.11	0.03	
G) Paenibacillus		0.02	0.02	0.07	0.29		0.35	0.25	0.84
G) Cohnella			0.01	0.01	0.16	0.02	0.03	0.02	
F) Planococcaceae	0.01			0.07		0.06	1.35	0.76	
UC) Planococcaceae	0.01			0.02		0.03	0.89	0.47	
G) Sporosarcina						0.03	0.28	0.12	
G) Pasteuriaceae Incertae Sedis				0.03			0.11	0.12	
C) Clostridia	0.64	0.54	0.21	0.41	0.10	4.22	0.63	0.88	0.29
UC) "Clostridia"	0.22	0.15	0.06	0.05	0.03	0.80	0.17	0.24	0.11
O) Clostridiales	0.41	0.34	0.15	0.36	0.06	3.36	0.44	0.64	0.18
UC) Clostridiales	0.36	0.24	0.04	0.25	0.03	1.88	0.14	0.24	0.06
F) Incertae Sedis XI	0.01	0.02	0.02	0.01		0.07	0.11	0.07	0.03
G) Sedimentibacter	0.01	0.02	0.02	0.01		0.07	0.11	0.07	0.03
F) Ruminococcaceae						0.43	0.01	0.00	0.00
UC) "Ruminococcaceae"						0.11			
G) Acetivibrio						0.13	0.00	0.00	0.00
G) Ruminococcaceae IS						0.19	0.00		
F) Peptococcaceae			0.03	0.03		0.17	0.03	0.05	0.01
F) Clostridiaceae	0.01	0.05		0.02		0.04	0.05	0.18	0.00
SF) Clostridiaceae 1	0.01	0.05		0.02		0.03	0.05	0.18	0.00
G) Clostridium	0.01	0.05		0.02		0.02	0.04	0.17	0.00
F) Incertae Sedis XV	0.02			0.05		0.22			
UC) Incertae Sedis XV	0.02			0.05		0.20			
F) Incertae Sedis XII						0.24	0.00		0.03
G) Fusibacter						0.24	0.00		0.03
P) Gemmatimonadetes	3.06	6.07	5.67	1.42	0.96	0.50	2.17	2.65	0.32
C) Gemmatimonadetes	3.06	6.07	5.67	1.42	0.96	0.50	2.17	2.65	0.32
O) Gemmatimonadales	3.06	6.07	5.67	1.42	0.96	0.50	2.17	2.65	0.32
F) Gemmatimonadaceae	3.06	6.07	5.67	1.42	0.96	0.50	2.17	2.65	0.32
G) Gemmatimonas	3.06	6.07	5.67	1.42	0.96	0.50	2.17	2.65	0.32
P) Chlamydiae	0.22	0.27	0.13	0.17		0.09		0.01	0.01
C) Chlamydiae	0.22	0.27	0.13	0.17		0.09		0.01	0.01
O) Chlamydiales	0.22	0.27	0.13	0.17		0.09		0.01	0.01

Appendix A cont'd

F) Parachlamydiaceae	0.12	0.22	0.06	0.11		0.01			
G) Parachlamydia	0.07	0.15	0.02	0.07		0.01			
P) Planctomycetes	1.76	1.39	0.93	1.57		1.13	0.30	0.30	0.05
C) Planctomycetacia	1.76	1.39	0.93	1.57		1.13	0.30	0.30	0.05
O) Planctomycetales	1.76	1.39	0.93	1.57		1.13	0.30	0.30	0.05
F) Planctomycetaceae	1.76	1.39	0.93	1.57		1.13	0.30	0.30	0.05
UC) Planctomycetaceae	0.81	0.51	0.43	0.64		0.28	0.10	0.12	0.01
G) Gemmata	0.39	0.46	0.19	0.31		0.05	0.15	0.13	
G) Planctomyces	0.28	0.05	0.05	0.20		0.11	0.02	0.01	
G) Blastopirellula	0.06		0.04	0.05		0.20			0.01
G) Pirellula	0.21	0.29	0.14	0.25		0.48	0.00	0.02	0.02
G) Isosphaera	0.01	0.07	0.09	0.12		0.01	0.02	0.02	
Domain Archaea						0.13			
P) Euryarchaeota						0.13			
C) Methanomicrobia						0.13			
O) Methanomicrobiales						0.11			
F) Methanomicrobiaceae						0.11			

Appendix A. Detailed classification of sequences of bacterial assemblages from chapter 4

1. P) Phylum, C) class, SC) subclass, O) order, SO) suborder, F) family, SF) subfamily, G) genus, and U) "unclassified" artificial taxa.
2. Classification is based on RDP classifier result at 50% threshold.
3. Taxons with maximum value of nine samples > 0.1% was shown in this table.
4. "0.00" indicates < 0.05% and > 0.001%.

