Long Term Impacts of a Genetically Engineered Microorganism (GEM) and Polycyclic Aromatic Hydrocarbons (PAHs) on Soil Bacterial Communities

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Xiaoci Ji

May 2013

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my advisor Dr. Jennifer DeBruyn, for the immeasurable support, help and guidance she provided me. Without her dedicated effort, I would not have achieved so much from this master program.

I would like to thank my committee members Dr. John Buchanan, Dr. Arnold Saxton, Dr. Mark Radosevich for their kindness, help, and advice in my courses and research.

Also, special thanks to Dan Williams and Scott Moser at Center for Environmental Biotechnology (UT) for lysimeter sampling and to Dr. Alice Layton and Dr. Steven Ripp at Center for Environmental Biotechnology (UT) for their guidance on this project.

In addition, I would like to give thanks to other students in our lab Kelly Cobaugh, Mikhail Androsov, Kathleen Hauther, Kristen Vaughn, and Mariam Fawaz (Radosevich lab), for their help in my project and the joy they brought into our lab work.

Special thanks to my parents for they always support my decisions, and encouraged me to finish whatever I started.

This work was funded by the Microbiology across Campuses Educational and Research Venture (M-CERV) and University of Tennessee Institute of Agriculture AgResearch.

Last but not least, I want to give my thanks to God, who found me and transformed me during the last two years. It's God who makes all things possible.

ABSTRACT

Microbes capable of polyaromatic hydrocarbon (PAH) biodegradation can be used to remediate soils contaminated with these persistent pollutants. To monitor *in situ* PAH-biodegradation, the bioluminescent bio-reporter *Pseudomonas fluorescens* HK44, containing a lux luminescent gene cassette inserted into its naphthalene degradation operon, was released into PAH-contaminated soil in lysimeters in 1996. Three treatments were imposed: strain HK44 mixed with PAH-contaminated soil (PAH+, HK44+; n=3); strain HK44 mixed with uncontaminated soil (PAH–, HK44+; n=2) and PAH-contaminated soil alone (PAH+, HK44–; n=1). The objective of this study was to assess the long term impacts of these treatments on the indigenous soil bacterial community structure in the lysimeters. In 2010, 14 years after experiment initiation, replicate soil cores were taken from each lysimeter. Soil bacterial community structures were determined by 454 pyrosequencing of 16S rRNA gene amplicons. Even though PAH concentrations fell below detectable levels within the first couple years of the lysimeter experiment, PAH+ lysimeters showed significantly higher soil organic matter content (1.30 ± 0.23%) than PAH– lysimeters (0.81 ± 0.08%). 16S rRNA gene libraries reveal that there was a change in the bacterial community structure in PAH+ compared to PAH– lysimeters: 9.59% of OTUs (operational taxonomic units) were shared between PAH+ lysimeters while only 4.08% were shared between PAH+ and PAH– lysimeters. Multivariate ordination and cluster analysis, and phylogenetic tree-based analysis indicate that communities fell into three clusters: lysimeter 1 and 2 (both PAH+, HK44+); lysimeter 4 (PAH+, HK44+) and 6 (PAH+, HK44–); and lysimeter 3 and 5 (both PAH–, HK44+). Therefore over the long term, the addition of PAHs was more influential on bacterial community structure than the introduction of a GEM.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 PAHs

Polycyclic aromatic hydrocarbons (PAHs) are a group of organic compounds characterized by a chemical structure consisting of two or more fused aromatic rings and/or pentacyclic molecules (Bamforth et al., 2005; Cerniglia, 1992). PAHs are a constituent of fossil fuels, derived from incomplete combustion of organic matter. They can be created through natural processes such as forest fires or anthropogenic activities such as coking processes in industrial production (Bamforth et al., 2005; Cerniglia, 1992; Ni Chadhain et al., 2006). Anthropogenic activities may cause PAH contamination in environments through a wide variety of ways. For example, during industrial activities, leakage, disposal or unexpected discharge of industrial effluents, fossil fuel and petroleum products could introduce high concentrations of PAHs to the environments (Bamforth et al., 2005). In the past 100 to 150 years, anthropogenic activities have become the dominate origin of PAHs contamination in the environment (Jones et al., 1989; Juhasz et al., 2000).

In general, PAHs have low water solubility, and PAH solubility decreases as the number of fused aromatic rings increases (Cerniglia, 1992; Juhasz et al., 2000). Thus, PAHs tend to be persistent in ecosystems, particularly in soils and sediments. This is a concern because PAHs are known to be harmful to human health (Bamforth et al., 2005), as they can be carcinogenic, mutagenic and/or teratogenic (Incardona et al., 2004; Kim et al., 2005). Because PAHs are toxic and persistent in the environment, they are a major concern, and approaches to remove PAHs and recover ecosystems are of great interest (Bamforth et al., 2005; Jones et al., 1989; Juhasz et al., 2000). Conventional soil remediation methods, such as excavation, do not remove PAHs

effectively. Previous studies have shown that in contaminated areas, PAH concentrations remained high even after excavation (Dionisi et al., 2004). Therefore biodegradation by microorganisms is a desirable approach for further clean up on site (Bamforth et al., 2005; Cerniglia, 1992; DeBruyn et al., 2007).

1.2 Biodegradation and Bioremediation

The purpose of environmental remediation is to reduce the pollutant in terms of concentration, mobility and/ or toxicity. Remediation technologies can be physical (e.g., excavation), chemical (e.g., chemical oxidation), and/or biological (e.g., bacterial degradation). Biodegradation is the decomposition of complex molecules into simpler molecules by organisms (Bamforth et al., 2005; Mueller et al., 1997). Microorganisms in particular have a ubiquitous distribution and are capable of metabolizing a wide variety of substrates. In almost all ecosystems, microorganisms are involved in a wide range of natural biogeochemical cycles and are able to degrade or transform hydrocarbon molecules. Therefore, these abilities have been utilized for bioremediation of environmental contaminants. Microbes with the ability to utilize the contaminants as energy or carbon sources are applied or promoted to grow in the polluted environment. In these processes, microbial metabolism may convert the contaminants to compounds that are less toxic, active or mobile. In general, biological remediation (bioremediation) costs less than conventional physical and chemical methods. In addition, bioremediation can take place in inaccessible environments (e.g., groundwater).

The efficiency of PAH biodegradation is affected by many factors. Microbial metabolism of PAHs can be controlled by environmental conditions such as temperature, oxygen, nutrient (N, P and K) availability, and accessibility of the PAHs. In addition, as the number of aromatic rings in a PAH molecule increases, its complexity increases and its solubility decreases, decreasing

bioavailability (Johnsen et al., 2005; Lors et al., 2010; Ni Chadhain et al., 2006). It is also important to take into account physiological characteristics of the PAH degrading microorganisms, their abilities to colonize soil environments, and their interaction with indigenous microbes (Johnsen et al., 2005). For example, increases in temperature can increase PAH solubility (Margesin et al., 2001) but decrease oxygen solubility, thereby limiting the growth of microorganisms (Bamforth et al., 2005). Thus, these interacting factors can impact PAH biodegradation rate by affecting the growth rate of PAH-degrading bacteria and PAH bioavailability.

Because addition of PAHs changes soil organic matter composition and quantity, the exposure of native soil microorganisms to PAHs changes the types and abundances of microorganisms present (i.e. community structure). Studies of soil and sediment bacterial communities during PAH exposure and degradation have provided some information on which types of bacteria are most abundant and likely most functionally important. For example, Vinas et al. (2005) investigated PAH impacts on bacterial community structure in soil from a creosote contaminated field with the initial total PAHs concentration around 2700 mg/kg soil. Three- and four-ring PAHs, including phenanthrene, anthracene, fluoranthene, pyrene, benzo(a)anthracene, chrysene, all had initial concentration over 100 mg/kg soil, and were the dominated PAH contaminates. The soil was treated with seven different microcosm conditions for 200 days. These treatments were: low moisture, autoclaved soil, addition of inorganic nutrients, addition of inorganic nutrients and biosurfactant, addition of inorganic nutrients and inoculation of a PAH-degrading consortium, and addition of inorganic nutrients and iron octoate (Vinas et al., 2005). They reported a rapid increase of heterotrophic bacteria populations during first 45 days after PAH degradation initiated and observed that dominate bacteria groups differ between PAH

degradation stages and different medium conditions (Vinas et al., 2005). During the early rapid degradation stage, bacteria belonging to class *α-Proteobacteria* (genera *Sphingomonas* and *Azospirillum*) dominated in all treatments. However, as PAH degradation progressed, the dominant bacterial groups in soil with and without nutrient addition diverged: in both treatments, the most abundant groups included *γ-Proteobacteria* (genus *Xanthomonas*) and *α-Proteobacteria* (genus *Sphingomonas*). In the non-nutrient addition treatment, *Bacteroidetes* phylum also dominated while in the nutrient addition treatment *β-Proteobacteria* group (genera *Alcaligenes* and *Achromobacter*) dominated (Vinas et al., 2005).

In addition, Vinas et al. (2005) found that soil moisture is a key factor modulating biodegradation rates. Soils with five different water contents (5%, 20%, 40%, 60% and 75% water holding capacity) were used to optimize biodegradation. In this microcosm experiment, highest biodegradation rate were observed with 40% and 60% WHC soils. This result shows that soil moisture content impacts PAH biodegradation, and that both too little and too much soil moisture can reduce PAH biodegradation rates.

In another study, Lors et al. (2010) performed a six month remediation of an abandoned coal tar contaminated soil using windrows. The initial total concentration of 16 PAHs in that soil was approximate 3000 mg/kg dry soil with 2-ring, 3-ring, and 4-ring PAHs accounting for 20%, 44%, and 28% of the total concentration respectively. Soil windrows were sampled on day 0, 44, 60, 92 and 182 and the community structure assessed by 16S rRNA gene fingerprinting using denaturing gradient gel electrophoresis (DGGE) (Lors et al., 2010).. In the bacterial community, *γ- Proteobacteria* had the highest relative abundance (62%). The next two most abundant groups were *α-Proteobacteria* (15%) and *β-Proteobacteria* (8%). Most PAH degraders belonged to *γ-Proteobacteria*. *β-Proteobacteria* were not detected until 3 months after this experiment started

indicating that this group tended to increase when PAHs concentrations were relatively low (Lors et al., 2010).

During an 88 day PAH biodegradation experiment, Margesin et al. (2000) monitored microbial activities by microbial counts, measurement of soil respiration, soil biomass and changes of several enzymes. Uncontaminated soil was mixed with 625 mg/kg of diesel oil and subjected to N, P, and K fertilizer addition treatments. Fertilizer was added to some treatments with different ratios, and in some treatments known hydrocarbon degradative bacteria were inoculated. PAH contaminated soil with organic nitrogen addition (urea) had higher respiration and biomass carbon (Margesin et al., 2000). In addition, microbial counts showed that the total number of heterotrophic bacteria remained consistent during the hydrocarbon degradation process, but the proportion of hydrocarbon degradation microorganisms increased (Margesin et al., 2000). These results indicated that while the total number of bacteria remained constant, the types and abundances of particular groups (i.e. community structure) shifted (Margesin et al., 2000).

In summary, PAH biodegradation in soil involves a consortium of bacterial groups, and as a result, PAH contamination generally causes shifts in bacterial community structure (Bouchez et al., 1995; Lors et al., 2010). Bacteria belonging to the *Proteobacteria* phylum are generally found to be the most active and abundant taxa in soil bacterial communities exposed to PAHs (Lors et al., 2010; Ni Chadhain et al., 2006; Zhang et al., 2011). In this phylum, several genera belonging to class *γ- Proteobacteria* (especially genus *Pseudomonas*) were found to be dominant and active during PAH degradation (Layton et al., 2012; Lors et al., 2010; Ni Chadhain et al., 2006). However, these studies on bacterial community structure focus on relatively short time periods (i.e. 200 days in Vinas et al., six month in Lors et al. 2010, 88 days in Margesin et

al. 2000). Our study presents a rare opportunity to examine the long term (> 10 year) changes in bacterial community structure when challenged with PAHs in a relatively large experimental scale.

1.3 HK44, a Genetically Modified Bio-reporter Used for Soil Remediation

A genetically engineered microorganism (GEM), *Pseudomonas fluorescens* HK44 (HK44) was used in this study for remediation and monitoring purposes (Heitzer et al., 1992). Genetically engineered microbes (GEMs) are microbes whose genetic material has been modified using genetic engineering techniques. Usually it involves introducing foreign or synthetic genes into the microorganism of interest. However, GEMs have been of increasing concern due to possible threats to the environment and public health. Since GEMs contain gene(s) from more than one organism and from other environments, there is concern of their impacts on the native ecosystem. Impacts of genetically engineered microbes are considered to be relatively slight since genetic engineering generally reduces the biological fitness compared with their parent wild-type (Lelley et al., 2003; Trogl et al., 2012). However, further studies of long term impacts of GEMs on ecosystems are needed to fully understand the risk associated with their environmental release.

HK44 is the first genetically engineered microorganism allowed for field testing in the U.S. for bioremediation purposes (Ripp et al., 2000a; Ripp et al., 2000b). *P. fluorescens* strain HK44 harbors the plasmid pUTK21 containing a tetracycline resistance marker (*tetA*) and introduced bacterial luminescence genes (*luxCDABE*) in a naphthalene degradation pathway gene (*nahG*) (King et al., 1990; Ripp et al., 2000a; Ripp et al., 2000b). This bacteria strain was isolated from a manufactured gas plant (MGP) soil (King et al., 1990) and was chosen because it displayed enhanced naphthalene degradation gene expression and PAH degradation rates (Lelley

et al., 2003). The plasmid pUTK21 was constructed based on pKA1, a natural plasmid in bacterium *P. fluorescens* 5R. This plasmid harbors the upper (*nahABCDEF*) and lower (*nahGHIJK*) naphthalene metabolism pathway (Trogl et al., 2012). The *luxCDABE* genes were derived from *Vibrio fischeri* strain MJ-1, which was isolated from the light organ of the fish *Monocentris japonicas* (King et al., 1990). The *luxA* and *luxB* genes code for α and β subunits of luciferase, an enzyme capable of aldehyde-dependent luminescence and the *luxCDE* genes code for aldehyde synthesis (Meighen, 1988). Thus, the *luxCDABE* genes enabled the host bacterium aldehyde-independent luminescence (Meighen, 1988). In summary, this genetically engineered microbe bioluminesces during degradation of selected two or three ring PAHs, and is used as a bioreporter of PAH biodegradation activity (Heitzer et al., 1992). HK44 was designed to monitor PAH bioremediation in soils by measuring the strength of bioluminescence it emitted during PAH degradation (Ripp et al., 2000a). Since HK44 needs more energy for the introduced foreign *lux* genes and has reduced fitness because of its artificially modified genetic materials, it is hypothesized to be less competitive than its parent wild strain, and whether or not it survives long term (in this study, approximately 14 years) is questionable. In addition, it is also unknown whether or not this microorganism might interact with or change native soil bacterial communities.

1.4 Microbial Community Composition Analysis

Microorganisms are involved in a wide range of critical processes in soil ecosystems, and are very important in maintaining ecosystem function. For example, microorganisms contribute to maintenance of soil structure, decomposition of complex molecules, and cycling of carbon, nitrogen, phosphorus etc. Decomposition can also include breaking toxins or contaminants down to simpler, potentially less toxic molecules (Garbeva et al., 2004). These activities are generally

carried out by a complex and interacting consortium of hundreds or thousands of different microbial groups. Thus, microbial community analysis is critical to understanding the functions of an ecosystem (Garbeva et al., 2004).

In the past, microbial diversity was determined based on cultivable bacteria. However, only about 1% of existing bacteria are cultivable (Borneman et al., 1997; Staley et al., 1985; Torsvik et al., 1990). In addition, morphological and physiological characters of microorganisms are ambiguous and do not provide sufficient information for microbial classification and the analyses of microbial community structure (Garbeva et al., 2004; Olsen et al., 1986; Zinger et al., 2012). As molecular biology developed, extraction of DNA directly from soil allowed researchers to get genetic information from those microorganisms that were uncultivable in laboratory. The small subunit ribosomal RNA (rRNA) gene is currently the most commonly used phylogenetic marker in determining the phylogenetic relationships of all organisms. rRNA genes are used as a phylogenetic indicator for several reasons: first, ribosomes are a critical element of protein synthesis, and therefore they exist in all organisms. Second, because they are essential, selection pressure is high and rRNA gene sequences and secondary structures are highly conserved. Therefore, neutral mutations that have accumulated in rRNA genes over time reflect the divergence time and phylogenic relationships of organisms. (Olsen et al., 1986; Olsen et al., 1993; Smit et al., 2007). Finally, the length of the rRNA genes make them appropriate for phylogenetic analysis (Olsen et al., 1993). The prokaryotic ribosome is made up of two subunits, the 50S large subunit and the 30S small subunit. The former one contains 5S rRNA and 23S rRNA, and the latter one contains 16S rRNA. 16S rRNA is longer than 5S and therefore provides more information. Meanwhile, its sequence change rate is less than 23S rRNA, and therefore could be a better indicator in large scale comparisons between bacteria species (Olsen et al.,

1993). In addition, 16S and 23S rRNA are functionally related to each other, and usually provide similar phylogeny information (Olsen et al., 1993). Thus, 16S rRNA gene is widely used as phylogeny index for prokaryotes.

There are a wide variety of DNA-based methods for investigating microbial community structure that use 16S rRNA (or other phylogenetic marker) genes. For example, in terminal restriction fragment length polymorphism (T-RFLP), 16S rRNA genes are amplified, and amplicons are digested by restriction enzymes and separated by length. Different sequences from different species will have different restriction sites; therefore this technique gives a fingerprint of the bacterial community structure. Temperature or denaturing gradient gel electrophoresis (TGGE or DGGE) methods are based on separating gene fragments of each library by their GC content to profile the bacterial community structures. These methods enable analysis of many samples, but the information provided by these methods is limited: they cannot identify the types of microorganisms present. In contrast, cloning and Sanger sequencing methods return specific 16S rRNA gene sequences. These methods require more work, but can identify the microorganisms. The major disadvantage of Sanger sequencing is the high cost. Recently, high throughput sequencing platforms (e.g. 454 pyrosequencing) have been developed. These sequencing methods use massively parallel approaches. Compared with conventional method (cloning and Sanger sequencing), these high throughput methods reduced the cost by over two orders of magnitude (Shendure et al., 2008). Thus, high throughput methods allow researchers to do the sequencing in larger scale. Although high throughput methods generally have shorter reads and less accuracy than Sanger sequencing, the sequencing data is usually enough to identify the microorganism species and downstream bioinformatics can help identify and remove erroneous sequences. In this study, we used 454 sequencing of 16 S rRNA gene fragments

allowing us to sequence more deeply and profile the bacterial communities more accurately than would be possible with Sanger sequencing.

Since a traditional biological species concept does not apply to prokaryotes, the term "Operational Taxonomical Unit" (OTU) is used to represent a terminal node in phylogenetic analysis. The commonly accepted basis of bacterial taxonomy standard is that bacterial strains with "…approximately 70% or greater DNA-DNA relatedness and with 5 °C or less ΔTm," and that "…phenotypic characteristics should agree with this definition…" (Wayne et al., 1987). In 1994, Stackebrandt and Goebel reported that for species that have 70% or more genomic DNA relatedness, the 16S rRNA gene sequence similarity was generally more than 97% (Stackebrandt et al., 1994). Therefore 97% similarity is commonly accepted and widely used as the species-level cutoff in 16S rRNA analysis.

Phylotype and OTU-based analyses assume the taxonomy represents the functional differences and group sequences by mathematical similarity of sequences. OTU-based analysis is a classification-independent analysis. In OTU-based analysis, the taxonomic information is not included. In this study, OTU-based analysis took 16S rRNA gene distance = 0.03. This distance allows sequences that have similarity $\geq$ 97% to form OTUs. So with distance = 0.03, these OTUs should represent species. In contrast, phylotype-based analysis is a classification-dependent analysis. This approach assumes the names of OTUs are meaningful. In this study, phylotype-based analysis is on genus level instead of species level to get more general taxonomic information. A third type of analysis was also performed: phylogeny-based analysis, which assumes that unique branch length to a specific phylogenetic terminal node is proportional to the functional differences. In this approach a phylogenetic tree is constructed and the sequences are

grouped by evolutionary similarity. In this approach, the UniFrac algorithm was used to compare bacterial communities based on phylogenetic information (Lozupone et al., 2005).

1. 5 Bioremediation Soil Lysimeter Study

In 1996, an intermediate scale experiment was designed for the purpose of assessing HK44 as an *in situ* bio-reporter of PAH degradation. Six lysimeters (4 m depth and 2.5 m diameter) were set up in October of 1996 (Ripp et al., 2000a). Lysimeter 1, 2, 4 were treated with PAH-contaminated soil and inoculated with HK44 (PAH+, HK44+); lysimeter 3 and 5 were treated with clean soil and inoculated with HK44 (PAH–, HK44+); and lysimeter 6 was treated with PAH-contaminated soil only (PAH+, HK44-) (Ripp et al., 2000a).

During the first two years of this experiment, *in situ* bioluminescence was monitored using fiber optics and photomultiplier tubes (PMT). At day 474, naphthalene concentrations were below 10 ppm in most depths of lysimeter 1, 2 and 6 (Layton et al., 2012; Ripp et al., 2000a). During this two year period, as long as the PAH contaminants and supplemental inorganic nutrients were added, bioluminescence emission were detectable (Layton et al., 2012). In addition, after two years, the measured concentration of HK44 was reduced from 1×10^6 colony forming units (cfu)/g soil to 1×10^3 cfu/g soil (Layton et al., 2012). Since these preliminary results suggested that HK44 survived this two year period, the lysimeters were sealed with steel lids and remained untouched until 2000 (Layton et al., 2012; Ripp et al., 2000a; Ripp et al., 2000b).

In 2000, four years after experiment initiation, all six soil lysimeters were unsealed and sampled. Ripp et al. (2000) reported that HK44 and its recombinant genes were still functionally active: bioluminescence was detected via fiber optics (approximately 166, 000 counts/ s) and

PMT (approximately 4300 counts/ s) (Layton et al., 2012; Ripp et al., 2000a). In addition, *luxA* gene mRNA was recovered from soil indicating that HK44 was still present and active (Layton et al., 2012). Thus, the researchers were interested in the impacts of HK44 and PAH contamination over a longer time period. The six lysimeters were then sealed and kept untouched until 2010.

On 2010, 14 years after initiation, these lysimeters were unsealed and sampled again. The soil samples were taken throughout the treatment zone, and were used in this study. Layton et al. (2012) used these same soil samples for some tests and reported that total heterotrophic bacteria concentrations decreased approximate 100 to 1000-fold from the original population, based on plate counts (Layton et al., 2012). In addition, DNA extracted from lysimeter 1, 2, and 4 were used in quantitative PCR targeting HK44 specific *luxA*, *tetA*, and *nahA* gene. *luxA* was quantified in only one sample; *tetA* was quantified in two samples and *nahA* gene was not detected, indicating that HK44 populations have declined and/or disappeared (Layton et al., 2012).

1.6 General Objectives and Hypothesis

PAHs have been found toxic to organisms. Because anthropogenic activities have led to PAH contamination of the environment, it is a major concern. The soil health is closely connected with the bacteria in it. Because aerobic microbial degradation is the primary method of PAH removal from most soils, much PAH remediation research focusses on the microorganisms. However, a limitation of most of these studies is that they only take place over relatively short terms (<1 year). Our study presents a rare opportunity to examine the long term changes in bacterial community structure after a one-time PAHs contamination event at a relatively large experimental scale. What's more, since HK44 is the first GEM permitted for field

testing, this study also provided a unique chance to evaluate the impacts of a GEM release on soil bacteria community. The objectives of this study were:

- To determine the long term impact of a one-time PAH contamination on the indigenous soil bacterial community.
- To determine the long term impact of a one-time release of a biodegradative, bioluminescent GEM (*Pseudomonas fluorescens* HK44) on indigenous soil bacterial communities.

This study tested the hypotheses that PAH contamination have a long term effect on soil bacterial communities, and that the impact of genetically engineered bioremediator HK44 on community structure is negligible in the long term. This study provided a rare chance to assess the impacts of these two factors on soil bacterial community over a long term (approximately fourteen years).

CHAPTER 2 METHODS & MATERIALS

2.1 Experimental Design and Soil Sampling

2.1.1 Experimental Setup

In October of 1996, under sponsorship by the Department of Energy's Natural and Accelerated Bioremediation Research (NABIR) program, *Pseudomonas fluorescens* HK44 was the first genetically engineered microorganism (GEM) in the United States allowed to be released into the field for soil bioremediation tests (Ripp et al., 2000a). Lysimeters were set up in October of 1996 by Ripp et al. (2000a). The 4 m deep and 2.5 m diameter lysimeters were made from epoxy-coated galvanized steel with a stainless steel lid, as shown in Figure 1. They were buried 3 m underground. Another lysimeter filled with water was used to adjust the water levels of these six lysimeters. Lysimeters were arranged around a central area containing computer equipment, monitoring equipment and necessary plumbing (Ripp et al., 2000a).