Appendix B1. Habitat-Lite two level scheme and its terms definition

Top level term	Definition
Aquatic	A habitat that is in or on water
Aquatic: Freshwater	A habitat that is in or on a body of water containing low concentrations of dissolved salts and other total dissolved solids (<0.5 grams dissolved salts per liter)
Aquatic: Marine	A habitat that is in or on a sea or ocean containing high concentrations of dissolved salts and other total dissolved solids (typically >35 grams dissolved salts per liter)
Terrestrial	A habitat that is on or at the boundary of the surface of the Earth
Air	The mixture of gases, roughly (by molar content/volume: 78% nitrogen, 20.95% oxygen, 0.93% argon, 0.038% carbon dioxide, trace amounts of other gases, and a variable amount [average around 1%] of water vapor), that surrounds the planet Earth
Fossil	The mineralized or otherwise preserved remains or traces (such as footprints) of animals, plants, and other organisms
Food	A substance, usually composed primarily of carbohydrates, fats, water and/or proteins, that can be eaten or drunk by an animal or human being for nutrition or pleasure
Organism-Associated	A habitat that is in or on a living thing
Extreme	A habitat having at least one environmental quality that tends towards either the largest or smallest element of the set. The physical or geochemical extreme conditions found in an extreme
Cultured	Cultured habitat is an controlled habitat created by humans through laboratory techniques usually for the purposes of preparing cell, organ, tissue and plant tissue cultures
Other	

Second level terms	Definition
soil	Any material within 2 m from the Earth's surface that is in contact with the atmosphere, with the exclusion of living organisms, areas with continuous ice not covered by other material, and water bodies deeper than 2 m
sediment	Sediment is an environmental substance comprised of any particulate matter that can be transported by fluid flow and which eventually is deposited as a layer of solid particles\non the bedor bottom of a body of water or other liquid
sludge	The residual semi-solid material left from domestic or industrial processes, or wastewater treatment processes
waste water	A habitat that is in or on a body of water containing low concentrations of dissolved salts and other total dissolved solids (<0.5 grams dissolved salts per litre)
hot spring	A spring that is produced by the emergence of geothermally-heated groundwater from the Earth's crust
hydrothermal vent	A fissure in the Earths's surface from which geothermally heated water issues
biofilm	A complex aggregation of microorganisms marked by the excretion of a protective and adhesive matrix; usually adhering to a substratum
microbial mat	

Table 1. **Definition of terms in Habitat-Lite version 0.4 (revised May 20, 2009).** A given habit might be described with one or more appropriate Top-level terms, and second level terms as appropriate (Hirschman *et al.*, 2008).

Appendix B2. Priori groups described by Habitat-Lite

Group	Numbers of samples	Habitat-Lite description
G 01	116	Terrestrial[1], Soil[2]
G 02	6	Extreme[1], Soil[2]
G 03	12	Terrestrial[1], Extreme[1], Soil[2]
G 04	16	Oragnism-Associated[1]
G 05	6	Freshwater[1], Waste water[2]
G 06	7	Freshwater[1], Sediment[2]
G 07	2	Fossil[1], Oragnism-Associated[1]
G 08	10	Marine[1], Sediment[2]
G 09	14	Cultured[1], Soil[2] or Sediment[2]
G 10	20	Extreme[1], Freshwater[1]Sediment[2]
G 11	2	Extreme[1], Microbial mat[2]

[1] Top level terms in Habitat-Lite

[2] Second level terms

Appendix B3. List of samples and their priori groups

Sample ID	Sampling description and location	Habitat_Lite Description	Groups
Cz_0D	PCB-contaminated soil under Austrian pine tree, Czech Republic	Terrestrial^1, Soil^2	G 01
Du_E22_7	Rhizosphere	Terrestrial^1, Soil^2	G 01
Du_E22_8	Rhizosphere	Terrestrial^1, Soil^2	G 01
Gh_BF 1	Bare follow plots (BF), replication1, Kpeve Agricultural Experimental Station (KAES) in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_BF2	BF, rep2, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_BF3	BF, rep3, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_BF4	BF, rep4, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_BFc	BF, composiite, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EbM1	Maize-elephant grass (*Pennisetum* sp) rotation with fallow residue burning plot (EbM), rep1, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EbM2	EbM, rep2, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EbM3	EbM, rep3, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EbM4	EbM, rep4, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EbMc	EbM, composite, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EfM1	Fertilized maize-elephant grass rotation with minimum tillage of fallow residue by hand slashing (EfM), rep1, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EfM2	EfM, rep2, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EfM3	EfM, rep3, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EfM4	EfM, rep4, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_EfMc	EfM, composite, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_Eu 1	Unmanaged elephant grass (**Eu**), rep1, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_Eu2	**Eu**, rep2, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01
Gh_Euc	**Eu**, composite, KAES in Volta Region, Ghana	Terrestrial^1, Soil^2	G 01

Appendix B3 cont'd

Gh_PM1	Maize-pigeon pea (*Cajanus cajan*) rotation with minimum tillage of fallow residue by hand slashing (**PM**), rep1, KAES in Volta Region, Ghana	Terrestrial[1], Soil[2]	G 01
Gh_PM2	**PM**, rep2, KAES in Volta Region, Ghana	Terrestrial[1], Soil[2]	G 01
Gh_PM3	**PM**, rep3, KAES in Volta Region, Ghana	Terrestrial[1], Soil[2]	G 01
Gh_PM4	**PM**, rep4, KAES in Volta Region, Ghana	Terrestrial[1], Soil[2]	G 01
HA_site1	Hawaii Mauna Kea permafrost_location1	Terrestrial[1], Soil[2]	G 01
HA_site2	Hawaii Mauna Kea permafrost_location2	Terrestrial[1], Soil[2]	G 01
HA_site3	Hawaii Mauna Kea permafrost_location3	Terrestrial[1], Soil[2]	G 01
Hi_50H_1	Kanchenjunga glacier (5000 m), rep1, slopes descending from Drohmo peak (6980 m) in Himalaya, Nepal (27° 48' 00" N and 88° 07' 01" E).	Terrestrial[1], Soil[2]	G 01
Hi_50H_2	slope at 5000 m, rep2 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_50H_4	slope at 5000 m, rep4 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_52H_1	slope at 5200 m, rep1 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_52H_2	slope at 5200 m, rep2 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_52H_3	slope at 5200 m, rep3 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_52H_4	slope at 5200 m, rep4 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_54H_1	slope at 5400 m, rep1 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_54H_2	slope at 5400 m, rep2 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_54H_3	slope at 5400 m, rep3 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_54H_4	slope at 5400 m, rep4 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_56H_1	slope at 5600 m, rep1 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_56H_2	slope at 5600 m, rep2 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_56H_3	slope at 5600 m, rep3 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_56H_4	slope at 5600 m, rep4 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_58H_1	slope at 5800 m, rep1 from Drohmo peak	Terrestrial[1], Soil[2]	G 01