As shown in Figure 1, from deep to shallow, each lysimeter was filled with a 31 cm layer of graded gravel (stratified bed), a 61 cm layer of coarse sand, a 92 cm layer of clean soil, a 92 cm layer designed for treatment, and a 61 cm uncontaminated soil layer on top, as shown in Figure 1. PAHs and / or HK44 were introduced into the treatment layer of each lysimeter (Figure 2). Therefore, lysimeter 1, 2, 4 were replicates (PAH+, HK44+); lysimeter 3 and 5 were replicates (PAH–, HK44+); and lysimeter 6 was a control (PAH+, HK44-).

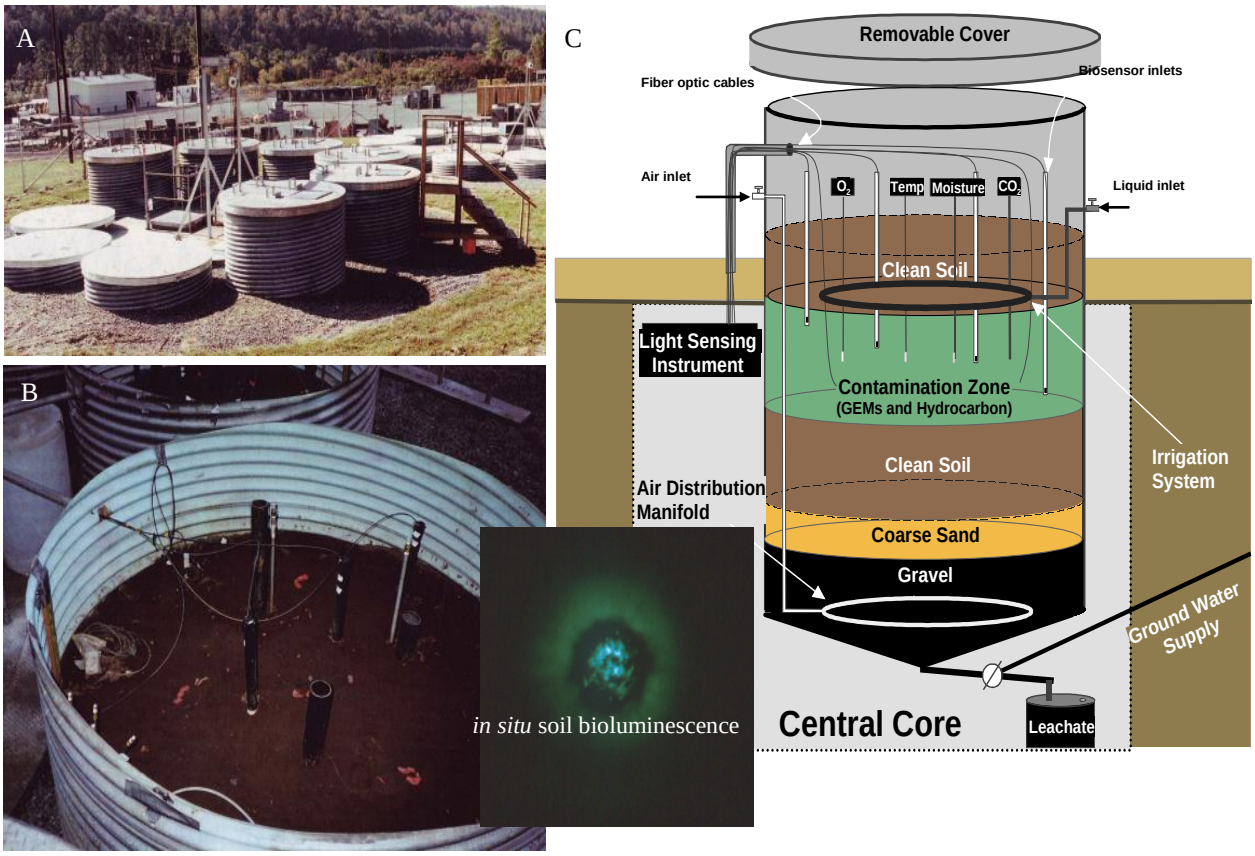

Figure 1. Soil lysimeters.

A) The lysimeter facility. B) Inside view of one of the six soil packed lysimeters. C) Representative schematic of one of the 4 m deep × 2.5 m diameter lysimeters used for the release of *P. fluorescens* HK44 (Sayler et al., 2001; Trogl et al., 2012).

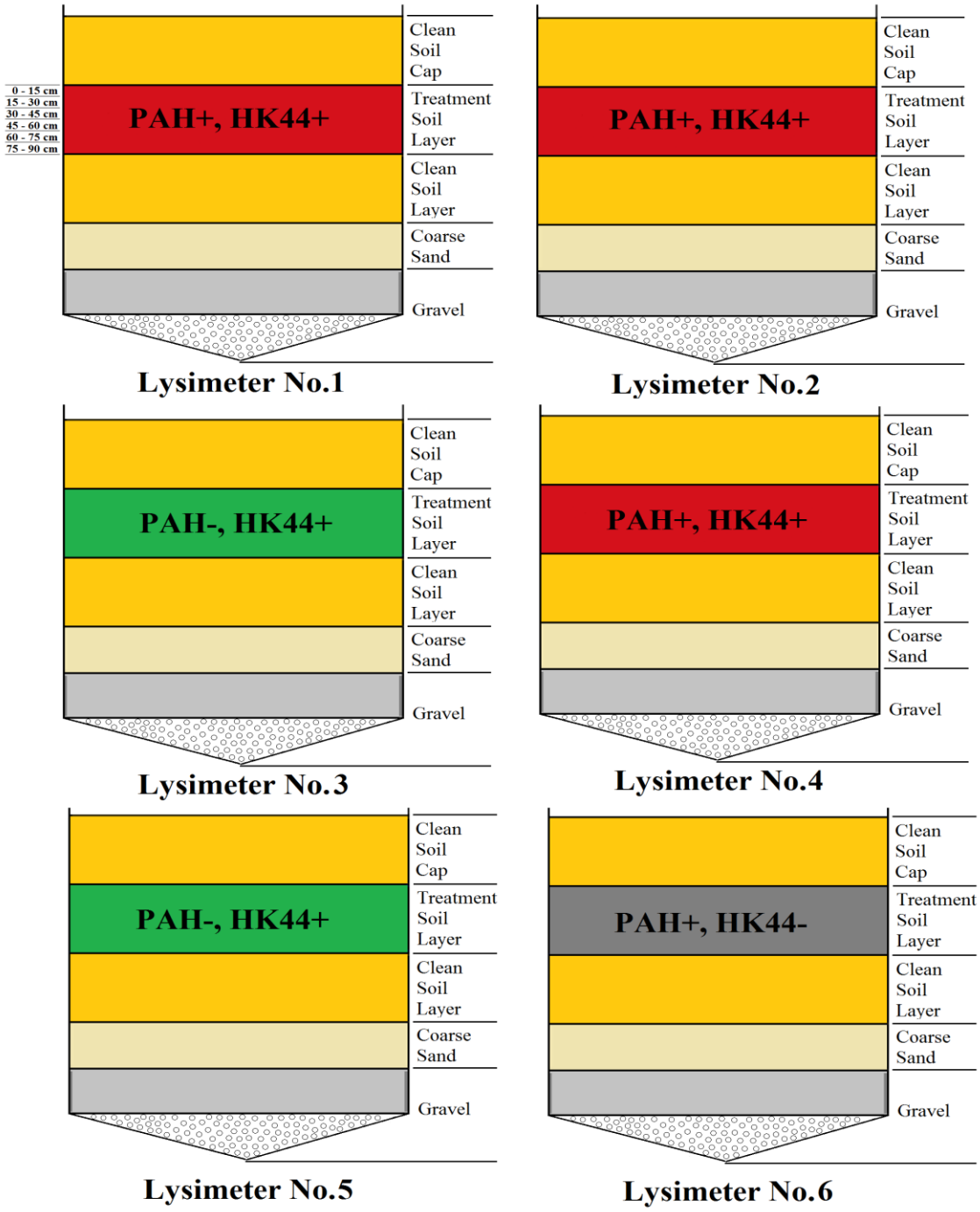

Figure 2. Experimental design: Treatments and Components of soil columns
The colors of treatment zone (red, green, grey) represent the treatment type PAH+, HK44+; PAH-, HK44+; PAH-+, HK44- respectively.

For the PAH+ group (lysimeter 1, 2, 4, 6), a mixture of PAHs was applied at a concentration of 1000 mg naphthalene/kg soil, 100 mg anthracene/kg soil, and 100 mg phenanthrene/kg soil. The design of PAH mixture was to establish contaminated conditions similar to the soils heavily contaminated by PAHs from manufactured gas plant (MGP), where the original host microorganism and plasmid of HK44 isolated from (Ripp et al., 2000a). Because of delays in receiving permission for HK44 release, the treated soil was stored under plastic tarpaulins for 260 days instead of the originally scheduled 100 days. Thus, more than 90% of PAHs were lost through desorption or volatilization. To compensate, an additional 24 kg naphthalene and 2.4 kg anthracene were dissolved in 833 L Exxon UNIVOLT 60 transformer oil, and on day 135 after the test started, lysimeter 1, 2, 4, and 6 each received 208 L oil irrigated directly above the treatment zone, which produced an approximate loading capacity of 1000 mg naphthalene/kg soil and 100mg anthracene/kg soil throughout the treatment zone. The gravity and a pressure produced by lowering the water table were utilized to pull the oil down into the treatment zone. Seven days later, 190 L of inorganic nutrient solution (2 g $NaNO_3$, 0.75 g KH_2PO_4, 0.003 g $FeCl_3$, 0.1 g $MgSO_4$, 0.005 g $CaCl_2$, and 0.25 gNa_2HPO_4 per L of water) was added to each of lysimeters 1, 2, 4, and 6 using the same method (Ripp, Nivens, Werner, et al., 2000).

2.1.2 Information of Soil Sampling and Soil Samples Storage

The lysimeter experiment began on October 30, 1996 (Ripp et al., 2000a). The soil sampling was conducted in June 2010 and is described in Layton et al. (2012). Six soil cores (2.3 cm diameter) were taken from each lysimeter. The 62 cm deep clean soil layer above the treatment zone was discarded, leaving soil cores from the 92 cm treatment zone. These soil cores were fractionated into 6 sections evenly by length. The treatment zone was 92 cm (36 inches)

deep, so each soil sample core was separated to six inch (15.3 cm) long sections. These sections will hence be referred to as 0 - 15 cm, 15 - 30 cm, 30 -45 cm, 45 - 60 cm, 60 - 75 cm, 75 - 90 cm. The soil lysimeters were buried along a hillside so soil water tensions may have been different outside each lysimeter. The lids of these lysimeters should been sealed, but when the lysimeters were opened on June 2010 for sampling, the researchers found that the soils in lysimeters buried on the lower side were wetter than the soils in lysimeters buried at higher locations. Lysimeter 2 and 3 had the highest moisture, and lysimeter 4, 5, 6 were relatively dry when they were sampled. Thus, in lysimeter 4, 5, and 6, deep soils were too solid to take out, so deep core samples were unavailable for lysimeters 4, 5 and 6 (Table 1). In addition, the soils were compacted with time, so these soil cores only reflect relative depth within each lysimeter.

All original soil samples and mixed samples were stored in sealed plastic bags to prevent moisture loss. After soil sampling, a small amount of each soil sample was taken for soil moisture content measurement, and the remainder was stored at 4 °C. On May 2011, a small amount of each soil sample was taken for DNA extraction and soil organic matter content measurement. For these measurements, the six core samples from the same lysimeter and same depth were mixed as a composite sample (Table 2). The rest were stored at 4 °C.

Table 1. Soil sampling of lysimeters' treatment zones in June 2010.

C1 to C6 refers to the replicate soil samples

Depth	Lysimeter 1	Lysimeter 2	Lysimeter 3	Lysimeter 4	Lysimeter 5	Lysimeter 6
Treatment	PAH+, HK44+	PAH+, HK44+	PAH-, HK44+	PAH+, HK44+	PAH-, HK44+	PAH+, HK44-
0 - 15 cm (6 inches)	C1-C6	C1-C6	C1-C6	C1-C6	C1-C4	C1-C6
15 -30 cm (12 inches)	C1-C6	C1-C6	C1-C6	C1-C6	C1-C5	C1-C6
30 - 45 cm (18 inches)	C1-C6	C1-C6	C1-C6	C1, C2, C4, C5	C1-C3, C5, C6	C1
45 - 60 cm (24 inches)	C1-C6	C1-C6	C1-C6	C1, C2, C4, C5	C1-C3, C5, C6	C1
60 - 75 cm (30 inches)	C1-C6	C1-C6	C1-C6	**NA**	C5, C6	**NA**
75 - 90 cm (36 inches)	C1-C6	C1-C6	C1-C6	**NA**	C6	**NA**

Table 2. Labels of composite soil samples of each lysimeter in each depth.

Depth	Lysimeter 1	Lysimeter 2	Lysimeter 3	Lysimeter 4	Lysimeter 5	Lysimeter 6
Treatment	PAH+, HK44+	PAH+, HK44+	PAH-, HK44+	PAH+, HK44+	PAH-, HK44+	PAH+, HK44-
0 - 15 cm (6 inches)	L1-06	L2-06	L3-06	L4-06	L5-06	L6-06
15 -30 cm (12 inches)	L1-12	L2-12	L3-12	L4-12	L5-12	L6-12
30 - 45 cm (18 inches)	L1-18	L2-18	L3-18	L4-18	L5-18	L6-18
45 - 60 cm (24 inches)	L1-24	L2-24	L3-24	L4-24	L5-24	L6-24
60 - 75 cm (30 inches)	L1-30	L2-30	L3-30	**NA**	L5-30	**NA**
75 - 90 cm (36 inches)	L1-36	L2-36	L3-36	**NA**	L5-36	**NA**

2.2 Measurement of Soil Parameters

2.2.1 Soil Moisture

Soil moisture content was determined gravimetrically. 5 g of each soil sample was dried to constant weight at 104°C, and weighted to calculate the soil moisture content.

$$Soil\ Gravimetric\ Water\ Content = \frac{Weight_{wet\ soil} - Weight_{dry\ soil}}{Weight_{dry\ soil}}$$

2.2.2 Soil Total Carbon, Total Nitrogen, C:N Ratio

For each lysimeter, soil samples from every other core section were analyzed for total carbon and total nitrogen. The soil samples were from sections 0 - 15 cm, 30 - 45 cm and 60 - 75 cm deep of each lysimeter (except for lysimeter 4 and 6, whose soil samples from 60 - 75 cm deep are not available). The 16 representative soil samples (mixture of soil cores from same lysimeter and same depth) were air dried and sieved through a No. 40 mesh (0.425 mm) and total carbon and total nitrogen were measured using Carlo Erba combustion CN analyzer at the UT Soil, Plant and Pest Center (Nashville, TN). All organic and inorganic substances in the sample were converted into combustion products by a complete and instantaneous "flash combustion" oxidation. The combustion gases were then passed through a reduction furnace and separated and detected by a thermal conductivity detector to determine concentrations of each individual gas. Total carbon and total nitrogen was reported as %C and %N. C:N ratio was calculated %C: %N.

2.2.3 Soil Organic Matter

A loss-on-ignition method was used to measure soil organic matter (SOM). The 32 composite soil samples were air dried and sieved through a No. 40 mesh (0.425 mm) before measurement. First, ceramic crucibles were heated to 400 °C for 2 hours to remove any moisture

or organic matter left in these containers. Next the crucibles were cooled to room temperature in two desiccators and weighed. Each crucible was labeled so that they were distinguishable later. Approximately 5 g of each soil sample were put into each crucible, and the total weight of each crucible and soil sample was measured. Then these soil samples were oven dried at 105°C for 24 hours to remove soil moisture. The samples were cooled to room temperature in a desiccator and weighed. Next, samples were heated in 400°C for 16 hours so that the organic matter was oxidized and volatilized. Afterwards, the samples were cooled down to room temperature in a desiccator and weighed. The equation used to calculate SOM was:

$$SOM = \frac{Weight_{soil\ after\ 105℃} - Weight_{soil\ after\ 400℃}}{Weight_{soil\ after\ 105℃}}$$

2.3 DNA Extraction and Preparation

2.3.1 DNA Extraction

Soil DNA of 32 composite soil samples was extracted using FastDNA® SPIN Kit for soil (QBiogene, Morgan Irvine, CA). Briefly, for each composite sample 0.5 g soil sample was added into a Lysing Matrix A tube that contained Cell Lysis Solution (CLS), and then homogenized in the FastPrep® Instrument for 40 seconds at speed 6.0 and centrifuged to pellet debris. Next, the supernatant was transferred to a new microcentrifuge tube with Binding Matrix and mixed. After 5 minutes of incubation, 600 µl suspensions of each sample was transferred to a SPIN™ Filter and centrifuged to separate pellet from suspension. The pellet containing DNA was re-suspended in SEWS-M solution and centrifuged through a column to clean DNA. Finally, the bound DNA was re-suspended in DES water and centrifuged to elute DNA. DNA extractions were stored at -20°C.

2.3.2 DNA Quantification

The soil total DNA concentration was measured using Nano drop spectrophotometer (Nanodrop® ND-1000) and the fluorometry-based Quant-it PicoGreen (Invitrogen) dsDNA Reagent and Kits. For the Quant-it PicoGreen assay, a standard curve was made using purified dsDNA with a series of concentrations (1000ng/ml, 100ng/ml, 10ng/ml, 1ng/ml, and 0ng/ml); soil DNA extractions were diluted to within the range of standard curve. Triplicate of each standard curve solution and soil DNA sample dilutions were measured. 100 µl of each sample was added into a well on 96-well microtiter plate. Then 100 µl of working solution of Quant-iT™ PicoGreen® reagent was mixed with each sample and incubated without light for about 2 to 5 minutes in room temperature. After incubation, the fluorescence of samples was measured using a fluorescence microplate reader (BioTek Synergy HT) at standard fluorescein wavelengths (excitation ~480 nm, emission ~520 nm). The fluorescence value of the reagent blank was subtracted from that of each sample and the standard curve was used to calculate DNA concentrations.

2.3.3 Checking for PCR Inhibition

PCR (polymerase chain reaction) was used to check for possible PCR inhibition. 16S rRNA genes were amplified using PCR with universal bacterial primers 338F and 926R (Burke et al., 2006). PCR reactions contained 12.5 µl 2× PCR Master Mix (Fermentas Life Science), 0.4 µM 338F primer and 0.4 µM 926 primer and 1 µl soil total DNA extractions, and adding PCR clean water (Fermentas Life Science) to make final reaction volume equal to 25 µl. The PCR reactions was performed on an Mastercycler® pro thermocycler (Eppendorf) using the following protocol: 94°C for 2 minutes, then 35 cycles of denaturing at 94°C for 15 seconds, annealing at 50°C for 30 seconds, and extending at 72°C for 30 seconds, and a final extension at 72°C for 1

minute after the 35 cycles. Samples were electrophoresed in 1.5% (w/v) agarose gel with Ethidium Bromide, and observed under UV light.

2.4 Quantification of Bacteria

2.4.1 Detection of PAH Degradation Genes

PCR (polymerase chain reaction) was used to determine if there were PAH degradation genes in the soil samples. Two pairs of primers targeting PAH degradation genes were used in PCR. One was DP1 (298 - 311) and DP2 (358 -375) primer set. This primer set was designed to target the Rieske domain sequences of PAH dioxygenase genes (Ni Chadhain et al., 2006), which widely present in PAH dioxygenase genes in bacteria. PCR reactions contained 12.5 µl 2X PCR Master Mix (Fermentas Life Science), 0.4 µM DP1 primer and 0.4 µM DP2 primer, 1 µl soil DNA extractions, and PCR clean water (Fermentas Life Science) to make final reaction volume of 25 µl. The PCR reaction was performed on an Mastercycler® pro thermocycler (Eppendorf) using the following protocol: 94°C for 5 minutes (hot start), then 40 cycles of denaturing at 94°C for 30 seconds, annealing at 48°C for 30 seconds, and extending at 72°C for 30 seconds, and a final extension at 72°C for 1 minute after the 40 cycles. Samples were then electrophoresed in 1.5% (w/v) agarose gel with Ethidium Bromide, and observed under UV light.

The other primer set used was Ac114F (171 - 189) and Ac596R (654 - 670) primer set, which targets the naphthalene dioxygenase genes from gram-negative bacteria (Stach et al., 2002). PCR reactions contained 12.5 µl 2X PCR Master Mix (Fermentas Life Science), 0.4 µM Ac114F primer and 0.4 µM Ac596R primer in the 25 µl reaction, and 1 µl soil total DNA extractions, and PCR clean water (Fermentas Life Science) to make final reaction volume equal to 25 µl. The PCR reactions was performed on an Mastercycler® pro thermocycler (Eppendorf)

using the following protocol: 94°C for 5 minutes (hot start), then 40 cycles of denaturing at 94°C for 30 seconds, annealing at 43°C for 30 seconds, and extending at 72°C for 30 seconds, and a final extension at 72°C for 15 minute after the 40 cycles. Samples were then electrophoresed in 1.5% (w/v) agarose gel with Ethidium Bromide, and observed under UV light.

2.4.2 Quantification of Bacteria 16S rRNA gene copies

qPCR was used to enumerate copies of bacterial 16S rRNA genes using universal bacterial primers 338F and 926R (Burke et al., 2006). Each qPCR reaction contained 12.5 µl Maxima® SYBR Green/Fluorescein qPCR Master Mix (2X) (Thermo Fisher Scientific Inc.), 0.4 µM 338F and 0.4 µM 926R primer in the 25 µl reaction, 5 µl diluted soil total DNA extractions. Nuclease-free water (Fermentas Life Science) was added to make final reaction volume equal to 25 µl. The DNA extraction of each sample was diluted to 2 dilutions within the range of standard curve: 20-fold and 100-fold dilution. For the ones whose 16S rRNA gene copies exceed upper level of standard curve, 500-fold and 1000-fold dilution were used. The PCR reactions was performed on an Mastercycler® pro thermocycler (Eppendorf) using the following protocol: 95°C for 10 minutes (hot start), then 40 cycles of denaturing at 95°C for 5 seconds, annealing at 45°C for 30 seconds, and extending at 72°C for 30 seconds. Standard curves were made from 10^3 to 10^8 copies / reaction.

Each template concentration had 3 replicates for a total of 6 PCRs per sample. If the difference between a measured $\log_{10}$ starting quantity and its replicates was greater than 0.2, it was excluded from mean and standard deviation calculations. If the difference between the two dilution means of the same sample was greater than 0.2, this sample was re-quantified. If the difference was still larger than 0.2, then all the measured bacteria abundances of this sample were averaged to get a mean.

2.5 Community Structure Analysis

2.5.1 16S rRNA Gene Sequencing

2.5.1.1 PCR Optimization

Amplicon libraries were prepared according to the 454 recommended protocols (Roche). DNA extracted from each composite soil sample was amplified with a primer that contains a 454-linker and a unique multiplex identifier (MID) tag in front of a regular 16S rRNA gene primer (338F). Thus, every sample's PCR products have a different barcoded 338F primer, and all samples use the same 926R primer. The primer sequences are shown in Appendix Table 14. PCR reactions were processed using the same conditions determined during initial PCR-inhibition checks. As a positive control, 16S rRNA genes from a mock community containing cloned 16S rRNA gene fragments of 49 known soil bacteria strains were amplified as well. The purpose of sequencing this community alongside was to help determine sequencing quality or PCR error rates. The 16S rRNA gene fragments of these strains were sequenced using Sanger reaction so that we know the “true sequences”, and mock library data from these two methods were compared to estimate the error rate after quality control steps, and optimize the quality control procedure. Also as a control, two of the lysimeter samples (L2-36 and L4-18) were picked to create libraries with two different sets of primers to determine the reproducibility of our library construction and sequencing.

PCR reactions were processed using the same conditions determined during initial PCR-inhibition checks. However, not all the DNA extracts yielded PCR products under these conditions. Amplification of some samples showed either low PCR product concentration or extremely high product concentration. PCR reactions proceed through three stages. First is the

exponential amplification stage, in which the amount of product doubles after every cycle. Next is the leveling stage. As the PCR proceeds, polymerase activity decreases and dNTPs become limiting: since deoxynucleoside triphosphates (dNTPs: where N is A, T, G, or C) are added at the same initial concentration, but the GC percentage of template DNA fragments can vary, using up dNTP(s) at different rates. The last stage is plateau, in which a certain reagent or material becomes depleted. Suzuki et al. (1996) showed that PCR bias occurs mainly on the late stage of PCR reaction (Suzuki et al., 1996). Thus, in our amplicon libraries, those whose electrophoresis result indicating extremely high PCR product concentration were suspected to have reached plateau stage, and were re-amplified with more stringent conditions so that it ended within exponential stage and the amplicons were representative of the initial template concentrations. Reactions yielding no PCR product through this protocol were re-amplified with less stringent PCR conditions to get a yield. The modifications to PCR cycle conditions included: 10-fold dilution of template DNA extraction to dilute possible PCR inhibitors; lowering annealing temperature; increasing or decreasing cycle number.