Appendix B3 cont'd

Hi_58H_2	slope at 5800 m, rep2 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_58H_3	slope at 5800 m, rep3 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_58H_4	slope at 5800 m, rep4 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_60H_1	slope at 6000 m, rep1 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_60H_2	slope at 6000 m, rep2 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_60H_3	slope at 6000 m, rep3 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
Hi_60H_4	slope at 6000 m, rep4 from Drohmo peak	Terrestrial[1], Soil[2]	G 01
IA	Iowa farm soil after corping, IA , USA	Terrestrial[1], Soil[2]	G 01
Je_A72_1	California	Terrestrial[1], Soil[2]	G 01
Je_A72_2	California	Terrestrial[1], Soil[2]	G 01
Je_A74_1	California	Terrestrial[1], Soil[2]	G 01
Je_A74_2	California	Terrestrial[1], Soil[2]	G 01
Je_A74_2	California	Terrestrial[1], Soil[2]	G 01
Je_A82_1	California	Terrestrial[1], Soil[2]	G 01
Je_A82_2	California	Terrestrial[1], Soil[2]	G 01
Je_A84_1	California	Terrestrial[1], Soil[2]	G 01
Je_A84_2	California	Terrestrial[1], Soil[2]	G 01
Je_G72_1	California	Terrestrial[1], Soil[2]	G 01
Je_G74_1	California	Terrestrial[1], Soil[2]	G 01
Je_G74_2	California	Terrestrial[1], Soil[2]	G 01
Je_G74_2	California	Terrestrial[1], Soil[2]	G 01
Je_G82_1	California	Terrestrial[1], Soil[2]	G 01
Je_G84_1	California	Terrestrial[1], Soil[2]	G 01
Je_G84_2	California	Terrestrial[1], Soil[2]	G 01
Mi_Ag_C1	MSU farm, East Lansing, corn	Terrestrial[1], Soil[2]	G 01
Mi_Ag_C2	MSU farm, East Lansing, corn	Terrestrial[1], Soil[2]	G 01
Mi_Ag_C3	MSU farm, East Lansing, corn	Terrestrial[1], Soil[2]	G 01
Mi_Ag_FC1	MSU farm, East Lansing, canola	Terrestrial[1], Soil[2]	G 01
Mi_Ag_FC2	MSU farm, East Lansing, canola	Terrestrial[1], Soil[2]	G 01
Mi_Ag_FC3	MSU farm, East Lansing, canola	Terrestrial[1], Soil[2]	G 01
Mi_Ag_SB1	MSU farm, East Lansing, soybean	Terrestrial[1], Soil[2]	G 01
Mi_Ag_SB2	MSU farm, East Lansing, soybean	Terrestrial[1], Soil[2]	G 01

Appendix B3 cont'd

Mi_Ag__SB3	MSU farm, East Lansing, soybean	Terrestrial[1], Soil[2]	G 01
Mi_Ag__SF1	MSU farm, East Lansing, sunflower	Terrestrial[1], Soil[2]	G 01
Mi_Ag__SF2	MSU farm, East Lansing, sunflower	Terrestrial[1], Soil[2]	G 01
Mi_Ag__SF3	MSU farm, East Lansing, sunflower	Terrestrial[1], Soil[2]	G 01
Mi_Ag__SW1	MSU farm, East Lansing, switchgrass	Terrestrial[1], Soil[2]	G 01
Mi_Ag__SW2	MSU farm, East Lansing, switchgrass	Terrestrial[1], Soil[2]	G 01
Mi_Ag__SW3	MSU farm, East Lansing, switchgrass	Terrestrial[1], Soil[2]	G 01
Mi_Fo__M1	East Lansing, deciduous forest	Terrestrial[1], Soil[2]	G 01
Mi_Fo__M2	East Lansing, deciduous forest	Terrestrial[1], Soil[2]	G 01
Mi_Fo__M3	East Lansing, deciduous forest	Terrestrial[1], Soil[2]	G 01
Mi_Fo__U1	Chatham, Upper Peninsula, MI, pine forest	Terrestrial[1], Soil[2]	G 01
Mi_Fo__U2	Chatham, Upper Peninsula, MI, pine forest	Terrestrial[1], Soil[2]	G 01
Mi_Fo__U3	Chatham, Upper Peninsula, MI, pine forest	Terrestrial[1], Soil[2]	G 01
Mi_Ro__C2R	Rose Township, MI, corn	Terrestrial[1], Soil[2]	G 01
Mi_Ro__C3R	Rose Township, MI, corn	Terrestrial[1], Soil[2]	G 01
Mi_Ro__C4R	Rose Township, MI, corn	Terrestrial[1], Soil[2]	G 01
Mi_Ro__FC2R	Rose Township, MI, canola	Terrestrial[1], Soil[2]	G 01
Mi_Ro__FC3R	Rose Township, MI, canola	Terrestrial[1], Soil[2]	G 01
Mi_Ro__FC4R	Rose Township, MI, canola	Terrestrial[1], Soil[2]	G 01
Mi_Ro__R2	Rose Township, MI, Trees	Terrestrial[1], Soil[2]	G 01
Mi_Ro__R3	Rose Township, MI, Trees	Terrestrial[1], Soil[2]	G 01
Mi_Ro__R4	Rose Township, MI, Trees	Terrestrial[1], Soil[2]	G 01
Mi_Ro__SB2R	Rose Township, MI, Soybean	Terrestrial[1], Soil[2]	G 01
Mi_Ro__SB3R	Rose Township, MI, Soybean	Terrestrial[1], Soil[2]	G 01
Mi_Ro__SB4R	Rose Township, MI, Soybean	Terrestrial[1], Soil[2]	G 01
Mi_Ro__SF2R	Rose Township, MI, Sunflower	Terrestrial[1], Soil[2]	G 01
Mi_Ro__SF3R	Rose Township, MI, Sunflower	Terrestrial[1], Soil[2]	G 01
Mi_Ro__SF4R	Rose Township, MI, Sunflower	Terrestrial[1], Soil[2]	G 01
Mi_Ro__SW2R	Rose Township, MI, Switchgrass	Terrestrial[1], Soil[2]	G 01
Mi_Ro__SW3R	Rose Township, MI, Switchgrass	Terrestrial[1], Soil[2]	G 01
Mi_Ro__SW4R	Rose Township, MI, Switchgrass	Terrestrial[1], Soil[2]	G 01