2.5.1.2 Sequencing Amplicon Library Preparation and Sequencing

After the optimal PCR conditions were determined, each DNA sample was amplified with three 50 µl reaction system and pooled together. PCR reaction contained 25 µl 2X Phusion® High-Fidelity PCR Master Mixes with high-fidelity buffer (Thermo Fisher Scientific Inc.), 0.4 µM unique barcoded forward primer, 0.4 µM 926 primer and 2 µl soil total DNA extractions, and PCR clean water (Fermentas Life Science) to make final reaction volume equal to 50 µl. The sequences of all barcoded primers used are listed in Table 14. The PCR reactions were performed on a Mastercycler® pro thermocycler (Eppendorf). L2-36" and L4-18" duplicate libraries were used as technical controls. The mock library was used as an internal control.

Ultimately, there were 35 16S rRNA gene libraries: 32 unique soil sample DNA libraries, 2 replicates of soil DNA, and the internal control library.

Amplicons from each sample were purified using the Wizard DNA Cleanup Kit (Promega) using the manufacturer's directions. This removed primers and other PCR artifacts. Purified PCR products were quantified using the Quant-it Pico Green dsDNA quantification kit as previously described. The amplicons were then diluted to the same concentration and pooled together. The pooled amplicons were sequenced using 454 sequencing method on the Genome Sequencer FLX System (Roche) by a professional technologist.

2.5.2 Sequencing Data Analysis

Sequencing data were filtered through a quality control pipeline using an open source, command line software package "MOTHUR" (Version 1.27.0). Mothur includes both a pipeline for quality control and error checking of sequences as well as a broad suite of tools for describing and statistically analyzing microbial populations (http://www.mothur.org/). The quality control pipeline identified and removed reads with errors. Errors in sequencing data mainly come from PCR bias, polymerase mistakes, and formation of chimera sequences. In Mothur, low quality sequences (had flow number less than 260 (~ 120basepair)) were detected and removed from each library. In this process, barcode & primer error, ambiguous bases, and sequences with homopolymers longer than eight nucleotides were identified. Then the primers & barcodes were trimmed off of the sequences. Next the retained sequences were aligned using the Silva bacterial database as a reference with a fixed starting position. The Silva database contains datasets of aligned small and large subunit rRNA sequences of Bacteria, Archaea and Eukarya (http://www.arb-silva.de/). The end position was selected where 90% of the sequences ended. Next, sequences within 2 bp difference of a more abundant sequence (considered to be PCR

errors) were detected and merged to the more abundant sequence. Then suspect chimera sequences were detected in each library using the UCHIME algorithm (Edgar et al., 2011). In this method, both the sample sequencing dataset and Silva database were used as references. The chimera checking program searched for 3-way alignment that made two segments of a query sequence most similar to two different reference sequences. For each one of these 3-way alignments, the chimeric score was calculated based on sequence similarity to the two parent sequences. A cutoff criterion was set so that all query sequences with chimeric score higher than that cutoff were considered chimeras. The detected chimeras were removed. The final step of quality control was detecting and removing "contaminants", non-bacterial sequences from Mitochondria, Chloroplast, Archaea, Eukarya, and unknown species. The Mothur codes used are included in Table 15.

After the quality control steps, retained sequencing data were clustered into OTUs by generating a distance matrix: since 97% similarity of 16S rRNA gene is a widely used criteria when classifying bacteria species, distance = 0.03 was used. This distance matrix was used to cluster sequences of each library into OTUs using the average neighbor hierarchical clustering algorithm. The number of times each OTU appeared in each library was calculated. In phylotype-based analysis, similar steps were done to generate distance matrix. The difference is that distance=1 was used so that the sequences in each library were grouped into OTUs at the genus level.

These 34 sequence libraries contained different numbers of sequences retained after the quality filtering steps. Rarefaction curves of each library were calculated, showing relationships between sampled sequence number and OTU number (richness). Since diversity and richness estimates are highly sensitive to library size, libraries were normalized to an equal size for both

OTU-based and phylotype-based distance matrices (Gihring et al., 2012). Since the sampling size of the smallest library is 323, 323 sequences from each library were randomly picked to form normalized libraries.

In phylogeny-based analysis, neighbor joining phylogenetic trees of each library were generated using an adapted clearcut program (Initiative for Bioinformatics and Evolutionary Studies (IBEST) at the University of Idaho). The unique branch length of each tree was used as the indicator of diversity.

2.5.2.1 α Diversity

α diversity estimates diversity and richness in a single library. Chao1 richness estimator (Colwell et al., 1994) was used to determine the minimum richness of each library. This calculator was used to generate rarefaction curves as well. Rarefaction predicts the number of OTUs that would be observed if fewer individuals were sampled. In community structure analysis, as more individuals are sampled, the increase in rate of richness depends on the species evenness in that community. Thus, rarefaction provides a way to compare the richness observed in the same sampling depth of all communities.

$$S_{chao1} = S_{obs} + \frac{n_1(n_1 - 1)}{2(n_2 + 1)}$$

Where,

S_{chao1} is the estimated richness;

S_{obs} means the observed number of species;

n_1 means the number of OTUs that have only one sequence;

n_2 means OTUs that have only two sequences.

To calculate the 95% confidence intervals, the variance were assumed to have a lognormal distribution:

$var(S_{chao1}) = \frac{n_1\,(n_1-1)}{2(n_2+1)} + \frac{n_1(2n_1-1)^2}{4(n_2+1)^2} + \frac{{n_1}^2 n_2(n_1-1)^2}{4(n_2+1)^4}$, when $n_1 > 0$ and $n_2 > 0$;

$var(S_{chao1}) = \frac{n_1\,(n_1-1)}{2(n_2+1)} + \frac{n_1(2n_1-1)^2}{4} + \frac{{n_1}^4}{4S_{chao1}}$, when $n_1 > 0$ and $n_2 = 0$;

$var(S_{chao1}) = S_{obs}\, e^{\left(\frac{-N}{S_{obs}}\right)}(1 - e^{\left(\frac{-N}{S_{obs}}\right)})$, when $n_1 = 0$ and $n_2 > 0$;

If $n_1 > 0$, then:

$$C = \exp\left(1.96\sqrt{\ln\left(1 + \frac{var(S_{chao1})}{(S_{chao1} - S_{obs})^2}\right)}\right)$$

$$LCI_{95\%} = S_{obs} + \frac{S_{chao1} - S_{obs}}{C}$$

$$UCI_{95\%} = S_{obs} + C(S_{chao1} - S_{obs})$$

Otherwise:

$$P = e^{\left(\frac{-N}{S_{obs}}\right)}$$

$$LCI_{95\%} = max\left(S_{obs}, \frac{S_{obs}}{1-P} - 1.96\left(\frac{S_{obs}P}{1-P}\right)^{\frac{1}{2}}\right)$$

$$UCI_{95\%} = \frac{S_{obs}}{1-P} + 1.96\left(\frac{S_{obs}P}{1-P}\right)^{\frac{1}{2}}$$

where,

LCI means Lower bound of confidence interval;

UCI means Upper bound of confidence interval.

Simpson diversity index is an estimator of community diversity. Since the value of Simpson diversity increases as the community diversity decrease, inverse Simpson diversity (1/D) was used in this study. The calculation formula of Simpson diversity index (Simpson, 1949) is showed below (http://www.mothur.org/wiki/Simpson):

$$D_{simpson} = \frac{\sum_{i=1}^{S_{obs}} n_i(n_i - 1)}{N(N-1)}$$

To calculate the lower and upper confidence intervals of Simpson diversity:

$$var(D_{simpson}) = \frac{\sum_{i=1}^{S_{obs}} \left(\frac{n_i}{N}\right)^3 - \left(\sum_{i=1}^{S_{obs}} \left(\frac{n_i}{N}\right)^2\right)^2}{0.25N}$$

$$LCI_{95\%} = D_{simpson} - 1.96\sqrt{var(D_{simpson})}$$

$$UCI_{95\%} = D_{simpson} + 1.96\sqrt{var(D_{simpson})}$$

Where,

S_{obs} is the number of observed OTUs;

n_i is the number of individuals in the ith OTU;

N is the total number of individuals in the community;

LCI means lower bound of confidence interval;

UCI means upper bound of confidence interval.

2.5.2.2 β Diversity

Analysis of β diversity investigates how communities compare to each other. In other words, the β diversity compares membership and structure between communities. Membership describes the OTUs present or absent in a community. Jaccard index was used to compare membership between communities in species and genus level. The Jaccard index (Lozupone et al., 2005) is calculated by (http://www.mothur.org/wiki/Jclass):

$$D_{Jaccard} = \frac{S_{AB}}{S_A + S_B - S_{AB}}$$

Where,

S_{AB} is the number of shared OTUs between communities A and B;

S_A is the number of OTUs in community A;

S_B is the number of OTUs in community B.

In phylogeny-based analysis, the unweighted UniFrac algorithm was used to compute the membership differences between bacterial communities. This algorithm was an extension of Jaccard index on phylogenetic information (Lozupone et al., 2005). The formula of unweighted UniFrac (U) is showed below (http://www.mothur.org/wiki/Unweighted_UniFrac_algorithm):

$$U = \frac{\sum_{i=1}^{N} l_i |A_i - B_i|}{\sum_{i=1}^{N} l_i max(A_i, B_i)}$$

Where,

N is the number of nodes in the tree;

l_i is the branch length between node i and its parent;

A_i equal to 0 if descendants of node i absent in communities A, and equals to 1 if descendants of node i present in communities A;

B_i equal to 0 if descendants of node i absent in communities B, and equals to 1 if descendants of node i present in communities B.

Structure describes both the presence (membership) and relative abundance of groups. Θ_{YC} (Yue et al., 2005) was use to compare the community structures. The calculation formula of Θ_{YC} is showed below (http://www.mothur.org/wiki/Thetayc):

$$D_{\Theta_{YC}} = 1 - \frac{\sum_{i=1}^{S_T} a_i b_i}{\sum_{i=1}^{S_T} (a_i - b_i)^2 + \sum_{i=1}^{S_T} a_i b_i}$$

Where,

S_T is the total number of OTUs in communities A and B;

a_i is the relative abundance of the ith OTU in community A;

b_i is the relative abundance of the ith OTU in community B.

In phylogeny-based analysis, the weighted UniFrac algorithm was used to compute the structure differences between bacterial communities. This algorithm was an extension of Θ_{YC} algorithm on phylogenetic information (Lozupone et al., 2007). The formula of weighted UniFrac (W) is showed below (http://www.mothur.org/wiki/Weighted_UniFrac_algorithm):

$$W = \frac{\sum_{i=1}^{N} I_i \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|}{\sum_{j=1}^{S} L_j}$$

Where,

N is the number of nodes in the phylogenetic tree;

S is the number of sequences represented by the tree;

I_i is the branch length between node i and its parent;

L_j is the total branch length from the root to the tip of the tree for sequence;

A_i and B_i are the number of sequences from communities A and B that descend from the node, respectively;

A_T and B_T are the total number of sequences from communities A and B, respectively.

For membership relationships, pairwise comparisons of shared OTU numbers between each two communities were done, and these comparisons were grouped by the treatments (PAH+ vs. PAH–, HK44+ vs. HK44-,) and within the same treatments. Therefore these pairwise comparisons were put into 3 groups: PAH+ vs. PAH–, HK44+ vs. HK44-, and comparisons within replicates. They were assigned to group number 1, 2, and 3, respectively. The means of each group were compared to determine the relative impacts of PAH addition, HK44 inoculation, and time effects on membership. The higher the shared OTU number, the more similar the memberships of two communities should be.

Calculating all pairwise Jaccard indices and Θ_{YC} distances generated a distance matrix reflecting the relationships between the 34 communities. The dimension number of these two distance matrices equals the number of communities compared. To make the result visual, non-metric multidimensional scaling (NMDS) was used to reduce the 34 dimensions to 2. In this study, NMDS used MOTHUR's NMDS command, which was adapted from the NMDS code in R (written by Sarah Goslee), which used the majorization algorithm (Borg et al., 2005).

In addition, the analysis of similarity (ANOSIM) test was used to measure statistically significant differences between two or more groups of samples. ANOSIM uses ranked dissimilarities between communities, so it is a distribution-free analogue of one-way analysis of variance (ANOVA) (Clarke, 1993). As with ANOVA, ANOSIM tests hypothesis about the similarity of user-defined groups. If the samples between groups have greater dissimilarities than those within each group, the defined groups will be considered significantly different from each other. In this study, each index (Jaccard, Θ_{YC}, unweighted UniFrac, weighted UniFrac) calculated the dissimilarities between each two samples. Jaccard and unweighted UniFrac algorithms calculated the membership dissimilarities between communities. Θ_{YC} and weighted UniFrac algorithm calculated the structure dissimilarities between communities. Two grouping approaches were used to test whether groups were significantly difference. In one approach, libraries were grouped by treatment (PAH + and PAH– groups; HK44+ and HK44– groups). In the other approach, libraries were grouped by soil moisture of lysimeters. Since soil moisture contents of these lysimeters were different, non-parametric significant test was used to determine the significant difference of moisture between lysimeters. Then the samples were grouped by moisture.

In OTU-based analysis and phylotype-based analysis, Jaccard distance matrices showed the membership relationships between libraries, and Θ_{YC} distance matrices showed the community structure relationships between libraries. In phylogeny-based analysis, unweighted UniFrac distance matrix showed membership relationships of libraries, and weighted UniFrac distance matrix showed structure relationships between libraries. These distance matrices were tested to determine if bacterial community membership and structure had significant difference

between chosen groups. In summary, ANOSIM was used to test if PAH, HK44 or soil moisture caused significant difference on soil bacterial community membership or structure.

2.5.3 Error Analysis

In this study, a positive control for library preparation and sequencing analysis was used to assess error rate of sequencing data (Schloss et al., 2011). Since this library was an artificially constructed soil bacterial community, it will be referred hereafter as "mock library". It was built using genomic DNA from 49 strains of typical soil bacteria isolated from the Kellogg Biological Station, Michigan, and provided by Dr. M. Radosevich, University of Tennessee. The 16S rRNA genes of these strains were cloned and individually sequenced using Sanger (chain terminator sequencing) sequencing to be used as reference sequences in error analysis. The mock library was prepared by mixing the genomic DNA of the 49 strains. In the 454 sequencing amplicon preparation, sequencing, and quality filtering, this mock library was amplified and treated the same as the other 34 soil sample libraries. These 454 sequences were taken as query sequences, and compared with the Sanger reference sequences. Based on the comparison of these two datasets, the accuracy of 454 pyrosequencing was estimated.

CHAPTER 3 RESULTS

3.1 Soil Properties

Soil total carbon, total nitrogen, C:N ratio, moisture content and organic matter content of the composite soil samples from each lysimeter were measured to evaluate whether the soil properties had significant difference within or between treatments (PAH+/–, HK44+/–). For each parameter measured, the mean of all depths within lysimeters of the same treatment were compared. Lysimeter 1, 2, and 4 (L1, 2, 4) were treated with PAHs (PAH+) and inoculated with HK44 (HK44+); lysimeter 3 and 5 (L3 & 5) had clean soil inoculated with HK44+ (PAH–, HK44+); lysimeter 6 (L6) was treated with PAHs but no HK44 (PAH+, HK44–).

3.1.1 Soil Total Carbon, Total Nitrogen, C:N Ratio, and Soil Organic Matter

Carbon and nitrogen are important carbon and energy sources for soil bacterial, so soil total carbon, total nitrogen and C:N Ratio were measured to evaluate if the soil nutrient composition was different within or between treatments. For each lysimeter, we analyzed soil samples from every other depth for soil total carbon, total nitrogen and C:N ratio measurement (Table 3). In addition, soil organic matter (SOM) contents of all 34 soil samples were measured to assess changes in soil properties. Soil organic matter was measured by loss on ignition (Figure 3). Since the deepest two soil samples in lysimeters 4 and 6 were not available, there were 16 samples in total. For each measurement (total C, total N, soil organic matter), the means of all 16 samples were used to test if the measured parameter had difference between six lysimeters or between treatments (PAH+ vs. PAH–; HK44+ vs. HK44–). A non-parametric Mann–Whitney U test was used to estimate whether there is significant difference between two groups; another non-parametric hypothesis test, Kruskal-Wallis one-way ANOVA, was used to assess if there

Table 3. Soil total carbon, total nitrogen and C:N ratios in the six lysimeters (L1 – L6). STD means standard deviation.

Depth (cm)	Measurement	PAH+, HK44+			PAH–, HK44+		PAH+, HK44–
		L1	L2	L4	L3	L5	L6
0 - 15	C %	0.388	0.233	0.158	0.044	0.027	0.099
0 - 15	N%	0.0068	0.0073	0.0080	0.0081	0.0066	0.0074
0 - 15	C:N Ratio	56.82	31.70	19.85	5.37	4.15	13.38
30 - 45	C %	0.303	0.363	0.488	0.019	0.017	0.292
30 - 45	N%	0.0104	0.0083	0.0135	0.0043	0.0075	0.0095
30 - 45	C:N Ratio	29.03	43.90	36.17	4.45	2.27	30.79
60 - 75	C %	0.345	0.341	NA	0.019	0.015	NA
60 - 75	N%	0.0143	0.0073	NA	0.0059	0.0057	NA
60 - 75	C:N Ratio	24.07	46.62	NA	3.26	2.60	NA
Mean ± STD of Lysimeter	C%	0.345 ± 0.04	0.312 ± 0.07	0.323 ± 0.23	0.027 ± 0.01	0.020 ± 0.01	0.195 ± 0.14
	N%	0.011 ± 0.004	0.008 ± 0.001	0.011 ± 0.004	0.006 ± 0.002	0.007 ± 0.001	0.008 ± 0.001
	C:N Ratio	36.64 ± 17.65	40.74 ± 7.95	28.01 ± 11.54	4.36 ± 1.05	3.00 ± 1.00	22.09 ± 12.31

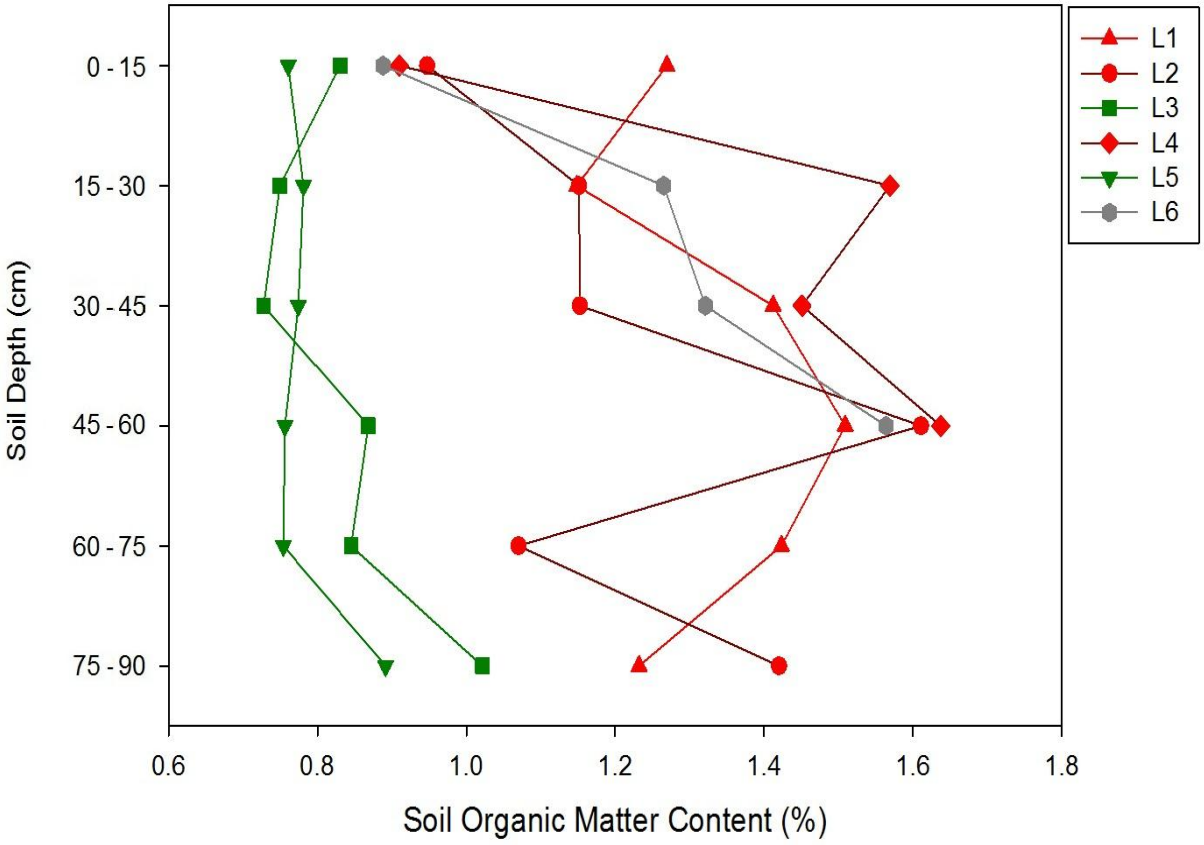

Figure 3. Soil Organic Matter (SOM) content in the six lysimeters.
Data points belong to different treatment groups are labeled by different color: treatment PAH+, HK44+ are red, treatment PAH–, HK44+ are green, treatment PAH+, HK44– are grey.

were significant differences between three or more groups. For the comparisons that showed difference between groups, the post hoc means comparison Tukey's studendized range (HSD) test using ranked data was done to determine the significantly different groups.

Compared to the PAH–, the PAH+ group had significantly higher soil total carbon contents (Mann–Whitney U test, Z= –3.2021, p=0.0014; HSD test, F=33.77, p<0.0001), significantly higher soil total nitrogen contents (Z= –2.2252, p=0.0261; HSD test, F=7.42, p=0.0165), significantly higher C:N ratio (Z= –3.1997, p=0.0014; HSD test, F=33.60, p<0.0001), and significantly higher SOM (Z= –4.5190, p<0.0001; HSD test, F=59.41, p<0.0001) compared to the PAH– group.

Comparisons between HK44+/– groups showed no significant difference of soil total carbon contents (Mann–Whitney U test, Z=0.0000, p=1.0000), soil total nitrogen contents (Z=0.5561, p=0.5781), C:N ratio (Z=0.0794, p=0.9367) or SOM (Z=1.0550, p=0.2914).

Comparisons between replicates (lysimeter 1, 2, and 4) showed no significant difference of soil total carbon contents (Kruskal-Wallis test, χ^2=0.2500, p=0.8825), soil total nitrogen (χ^2=1.1526, p=0.5620), C:N ratio (χ^2=1.4444, p=0.4857) or SOM (χ^2=1.7333, p=0.4204).

Comparisons between replicate lysimeter 3 and 5 showed no significant difference in soil total carbon content (Mann–Whitney U test, Z=0.8856, p=0.3758), soil total nitrogen content (Z=0.0000, p=1.0000), C:N ratio (Z=1.3093, p=0.1904) or SOM (Z=0.1834, p=0.8545).

In summary, the soil total carbon, total nitrogen, C:N ratio and organic matter content were significantly higher in PAH+ treatments compared to the PAH– treatment. Comparisons between HK44 +/– and between replicates showed no significant difference. Since PAHs are

hydrocarbons, and lysimeters treated with PAHs also received an inorganic nutrient solution, these results show that after 14 years, soil nutrition differences were still altered.

3.1.2 Soil Moisture

In this study soil moisture content was not a treatment, but a controlled factor, so soil moisture should be similar across these six lysimeters. The soil moisture content of each soil cores was measured to estimate how well the controlled soil parameters were maintained during this experiment. Soil moisture content of all the 32 soil samples from each soil depth and each lysimeter were measured, and data was reported as % soil moisture. Soil samples from each of the replicate soil cores were measured separately (Figure 4).

There is significant difference of soil moisture between these lysimeters (Kruskal-Wallis test, χ^2=104.4302, p<0.0001). Between these lysimeters, the significant differences of soil moisture are (HSD test, F=56.34, p<0.0001): lysimeter 2 and 3 > lysimeter 1 > lysimeter 5 > lysimeter 4 and 6.

Soil moisture content also changed with depth: the moisture increased with depth. Analysis of each lysimeter by depth showed that soil from lysimeter 1, 2, 3 and 4 had higher soil moisture content as depth increased, but soil moisture in lysimeter 5 and 6 showed no significant difference along the depth.

These results indicate that the soil moisture across these 6 lysimeters was not consistent, and this should be taking into consideration when analyzing possible factors affecting the bacterial community.

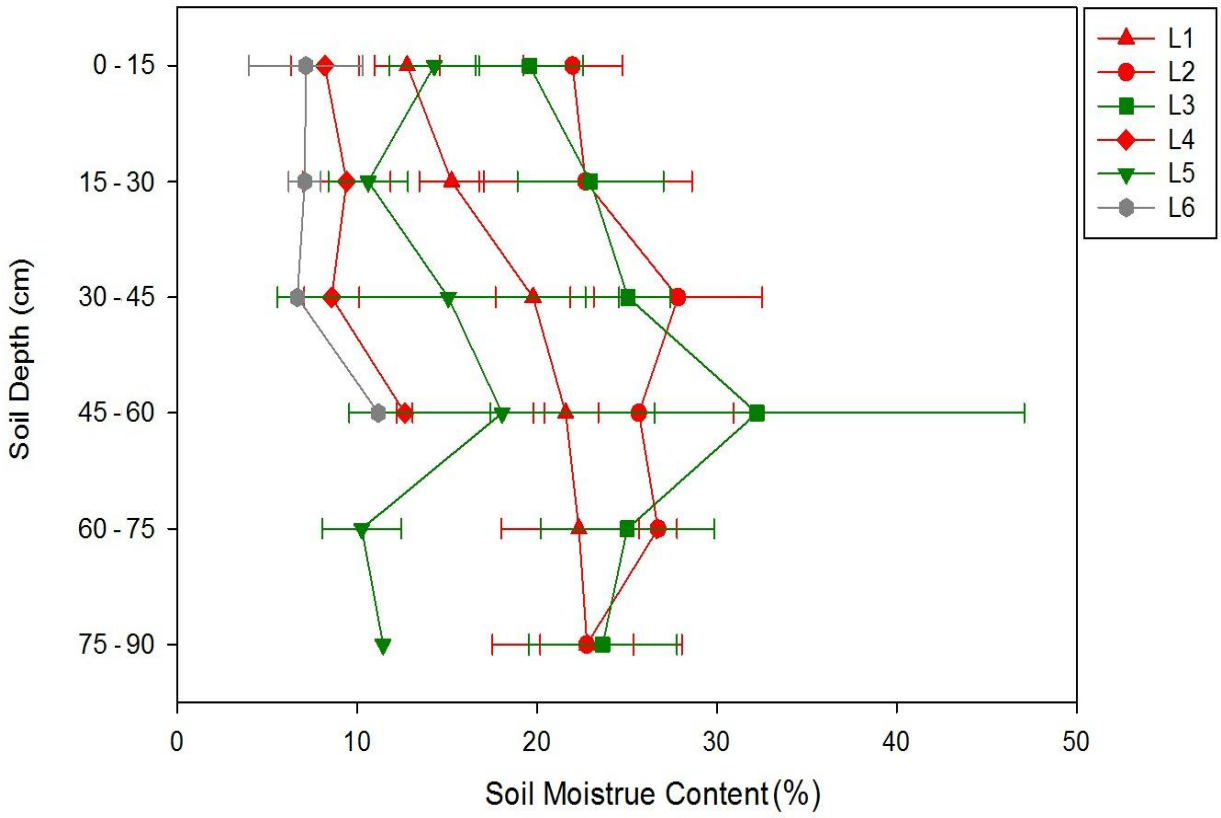

Figure 4. Percent soil moisture content in the six lysimeters.
Points are the means of replicate soil cores (n=6 if available; some sample from L4, 5, and 6 were unavailable) from each lysimeter. Error bars are standard deviations of each mean. PAH+, HK44+ are red, treatment PAH–, HK44+ are green, treatment PAH+, HK44– are grey.

3.2 Detection of Bacterial PAH Degradation Genes

In this study, PAH degradation genes were not amplifiable by PCR (using primers Rieske_f and Rieske_r (Ni Chadhain et al., 2006), Ac114F and Ac596R (Lozada et al., 2008) targeting PAH dioxygenase genes) (data not shown). This indicated that PAH degradation genes targeted by these primer sets were below a detectable concentration in the soil.

3.3 Abundances of Bacteria

Bacterial abundances were estimated by quantification of 16S rRNA genes by qPCR. They are expressed as copies of 16S rRNA gene per gram dry weight soil (Table 4).

The PAH+ group showed significantly higher abundances of 16S rRNA genes than the PAH– group (Mann–Whitney U test, Z= –2.3939, p=0.0167; HSD test, F=6.94, p=0.0132).

The16S rRNA gene concentrations between HK44+/– treatments showed no significant differences (Mann–Whitney U test, Z= –1.2251, P=0.2205; HSD test, F=1.60, P=0.2154).

There is no significant difference of 16S rRNA gene concentrations within each treatment (PAH+, HK44+ group: Kruskal-Wallis test, χ^2=0.7537, P=0.6860; PAH–, HK44+ group: Mann–Whitney U test, Z= –1.2010, P=0.2298).

Table 4. Bacteria abundances in each of the six lysimeters.
The bacteria abundance of each sample is shown as the mean ± standard deviation of replicate PCRs (n=6) on composite soil samples

Sample Depth (cm)	Bacteria 16S rRNA gene copies/g dry soil (Mean ± Standard Deviation)					
	L1	L2	L3	L4	L5	L6
0 – 15	4.50E+08 ± 9.23E+01	9.02E+08 ± 1.37E+02	2.38E+11 ± 1.27E+02	6.99E+09 ± 8.81E+01	1.00E+11 ± 9.57E+01	3.32E+08 ± 9.84E+01
15 – 30	3.22E+10 ± 1.06E+02	1.43E+11 ± 1.06E+02	6.61E+08 ± 1.07E+02	4.67E+10 ± 1.11E+01	9.67E+08 ± 6.79E+02	4.04E+09 ± 9.74E+01
30 – 45	1.17E+11 ± 9.94E+01	2.04E+11 ± 4.29E+02	5.18E+09 ± 6.43E+02	3.37E+10 ± 1.11E+02	7.14E+09 ± 1.64E+03	1.19E+10 ± 8.92E+01
45 – 60	4.78E+11 ± 1.47E+02	3.18E+12 ± 2.01E+02	6.56E+08 ± 1.19E+02	2.41E+11 ± 2.68E+02	1.32E+08 ± 1.08E+02	1.06E+10 ± 1.03E+02
60 – 75	3.63E+11 ± 2.64E+02	8.68E+09 ± 1.24E+02	1.67E+10 ± 1.20E+02	NA	1.02E+07 ± 9.04E+01	NA
75 – 90	4.05E+11 ± 1.45E+02	1.88E+12 ± 1.72E+02	2.41E+10 ± 1.14E+02	NA	3.65E+08 ± 9.64E+01	NA
Mean ± STD of Lysimeter	1.41E+11 ± 8.09E+02	2.28E+11 ± 1.52E+03	6.97E+09 ± 7.58E+02	5.87E+10 ± 4.13E+02	1.40E+09 ± 1.86E+03	3.61E+09 ± 4.18E+02

3.4 Quality Control and Error Analysis of Sequencing Data

3.4.1 Optimization of PCR Conditions

Since the primer set used to build each library had a unique sequence (due to barcoding sequences), optimal PCR conditions for each sample could be different, and was optimized for each template – primer combination. The optimal PCR conditions for each sample are showed in Table 5. Primer sequences are in Table 14. These libraries had different concentrations after amplification (Table 6). They were diluted to the same concentration as the lowest one (9×10^8 molecules/µl) and pooled together. The pooled sample was used for 454 pyrosequencing.

3.4.2 Error Analysis

After optimization, error analysis of mock library showed that the final error rate is 6.4%. The final mock library contains 150 unique sequences. 116 of the 150 sequences were found to be combination of 2 unique sequences that used to construct the mock library. Thus, 116 chimeras were not detected in the quality control pipeline; and 33 of the 150 sequences in the mock library were identical to the sequences that used to construct the mock library.

OTU-based analysis grouped these 150 sequences into 90 OTUs (species level). Since the smallest library of all 35 libraries (34 sample libraries and the mock library) had 323 sequences, all libraries were normalized to that size (323 sequences per library) by random subsampling. The subsampled mock library had (35. 1 ± 3.8) OTUs in species level, while comparison with the real mock library (sequenced using Sanger method) showed that this 454 sequenced mock library retained 24 OTUs in species level.

Table 5. Optimal barcoded PCR conditions for 34 soil DNA samples and a mock library. The mock library is an artificial bacteria DNA library. The cells marked in grey are the libraries with duplicates (same soil DNA sample amplified with two different barcoded primer sets).

Sample	Barcode	Cycle Number	Annealing Temperature	Dilution
L1-06	AA	35	45	No
L1-12	AB	30	50	No
L1-18	AC	30	50	No
L1-24	AD	35	45	10-fold
L1-30	AE	30	50	No
L1-36	BB	35	50	No
L2-06	AG	37	50	No
L2-12	AL	30	50	No
L2-18	AH	30	50	No
L2-24	AI	30	50	No
L2-30	AK	35	50	No
L2-36	AJ	30	50	No
L2-36	BG	30	50	No
L3-06	AM	35	50	No
L3-12	AN	30	50	No
L3-18	AO	40	45	No
L3-24	AP	35	50	No
L3-30	AQ	30	50	No
L3-36	AR	40	45	No
L4-06	AS	35	45	No
L4-12	AT	35	50	No
L4-18	AU	30	50	No
L4-18	BI	30	50	No
L4-24	AV	35	50	No
L5-06	AW	30	50	No
L5-12	AX	40	50	No
L5-18	AY	40	50	No
L5-24	BJ	40	45	No
L5-30	AF	40	45	No
L5-36	BH	40	50	No
L6-06	BC	30	50	No
L6-12	BD	35	50	No
L6-18	BE	30	50	No
L6-24	BF	40	45	10-fold
Mock Library	BK	30	50	No

Table 6. Concentrations of amplicon libraries before dilution and pooling. The molecules/µl column is the calculated DNA concentrations of each amplicon library.

Sample Name	Nano Drop Concentration (ng/µl)	260/280	260/230	PicoGreen Real Concentration (ng/µl)	Molecules/µl
L1-06	6.87	2.95	0.04	3.86	5.66E+09
L1-12	16.55	1.78	0.65	4.25	6.23E+09
L1-18	15.86	1.96	0.90	3.53	5.17E+09
L1-24	5.32	2.17	0.03	1.06	1.56E+09
L1-30	35.34	33.70	0.60	12.59	1.85E+10
L1-36	54.22	4.27	0.82	21.44	3.14E+10
L2-06	9.77	2.17	0.12	6.21	9.10E+09
L2-12	11.79	2.26	0.15	6.72	9.84E+09
L2-18	25.25	1.78	0.24	9.57	1.40E+10
L2-24	11.15	2.10	0.08	3.25	4.76E+09
L2-30	4.29	2.53	0.06	0.70	1.02E+09
L2-36 **A**	15.73	1.97	0.33	4.14	6.07E+09
L2-36 **B**	13.21	2.24	0.16	5.21	7.63E+09
L3-06	6.65	2.71	0.05	1.12	1.64E+09
L3-12	3.01	3.24	0.05	0.95	1.40E+09
L3-18	17.95	2.09	0.17	13.32	1.95E+10
L3-24	5.75	3.88	0.04	3.25	4.76E+09
L3-30	4.83	3.14	0.05	2.91	4.26E+09
L3-36	11.86	2.29	0.10	6.55	9.59E+09
L4-06	7.40	2.39	0.07	2.63	3.85E+09
L4-12	6.36	2.59	0.11	2.91	4.26E+09
L4-18 **A**	7.26	2.48	0.03	3.69	5.41E+09
L4-18 **B**	7.88	3.51	0.04	4.48	6.56E+09
L4-24	10.43	2.34	0.12	3.25	4.76E+09
L5-06	7.96	2.45	0.10	3.86	5.66E+09
L5-12	4.59	5.62	0.02	0.62	9.03E+08
L5-18	7.33	5.72	0.02	2.97	4.35E+09
L5-24	5.37	4.57	0.08	2.13	3.12E+09
L5-30	6.03	2.20	0.06	2.07	3.03E+09
L5-36	9.22	3.03	0.07	2.63	3.85E+09
L6-06	16.33	2.03	0.17	11.81	1.73E+10
L6-12	19.09	2.18	0.12	11.59	1.70E+10
L6-18	7.00	2.46	0.02	1.29	1.89E+09
L6-24	6.95	2.38	0.03	0.78	1.15E+09
Mock Library	66.14	1.91	2.11	23.06	3.38E+10

3.5 Community Diversity of Each Bacterial Library

3.5.1 α Diversity of 34 Libraries

3.5.1.1 OTU-based Analysis

The OTU-based analysis grouped sequences with sequencing similarity ≥ 97% into one OTU. The rarefaction curves of each sequencing library were calculated based on OTU analysis (Figure 5). From Figure 5 we could find that as sampled sequence number increased, the richness of each library increased. Since library sizes were variable, rarefaction curves ended at different sampling depth.

After normalization to the same number of sequences per library (n=323) by random subsampling, the Chao1 richness estimate and inverse Simpson diversity of each library was calculated, and was reported as means and 95% confidence intervals of each mean (Figures 6 and 7). The community richness in each lysimeter showed no significant difference by depth (Figure 6). Since soils in these lysimeters were compressed, comparison of the same depth across 6 lysimeters is not applicable. Thus, for the rest of the statistical analysis of bacterial community richness and diversity, the mean of the 6 depths of the same lysimeter was used.

Given $\alpha = 0.05$, there is no significant difference of mean richness (Chao1) between 6 lysimeters (Kruskal-Wallis test, $\chi^2=3.1841$, $p=0.6716$). The inverse Simpson index is an estimate of the diversity of each bacterial community. As the soil depth increases, the community diversity of every lysimeter showed no significant difference along depth (Figure 7). Since soils in these lysimeters were compressed, comparison of the same depth across 6 lysimeters is not

applicable. In addition, given $\alpha = 0.05$, there is no significant difference of mean diversity (inverse Simpson index values) between 6 lysimeters (Kruskal-Wallis test, χ^2=1.8115, p=0.8746).

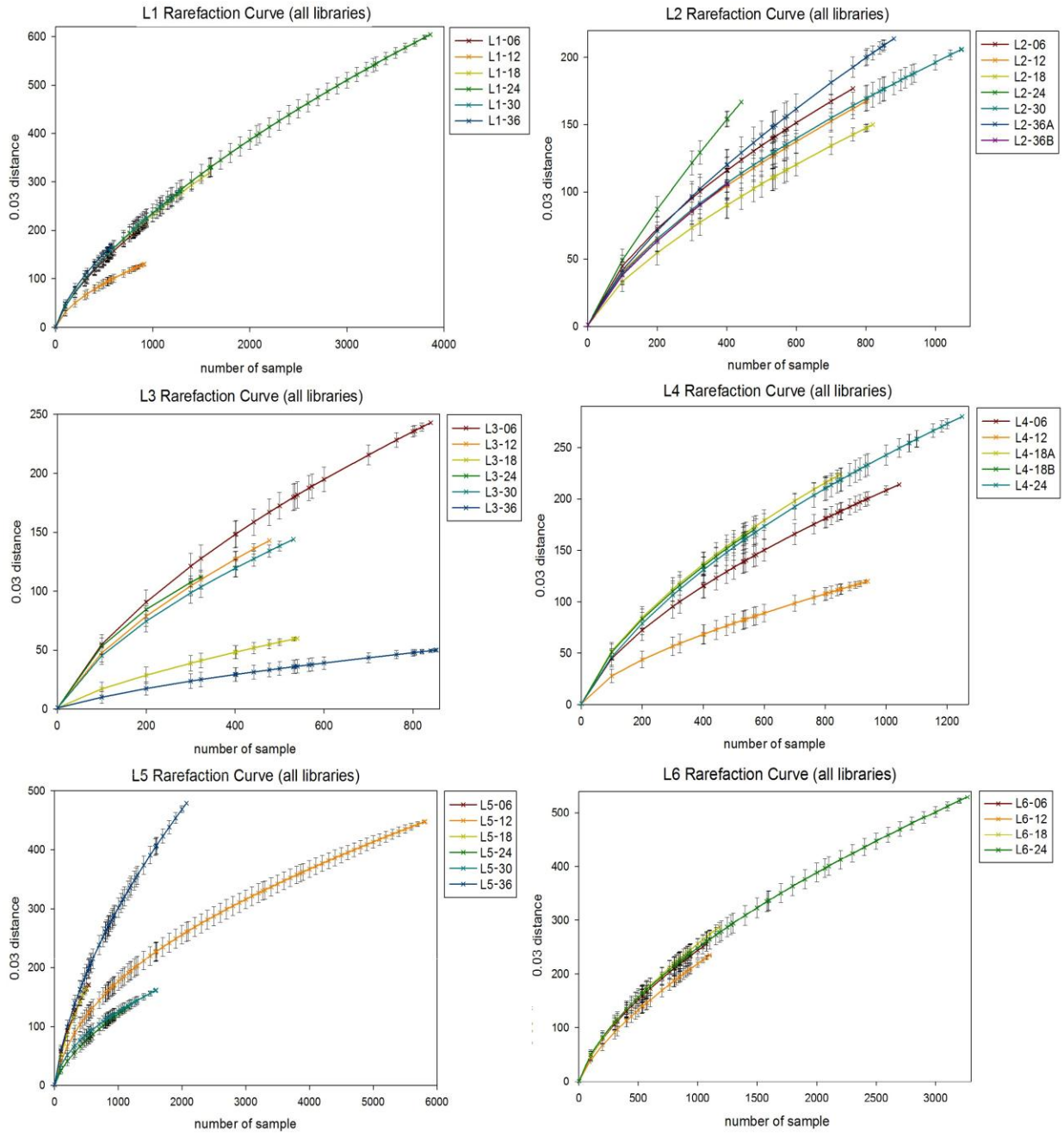

Figure 5. Rarefaction curves of each soil sample library from OTU-based analysis. The rarefaction curves are grouped by lysimeter, and colored by soil sample depth. Duplicate libraries are assigned to different color: for L2 and L4, "A" and "B" are the duplicate libraries from a single sample.

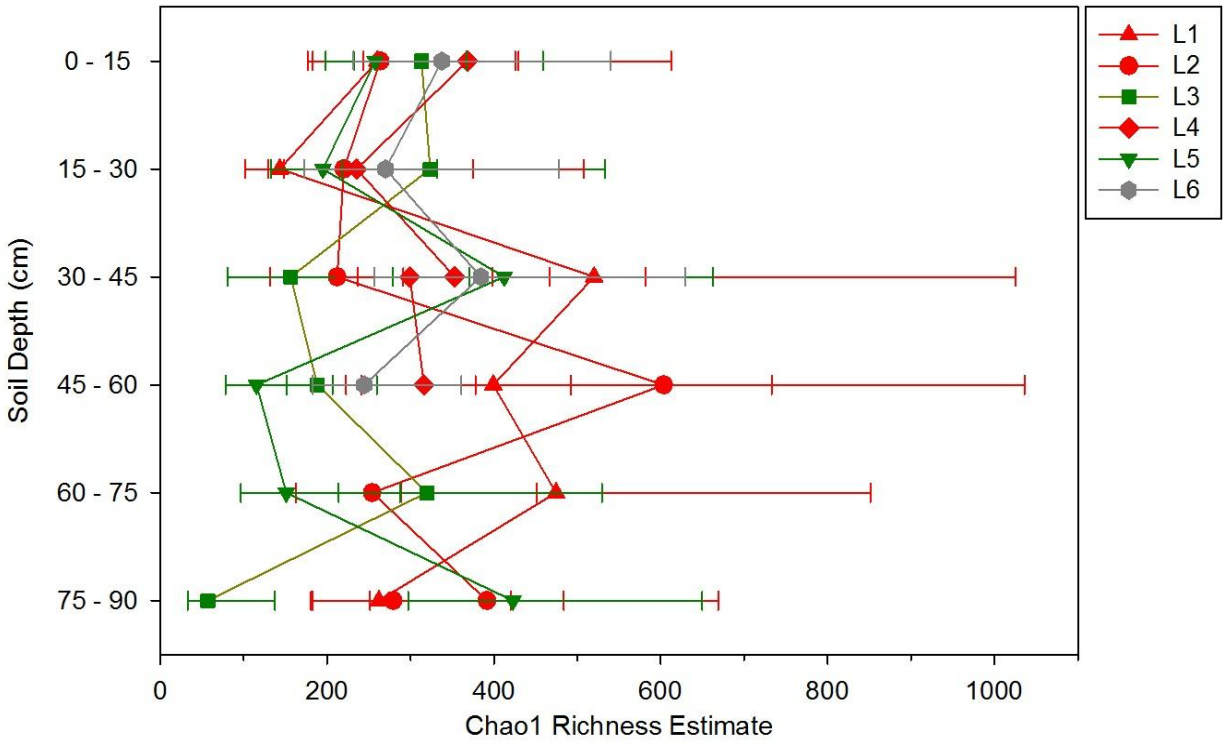

Figure 6. Species level richness estimate of all 34 soil bacterial DNA libraries at size=323 sequences. This is the size of smallest library before sub-sampled. Each data point represent the Chao1 richness estimate value of a specific library at sample size = 323, and error bar represent calculated upper and lower confidence interval of Chao1 richness in each library. Data points belong to different treatment groups are labeled by different color: treatment PAH+, HK44+ are red, treatment PAH-, HK44+ are green, treatment PAH+, HK44- are grey.

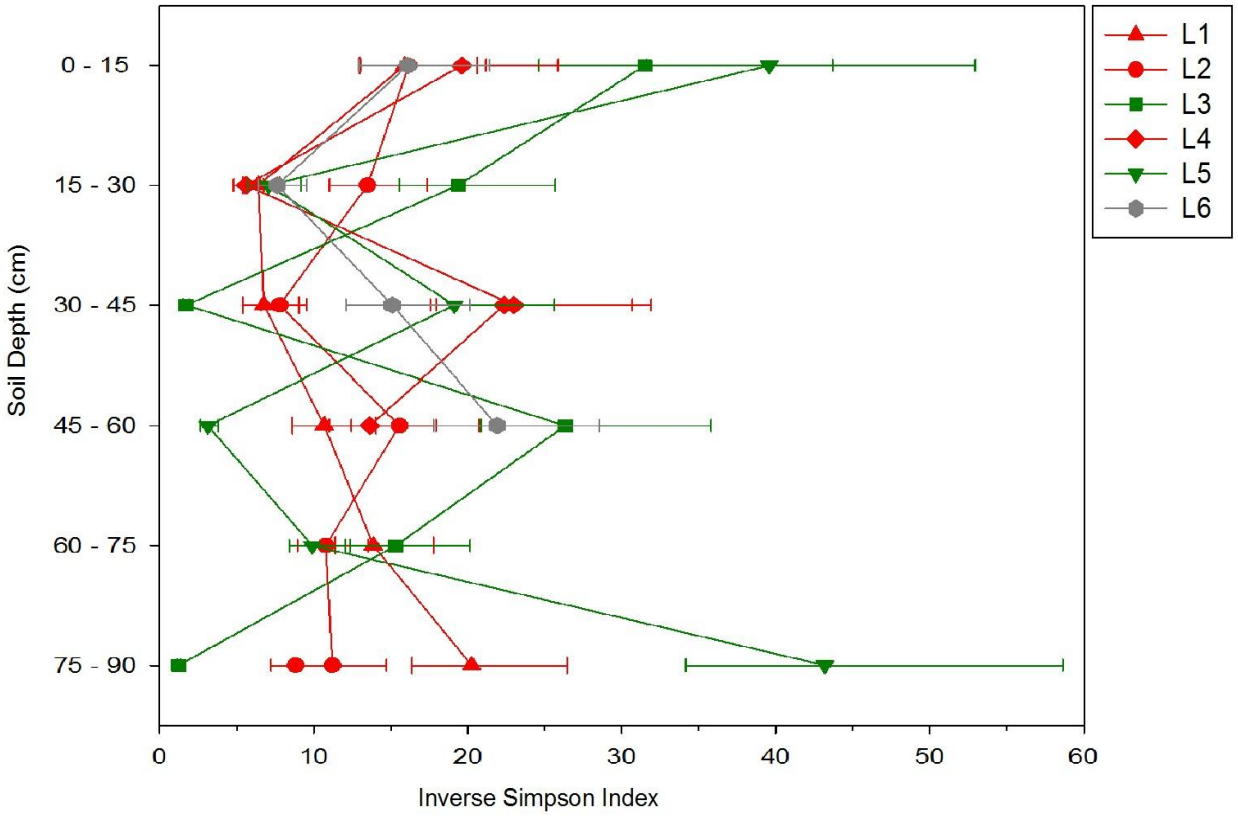

Figure 7. Species level inverse Simpson diversity of 34 soil bacterial DNA libraries at size=323 sequences. This is the size of smallest library before sub-sampled. Each data point represents the inverse Simpson index value and error bars represent calculated upper and lower confidence interval of inverse Simpson diversity index in each library. Data points belong to different treatment groups are labeled by different color: treatment PAH+, HK44+ are red, treatment PAH-, HK44+ are green, treatment PAH+, HK44- are grey.