Appendix B3 cont'd

OH_E1x1a	Ohio	Terrestrial[1], Soil[2]	G 01
OH_E1x1b	Ohio	Terrestrial[1], Soil[2]	G 01
OH_E1x6a	Ohio	Terrestrial[1], Soil[2]	G 01
OH_E1x6b	Ohio	Terrestrial[1], Soil[2]	G 01
OH_E2x1a	Ohio	Terrestrial[1], Soil[2]	G 01
OH_E2x1b	Ohio	Terrestrial[1], Soil[2]	G 01
OH_E2x6a	Ohio	Terrestrial[1], Soil[2]	G 01
OH_E2x6b	Ohio	Terrestrial[1], Soil[2]	G 01
Pi_0D	PCB-contaminated sandy soil, Picatinny arsenal, NJ, US	Terrestrial[1], Soil[2]	G 01
Si_1OO__120_11	Siberia	Extreme[1], Soil[2]	G 02
Si_1S___40_10	Siberia	Extreme[1], Soil[2]	G 02
Si_1S___40_7	Siberia	Extreme[1], Soil[2]	G 02
Si_2__3M_21	Siberia	Extreme[1], Soil[2]	G 02
Si_2__3M_24	Siberia	Extreme[1], Soil[2]	G 02
Si_5__1OT_14	Siberia	Extreme[1], Soil[2]	G 02
Ant_AD10	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
Ant_AD11	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
Ant_IC1	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
Ant_IC2	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
Ant_ID1	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
Ant_ID2	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
Ant_QC7	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
Ant_QC8	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
Ant_QD7	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
Ant_QD8	Antartica	Terrestrial[1], Extreme[1], Soil[2]	G 03
St_Av1	Spitsbergen	Terrestrial[1], Extreme[1], Soil[2]	G 03
St_Av2	Spitsbergen	Terrestrial[1], Extreme[1], Soil[2]	G 03
Pig_Dom	Pig feces,	Oragnism-Associated[1]	G 04
Pig_FO26	Pig feces,	Oragnism-Associated[1]	G 04
Pig_FO31	Pig feces,	Oragnism-Associated[1]	G 04
Pig_FO32	Pig feces,	Oragnism-Associated[1]	G 04
Pig_FO35	Pig feces,	Oragnism-Associated[1]	G 04
Pig_FO37	Pig feces,	Oragnism-Associated[1]	G 04
Pig_F1O4	Pig feces,	Oragnism-Associated[1]	G 04
Pig_F2	Pig feces,	Oragnism-Associated[1]	G 04
Pig_F3 11	Pig feces,	Oragnism-Associated[1]	G 04

Appendix B3 cont'd

Pig_F3 12A	Pig feces,	Oragnism-Associated[1]	G 04
Pig_F3 12B	Pig feces,	Oragnism-Associated[1]	G 04
Pig_F3 13	Pig feces,	Oragnism-Associated[1]	G 04
Pig_F6	Pig feces,	Oragnism-Associated[1]	G 04
Pig_001	Pig feces,	Oragnism-Associated[1]	G 04
Pig_002	Pig feces,	Oragnism-Associated[1]	G 04
Pig_003	Pig feces,	Oragnism-Associated[1]	G 04
WWT___01	Urgary	Freshwater[1], Waste water[2]	G 05
WWT___02	Urgary	Freshwater[1], Waste water[2]	G 05
WWT___03	Urgary	Freshwater[1], Waste water[2]	G 05
WWT___04	Urgary	Freshwater[1], Waste water[2]	G 05
WWT___05	Urgary	Freshwater[1], Waste water[2]	G 05
WWT___06	Urgary	Freshwater[1], Waste water[2]	G 05
Mi_RR0D	PCB-contaminated sediment, River Raisin, MI, US	Freshwater[1], Sediment[2]	G 06
WA__DoH	Washington	Freshwater[1], Sediment[2]	G 06
WA__Han01	Columbina river, Washington	Freshwater[1], Sediment[2]	G 06
WA__Han02	Columbina river, Washington	Freshwater[1], Sediment[2]	G 06
WA__Han03	Columbina river, Washington	Freshwater[1], Sediment[2]	G 06
WA__Han04	Columbina river, Washington	Freshwater[1], Sediment[2]	G 06
WA__Han05	Columbina river, Washington	Freshwater[1], Sediment[2]	G 06
Mam__A2	Siberia	Fossil[1], Oragnism-Associated[1]	G 07
Mam__Ce	Siberia	Fossil[1], Oragnism-Associated[1]	G 07
Adria	Marine sediment from the Northern Adriatic sea, Gulf of Trieste (45°33'N 13°37E)	Marine[1], Sediment[2]	G 08
BC180	Barrow Canyon (BC, 186 m depth, 71.607N 156.214W) from the Alaskan maritime in the Chuckchi Sea	Marine[1], Sediment[2]	G 08
EHS	East Hanna Shoal (EHS, 160 m depth, 72.637N 158.667W) from the Alaskan maritime in the Chuckchi Sea	Marine[1], Sediment[2]	G 08
FL_10	Florida Bay 10 (FL10, 25.025N 80.681W)	Marine[1], Sediment[2]	G 08
FL_11	Florida Bay 11 (FL11, 24.913N 80.938W)	Marine[1], Sediment[2]	G 08
FL_9	Florida Bay 9 (FL9, 25.177N	Marine[1], Sediment[2]	G 08