3.5.1.2 Phylotype-based Analysis

For the phylotype-base analysis, sequences belonging to the same genus were grouped. In Figure 8 rarefaction curves of each sequencing library calculated based on phylotype are showed by lysimeter. As sampled sequence number increases, the richness of each library tends to increase. Since library sizes are different before normalizing these 35 libraries, these rarefaction curves end at different sampling depth.

The Chao1 richness estimate and inverse Simpson index of each library at a sample size of 323 are shown in Figure 9 and 10. Richness is consistent along depth (Figure 9). Given $\alpha = 0.05$, there is no significant difference of richness (Chao1 values) across 6 lysimeters (Kruskal-Wallis test, $\chi^2=7.5761$, $p=0.1812$).

Figure 10 showed the inverse Simpson diversity (1/D) index of all libraries at sample size = 323. Given $\alpha = 0.05$, the bacterial community diversities had no significant differences between these 6 lysimeters (inverse Simpson index values) across 6 lysimeters (Kruskal-Wallis test, $\chi^2=3.9449$, $p=0.5574$).

Phylotype analysis showed the genera with abundances > 1% of total sequences belong to *Acidobacteria* phylum, *Firmicutes* phylum, *Proteobacteria* phylum and an unclassifiable phylum (Table 7). The abundances distribution of these genera was also shown in Table 7. Red means high abundance, yellow means medium abundance, green means low abundance.

Because HK44 belonged to the *Pseudomonas* genus, distribution of *Pseudomonas* genus was of concern to us, and was summarized (Table 8). Among the most abundant genera, there is one genus unclassifiable at the phylum level. This genus showed similar distribution to a genus in *Bradyrhizobiaceae* family, and they both showed extremely high abundances in sample L5-12

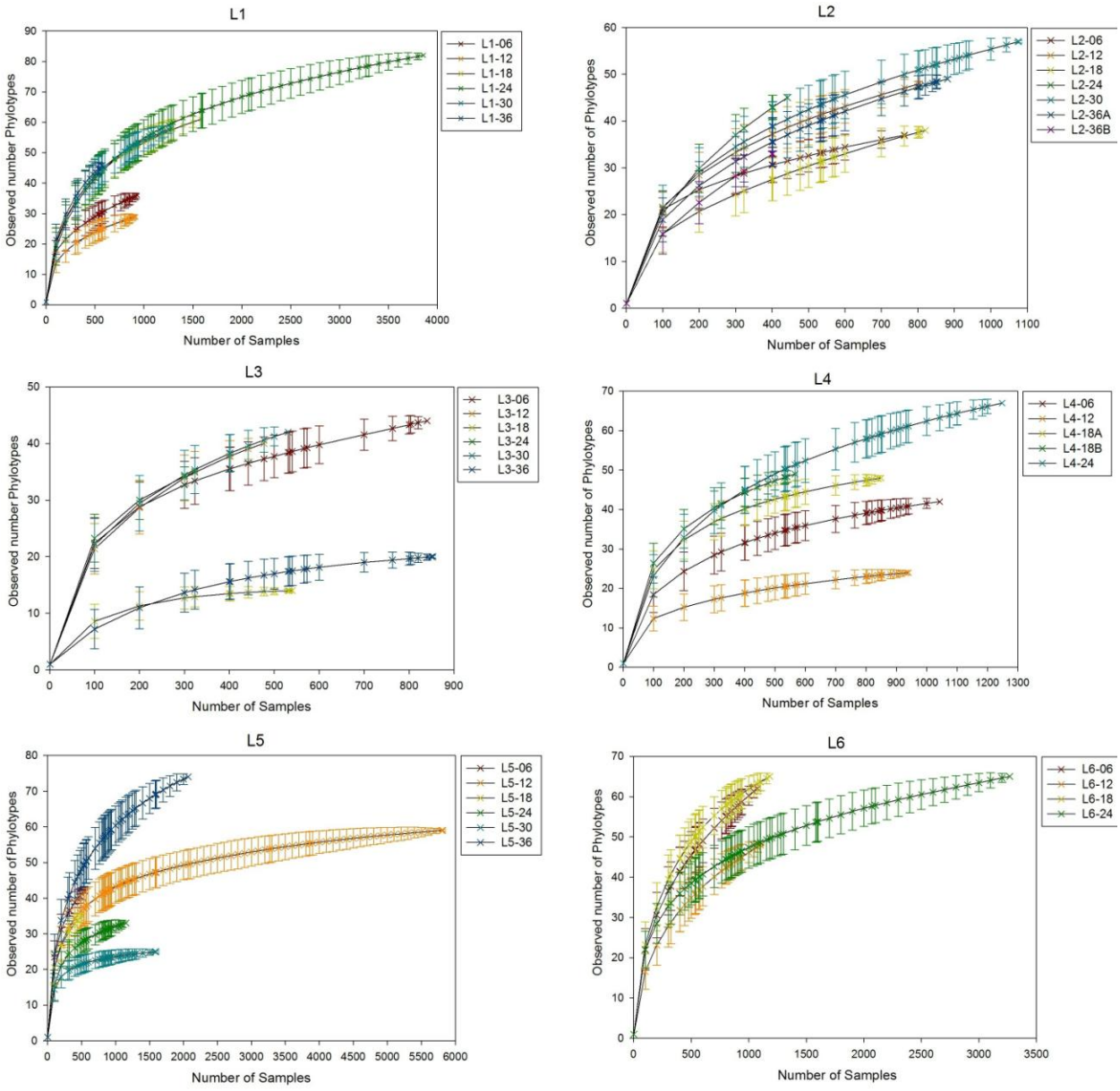

Figure 8. Rarefaction curves of each soil sample library from phylotype-based analysis. In this analysis, sequences of each library were grouped in genus level. The rarefaction curves are grouped by lysimeter. In each lysimeter, each sample was labeled by a different color. Duplicate libraries (L2-36A & L2-36B, L4-18A & L4-18B) are assigned to different colors as well.

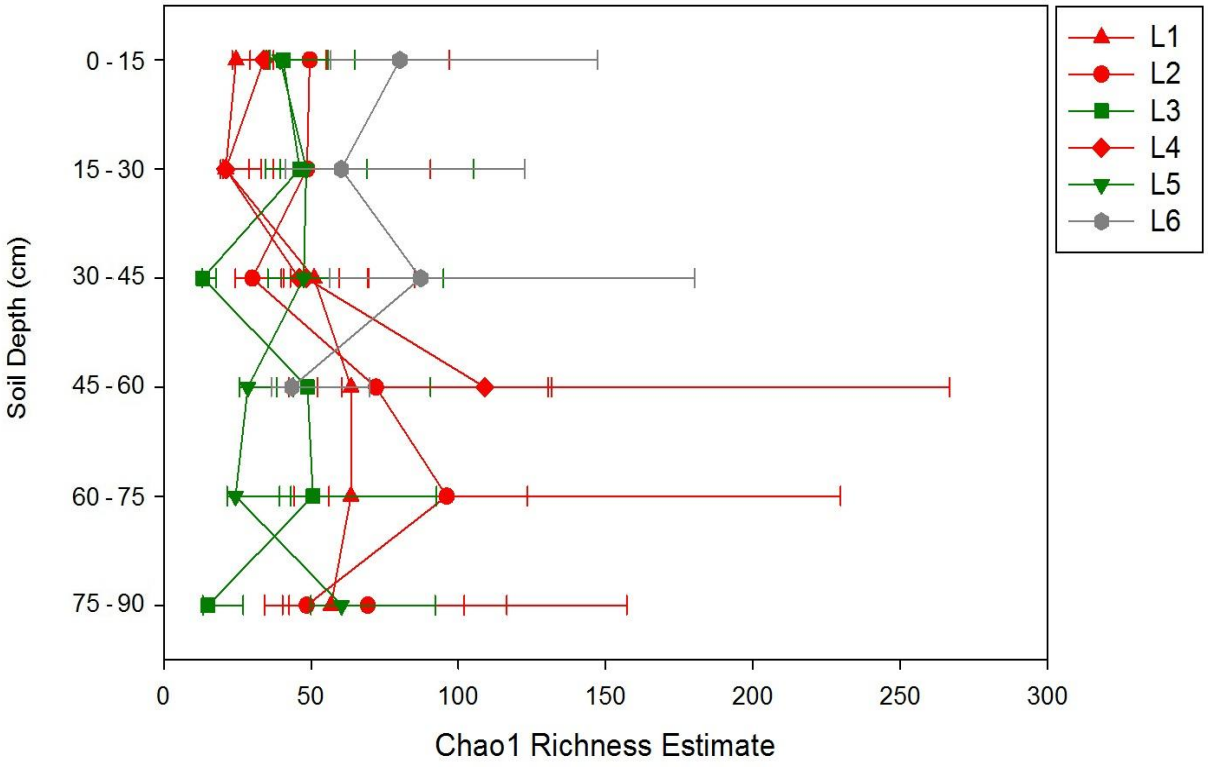

Figure 9. Genus level richness estimate of 34 soil bacterial DNA libraries at size=323 sequences. This is the size of smallest library before sub-sampled. Each data point represents the Chao1 richness estimate of a library at this sample size. Data points belong to different treatment groups are labeled by different color: treatment PAH+, HK44+ are red, treatment PAH-, HK44+ are green, treatment PAH+, HK44- are grey.

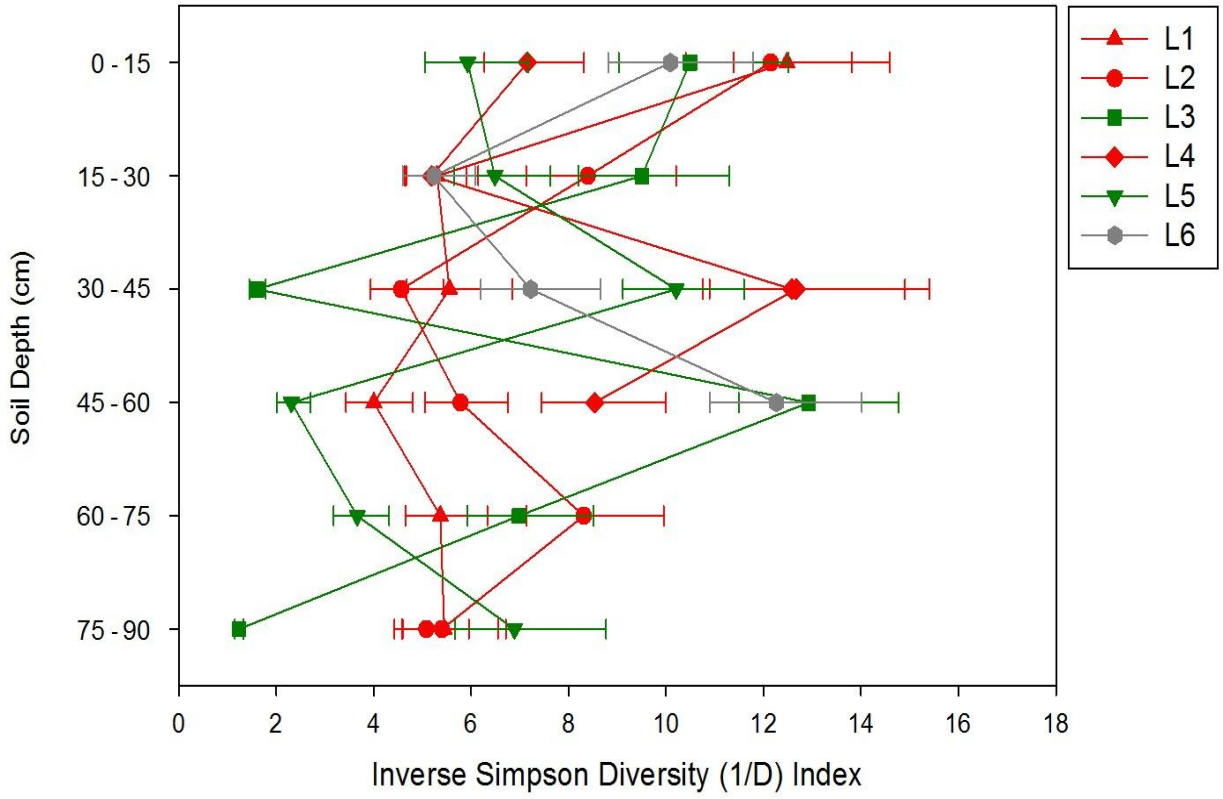

Figure 10. Genus level inverse Simpson diversity of 34 soil bacterial DNA libraries at size=323 sequences. This is the size of smallest library before sub-sampled. Each data point represent the inverse Simpson index value and error bar represent calculated upper and lower confidence intervals of inverse Simpson diversity indices. Data points belonging to different treatment groups are labeled by different color: treatment PAH+, HK44+ are red, treatment PAH-, HK44+ are green, treatment PAH+, HK44- are grey.

Table 7. Distribution of abundant bacteria genera (sequence abundance > 1% of the total) in each normalized library.

All libraries were normalized to 323 sequences per library by randomly subsampling. The sample names are labeled by treatment: PAH+, HK44+ are the red ones, PAH-, HK44+ are the green ones, PAH+, HK44- are the grey ones. The bacteria genera with abundance > 1% are listed by taxonomy. The "UC" means unclassifiable in that level.

OTU#	015	001	007	011	006	005	036	020	009	150	004	002	038	023	030	010	076
Domain	Bacteria																
Phylum	*Acidobacteria*			*Firmicutes*	*Proteobacteria*												UC
Class	*Acidobacteria*			*Bacilli*	*α-proteobacteria*						*β-proteobacteria*				*γ-proteobacteria*	UC	UC
	Gp1	*Gp2*	*Gp3*														
Order	Order incertae sedis			*Bacillales*	*Sphingomonadales*	*Rhizobiales*			*Rhodospirillales*	UC	*Burkholderiales*			UC	UC	UC	UC
Family	Family incertae sedis			UC	*Sphingomonadaceae*	*Bradyrhizobiaceae*		UC	*Acetobacteraceae*	UC	*Comamonadaceae*	*Oxalobacteraceae*	UC	UC	UC	UC	UC
Genus	*Gp1*	*Gp2*	*Gp3*	UC	UC	*Bradyrhizobium*	UC	UC	UC	UC	UC	UC	UC	UC	UC	UC	UC
% abundance	15.5	11.6	11.3	11.1	10.0	4.9	3.3	2.8	2.7	2.1	1.9	1.7	1.6	1.5	1.5	1.1	1.1
L1-06	13	16	2	32	42	16	34	0	43	0	1	1	6	25	0	0	0
L1-12	66	75	106	20	11	10	8	0	0	0	2	0	8	0	4	0	1
L1-18	37	20	122	35	20	13	8	0	0	2	5	5	8	2	5	0	4
L1-24	146	12	21	32	33	16	2	1	0	5	5	4	2	2	3	3	2
L1-30	107	17	74	40	11	4	1	1	1	0	8	6	7	3	9	2	1
L1-36	130	23	12	49	13	7	0	2	2	0	10	1	3	10	11	0	0
L2-06	11	45	67	11	20	18	10	1	23	0	15	0	18	13	2	1	4
L2-12	43	23	31	13	81	13	4	4	3	5	4	1	3	5	1	15	27
L2-18	123	12	74	22	29	5	8	4	1	1	5	0	0	3	5	0	14
L2-24	77	7	22	60	9	8	1	3	1	0	94	4	0	1	6	1	1
L2-30	59	41	73	17	20	8	5	7	4	1	3	8	0	8	0	5	21
L2-36A	105	11	46	54	26	12	1	6	1	6	13	4	0	2	3	0	3
L2-36B	114	20	48	61	24	4	2	8	0	0	10	1	0	4	2	0	1
L4-06	33	59	58	22	10	16	76	2	2	7	2	8	5	1	5	0	0
L4-12	46	79	110	10	16	22	13	0	0	4	1	0	6	1	0	0	0
L4-18A	30	64	24	14	18	22	27	0	3	27	1	12	7	3	0	0	0
L4-18B	29	58	17	22	18	21	18	0	2	31	0	12	3	3	1	0	0
L4-24	10	49	53	19	0	22	16	0	1	67	6	29	2	0	2	0	2
L3-06	93	5	0	43	41	6	1	1	0	1	0	6	8	7	16	19	0
L3-12	17	28	2	43	12	69	0	3	21	0	0	2	11	28	6	41	4
L3-18	0	6	3	2	261	1	0	4	3	0	1	0	0	17	0	21	2
L3-24	25	43	6	45	4	32	0	2	27	0	4	0	21	2	24	1	0
L3-30	50	36	2	21	5	11	1	8	102	1	2	2	11	3	5	11	12
L3-36	2	3	1	2	293	1	0	2	1	0	0	0	0	7	0	2	0
L5-06	97	13	4	72	1	19	13	2	0	0	1	0	13	1	8	0	0
L5-12	23	96	8	73	2	42	1	6	0	1	2	0	6	1	1	0	0
L5-18	24	45	5	33	10	3	8	49	48	0	2	2	3	5	2	0	1
L5-24	5	23	3	5	6	0	1	194	2	0	1	8	0	2	4	4	0
L5-30	45	52	10	140	6	7	4	0	0	0	1	0	16	2	7	0	5
L5-36	31	15	6	112	4	33	0	0	0	2	0	32	4	0	11	0	2
L6-06	25	68	54	38	19	15	37	0	0	2	1	6	2	0	6	0	4
L6-12	35	102	72	9	13	34	26	0	1	3	1	0	2	0	7	0	1
L6-18	12	63	76	19	18	19	21	0	0	25	3	9	1	0	4	0	6
L6-24	40	50	25	34	3	11	14	0	0	44	0	29	0	1	2	0	0

Table 8. Distribution of *Pseudomonas* genus in each normalized library.
All libraries were normalized to 323 sequences per library by randomly subsampling. The sample names are labeled by treatment: PAH+, HK44+ are the red ones, PAH-, HK44+ are the green ones, PAH+, HK44- are the grey ones. Although sequence abundance of *Pseudomonas* genus was relatively low (0.7% of the total sequences), its distribution is of concern to us since HK44 belonging to *Pseudomonas* genus.

OTU#	032	% in normalized library	% in each whole library
Domain	Bacteria		
Phylum	Proteobacteria		
Class	γ-proteobacteria		
Order	Pseudomonadales		
Family	Pseudomonadaceae		
Genus	Pseudomonas		
% abundance	0.7		
L1-06	19	5.882	0.041
L1-12	0	0.000	0.000
L1-18	0	0.000	0.000
L1-24	1	0.310	0.001
L1-30	1	0.310	0.001
L1-36	4	1.238	0.010
L2-06	12	3.715	0.033
L2-12	8	2.477	0.032
L2-18	2	0.619	0.004
L2-24	1	0.310	0.007
L2-30	14	4.334	0.037
L2-36A	11	3.406	0.020
L2-36B	6	1.858	0.022
L4-06	0	0.000	0.000
L4-12	0	0.000	0.000
L4-18A	0	0.000	0.000
L4-18B	0	0.000	0.000
L4-24	0	0.000	0.000
L3-06	0	0.000	0.000
L3-12	0	0.000	0.000
L3-18	0	0.000	0.000
L3-24	0	0.000	0.000
L3-30	0	0.000	0.000
L3-36	0	0.000	0.001
L5-06	0	0.000	0.000
L5-12	0	0.000	0.000
L5-18	0	0.000	0.000
L5-24	1	0.310	0.014
L5-30	0	0.000	0.000
L5-36	0	0.000	0.000
L6-06	0	0.000	0.000
L6-12	0	0.000	0.000
L6-18	0	0.000	0.000
L6-24	0	0.000	0.000

(the 15 − 30 cm sample from lysimeter 5) suggesting that the unclassifiable genus may be another *Bradyrhizobiaceae*. Strain HK44 belongs to *Pseudomonas* genus, therefore we specifically examined the abundance of Pseudomonas sequences in the libraries. *Pseudomonas* was only present in lysimeter 1, 2, 3 and 5, all of in which were HK44+. Although lysimeter 4 had HK44 inoculation, none of the sequences from L4 were classified as *Pseudomonas* genus. Thus, even though the present of *Pseudomonas* genus does not necessarily indicates the existence of HK44, the absence of *Pseudomonas* in L4 suggests that the HK44 (and other *Pseudomonas*) were below the detection limit.

3.5.1.3 Phylogeny-based Analysis

Table 9 is the summary of soil bacterial community diversities based on phylogeny. In phylogeny-based analysis, α diversity of each library was calculated as the total of the unique branch length from each library in the phylogenetic tree constructed by sequencing data from all these libraries. Given $\alpha = 0.05$, there is no significant difference of phylogeny-based diversity estimates across the 6 lysimeters (Kruskal-Wallis test, $\chi^2=3.8112$, $p=0.5769$).

Table 9. Bacteria phylogeny diversities of each soil sample.
From left to right are samples from lysimeter 1, 2, 3, 4, 5 and 6.

Sample Depth (cm)	Phylogeny Diversity					
	L1	L2	L3	L4	L5	L6
0 - 15	6.298	7.172	3.728	6.649	2.811	2.234
15 - 30	4.697	1.525	8.270	3.852	8.079	2.291
30 - 45	1.951	6.657	0.947	1.921	8.755	2.191
45 - 60	8.888	9.572	8.662	7.811	3.849	2.360
45 - 60	NA	NA	NA	7.191	NA	NA
60 - 75	8.734	1.583	7.376	NA	1.774	NA
75 - 90	2.901	2.771	0.356	NA	9.738	NA
75 - 90	NA	7.657	NA	NA	NA	NA
Mean	5.578	5.277	4.890	5.485	5.834	2.269
Standard Deviation	±2.917	±3.256	±3.723	±2.502	±3.417	±0.073

3.5.1.4 Correlation of α Diversity Estimated in OTU-based, Phylotype-based and Phylogeny-based Analyses and Soil Properties

Alpha diversity of a bacterial community includes richness and diversity. In this study, bacterial community diversity of each sample was estimated in three analyses (OTU-based, phylotype-based and phylogeny-based). In addition, bacterial community richness of each sample was estimated in OTU-based and phylotype-based analyses. The correlations between each two diversity index and between the richness estimates in two taxonomic levels were calculated using Spearman correlation coefficients (ρ) (Table 10). Spearman's rank correlation coefficient is a non-parametric measure of the correlation between two variables. Figure 11 showed the scatter plots of these pairwise comparisons. Given $\alpha = 0.05$, results showed that the community diversities from OTU-based and phylotype-based analyses were strongly correlated ($\rho = 0.525$, $p = 0.002$). The community richness estimates from OTU-based and phylotype-based analysis were linear correlated as well ($\rho = 0.537$, $p = 0.001$). The phylogeny-based diversity was not positively correlated with the other two diversity indices (for phylogeny diversity and OTU-based diversity: $\rho = 0.308$, $p = 0.09$; for phylogeny diversity and phylotype-based diversity: $\rho = 0.265$, $p = 0.14$).

In this study, environmental variables measured included soil moisture, soil total carbon, total nitrogen, C:N ratio, and soil organic matter content. Soil bacterial community variables measured included bacterial abundance, bacterial community richness, community diversity estimates. The Spearman correlation coefficient between each pair of these measured variables were calculated, and the null hypothesis that there is no significant correlation (H_0: $\rho = 0$) was tested. Given $\alpha = 0.05$, the pairs that had significant linear correlation are shown in Table 10. The variables: total carbon, total nitrogen, C:N ratio and soil organic matter were found had strong

correlations between each other. Bacteria abundance, bacterial community richness had relatively strong correlations with soil total nitrogen and soil organic matter content.

In summary, all three analysis approaches (species-level, genus-level and phylogenies) agreed there were no significant differences between the six lysimeters in both richness and diversity of the soil bacterial communities. However, looking at Table 7, it can be seen in the genus level analysis the distributions of most abundant (sequence abundance >1% of total) genera were not consistent between treatments. We will discuss more about the differences in the β diversity analysis section.

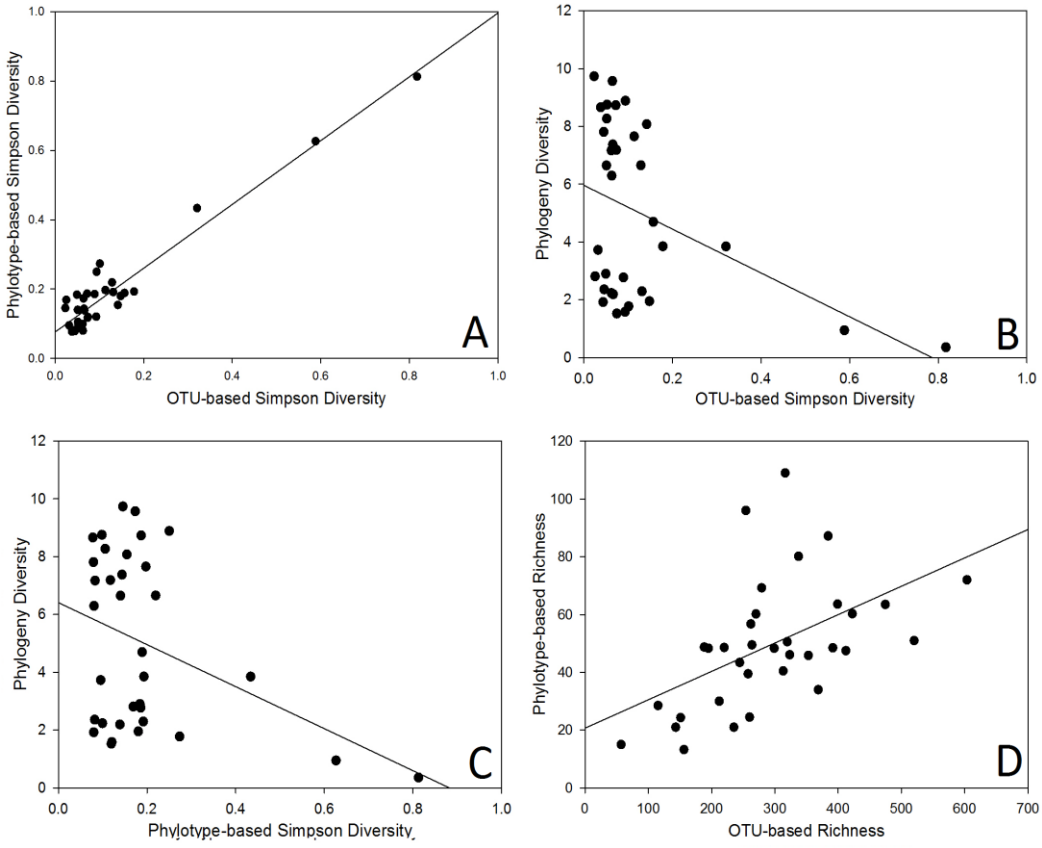

Figure 11. (A) The correlation of OTU-based diversity and phylotypes-based diversity. (B) The correlation of OTU-based diversity and phylogeny-based diversity. (C) The correlation of phylotypes-based diversity and phylogeny-based diversity. (D) The correlation of OTU-based richness and phylotypes-based richness.
The spearman correlation coefficient values and p values of each pair of variables are listed in Table 10.

Table 10. Spearman correlation coefficients of bacterial community α diversity estimates and soil properties.

Variable 1	Variable 2	Spearman correlation coefficient	Fisher's z	95% Confidence interval		p value for H_0: No correlation (ρ=0)
				Upper	Lower	
OTU-based Diversity	Phylotype-based Diversity	0.74	0.95	0.53	0.87	<.0001
OTU-based Richness	Phylotype-based Richness	0.66	0.79	0.40	0.82	<.0001
Soil Organic Matter	C:N Ratio	0.90	1.48	0.74	0.97	<.0001
Total Carbon	C:N Ratio	0.92	1.62	0.79	0.97	<.0001
Soil Organic Matter	Bacterial Abundance	0.62	0.72	0.34	0.80	0.0001
Soil Organic Matter	Total Nitrogen	0.78	1.06	0.47	0.92	0.0001
Soil Organic Matter	Total Carbon	0.76	1.01	0.43	0.91	0.0003
OTU-based Richness	Total Nitrogen	0.70	0.87	0.31	0.89	0.002
Bacterial Abundance	Total Nitrogen	0.63	0.74	0.20	0.86	0.008
Phylogeny-based Diversity	OTU-based Richness	0.46	0.50	0.13	0.70	0.008
Total Nitrogen	C:N Ratio	0.62	0.73	0.18	0.85	0.009
OTU-based Diversity	OTU-based Richness	0.42	0.44	0.08	0.67	0.017
Phylotype-based Richness	Soil Organic Matter	0.41	0.43	0.07	0.66	0.020
OTU-based Diversity	Phylogeny-based Diversity	0.37	0.38	0.02	0.63	0.038
OTU-based Richness	Soil Organic Matter	0.36	0.38	0.01	0.63	0.043

3.5.2 β Diversity

The purpose of β diversity analysis was to estimate the difference (membership and structure) between communities.

3.5.2.1 Comparison of Bacterial Community Membership: Shared OTUs

Table 11 is a summary of all the average percent shared OTUs between each pair of lysimeters. The shared OTU percentages distribution of each comparison is shown in Figure 12.

$$\text{The percent shared OTU} = \frac{\text{Shared OTU number between these two libraries}}{\text{Total OTU number of these two libraries}} \times 100\%$$

First, sequences of the normalized libraries that belong to the same lysimeter (i.e. 6 depths) were merged together to create one library for that lysimeter. Next, each pair of lysimeter libraries was compared to get shared OTU numbers between each pair. The percent shared OTUs of each pair was then calculated based on the equation above. Finally, the comparisons were grouped into comparisons between each treatment condition: comparison 1 is the percent of shared OTUs between PAH+/–; comparison 2 is the percent of shared OTUs between HK44+/ –; comparison 3 is the percent of shared OTUs between replicates. Given $\alpha = 0.05$, replicates had significantly higher percent of shared OTUs than comparison 1 (between PAH+/–), but there is no significant difference of the percent of shared OTUs between comparison 1 and 2, or between comparisons 2 and 3 (Mann–Whitney U test, χ^2=7.0330, p=0.0297; HSD test, F=7.08, p=0.0122).These results indicate that PAH+ and PAH− lysimeters had the greatest differences in the bacteria species present in each community. However, in this analysis of presence/absence of OTUs, the effects of HK44 inoculation on soil bacterial community membership was not distinguishable from time effects seen between replicates.

Table 11. Summary of Shared OTUs between treatments.
Group 1 is the percent of shared OTUs between PAH+ and PAH- treatment; group2 is the percent of shared OTUs between HK44+ and HK44- treatment; group3 is the percent of shared OTUs between PAH+, HK44- and PAH-, HK44+ treatment; group 4 is the percent of shared OTUs between each two lysimeters of the same treatment.

Comparison 1: PAHs			Comparison 2: HK44			Comparison 3: Replicates		
PAH+	PAH-	% Shared	HK44+	HK44-	%Shared	Replicates		%Shared
L1	L3	9.91	L1	L6	12.21	L1	L2	13.11
L1	L5	10.40	L2	L6	8.95	L1	L4	12.73
L2	L3	6.31	L4	L6	9.30	L2	L4	9.97
L2	L5	5.74	–	–	–	L3	L5	13.56
L4	L3	5.79	–	–	–	–	–	–
L4	L5	7.95	–	–	–	–	–	–
Average		7.68	Average		10.15	Average		12.34

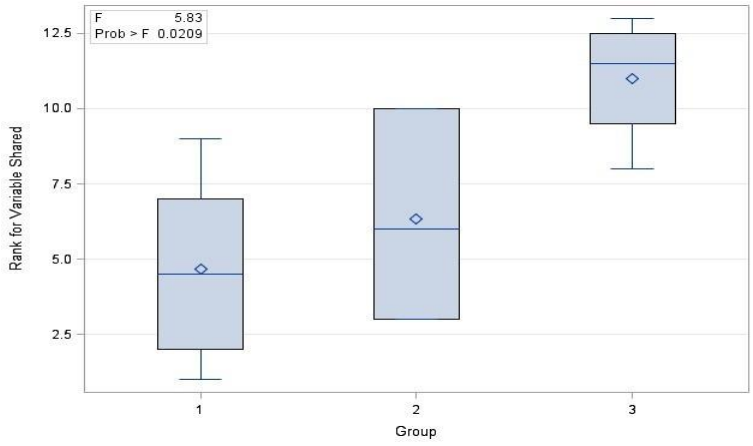

Figure 12. Distribution of shared OTUs numbers by group.

3.5.2.2 Comparison of Bacterial Community Structures by Hierarchical Clustering

The phylogram showed in Figure 13 shows the relative relationships between 34 soil sequencing libraries using Θ_{YC} distances in species level (OTU-based). The similarities between these libraries were calculated by Θ_{YC} index, which takes into account both shared OTU numbers between each two libraries, and the abundances of each OTU (Yue et al., 2005). The phylogram tree was generated using a UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm. This tree indicates that samples from lysimeter1 and 2, 4 and 6, 3 and 5 tend to have more similar bacteria community structures to each other. Lysimeter 1, 2, 4 are PAH+ & HK44+, lysimeter 3 and 5 are PAH+ & HK44 -, and lysimeter 6 is PAH– & HK44+. As previously mentioned, the soil moisture content of lysimeter 4 and lysimeter 6 were lower than the others. In addition, the replicate libraries (L2-36A & L2-36B, L4-18A & L4-18B) have the closest distances to each other, indicating good technical reproducibility of our amplicon library construction and sequencing.

The phylogram showed in Figure 14 shows the relative relationships between 34 soil sequencing libraries using Θ_{YC} distances in genus level (phylotype-based). The phylotype-based distance between these libraries were calculated by Θ_{YC} index, take into account both shared phylotype numbers between each two libraries, and the abundances of each phylotype. Although this phylogram tree is different from OTU-based phylogram tree (Figure 14), it indicates the same relationship tendency: samples from lysimeter1 and 2, 4 and 6, 3 and 5 tend to have more similar bacteria community structures to each other. In addition, the replicate libraries (L2-36A & L2-36B, L4-18A & L4-18B) have closest distances to each other indicating good technical reproducibility.

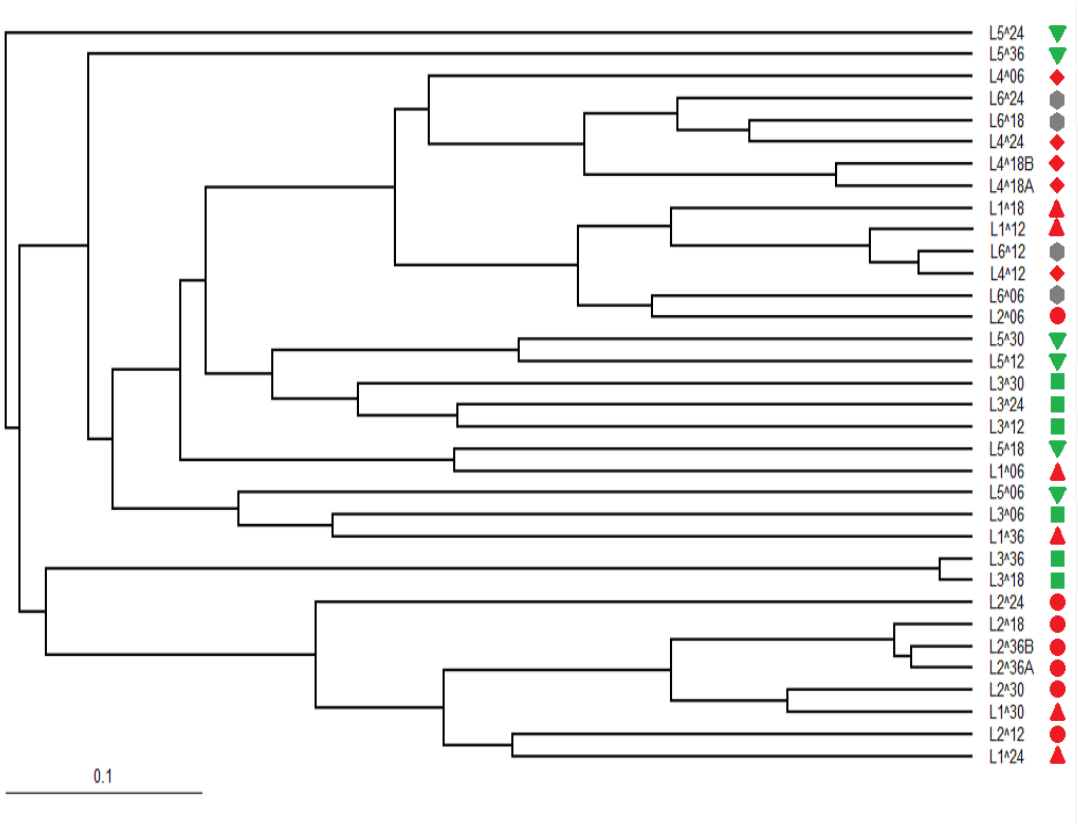

Figure 13. Phylogram of 34 soil bacterial communities (Θ_{YC} distances of species, distance=0.03). This phylogram was generated byUPGMA (Unweighted Pair Group Method with Arithmetic Meanclustering) algorithm. Sequencing data was grouped by distance=0.03. Data points belonging to different treatment groups are labeled by colors: treatment PAH+, HK44+ are red, treatment PAH-, HK44+ are green, treatment PAH+, HK44- are grey.

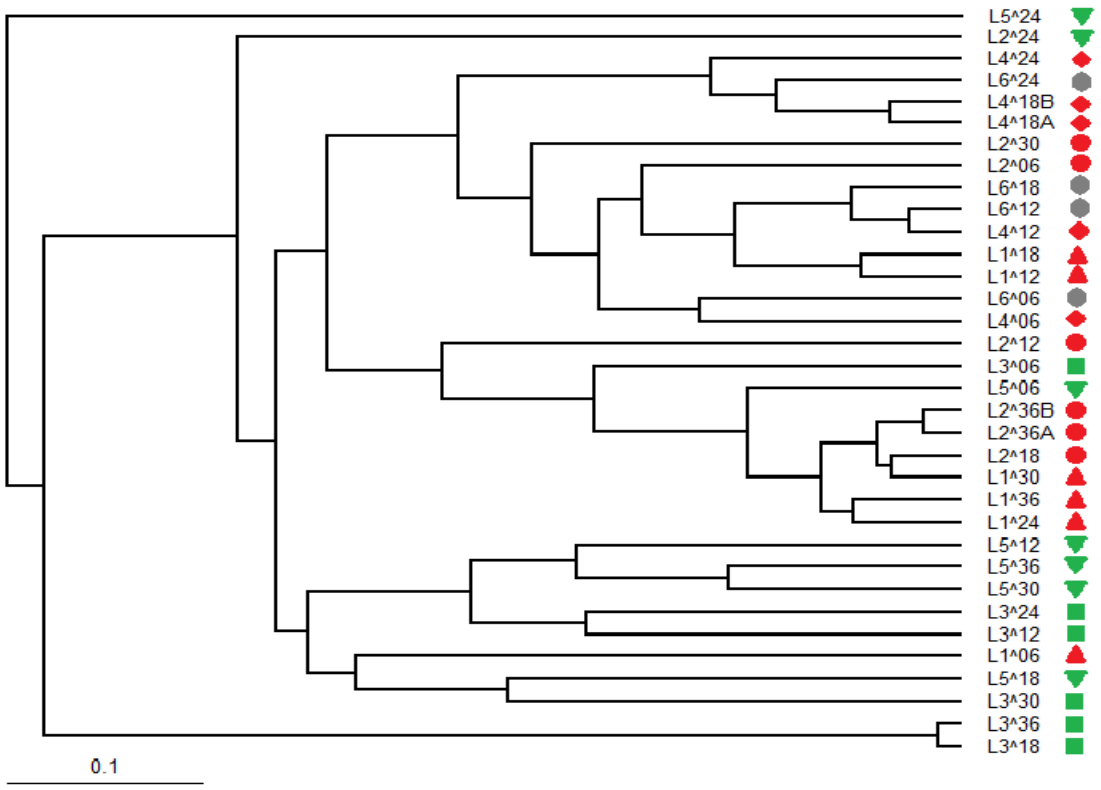

Figure 14. Phylogram of 34 soil bacterial communities representing Θ_{YC} distances of genera. This phylogram was generated by a UPGMA (Unweighted Pair Group Method with Arithmetic Meanclustering) algorithm. Data points belonging to different treatment groups are labeled by different color: treatment PAH+, HK44+ are red, treatment PAH-, HK44+ are green, treatment PAH+, HK44- are grey.

3.5.2.3 Comparison of Bacterial Community Structures using Ordination

Distance matrices were calculated using Jaccard index and Θ_{YC} calculator in both OTU-based (species level) analysis and phylotype-based (genus level) analysis. In phylogeny-based analysis, distance matrices were calculated using unweighted UniFrac algorithm and weighted UniFrac algorithm. These two algorithms are the phylogeny-based extensions of Jaccard and Θ_{YC} algorithms. Jaccard and unweighted UniFrac distance matrices both document membership relationships between samples while Θ_{YC} and weighted UniFrac document the structure (membership and relative abundance) relationships between samples. Since 34 sample libraries were analyzed, the distance matrices all had 34 dimensions. Non-metric multidimensional scaling (NMDS) method was used to reduce the number of dimensions for visualization in 2 dimensional plots. In this method, the two components that loaded most variances between these samples were found and retained. Stress of each NMDS plot was calculated to assess the mismatch between the 34 dimension and 2 dimension relationships: the stress describes how well the distances in the 2D plots represent the original distances in 34 dimension matrix, with a lower stress indicating a better representation.

NMDS 2 dimensional plots of each distance matrix are shown in Figure 15 – 20. Given $\alpha = 0.05$, analysis of similarity (ANOSIM) showed that PAH+/– groups were significantly different from each other in each distance matrix, and there was no significant difference between HK44+/– groups (Table 12). The phylogram trees (Figure 13 and 14) showed the same clustering tendency. In addition, since soil moisture contents of these six lysimeters were significantly different (L2 & 3 > L1 > L5 > L4 & 6), these communities were assigned to 4 soil moisture levels based on their soil moisture content: L2 & 3 are level 1, L1 is level2, L5 is level 3, L4 & 6 are level 4. ANOSIM suggested that within PAH+ group, community memberships and structures had a significant difference between different moisture levels (Table 13).

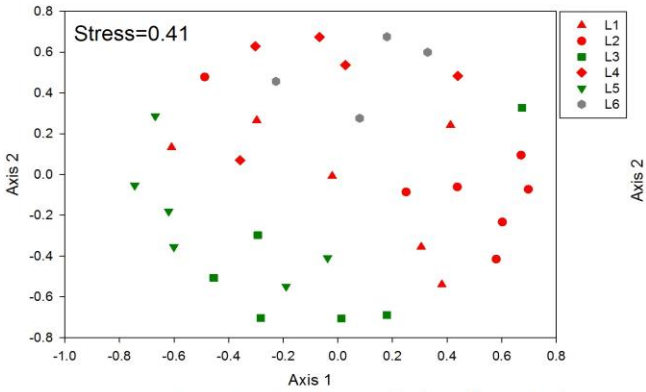

Figure 18. NMDS 2-dimentional plot of Jaccard distance (OTU-based analysis)

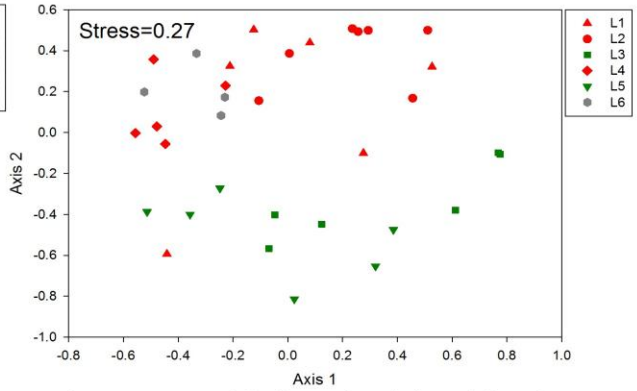

Figure 19. NMDS 2-dimentional plot of Θ_{YC} distance (OTU-based analysis)

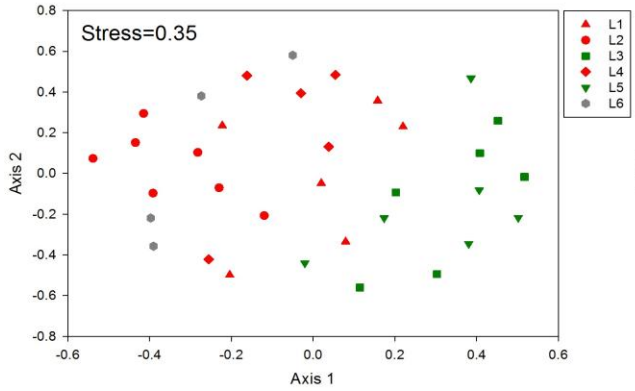

Figure 20. NMDS 2-dimentional plot of Jaccard distance (phylotype-based analysis)

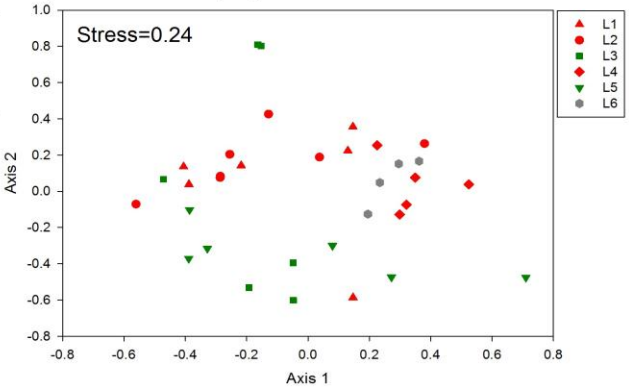

Figure 21. NMDS 2-dimentional plot of Θ_{YC} distance (phylotype-based analysis)

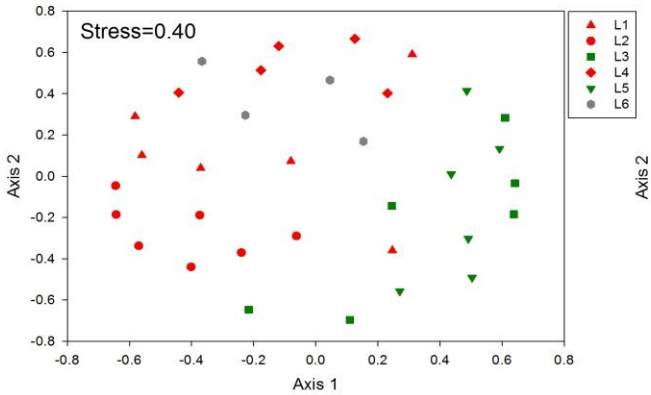

Figure 22. NMDS 2-dimentional plot of unweighted UniFrac algorithm calculated distances (phylogeny-based analysis)

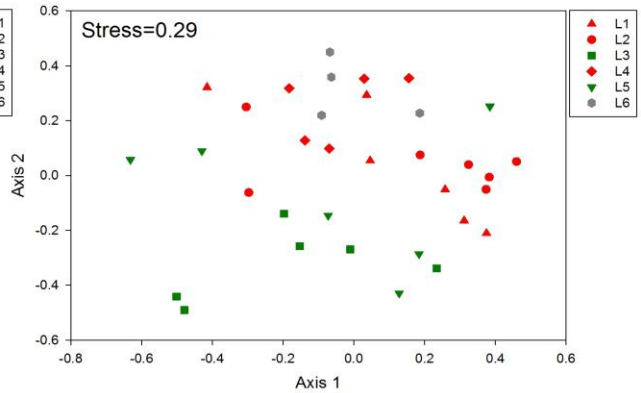

Figure 23. NMDS 2-dimentional plot of weighted UniFrac algorithm calculated distances (phylogeny-based analysis)

Figure 15, 16, 17, 18, 19, 20. Side by side comparisons of NMDS plots generated by 3 approaches. In each approach (OTU-based, phylotype-based, phylogenetic-based), distance of Jaccard index and Θ_{YC} were calculated. The figures in the first row are from OTU-based analysis, the ones on the second row are from phylotype-based analysis, and the ones in the third row are from phylogeny-based analysis. The ones on the left column showed the relationships of community memberships, and the ones on the right column showed the relationships of community structures. Data points belonging to different treatment groups are labeled by different color: treatment PAH+, HK44+ are red, treatment PAH-, HK44+ are green, treatment PAH+, HK44- are grey.

Table 12. ANOSIM tests of community membership and structure relationships between PAH+/– and between HK44+/– groups.
The "*" in p-value column indicates significant differences at α=0.05.

Analysis	Calculator	comparison	R-value	P-value
OTU-based	Jaccard Index	PAH+/-	0.692	<0.001*
		HK44+/-	0.031	0.399
OTU-based	Θ_{YC}	PAH+/-	0.487	<0.001*
		HK44+/-	-0.145	0.809
Phylotype-based	Jaccard Index	PAH+/-	0.445	<0.001*
		HK44+/-	0.176	0.057
Phylotype-based	Θ_{YC}	PAH+/-	0.389	<0.001*
		HK44+/-	-0.063	0.627
Phylogeny-based	Unweighted UniFrac	PAH+/-	0.594	<0.001*
		HK44+/-	0.086	0.212
Phylogeny-based	Weighted UniFrac	PAH+/-	0.381	<0.001*
		HK44+/-	-0.027	0.563

Table 13. ANOSIM tests of community membership and structure relationships between samples that had different soil moisture.
The "*" in p-value column indicates significant differences at α=0.05.