Appendix B3 cont'd

	80.490W)		
GM1	(800 m depth, 26.404N 96.064W) in the Gulf of Mexico	Marine[1], Sediment[2]	G 08
JF	West of the Juan de Fuca Ridge (JF, 3869 m depth, 46.783N 133.667W) n the Pacific Ocean	Marine[1], Sediment[2]	G 08
ST_2		Marine[1], Sediment[2]	G 08
WA_Coast	Washington Margin (WM, 1138 m depth, 46.575N 124.822W) n the Pacific Ocean	Marine[1], Sediment[2]	G 08
Cz_14D_SIP	PCB- and biphenyl-degrading population form PCB-contaminated soil under Austrian pine tree at 14 days incubation with 13C-biphenyl	Cultured[1], Soil[2] or Sediment[2]	G 09
Cz_4D_SIP	PCB- and biphenyl-degrading population form PCB-contaminated soil under Austrian pine tree at 4 days incubation with 13C-biphenyl	Cultured[1], Soil[2] or Sediment[2]	G 09
Mi_RR14D_SIP	PCB- and biphenyl-degrading population form PCB-contaminated River Raisin sediment at 14 days incubation with 13C-biphenyl	Cultured[1], Soil[2] or Sediment[2]	G 09
Mi_RR14Ds_SIP	PCB- and biphenyl-degrading population form PCB-contaminated River Raisin sediment at 14 days incubation with 13C-biphenyl with slurry	Cultured[1], Soil[2] or Sediment[2]	G 09
Mi_RR3D SIP	PCB- and biphenyl-degrading population form PCB-contaminated River Raisin sediment at 3 days incubation with 13C-biphenyl	Cultured[1], Soil[2] or Sediment[2]	G 09
Pi_14D SIP	PCB- and biphenyl-degrading population form PCB-contaminatedPicatinny sandy soil at 14 days incubation with 13C-biphenyl	Cultured[1], Soil[2] or Sediment[2]	G 09
St_AN1_IN	Spitsbergen	Cultured[1], Soil[2] or Sediment[2]	G 09
St_AN2_IN	Spitsbergen	Cultured[1], Soil[2] or Sediment[2]	G 09
St_anN1_IN	Spitsbergen	Cultured[1], Soil[2] or Sediment[2]	G 09
St_anN2_IN	Spitsbergen	Cultured[1], Soil[2] or Sediment[2]	G 09

Appendix B3 cont'd

St_O1_IN	Spitsbergen	Cultured[1], Soil[2] or Sediment[2]	G 09
St_O2_IN	Spitsbergen	Cultured[1], Soil[2] or Sediment[2]	G 09
St_ON1_IN	Spitsbergen	Cultured[1], Soil[2] or Sediment[2]	G 09
St_ON2_IN	Spitsbergen	Cultured[1], Soil[2] or Sediment[2]	G 09
FRC1	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC10	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC11	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC12	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC13	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC14	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC15	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC16	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC17	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC18	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC2	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC20	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC22	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC23	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC24	FRC	Extreme1,Freshwater[1]Sedi	G 10

Appendix B3 cont'd

		ment[2]	
FRC25	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC4	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC5	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC6	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
FRC9	FRC	Extreme1,Freshwater[1]Sediment[2]	G 10
Du_17_1	DUSEL	Extreme[1], Microbial mat[2]	G 11
Du_17_2	DUSEL	Extreme[1], Microbial mat[2]	G 11

Appendix B4. Confusion table of priori groups and bacterial assemblage' clusters by average distance clustering