Analysis	Calculator	comparison		R-value	P-value
OTU-based	Jaccard Index	L2 & 3	L1	0.034877	0.333
		L2 & 3	L5	0.406672	<0.001*
		L2 & 3	L4 & 6	0.481332	<0.001*
		L1	L5	0.719444	0.003*
		L1	L4 & 6	0.533406	<0.001*
		L5	L4 & 6	0.876906	<0.001*
OTU-based	Θ_{YC}	L2 & 3	L1	-0.13565	0.924
		L2 & 3	L5	0.275986	0.028*
		L2 & 3	L4 & 6	0.232419	0.011*
		L1	L5	0.359259	0.011*
		L1	L4 & 6	0.453159	<0.001*
		L5	L4 & 6	0.631082	0.001*
Phylotype-based	Jaccard Index	L2 & 3	L1	0.032396	0.329
		L2 & 3	L5	0.237662	0.024*
		L2 & 3	L4 & 6	0.447293	<0.001*
		L1	L5	0.269444	0.014*
		L1	L4 & 6	0.205882	0.058
		L5	L4 & 6	0.563907	<0.001*
Phylotype-based	Θ_{YC}	L2 & 3	L1	-0.10201	0.821
		L2 & 3	L5	0.181969	0.083
		L2 & 3	L4 & 6	0.282351	<0.001*
		L1	L5	0.275926	0.014*
		L1	L4 & 6	0.567901	<0.001*
		L5	L4 & 6	0.733479	<0.001*
Phylogeny-based	Unweighted UniFrac	L2 & 3	L1	0.029777	0.362
		L2 & 3	L5	0.303832	0.006*
		L2 & 3	L4 & 6	0.402759	<0.001*
		L1	L5	0.538889	0.009*
		L1	L4 & 6	0.431373	0.003*
		L5	L4 & 6	0.849673	<0.001*
Phylogeny-based	Weighted UniFrac	L2 & 3	L1	-0.07113	0.668
		L2 & 3	L5	0.252826	0.042*
		L2 & 3	L4 & 6	0.333333	0.001*
		L1	L5	0.174074	0.079
		L1	L4 & 6	0.481481	<0.001*
		L5	L4 & 6	0.712418	<0.001*

CHAPTER 4 DISCUSSION

The purpose of this study is to assess long term (approximate14 years) impact of a one-time PAH addition and inoculation of the PAH-degrading bioluminescent bioreporter HK44 inoculation on soil bacterial communities. Previous studies reported that PAHs addition can cause shifts in bacterial communities in terms of both membership (taxa present) and structure (membership and relative abundance of taxa). Since a subset of the resident soil microbes can use PAHs as a growth substrate, the addition of PAHs brings selective pressure into soil bacterial community (Trogl et al., 2012). However, most studies investigating the impact of PAHs on bacterial communities cover only relatively short terms (> 1 year). In addition to the impact of PAHs, this study also evaluated the effect of HK44 inoculation. Because this was the first genetically engineered microorganism (GEM) permitted for field testing (Ripp et al., 2000a; Ripp et al., 2000b), this study provided a rare chance to evaluate the long term impact of both PAHs and GEM on bacterial community.

4.1 Comparison of Sequencing Data Analyses Approaches

The OTU-based and phylotype-based analyses were based on mathematical similarities of sequences in species level and genus level respectively. Phylogeny-based analysis was based on phylogeny relatedness of sequences. In each approach, two major questions were addressed: what was occurring in a single community (α diversity), and how communities compared with each other (β diversity). Since these approaches analyzed bacterial communities from different aspects, comparison of the results of all three approaches could help us in understanding these two questions.

4.1.1 Comparison of α Diversity Analyses Results

Estimates of richness were calculated with Chao1 richness estimator, and done for communities at both the species level (OTU-based analysis) and genus level (phylotype-based analysis). Significance tests suggested that bacteria community richness and diversity had no significant differences between lysimeters in both species and genus level. Significance test of phylogeny diversity showed no significant difference between lysimeters as well.

In general, phylotype-based (genus level analysis) estimates of richness was lower than OTU-based (species level analysis) estimates of richness. Meanwhile, phylotype-based diversity indices were higher than OTU-based diversity indices. In addition, the community richness values were more consistent to each other in genus level than in species level (Figure 6 and Figure 9). Although statistical tests of both levels agreed with each other, richness estimates in genus level showed more general and intuitive results. This is possibly because in genus level analysis the sequences belonging to multiple species may be grouped into a single genus. The phylogeny-based analysis assumes branch lengths of phylogenetic tree are proportional to function. Thus, this is another aspect of the diversity of each bacterial community. In this study, the phylogeny diversity of communities agreed with the OTU and phylotype analyses and had no significant difference between lysimeters.

Richness estimates and diversity indices were strongly correlated for OTU-based approach and phylotype-based approach. This result is not surprising since both of them were based on the mathematical similarities of bacteria sequences. In contrast, the phylogeny-based diversity was only significantly correlated with OTU-based diversity and richness estimates, and these correlations were not as strong as the ones between OTU-based and phylotype-based estimates. In the phylogeny-based analysis, all the sequences were used to construct the

phylogenetic tree, and the unique branch length of every sequence is used to determine the phylogeny-based diversity. Thus, this analysis accounted more detailed information than the species level analysis (OTU-based analysis). This may explained why phylogeny-based diversity had weak correlation with OTU-based diversity and richness, but showed no significant correlation with the more general phylotype-based diversity or richness.

4.1.2 Comparison of β Diversity Analyses Results

Cluster analysis of OTU-based and Phylotype-based Θ_{YC} distance matrix were performed and phylograms generated to visualize the clustered relationships of community structures in species and genus level. These analyses revealed similar relationships of community structures. Both of them showed that L5-24″ library has the greatest distance from the others. They both agreed that most of the communities from L1 and L2 samples were more similar, those of L4 and L6 were more similar, and those of L3 and L5 were more similar. Clustering at the genus level showed clearer clustering results compared to the species level. This less noisy clustering is more likely to be the result of analysis done on a more general taxonomic level.

Ordination (NMDS) was also used to visualize the distance matrices. The distance matrices reflecting community membership relationships and structure relationships were generated by three approaches: in species level, in genus level, and in phylogeny-based analysis. In other words, the membership relationships addressed the question of similarity of the species/ genera present in each of two libraries. After ordination, the communities were not distributed in the same way as with the cluster analysis, but they did have similar clustering tendencies. ANOSIM results showed that for all six analysis approaches, the PAH+ and PAH– formed significantly difference clusters, while the HK44+ and HK44- clusters showed no significant

difference. Even though the distributions of data points in these plots were not necessarily identical to each other, the statistical clustering results are consistent.

4.2 Influence of PAHs and HK44 Inoculation on Soil Bacterial Communities

Previous studies stated that PAH contamination had impacts on soil bacterial communities. However, most studies have focused on soils in which the PAH concentrations were still detectable (Andreoni et al., 2004; Lors et al., 2010; Ni Chadhain et al., 2006; Vinas et al., 2005). In this study, the PAHs were depleted after the first two years, so this is a unique chance to analyze how the soil bacterial community had shifted a long time (14 years) after a one-time PAH contamination and bioremediation (Ripp et al., 2000a).

In this study, soil properties measured included soil moisture, total carbon, total nitrogen, C:N ratio and organic matter content. These variables could help us understand PAHs impacts on soil quality (Margesin et al., 2000). Since bacterial growth is impacted by many environmental factors, we analyzed how these soil properties correlated to changes in the soil bacterial community (DeBruyn et al., 2011; Garbeva et al., 2004; Johnsen et al., 2005; Lors et al., 2010; Ni Chadhain et al., 2006).

The soils that were initially PAH contaminated (PAH+ group) had significantly higher values of SOM, total carbon, total nitrogen and C:N ratio. The initial naphthalene concentration in lysimeter 1, 2, 4 and 6 was 1000 ppm, but extended storage caused the loss of more than 90% of the PAHs. So at day 135, naphthalene and anthracene were added into lysimeter 1, 2, 4 and 6 to restore these PAHs concentrations in PAH+ treated lysimeters. Assuming uniform distribution, this should produce approximately 1000 ppm naphthalene and 100 ppm anthracene in these 4 lysimeters. Naphthalene concentration was measured at day 474 in lysimeter1, 2 and 6. At that

time, most parts of lysimeter 1 showed the naphthalene concentrations below 20 ppm (Ripp et al., 2000a), so it can be assumed that PAHs were depleted at that time. The SOM, total carbon, total nitrogen and C:N ratio values were positively correlated to each other (Table 10). Since PAHs are organic compounds, this result suggests that the addition of PAHs may contribute to this difference in SOM content, even though the addition of PAHs was 14 years ago and the parent PAH compounds are no longer detectable.

In addition, PAH+ group showed significantly higher bacterial abundances than the PAH– group. This result suggested that the addition of PAH ultimately increased the bacteria abundance in long term. Since PAHs could be considered an extra nutrient source that not all bacteria could utilize, PAH addition could be a selective pressure in soil bacterial community shifting. Margesin et al. (2000) reported that the biomass C, N and proportion of hydrocarbon degradation bacteria increased rapidly after hydrocarbon addition (Margesin et al., 2000). Andreoni et al. (2004) reported that soil with a medium-term exposure (<3 years) to PAHs showed higher biodiversity than soil with a long-term exposition (>50 years) to PAHs and uncontaminated soil (Andreoni et al., 2004). In addition, to compensate for the unexpected loss of PAHs, naphthalene and anthracene were dissolved in transfer oil and added into lysimeter 1, 2, 4, and 6 in day 135. Each of these lysimeters received 208 L transfer oil along with the PAHs. The transfer oil was mainly constituted of organic compounds, which likely served as carbon and energy sources for some bacteria. Thus, the one-time addition of PAHs may have promoted increased microbial growth in soil at the beginning, and causing an increase in soil total carbon, total nitrogen and organic matter in long term.

Quantification of PAHs degradation genes showed that these genes were no longer detectable after 14 years. Likewise, Layton et al. (2012) tested the same soil sample from

lysimeter 1, 2, and 4, and reported that *nahA* (the naphthalene degradation gene of HK44) gene in these samples was below detectable level (Layton et al., 2012). These results suggest that bacteria community no longer has large populations of PAH-degrading microorganisms. However, PAH contaminations did result in higher bacteria abundances in long term. Since the SOM contents in PAH+ treated lysimeters were still higher than that in the PAH– treated soils, the bacteria in PAH+ treated lysimeters still had more abundant carbon resource, and therefore lead to higher bacteria abundances. In addition, the membership and structure of bacterial community in PAH+ and PAH– groups were significantly different. These results show that a one-time PAH contamination has had an impact on bacterial communities even though the parent contaminants were degraded long ago.

The influence of inoculation of the bioluminescent bioreporter HK44 on soil properties and bacterial communities was evaluated by comparing HK44+ and HK44- lysimeters. Between HK44+/– treatments, there was no significant difference in soil total carbon, soil organic matter, bacteria abundance, community richness, community diversity, community membership or community structures.

Taxa results showed that in PAH+, HK44– treated soil lysimeter (L6), the genera with abundances > 1% of total sequences belonging to *Acidobacteria* phylum, *Firmicutes* phylum, *Proteobacteria* phylum and an unclassifiable phylum (table 7). Vinas et al. (2005) also reported that *Proteobacteria* were dominant in PAH-contaminated soil (Vinas et al., 2005). Other researchers have isolated bacteria strains with PAH degradation capacities belonging to these three phyla as well (Haritash et al., 2009; Lors et al., 2010). These result suggested that in PAH contaminated soil, even if the PAH degradative bacteria were not added initially, the growth of indigenous PAH degradative bacteria strains was likely promoted. Meanwhile, in PAH+, HK44+

treated lysimeters (L1, 2, 4), HK44, selected for its efficiency in PAH degradation, was inoculated before the native bacteria species would have had time to respond to the added PAHs. If there were indigenous PAH degradative bacteria in the soil, they would have been in much lower population concentrations than HK44, since the soil packed was not initially PAH contaminated soil. Thus, in the beginning of this study, HK44 strain should have had a numerical and metabolic advantage over the indigenous bacteria. However, ANOSIM analysis of community membership and structure relationships between PAH+, HK44– group and PAH+, HK44+ group indicate that HK44 inoculation did not have significantly impact on membership or structure of soil bacterial communities (Figure 15 – 20) or on any of the soil chemical properties analyzed. Layton et al. (2012) also reported that HK44 was not recovered from 2010 soil samples (Layton et al., 2012). In summary, release of HK44 into field did not change the indigenous microorganisms or soil properties over the long term.

4.4 Other Factors Contributing to Differences between Lysimeters

In this experiment, soil moisture was intended to be controlled. However, these lysimeters were buried along a slope. In addition, during the 14 years long experiment, the lids of these lysimeters seemed not absolutely sealed. As a result of these unexpected factors, the soil moisture content was not the same between the 6 lysimeters. Soil moisture of these lysimeters showed significant differences: L2 & L3 > L1 > L5 > L4 &6.

ANOSIM analysis of community membership relationships and community structure relationships showed that within PAH+ group, communities that belong to L1 and L4 and 6 were significantly different from each other (Table13). The phylograms of community membership and structure relationships also showed that within the PAH+ group (L1, 2, 4, 6), the communities belong to drier soil lysimeters (L4 and 6) tend to be more similar to each other than

to the communities that belong to the wetter lysimeters (L1 and 2). However, in this study, soil moisture showed no significant linear correlation with bacterial community richness, diversity, or bacteria abundance (Table10). So the soil moisture may have contributed to the dissimilarities between bacterial community structures, but did not significantly affect community diversity, richness.

Studies showed that soil moisture can be related to soil bacterial community shifts (DeBruyn et al., 2011; Vinas et al., 2005). In this study, since PAH contamination seemed to have the dominant impact on bacterial communities, while soil moisture appears to be a secondary influence. In addition, the soil moisture changes over time were not available, so it is unknown when the soil moisture content of these lysimeters become dissimilar. It is possible that soil moisture varied only a few years or even a few months before these lysimeters were sampled in 2010. As a result, the impact of soil moisture changes may be underestimated in this study.

In summary, initial PAH contamination the greatest impact on bacterial community, even though the contaminants were absent for over ten years. The long term (>10 years) impact of the GEM HK44 inoculation was not observable. Soil moisture seems to have impacted on bacterial community structure as a secondary driving force.

REFERENCES

Andreoni, V., Cavalca, L., Rao, M. A., Nocerino, G., Bernasconi, S., Dell'Amico, E., . . . Gianfreda, L. (2004). Bacterial communities and enzyme activities of PAHs polluted soils. [Article]. *Chemosphere, 57*(5), 401-412. doi: 10.1016/j.chemosphere.2004.06.013

Bamforth, S. M., & Singleton, I. (2005). Bioremediation of polycyclic aromatic hydrocarbons: current knowledge and future directions. *Journal of Chemical Technology and Biotechnology, 80*(7), 723-736. doi: 10.1002/jctb.1276

Borg, I., & Groenen, P. J. F. (2005). Modern multidimensional scaling theory and applications, from http://site.ebrary.com/id/10171042

Borneman, J., & Triplett, E. W. (1997). Molecular microbial diversity in soils from eastern Amazonia: Evidence for unusual microorganisms and microbial population shifts associated with deforestation. [Article]. *Applied and Environmental Microbiology, 63*(7), 2647-2653.

Bouchez, M., Blanchet, D., & Vandecasteele, J. P. (1995). Degradation of polycyclic aromatic hydrocarbons by pure strains and by defined strain associations - inhibition phenomena and cometabolism. [Article]. *Applied Microbiology and Biotechnology, 43*(1), 156-164.

Burke, D. J., Kretzer, A. M., Rygiewicz, P. T., & Topa, M. A. (2006). Soil bacterial diversity in a loblolly pine plantation: influence of ectomycorrhizas and fertilization. [Article]. *Fems Microbiology Ecology, 57*(3), 409-419. doi: 10.1111/j.1574-6941.2006.00125.x

Cerniglia, C. E. (1992). Biodegradation of polycyclic aromatic hydrocarbons. [article]. *Biodegradation, 3*(2-3), 351-368. doi: 10.1007/bf00129093

Clarke, K. R. (1993). Nonparametric multivariate analyses of changes in community structure. [Article; Proceedings Paper]. *Australian Journal of Ecology, 18*(1), 117-143. doi: 10.1111/j.1442-9993.1993.tb00438.x

Colwell, R. K., & Coddington, J. A. (1994). ESTIMATING TERRESTRIAL BIODIVERSITY THROUGH EXTRAPOLATION. [Review]. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 345*(1311), 101-118. doi: 10.1098/rstb.1994.0091

DeBruyn, J. M., Chewning, C. S., & Sayler, G. S. (2007). Comparative quantitative prevalence of Mycobacteria and functionally abundant nidA, nahAc, and nagAc dioxygenase genes in coal tar contaminated sediments. *Environmental Science & Technology, 41*(15), 5426-5432. doi: 10.1021/es070406c

DeBruyn, J. M., Nixon, L. T., Fawaz, M. N., Johnson, A. M., & Radosevich, M. (2011). Global Biogeography and Quantitative Seasonal Dynamics of Gemmatimonadetes in Soil. [Article]. *Applied and Environmental Microbiology, 77*(17), 6295-6300. doi: 10.1128/aem.05005-11

Dionisi, H. M., Chewning, C. S., Morgan, K. H., Menn, F. M., Easter, J. P., & Sayler, G. S. (2004). Abundance of dioxygenase genes similar to Ralstonia sp strain U2 nagAc is correlated with naphthalene concentrations in coal tar-contaminated freshwater sediments. [Article]. *Applied and Environmental Microbiology, 70*(7), 3988-3995. doi: 10.1128/aem.70.3.3988-3995.2004

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. [Article]. *Bioinformatics, 27*(16), 2194-2200. doi: 10.1093/bioinformatics/btr381

Garbeva, P., van Veen, J. A., & van Elsas, J. D. (2004). Microbial diversity in soil: Selection of microbial populations by plant and soil type and implications for disease suppressiveness. [Review]. *Annual Review of Phytopathology, 42*, 243-270. doi: 10.1146/annurev.phyto.42.012604.135455

Gihring, T. M., Green, S. J., & Schadt, C. W. (2012). Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. [Letter]. *Environmental Microbiology, 14*(2), 285-290. doi: 10.1111/j.1462-2920.2011.02550.x

Haritash, A. K., & Kaushik, C. P. (2009). Biodegradation aspects of Polycyclic Aromatic Hydrocarbons (PAHs): A review. [Review]. *Journal of Hazardous Materials, 169*(1-3), 1-15. doi: 10.1016/j.jhazmat.2009.03.137

Heitzer, A., Webb, O. F., Thonnard, J. E., & Sayler, G. S. (1992). Specific and quantitative assessment of naphthalene and salicylate bioavailability by using a bioluminescent catabolic reporter bacterium. [Article]. *Applied and Environmental Microbiology, 58*(6), 1839-1846.

Incardona, J. P., Collier, T. K., & Scholz, N. L. (2004). Defects in cardiac function precede morphological abnormalities in fish embryos exposed to polycyclic aromatic hydrocarbons. [Article]. *Toxicology and Applied Pharmacology, 196*(2), 191-205. doi: 10.1016/j.taap.2003.11.026

Johnsen, A. R., Wick, L. Y., & Harms, H. (2005). Principles of microbial PAH-degradation in soil. *Environmental Pollution, 133*(1), 71-84. doi: 10.1016/j.envpol.2004.04.015

Jones, K. C., Stratford, J. A., Tidridge, P., Waterhouse, K. S., & Johnston, A. E. (1989). Polynuclear aromatic-hydrocarbons in an agricultural soil - long-term changes in profile distribution. [Article]. *Environmental Pollution, 56*(4), 337-351. doi: 10.1016/0269-7491(89)90079-1

Juhasz, A. L., & Naidu, R. (2000). Bioremediation of high molecular weight polycyclic aromatic hydrocarbons: a review of the microbial degradation of benzo a pyrene. [Review]. *International Biodeterioration & Biodegradation, 45*(1-2), 57-88. doi: 10.1016/s0964-8305(00)00052-4

Kim, J. H., Yamaguchi, K., Lee, S. H., Tithof, P. K., Sayler, G. S., Yoon, J. H., & Baek, S. J. (2005). Evaluation of polycyclic aromatic hydrocarbons in the activation of early growth response-1 and peroxisome proliferator activated receptors. [Article]. *Toxicological Sciences, 85*(1), 585-593. doi: 10.1093/toxsci/kfi118

King, J. M. H., Digrazia, P. M., Applegate, B., Burlage, R., Sanseverino, J., Dunbar, P., . . . Sayler, G. S. (1990). Rapid, sensitive bioluminescent reporter technology for naphthalene exposure and biodegradation. [Article]. *Science, 249*(4970), 778-781. doi: 10.1126/science.249.4970.778

Layton, A. C., Smartt, A. E., Chauhan, A., Ripp, S., Williams, D. E., Burton, W., . . . Sayler, G. S. (2012). Ameliorating risk: culturable and metagenomic monitoring of the 14 year decline of a genetically engineered microorganism at a bioremediation field site. [article]. *Journal of Bioremediation and Biodegradation, 3*(S1), 009.

Lelley, T., Balázs, E., & Tepfer, M. (2003). *Ecological Impact of GMO Dissemination in Agro-Ecosystems*.

Lors, C., Ryngaert, A., Perie, F., Diels, L., & Damidot, D. (2010). Evolution of bacterial community during bioremediation of PAHs in a coal tar contaminated soil. [Article]. *Chemosphere, 81*(10), 1263-1271. doi: 10.1016/j.chemosphere.2010.09.021

Lozada, M., Mercadal, J. P. R., Guerrero, L. D., Di Marzio, W. D., Ferrero, M. A., & Dionisi, H. M. (2008). Novel aromatic ring-hydroxylating dioxygenase genes from coastal marine sediments of Patagonia. [Article]. *Bmc Microbiology, 8*. doi: 10.1186/1471-2180-8-50

Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. [Article]. *Applied and Environmental Microbiology, 71*(12), 8228-8235. doi: 10.1128/aem.71.12.8228-8235.2005

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. [Article]. *Applied and Environmental Microbiology, 73*(5), 1576-1585. doi: 10.1128/aem.01996-06

Margesin, R., & Schinner, F. (2001). Biodegradation and bioremediation of hydrocarbons in extreme environments. [Review]. *Applied Microbiology and Biotechnology, 56*(5-6), 650-663.

Margesin, R., Zimmerbauer, A., & Schinner, F. (2000). Monitoring of bioremediation by soil biological activities. [Article]. *Chemosphere, 40*(4), 339-346. doi: 10.1016/s0045-6535(99)00218-0

Meighen, E. A. (1988). Enzymes And Genes From The lux Operons Of Bioluminescent Bacteria. [Review]. *Annual Review of Microbiology, 42*, 151-176. doi: 10.1146/annurev.micro.42.1.151

Mueller, J. G., Devereux, R., Santavy, D. L., Lantz, S. E., Willis, S. G., & Pritchard, P. H. (1997). Phylogenetic and physiological comparisons of PAH-degrading bacteria from geographically diverse soils. [Article]. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology, 71*(4), 329-343. doi: 10.1023/a:1000277008064

Ni Chadhain, S. M., Norman, R. S., Pesce, K. V., Kukor, J. J., & Zystra, G. J. (2006). Microbial dioxygenase gene population shifts during polycyclic aromatic hydrocarbon biodegradation. *Applied and Environmental Microbiology, 72*(6), 4078-4087. doi: 10.1128/aem.02969-05

Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R., & Stahl, D. A. (1986). Microbial ecology and evolution - a ribosomal-RNA approach. [Review]. *Annual Review of Microbiology, 40*, 337-365. doi: 10.1146/annurev.micro.40.1.337

Olsen, G. J., & Woese, C. R. (1993). Ribosomal-RNA - a key to phylogeny. [Article]. *Faseb Journal, 7*(1), 113-123.

Ripp, S., Nivens, D. E., Ahn, Y., Werner, C., Jarrell, J., Easter, J. P., . . . Sayler, G. S. (2000a). Controlled field release of a bioluminescent genetically engineered microorganism for bioremediation process monitoring and control. *Environmental Science & Technology, 34*(5), 846-853. doi: 10.1021/es9908319

Ripp, S., Nivens, D. E., Werner, C., & Sayler, G. S. (2000b). Bioluminescent most-probable-number monitoring of a genetically engineered bacterium during a long-term contained field release. [Article]. *Applied Microbiology and Biotechnology, 53*(6), 736-741.

Sayler, G. S., Fleming, J. T., & Nivens, D. E. (2001). Gene expression monitoring in soils by mRNA analysis and gene lux fusions. [Review]. *Current Opinion in Biotechnology, 12*(5), 455-460. doi: 10.1016/s0958-1669(00)00245-7

Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. [Article]. *Plos One, 6*(12). doi: e27310

10.1371/journal.pone.0027310

Shendure, J., & Ji, H. L. (2008). Next-generation DNA sequencing. [Review]. *Nature Biotechnology, 26*(10), 1135-1145. doi: 10.1038/nbt1486

Simpson, E. H. (1949). MEASUREMENT OF DIVERSITY. [Letter]. *Nature, 163*(4148), 688-688. doi: 10.1038/163688a0

Smit, S., Widmann, J., & Knight, R. (2007). Evolutionary rates vary among rRNA structural elements. [Article]. *Nucleic Acids Research, 35*(10), 3339-3354. doi: 10.1093/nar/gkm101

Stach, J. E. M., & Burns, R. G. (2002). Enrichment versus biofilm culture: a functional and phylogenetic comparison of polycyclic aromatic hydrocarbon-degrading microbial communities. [Article]. *Environmental Microbiology, 4*(3), 169-182. doi: 10.1046/j.1462-2920.2002.00283.x

Stackebrandt, E., & Goebel, B. M. (1994). A place for DNA-DNA reassociation and 16S ribosomal-RNA sequence-analysis in the present species definition in bacteriology. [Note]. *International Journal of Systematic Bacteriology, 44*(4), 846-849.

Staley, J. T., & Konopka, A. (1985). Measurement of insitu activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. [Review]. *Annual Review of Microbiology, 39*, 321-346. doi: 10.1146/annurev.micro.39.1.321

Suzuki, M. T., & Giovannoni, S. J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. [Article]. *Applied and Environmental Microbiology, 62*(2), 625-630.

Torsvik, V., Goksoyr, J., & Daae, F. L. (1990). High diversity in DNA of soil bacteria. [Article]. *Applied and Environmental Microbiology, 56*(3), 782-787.

Trogl, J., Chauhan, A., Ripp, S., Layton, A. C., Kuncova, G., & Sayler, G. S. (2012). Pseudomonas fluorescens HK44: Lessons Learned from a Model Whole-Cell Bioreporter with a Broad Application History. *Sensors, 12*(2), 1544-1571. doi: 10.3390/s120201544

Vinas, M., Sabate, J., Espuny, M. J., & Solanas, A. M. (2005). Bacterial community dynamics and polycyclic aromatic hydrocarbon degradation during bioremediation of heavily creosote-contaminated soil. [Article]. *Applied and Environmental Microbiology, 71*(11), 7008-7018. doi: 10.1128/aem.71.11.7008-7018.2005

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., . . . Truper, H. G. (1987). Report of the ad-hoc-committee on reconciliation of approaches to bacterial systematics. [Editorial Material]. *International Journal of Systematic Bacteriology, 37*(4), 463-464.

Yue, J. C., & Clayton, M. K. (2005). A similarity measure based on species proportions. [Article]. *Communications in Statistics-Theory and Methods, 34*(11), 2123-2131. doi: 10.1080/sta-200066418

Zhang, S. Y., Wang, Q. F., Wan, R., & Xie, S. G. (2011). Changes in bacterial community of anthracene bioremediation in municipal solid waste composting soil. [Article]. *Journal of Zhejiang University-Science B, 12*(9), 760-768. doi: 10.1631/jzus.B1000440

Zinger, L., Gobet, A., & Pommier, T. (2012). Two decades of describing the unseen majority of aquatic microbial diversity. [Review]. *Molecular Ecology, 21*(8), 1878-1896. doi: 10.1111/j.1365-294X.2011.05362.x

APPENDIX

Table 14. Primer set used for each soil DNA sample for building 16S rRNA gene amplicon libraries. The first column is the names of each soil composite sample; the second column is the name of barcoded primer used; the last four columns are the sequences of 454 primers, key, MID and forward/reverse primer for amplification.

Sample	Primer	454 Primer A (forward)/B(reverse)	Key	MID	Template specific (338F)
L1-06"	338F-Ti-AA	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TTGGCGTA	ACTCCTACGGGAGGCAGCAG
L1-12"	338F-Ti-AB	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TGCTGTTC	ACTCCTACGGGAGGCAGCAG
L1-18"	338F-Ti-AC	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TGAGGTGA	ACTCCTACGGGAGGCAGCAG
L1-24"	338F-Ti-AD	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TCCTTGCA	ACTCCTACGGGAGGCAGCAG
L1-30"	338F-Ti-AE	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TATAGCGC	ACTCCTACGGGAGGCAGCAG
L1-36"	338F-Ti-BB	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GGCCTAAT	ACTCCTACGGGAGGCAGCAG
L2-06"	338F-Ti-AG	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TAGGCGTT	ACTCCTACGGGAGGCAGCAG
L2-12"	338F-Ti-AL	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TGGACAAG	ACTCCTACGGGAGGCAGCAG
L2-18"	338F-Ti-AH	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TCAGGAGA	ACTCCTACGGGAGGCAGCAG
L2-24"	338F-Ti-AI	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TCCAACGA	ACTCCTACGGGAGGCAGCAG
L2-30"	338F-Ti-AK	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TCTGGTGA	ACTCCTACGGGAGGCAGCAG
L2-36"	338F-Ti-AJ	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TCGAGTAC	ACTCCTACGGGAGGCAGCAG
L2-36"	338F-Ti-BG	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GCAACCAT	ACTCCTACGGGAGGCAGCAG
L3-06"	338F-Ti-AM	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TGTCCTCA	ACTCCTACGGGAGGCAGCAG
L3-12"	338F-Ti-AN	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TTAAGGCC	ACTCCTACGGGAGGCAGCAG
L3-18"	338F-Ti-AO	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TTCGAAGC	ACTCCTACGGGAGGCAGCAG
L3-24"	338F-Ti-AP	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TGACACTG	ACTCCTACGGGAGGCAGCAG
L3-30"	338F-Ti-AQ	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TCTCTGTG	ACTCCTACGGGAGGCAGCAG
L3-36"	338F-Ti-AR	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TCACGACA	ACTCCTACGGGAGGCAGCAG
L4-06"	338F-Ti-AS	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TACCAACG	ACTCCTACGGGAGGCAGCAG
L4-12"	338F-Ti-AT	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TGCATCCT	ACTCCTACGGGAGGCAGCAG
L4-18"	338F-Ti-AU	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GTTGAGCT	ACTCCTACGGGAGGCAGCAG
L4-18"	338F-Ti-BI	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GAGTTCAC	ACTCCTACGGGAGGCAGCAG
L4-24"	338F-Ti-AV	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GTGTGACT	ACTCCTACGGGAGGCAGCAG
L5-06"	338F-Ti-AW	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GTGAAGTG	ACTCCTACGGGAGGCAGCAG
L5-12"	338F-Ti-AX	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GTCTCAGA	ACTCCTACGGGAGGCAGCAG
L5-18"	338F-Ti-AY	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GTAGTCGT	ACTCCTACGGGAGGCAGCAG
L5-24"	338F-Ti-BJ	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GAGAACAC	ACTCCTACGGGAGGCAGCAG
L5-30"	338F-Ti-AF	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	TACGGCTA	ACTCCTACGGGAGGCAGCAG
L5-36"	338F-Ti-BH	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GATCGTTG	ACTCCTACGGGAGGCAGCAG
L6-06"	338F-Ti-BC	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GGATAAGC	ACTCCTACGGGAGGCAGCAG
L6-12"	338F-Ti-BD	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GCTTCCTA	ACTCCTACGGGAGGCAGCAG
L6-18"	338F-Ti-BE	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GCTACCTT	ACTCCTACGGGAGGCAGCAG
L6-24"	338F-Ti-BF	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GCATGGTT	ACTCCTACGGGAGGCAGCAG
Mock	338F-Ti-BK	CCATCTCATCCCTGCGTGTCTCCGAC	TCAG	GACTCTGA	ACTCCTACGGGAGGCAGCAG
All	Ti-926R	CCTATCCCCTGTGTGCCTTGGCAGTC	TCAG	—	CCGTCAATTCMTTTRAGT

Table 15. Mothur Codes used in this study and annotations.
All codes came from the website of Mothur software (http://www.mothur.org), which was initiated by Dr. Patrick Schloss and his team in the Department of Microbiology & Immunology at The University of Michigan. The annotations listed in this table also based on the explanations on this website. More detailed annotations and more mothur commands can be found in www.mothur.org as well.

Section	Step	Annotation	Mothur Code
Reducing Sequencing Error	1	Extract fasta, qual and flow files generated from the two halves of the 454 sequencing plate. **[NOTE]** These two sffinfo files cannot be combined before extraction.	sffinfo (sff=seq1.sff)
			sffinfo (sff=seq2.sff)
	2	Read the flowgram files, separate sequences according to primer and barcodes; remove the sequence that had > 2 mismatches of primer and/or > 1 mismatch of barcode sequence; trim all sequences to 260 flows (~120 base pair). User should design the oligos file. It indicates sequences with which barcode(s) should be assigned into which group. **[NOTE]** Group name defined in oligos file should not contain "-" because MOTHUR uses "-" to separate files.	trim.flows(flow=seq1.flow, oligos=lysim.oligos, pdiffs=2, bdiffs=1, maxhomop=8, minflows=260, maxflows=260)
			trim.flows(flow=seq2.flow, oligos=lysim.oligos, pdiffs=2, bdiffs=1, maxhomop=8, minflows=260, maxflows=260)
	3	(Adaptation of Chris Quince's PyroNoise algorithm) Correct flowgrams, translate flowgrams to DNA sequences.	shhh.flows(fasta=seq1.flow.files)
			shhh.flows(fasta=seq2.flow.files)
	4	(Used 2 processors to run this command) Trim off primer and barcode sequences; use the barcode information to generate a group file (tell which sequence is which); split fasta file by provided oligos file; remove sequences that had ambiguous base pair, >8 homopolymers, >1 mismatch of barcode sequence and/or > 2 mismatch of primer sequence.	trim.seqs(fasta=seq1.shhh.fasta, oligos=lysim.oligos, maxambig=0, maxhomop=8, bdiffs=1, pdiffs=2, processors=2)
			trim.seqs(fasta=seq2.shhh.fasta, oligos=lysim.oligos, maxambig=0, maxhomop=8, bdiffs=1, pdiffs=2, processors=2)
	5	Merge files of the two halves together.	merge.files(input=seq1.shhh.trim.fasta-seq2.shhh.trim.fasta, output=seq.shhh.trim.fasta)
			merge.files(input=seq1.shhh.trim.names-seq2.shhh.trim.names, output=seq.shhh.trim.names)

Table 15 Continued.

Section	Step	Annotation	Mothur Code
			merge.files(input=seq1.shhh.groups-seq2.shhh.groups, output=seq.shhh.groups)
	6	Remove sequences that belong to libraries named "4^21W", "4^21C", "5^12W", and "5^12C" (these are Kelly's libraries).	remove.groups(fasta=seq.shhh.trim.fasta, name=seq.shhh.trim.names, group=seq.shhh.groups, groups=4^21W-4^21C-5^12W-5^12C)
Processing Improved Sequences	7	Returns only unique sequences found in fasta file (did not remove any sequence, just excluded them for the alignment to reduce the processing time of alignment).	unique.seqs(fasta=seq.shhh.trim.pick.fasta, name=seq.shhh.trim.pick.names)
	8	Take my unique sample sequences of fasta file as candidate sequences, align them with a template alignment file.	align.seqs(candidate=seq.shhh.trim.pick.unique.fasta, template=silva.bacteria.fasta, processors=2)
	9	Remove sequences that started after position 6428 of the alignment; optimize end position: remove sequences that ended before 90% of the sequences ended.	screen.seqs(fasta=seq.shhh.trim.pick.unique.align, name=seq.shhh.trim.pick.unique.names, group=seq.shhh.pick.groups, start=6428, optimize=end, criteria=90, processors=2)
	10	Remove columns that only contains gap character "." and any columns that contains at least one "." character.	filter.seqs(fasta=seq.shhh.trim.pick.unique.good.align, vertical=T, trump=., processors=2)
	11	After some columns were removed, some sequences may be redundant in this aligned region; check and returns only the unique sequences.	unique.seqs(fasta=seq.shhh.trim.pick.unique.good.filter.unique.fasta, name=seq.shhh.trim.pick.unique.good.names)
	12	checking sequences that were likely originated from pyrosequencing errors; sequences were ranked by abundances, and if rare sequences were within some threshold of original sequences, they were then merged with the more abundant one.	pre.cluster(fasta=seq.shhh.trim.pick.unique.good.filter.unique.fasta, name=seq.shhh.trim.pick.unique.good.filter.names, group=seq.shhh.pick.good.groups)

Table 15 Continued.

Section	Step	Annotation	Mothur Code
Detect and remove Chimeras	13	Reads fasta file; searching for a 3-way alignment of a query sequence and two parent sequences so that one segment of the query sequence was more similar to parent sequence A and the other segment more similar to parent sequence B; a score was calculated from this alignment and if the score > decided cutoff, the query sequence was considered a chimera; more abundant sequences within the same library were used as reference sequences	chimera.uchime(fasta=seq.shhh.trim.pick.unique.good.filter.unique.precluster.fasta, name=seq.shhh.trim.pick.unique.good.filter.unique.precluster.names, group=seq.shhh.pick.good.groups, processors=2)
	14	Remove detected chimera sequences	remove.seqs(accnos=seq.shhh.trim.pick.unique.good.filter.unique.precluster.uchime.accnos, fasta=seq.shhh.trim.pick.unique.good.filter.unique.precluster.fasta, name=seq.shhh.trim.pick.unique.good.filter.unique.precluster.names, group=seq.shhh.pick.good.groups)
Detect and remove "Contaminants"	15	Identify sequences taxonomic information using a reference database (fasta formatted) and taxonomy file	classify.seqs(fasta=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.fasta, name=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.names, group=seq.shhh.pick.good.pick.groups,template=trainset9_032012.pds.fasta, taxonomy=trainset9_032012.pds.tax, cutoff=80, processors=2)
	16	Remove sequences classified as "Mitochondria" "Chloroplast" "Archaea" "Eukarya" or "unknown"	remove.lineage(fasta=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.fasta, name=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.names, group=seq.shhh.pick.good.pick.groups, taxonomy=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pds.taxonomy, taxon=Mitochondria-

Table 15 Continued.

Section	Step	Annotation	Mothur Code
			Chloroplast-Archaes-Eukarya-unknown)]
Count the Number of Sequences in each group	17	Count the Number of Sequences in each group	count.groups(group=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pick.groups)
Error Analysis **[NOTE]** Error analysis, OUT-based analysis, phylotype-based analysis and phylogeny-based analysis do not have to be done in a specific order: these 4 analysis all use the quality filtered data.	18	Align the Sanger sequencing data of mock library using reference database	align.seqs(fasta=Mock.fasta, reference=silva.bacteria.fasta, processors=2
	19	Filter the aligned Sanger mock sequences using the filter file (seq.filter) generated in step 10.	filter.seqs(fasta=Mock.align, hard=seq.filter)
	20	Get the 454 mock sequences from quality filtered data set	get.groups(fasta=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pick.fasta , name=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pick.names, group=seq.shhh.pick.good.pick.pick.groups , groups=Mock,)
	21	Calculate error rate by comparing 454 mock aligned sequences with reference alignment data (Sanger aligned mock sequences); "final.pick.error.chimera" was generated in this step: we could found the sequences that had two parents -- the chimeras that was not detected in step 13	seq.error(fasta=final.pick.fasta, name=final.pick.names, reference=Mock.filter.fasta, processors=2)
	22	Last step generated a file names "final.pick.error.summary". Copy the second column content in this file, except for "reference" row, save this as a "mock.accnos" file. In this step, this file is used to pick the sequences that both Sanger mock library and 454 mock library contained to compare.	get.seqs(accnos=mock.accnos, fasta=Mock.filter.fasta)

Table 15 Continued.

Section	Step	Annotation	Mothur Code
	23	Use the picked mock data set (contained the unique sequences that Sanger mock and 454 mock shared) generated in last step, calculate uncorrected pairwise distance between these sequences; output a phylip-formatted distance matrix	dist.seqs(fasta=mock.filter.pick.fasta, output=lt)
	24	The same operation; calculate the distance matrix for 454 mock data set; any distance (sequence similarity) larger than 0.1 will be exclude from this distance matrix	dist.seqs(fasta=final.pick.fasta, cutoff=0.10)
	25	Assign these shared sequences into OTUs; with a cutoff = 0.03, if the sequences similarity > 97%, they will be assigned into the same OUT	cluster(phylip=mock.filter.pick.phylip.dist)
	26	The same operation, but uses column-formatted distance matrix instead of phylip-formatted distance matrix	cluster(column=final.pick.dist, name=final.pick.names)
	27	Summarize the number of observed OTUs using the list file generated in last step. Thus, we could get the OTU number of 454 mock sequence dataset in species level (label=0.03 means sequences with 97% or higher similarity will be assigned into the same OTU)	summary.single(calc=sobs, label=0.03, list=final.pick.an.list)
	28	Since the 34 sample sequencing libraries were subsampled to 323 sequences/library, here I did the same for mock to see how many OTUs retained in subsampled 454 mock library	summary.single(calc=sobs, label=0.03, list=final.pick.an.list, subsample=323)
Exclude Mock Library From Downstream Analysis	29	Remove Mock Library From 454 Sequencing Dataset **[NOTE]** Mock library is NOT a real sample library, and should not be analyzed with sample libraries	remove.groups(fasta=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pick.fasta, name=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pick.names, taxonomy=seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pds.pick.taxonomy, group=seq.shhh.pick.good.pick.pick.groups, groups=mock)

Table 15 Continued.

Section	Step	Annotation	Mothur Code
Rename the quality filtered files	30	The quality filtering process ended in step 16. The final files had extremely long names because in each step a ".step label" was added behind the file name to remind the user which command had been used to treat this file. These final files were renamed to final.taxonomy, final.fasta, final.names and final.groups.	system(copy seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pds.pick.taxonomy final.taxonomy)
			system(copy seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pick.fasta final.fasta)
			system(copy seq.shhh.trim.pick.unique.good.filter.unique.precluster.pick.pick.names final.names)
			system(copy seq.shhh.pick.good.pick.pick.groups final.groups)
Prepare inputs of OTU-based Analysis	31	Calculated distances between sequences for the quality filtered 454 sequencing dataset, and generate a distance matrix. Initial cutoff was set to 0.15 to avoid unnecessary calculation and save time. The automatically adjusted cutoff of the output was 0.0552914	dist.seqs(fasta=final.fasta, cutoff=0.15, processors=2)
	32	Cluster these sequences into OTUs based on the distance matrix generated in last step.	cluster(column=final.dist, name=final.names, processors=2)
	33	Summarize the number of times each OTU showed up in each sample library. With label = 0.03, the output file only contents information in this distance level (species level).	make.shared(list=final.an.list, group=final.groups, label=0.03)
	34	Since my smallest sample library only contained 323 sequences, all libraries were normalized to this size (n=323) by randomly subsampling.	sub.sample(shared=final.an.shared, size=323)
	35	Get the taxonomic information of each OTU. **[NOTE]** In MOTHUR, the taxonomy files of reference database only goes to genus level; you may manually add the species level information into the taxonomy file and may be able to get sequences classified in species level.	classify.otu(list=final.an.list, name=final.names, taxonomy=final.taxonomy, group=final.groups, label=0.03)

Table 15 Continued.

Section	Step	Annotation	Mothur Code
OTU-based Analysis of α diversity	36	(Calc= means calculate: calculate Chao1 richness estimates and inverse Simpson diversity index of each library as the library size = 10 sequences to all sequences. Freq = defined the scale of library size increase (i.e. freq = 10 means calculate the richness and diversity of library sequence number = 10, 20, 30...till the number equals the number of all sequences in that library).	collect.single(shared=final.an.0.03.subsample.0.03.pick.shared, calc=chao-invsimpson, freq=10)
	37	Calculate the rarefaction of each sample library: the rarefaction describes the number of OTUs observed as a function of sampling depth (the number of OTUs in that has been sampled in a specific environment) [NOTE] rarefaction is a measure of community diversity instead of richness;	rarefaction.single(shared=final.an.0.03.pick.shared, freq=100);
	38	Summarize the number of sequences, the sample coverage, the number of observed OTUs, and the inversed Simpson diversity estimate. Could adding more calculators in the "calc=" parameter)	summary.single(calc=nseqs-coverage-sobs-invsimpson, shared = final.an.0.03.pick.shared)
	39	Calculate the shared OUT numbers between each two of the four libraries listed in group= parameter. The same command was used to get the shared OTU numbers between each pair of the 34 sample libraries.	venn(groups=L1^06-L1^12-L1^18-L1^24, shared = final.an.0.03.subsample.0.03.pick.shared)
OTU-based Analysis of β diversity	40	Generate dendrogram that describes the similarities between each two sample libraries. Calculators used: Θ_{YC} calculator and Jaccard index. This step will generate a .tre file.	tree.shared(shared=final.an.0.03.subsample.0.03.pick.shared, calc=thetayc-jclass)
	43	Generate distance matrix that presents the similarities between each two sample libraries. Calculators used: Θ_{YC} calculator and Jaccard index	dist.shared(shared=final.an.0.03subsample.0.03.pick.shared, calc=thetayc-jclass)
	44	Use non-metric multidimensional scaling to preserve the distances between samples in Θ_{YC} matrix in a user defined number of dimensions (default dimension number is 2)	nmds(phylip=final.an.0.03subsample.0.03.pick.thetayc.0.03.lt.dist)

Table 15 Continued.

Section	Step	Annotation	Mothur Code
	45	Use non-metric multidimensional scaling to preserve the distances between samples in Jaccard matrix in a user defined number of dimensions (default dimension number is 2)	nmds(phylip=final.an.0.03subsample.0.03.pick.jclass.0.03.lt.dist)
	47	Use analysis of similarity (ANOSIM) to test if community memberships of samples (Jaccard index) are significantly different by PAH+/- or by HK44+/-;	anosim(phylip=final.an.0.03.subsample.0.03.pick.jclass.0.03.square.dist, design=treatment4.design)
	48	Use analysis of similarity (ANOSIM) to test if community structures of samples (Θ_{YC}) are significantly different by PAH+/- or by HK44+/-;	anosim(phylip=final.an.0.03.subsample.0.03.pick.thetayc.0.03.square.dist, design=treatment4.design)
	49	The same operation as last two steps, but grouping the libraries by soil moisture (use a different design file).	
Prepare Inputs of Phylotype-based Analysis	50	Assign the sequences by taxonomy.	phylotype(taxonomy=final.taxonomy, name=final.names, label=1)
	51	Summarize the number of times each OTU showed up in each sample library. With label = 1, the output file only contents information in to genus level.	make.shared(list=final.tx.list, group=final.groups, label=1)
	52	Normalize all libraries to equal sample size (n=323)	sub.sample(shared=final.tx.shared, size=323)
Phylotype-based Analysis of α diversity and β diversity	53	Get the taxonomy of each OTU (in this approach, a OTU represent a genus)	classify.otu(list=final.tx.list, name=final.names, taxonomy=final.taxonomy, label=1)
Prepare Inputs of Phylogeny-based Analysis	54	The same process as OUT-based analysis, but use the .shared file prepared for phylotype-based analysis (final.tx.1.subsample.shared)	
	55	Normalize all libraries to equal sample size (n=323)	sub.sample(fasta=final.fasta, name=final.names, group=final.groups, persample=T, size=323)
	56	Generate distance matrix of sequences	dist.seqs(fasta=final.fasta, output=lt, processors=2)
Phylogeny-based Analysis of α diversity	57	Construct phylogenetic tree based on the distance matrix generated in last step. **[NOTE]** Some computers may not able to run this command. The reason is still unknown.	clearcut(phylip=final.subsample.phylip.dist)
Phylogeny-based Analysis of β diversity	58	Calculate phylogeny diversity of each library using the tree generated in last step.	phylo.diversity(tree=final.phylip.tre, name=final.names, group=final.groups, rarefy=T)

Table 15 Continued.

Section	Step	Annotation	Mothur Code
	59	Generate distance matrix that describes the community membership similarities between each two libraries.	unifrac.unweighted(tree=final.phylip.tre, name=final.names, group=final.groups, distance=lt, processors=2, random=F, subsample=323)
	60	Generate distance matrix that describes the community structure similarities between each two libraries.	unifrac.weighted(tree=final.phylip.tre, name=final.names, group=final.groups, distance=lt, processors=2, random=F, subsample=323)
	61	Use non-metric multidimensional scaling to preserve the distances between samples in unweighted matrix in 2 dimensions	nmds(phylip=final.subsample.phylip.tre1.unweighted.phylip.dist)
	62	Use non-metric multidimensional scaling to preserve the distances between samples in weighted matrix in 2 dimensions.	nmds(phylip=final.subsample.phylip.tre1.weighted.phylip.dist)
	63	Use analysis of similarity test whether PAH or HK44 caused significant difference of community membership between libraries	anosim(phylip= final.subsample.phylip.tre1.unweighted.phylip.dist, design=treatment4.design)
	64	Use analysis of similarity test whether PAH or HK44 caused significant difference of community structure between libraries	anosim(phylip= final.subsample.phylip.tre1.weighted.phylip.dist, design=treamtent.design)
	65	The same operations as last two steps, but group the libraries by soil moisture (use a different design file).	

VITA

Xiaoci Ji was born in 1988. She grew up in Beijing, China, and graduated from China Agricultural University in June 30th, 2010 with a Bachelor of Science degree in Biology Science. She graduated from University of Tennessee, Knoxville in December 15th, 2012 with a Master of Science Degree in Biosystems Engineering Technology and a minor in Statistics. She was a graduate research assistant in the Department of Biosystems Engineering and Soil Science of UTK from January 2011 to December 2012.