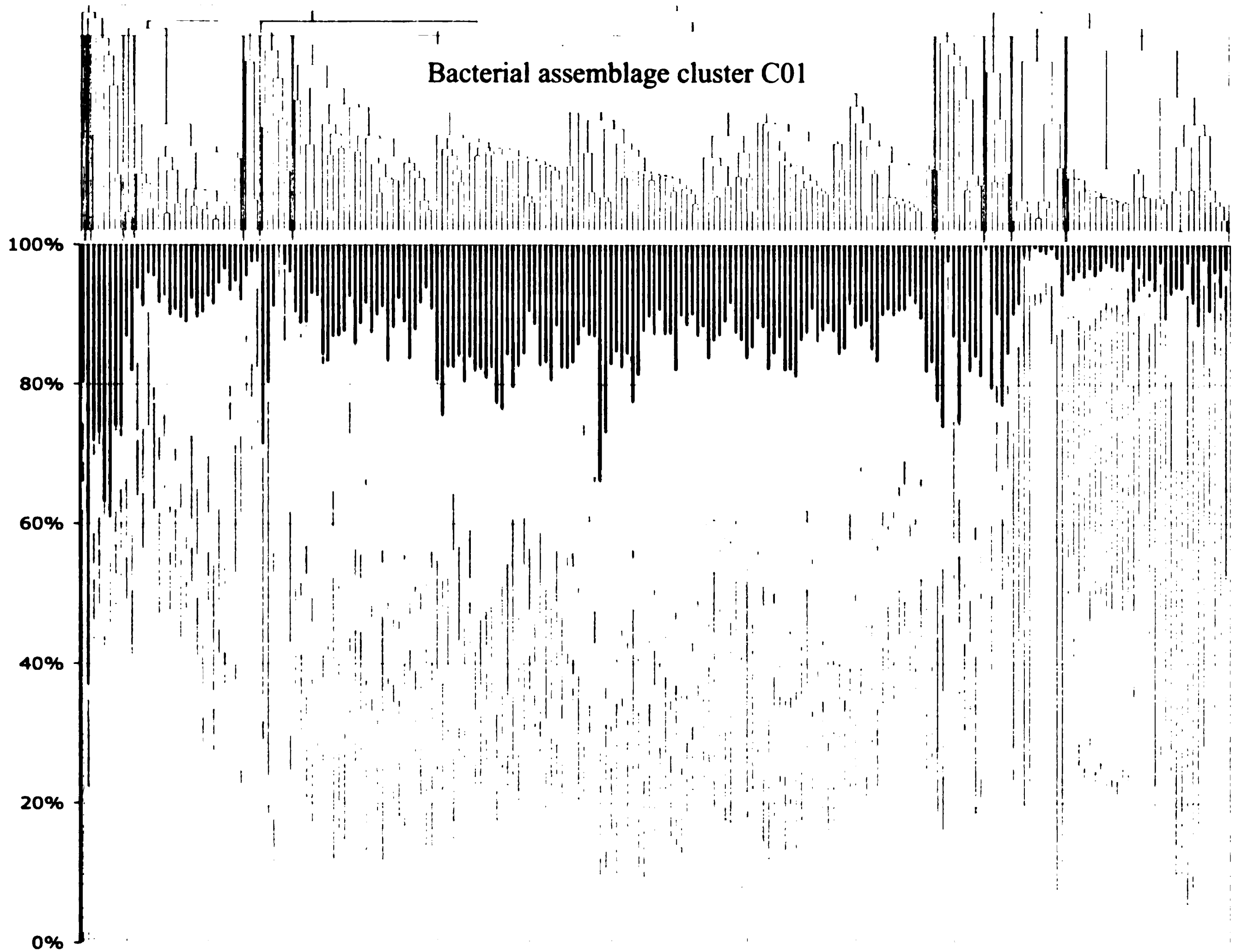
	G01	G02	G03	G04	G05	G06	G07	G08	G09-1*	G09-2*	G10	G11	Sum
C01	**114**								4				118
C02		**6**											6
C03			**12**	14	1			3					30
C04					**5**								5
C05	2					6			1				9
C06						1							1
C07							**2**		1				3
C08								**6**					6
C09								1					1
C10				2						**8**			10
C11											**20**		20
C12												2	2
Sum	116	6	12	16	6	7	2	10	6	8	20	2	211

* G09 were separated in two sub-groups: G09-1, PCB- and biphenyl-utilizing population (Chapter 4) and G09-2, various enrichments of bacterial community from Spitsbergen permafrost soil.

Table B4.1. Confusion table of priori groups and bacterial assemblage' clusters by average distance clustering

Priori groups (G01-G11), assigned based on Habitat-Lite based description, (Appendix B.1.) were compared to bacterial assemblage clusters (C01-C12) using Q-mode average clustering based on 1- Chao's corrected *Sørensen* similarities with at 97% OTU matrix. Most priori groups and bacterial assemblage clusters were correlated to each other with a few exceptions. It means that similar bacterial assemblage is present in the same microbial habitats, congruent with habitat-description.

Appendix B5. Bacterial Assemblage Clustering



■ Proteobacteria
■ Chloroflexi
□ TM7
Deinococcus-Thermus
■ Deferribacteres
□ OP10
■ SR1
■ Thermodesulfobacteria
■ unclassified_Bacteria

■ WS3
□ Actinobacteria
Gemmatimonadetes
BRC1
■ Fusobacteria
■ Nitrospira
■ Tenericutes
■ OP11

Bacteroidetes
■ Firmicutes
Acidobacteria
Chlamydiae
■ Lentisphaerae
■ Aquificae
■ Thermomicrobia
■ Euryarchaeota

■ Verrucomicrobia
□ Planctomycetes
OD1
Spirochaetes
□ Fibrobacteres
□ Cyanobacteria
■ Dictyoglomi
■ Crenarchaeota

Terresterial soil habitat (G01) was almost congruent with bacterial assemblage clustering (C01) (Appendix B4) with three notable indicator phyla: *Acidobacteria*, *Verrucomicrobia*, and *Gemmatimonadetes*.

Appendix B6. Indicator Species of Selected Priori Groups

Group 01 (Terrestrial^1, Soil^2)

Bradyrhizobium

Xiphinematobacteriaceae_genera_incertae_sedis

Gemmatimonas

***Acidobacteria* Gp3**

***Acidobacteria* Gp4**

***Acidobacteria* Gp5**

***Acidobacteria* Gp6**

***Acidobacteria* Gp7**

Unclassified Micromonosporaceae

Group 02 (Extreme^1, Soil^2)

Psychrobacter

Carboxydocella

Exiguobacterium

Group 08 (Marine^1, Sediment^2)

Jannaschia

Pelobacter

Desulfuromusa

Desulfosarcina

Desulfatibacillum

Desulfococcus

Desulforhopalus

Owenweeksia

Rubritalea

Acidobacteria Gp9

Acidobacteria Gp21

Acidobacteria Gp26

Caldithrix

Unclassified *Myxococcales*

Unclassified *Desulfuromonaceae*

Q-value < 0.05 (false discovery rate significant value)

To find indicator species that represents a specific habitat priori group, RDP classifier based taxonomy-bins at 50% threshold were used in function "duleg" (Dufrene-Legendre indicator species analysis in R package "labdsv"), which considers the occurrence frequency, and the relative abundance. Priori G01 contains the member of the phyla *Acidobacteria*, *Verrucomicrobia*, and *Gemmatimonadetes*, which were often found exclusively in soil habitats. Priori G02, contains 6 Siberian permafrost soils, has as indicator species, *Psychrobacter* and *Exiguobacterium*, which can grow at temperatures as low as -10 and -5 °C. *Exiguobacterium* spp. and *Psychrobacter* spp. abundance in these sites also were measured by Q-PCR amplification (Rodrigues *et al*, 2009).

Appendix B7. Functional Diversity Measures

INTRODUCTION

Functional diversity is "the value and range of the functional traits of the organisms in a given ecosystem" by definition of Tilman (2001). The distribution of trait values can be characterized through the average trait value, i.e. community-weighted mean (CWM) trait value (Violle *et al.*, 2007), which is an indicator of functional biodiversity and reflects the "average" trait value of dominant species in a community.

METHOD

Calculating community-weighted mean (CWM) trait value:

$$\mathrm{CWM} = \sum_{i}^{n} P_i * \mathrm{Trait}$$

where pi is the relative contribution of species i to the community, and trait i is the trait value of species i. The total number of species included in the calculation is "n".

I measured the bacterial traits using genomic information: gene copy numbers of each COG and KEGG categories obtained from complete (782 genomes) and draft genome (502 genomes) projects (total of 1284 genomes). I randomly selected the representative genomes of each genus and then match the genera names to taxonomy-bin names of RDP classifier results. Thus, 236 taxonomy-bins defined by RDP classifier at 50% threshold (for COG; 226 for KEGG) were given the assigned traits and gene numbers in the COG or KEGG categories. There are two assumptions for this analysis: 1) higher copy number in a gene category means possibly more diverse functions, and 2) the intraspecies variances within genera are small. Priori groups were aligned by Habitat-Lite definition (Appendix B2, and B3).

Priori Group	A cumulated relative abundance of species included in the calculation
G01	10.1
G02	43.4
G03	44.7
G04	18.1
G05	38.0
G06	11.0
G07	64.7
G08	20.3
G09-1	45.1
G09-2	25.7
G10	21.0
G11	17.9

Table. B8.1. A cumulated relative abundance of species included in the calculations.

RESULTS

Current genomes information only covers in the range of 10-65 % of the genera in 211 bacterial (community) assemblages. Though the lowest coverage, which is in priory group G01, has the highest CWM value in most of the COG categories. This means soil might possess highly divergent traits that reflects complexity of soil ecological niches. G03 (Antarctic rhizosphere) and G04 (animal feces) had the lowest CWM values in categories involved in metabolism and energy production. Surprisingly, three COG categories - replication and repair, translation, and transcription - showed constant CWM values regardless of priori groups. This means that the house-keeping genes, essential to sustaining bacterial live, is consistently present in all environments at the same level. It also supports the validity of this approach.

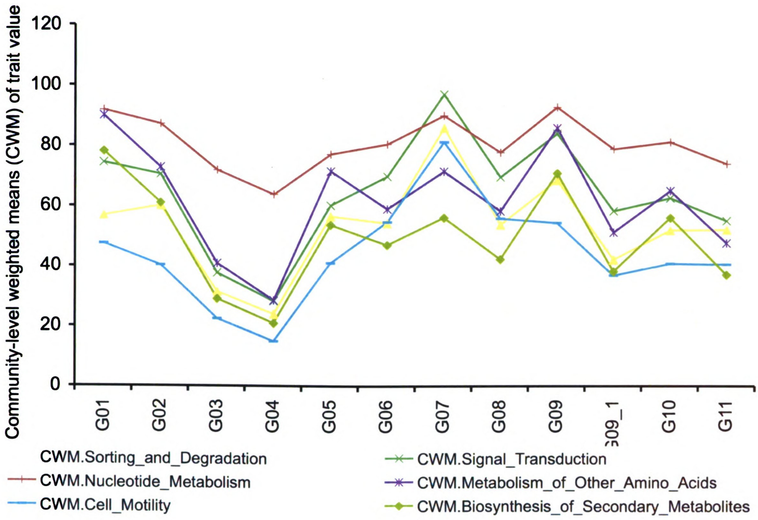


Figure B7.1. COG categories with CWM values by priori groups

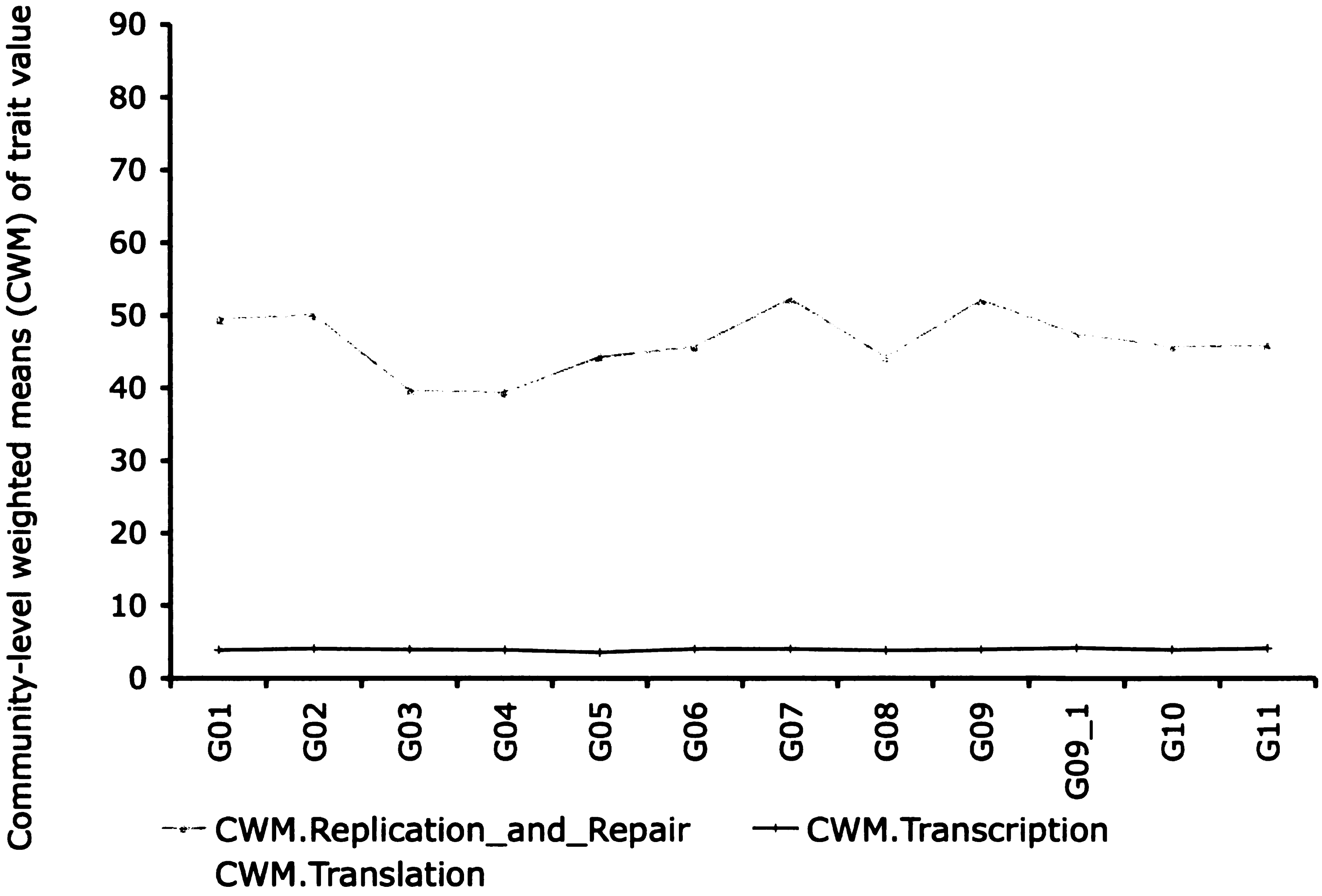


Figure B7.3 Constant CMW value in all groups.

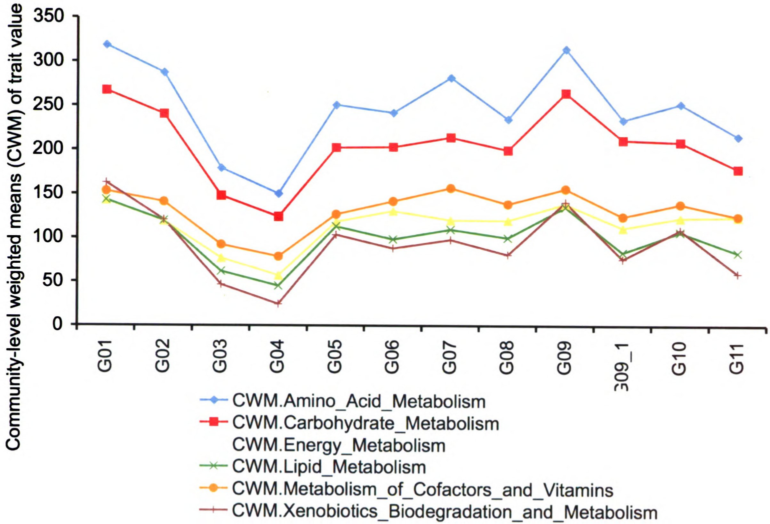


Figure B7.2 COG categories with CWM values by priori groups

REFERENCE

Rodrigues DF, da C Jesus E, Ayala-Del-Río HL, Pellizari VH, Gilichinsky D, Sepulveda-Torres L, Tiedje JM (2009) Biogeography of two cold-adapted genera: *Psychrobacter* and *Exiguobacterium*. *ISME J* **3**:658-665

Tilman D (2001) Functional diversity. – In: Levin, S. A. (ed.), Encyclopedia of biodiversity. Academic Press, pp. 109–120

Violle C, Navas ML, Vile D, Kazakou E, Fortunel C, Hummel I, Garnier E (2007) Let the concept of trait be functional! *Oikos* **116**:882-892